CONTENTS

STAT 151 Lab Manual in R

PREFACE

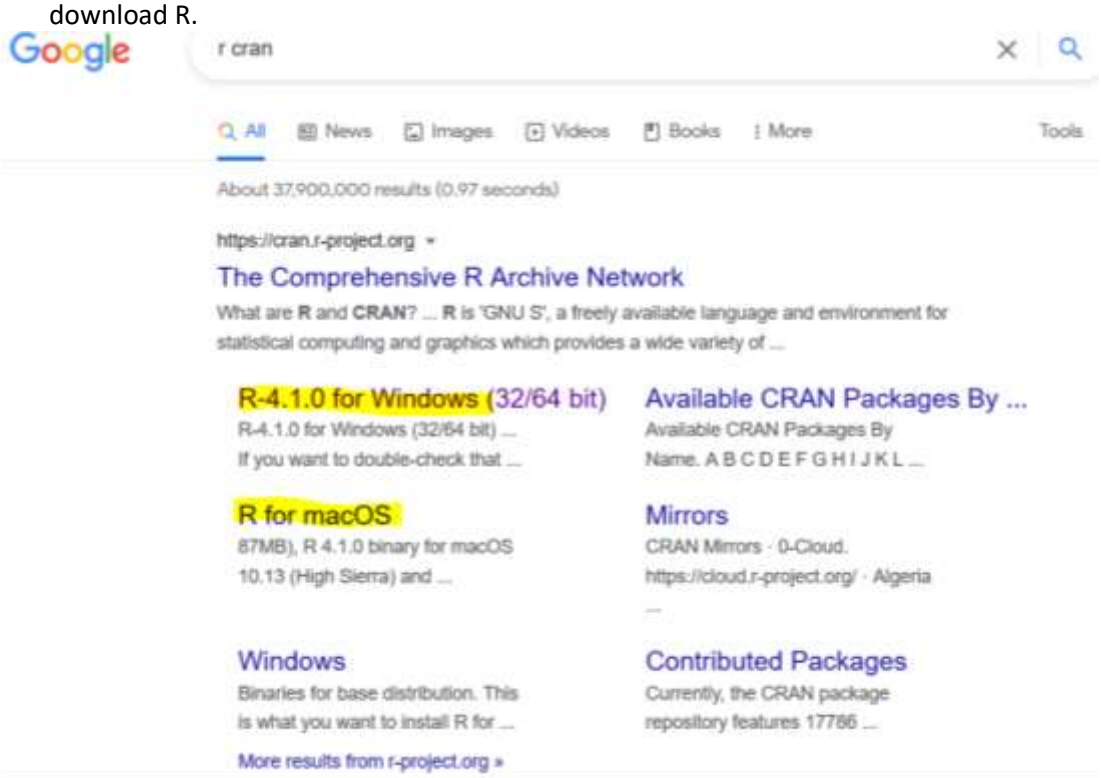
This lab manual was prepared for the lab component of the online STAT 151 course offered at MacEwan University. R is an open-source implementation of the S language. It works on multiple computing platforms and can be freely downloaded. This lab manual introduces how to conduct descriptive statistics and inferential statistics using R and R commander (an R package). Descriptive statistics include drawing figures such as histogram, boxplot, normal Q-Q plot, scatter plot and obtaining statistical summaries such as mean, median, standard deviation, and quartiles. Inferential statistics cover one-sample z test and interval, one-sample t test and t interval, two-sample t test and t interval, one-proportion z test and interval, two-proportion z test and interval, chi-square tests, one-way ANOVA F test, and simple linear regression. This lab manual also illustrates how to obtain probabilities and cumulative probabilities and quantiles based on binomial distributions and normal distributions.

LAB 1 DOWNLOAD AND INSTALL R AND R COMMANDER

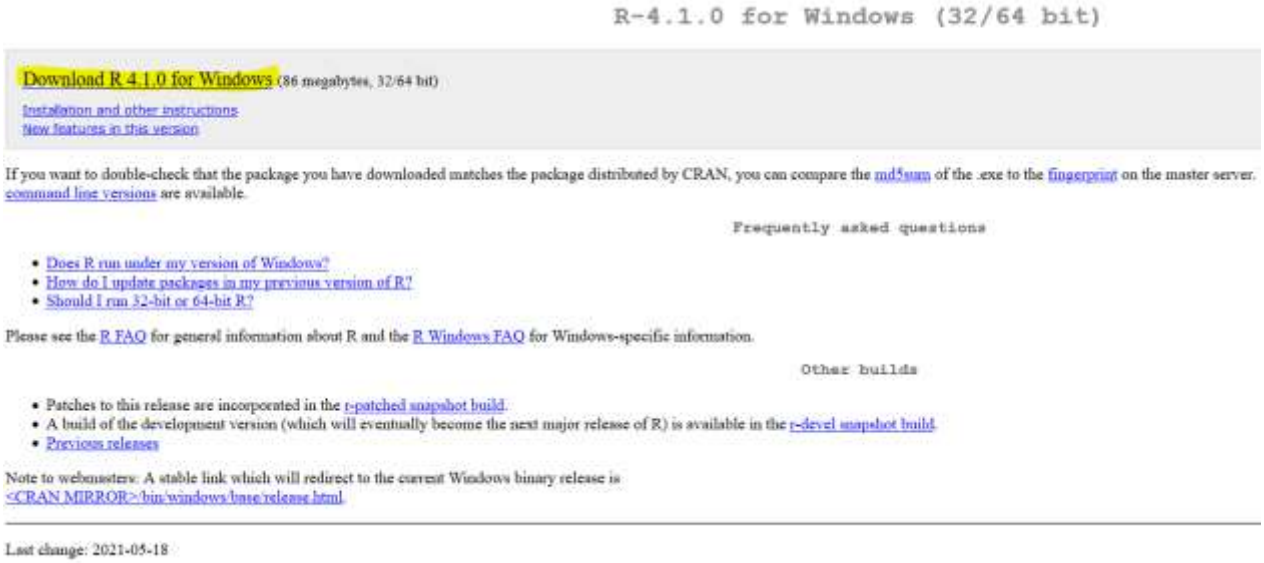
1. 1 DOWNLOAD AND INSTALL R

You can google the downloading website:

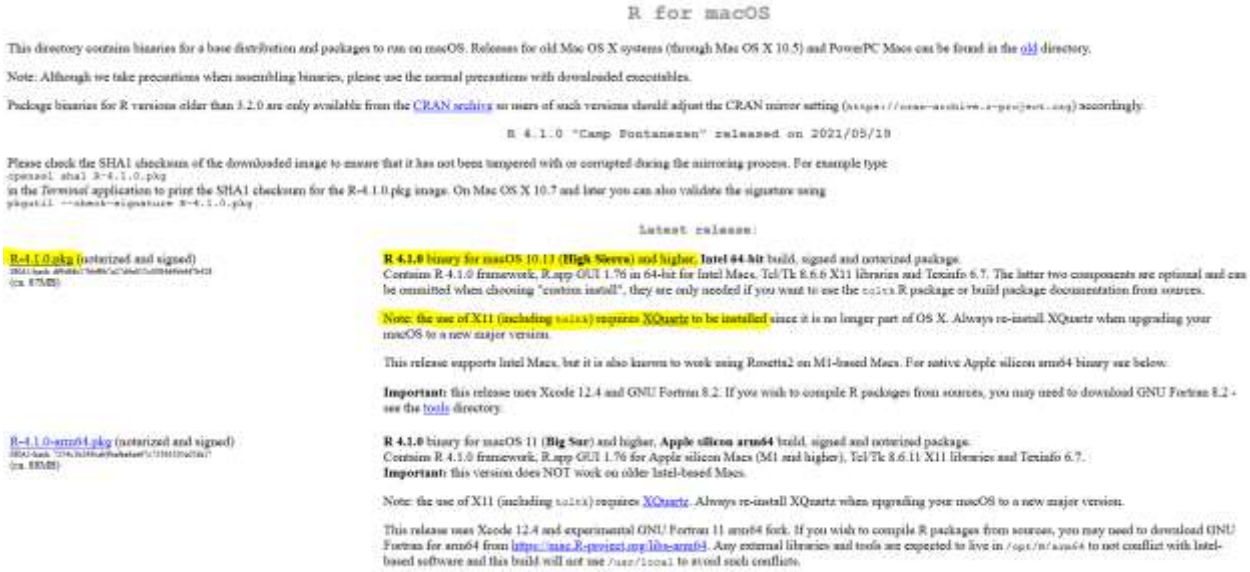
1. Visit https://www.google.com and search for “r cran”. The first item retrieved is the website to download R.



2. Click “Windows” if you have a windows machine or “R for Mac OS X” if you have a Mac machine.
3. **In general, it is the best to install the most current version of R.**
    a. For Windows users, click “Download R 4.1.0 for Windows”.



b. For Mac users, click “R-4.1.0.pkg”. Make sure that you install XQuartz at https://www.xquartz.org/ as well. You could find it in “Applications→Utilities” after installation.



R for macOS

This directory contains binaries for a base distribution and packages to run on macOS. Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the old directory.

Note: Although we take precautions when assembling binaries, please use the normal precautions with downloaded executables.

Package binaries for R versions older than 3.2.0 are only available from the CRAN archive so users of such versions should adjust the CRAN mirror setting (https://cran-archive.r-project.org) accordingly.

R 4.1.0 "Camp Pontanezen" released on 2021/05/18

Please check the SHA1 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type
`openssl sha1 R-4.1.0.pkg`
in the *Terminal* application to print the SHA1 checksum for the R-4.1.0.pkg image. On Mac OS X 10.7 and later you can also validate the signature using
`pkgutil --check-signature R-4.1.0.pkg`

Latest release:

R-4.1.0.pkg (notarized and signed)
(ca. 87MB)

R 4.1.0 binary for macOS 10.13 (**High Sierra**) and higher, **Intel 64-bit** build, signed and notarized package.
Contains R 4.1.0 framework, R.app GUI 1.76 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 6.7. The latter two components are optional and can be omitted when choosing "custom install", they are only needed if you want to use the `tcltk` R package or build package documentation from sources.

Note: the use of X11 (including `tcltk`) requires XQuartz to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your macOS to a new major version.

This release supports Intel Macs, but it is also known to work using Rosetta2 on M1-based Macs. For native Apple silicon arm64 binary see below.

Important: this release uses Xcode 12.4 and GNU Fortran 8.2. If you wish to compile R packages from sources, you may need to download GNU Fortran 8.2 - see the tools directory.

R-4.1.0-arm64.pkg (notarized and signed)
(ca. 88MB)

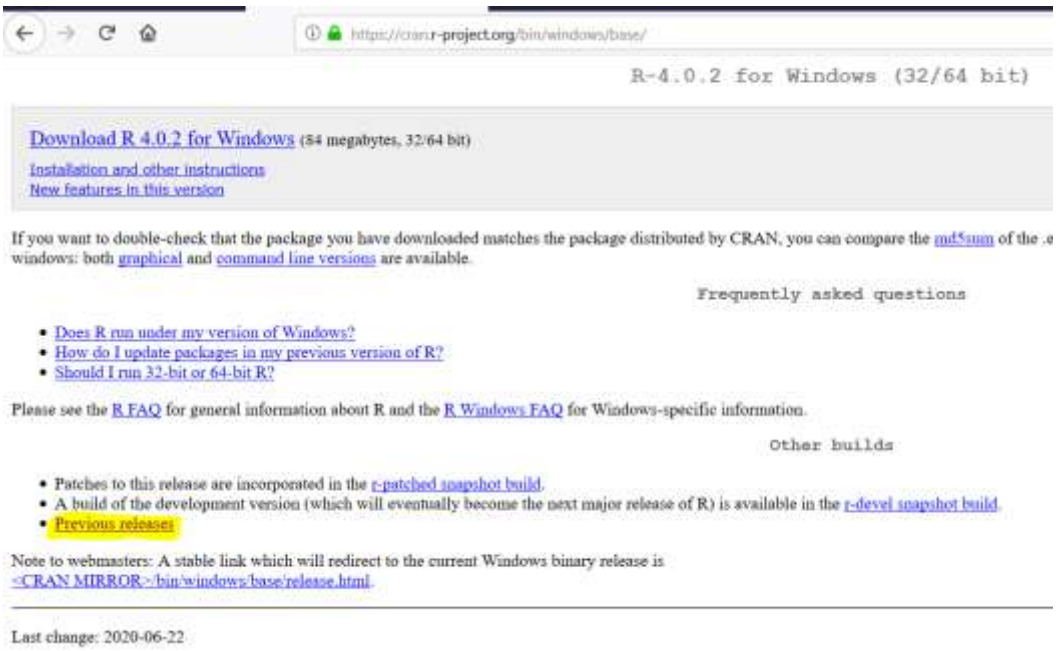
R 4.1.0 binary for macOS 11 (**Big Sur**) and higher, **Apple silicon arm64** build, signed and notarized package.
Contains R 4.1.0 framework, R.app GUI 1.76 for Apple silicon Macs (M1 and higher), Tcl/Tk 8.6.11 X11 libraries and Texinfo 6.7.
Important: this version does NOT work on older Intel-based Macs.

Note: the use of X11 (including `tcltk`) requires XQuartz. Always re-install XQuartz when upgrading your macOS to a new major version.

This release uses Xcode 12.4 and experimental GNU Fortran 11 arm64 fork. If you wish to compile R packages from sources, you may need to download GNU Fortran for arm64 from https://mac.R-project.org/libs-arm64. Any external libraries and tools are expected to live in `/opt/R/arm64` to not conflict with Intel-based software and this build will not use `/usr/local` to avoid such conflicts.

c. Please refer to Dr. John ‘s guidelines for trouble shooting at https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html.

4. If the most current release does not work well with the R commander package “Rcmdr” or the operating system of your machine does not support the most current release, you could install one of the earlier releases. For example, here are the steps to install previous release R 3.6.3 for Windows and R 3.3.3 for Mac instead.
   a. For Windows users, click “Previous releases” to get an earlier version of R. Choose “R 3.6.3 (February, 2020)”. And then click “Download R 3.6.3 for Windows”



https://cran.r-project.org/bin/windows/base/

R-4.0.2 for Windows (32/64 bit)

Download R 4.0.2 for Windows (84 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .e
windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.html.

Last change: 2020-06-22

https://cran.r-project.org/bin/windows/base/old/

Previous Releases of R for Windows

This directory contains previous binary releases of R for Windows.

The current release, and links to development snapshots, are available here. Source code for these releases and others is available through the main CRAN page.

In this directory:

R 4.0.2 (June, 2020)
R 4.0.1 (June, 2020)
R 4.0.0 (April, 2020)
R 3.6.3 (February, 2020)
R 3.6.2 (December, 2019)
R 3.6.1 (July, 2019)
R 3.6.0 (April, 2019)
R 3.5.3 (March, 2019)
R 3.5.2 (December, 2018)
R 3.5.1 (July, 2018)
R 3.5.0 (April, 2018)
R 3.4.4 (March, 2018)
R 3.4.3 (November, 2017)
R 3.4.2 (September, 2017)
R 3.4.1 (June, 2017)
R 3.4.0 (April, 2017)
R 3.3.3 (March, 2017)
R 3.3.2 (October, 2016)
R 3.3.1 (June, 2016)
R 3.3.0 (April, 2016)
R 3.2.5 (April, 2016)
R 3.2.4 (March, 2016)
R 3.2.3 (December, 2015)
R 3.2.2 (August, 2015)
R 3.2.1 (June, 2015)
R 3.2.0 (April, 2015)
R 3.1.3 (March, 2015)

https://cran.r-project.org/bin/windows/base/old/3.6.3/

R-3.6.3 for Windows (32/64 bit)

Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

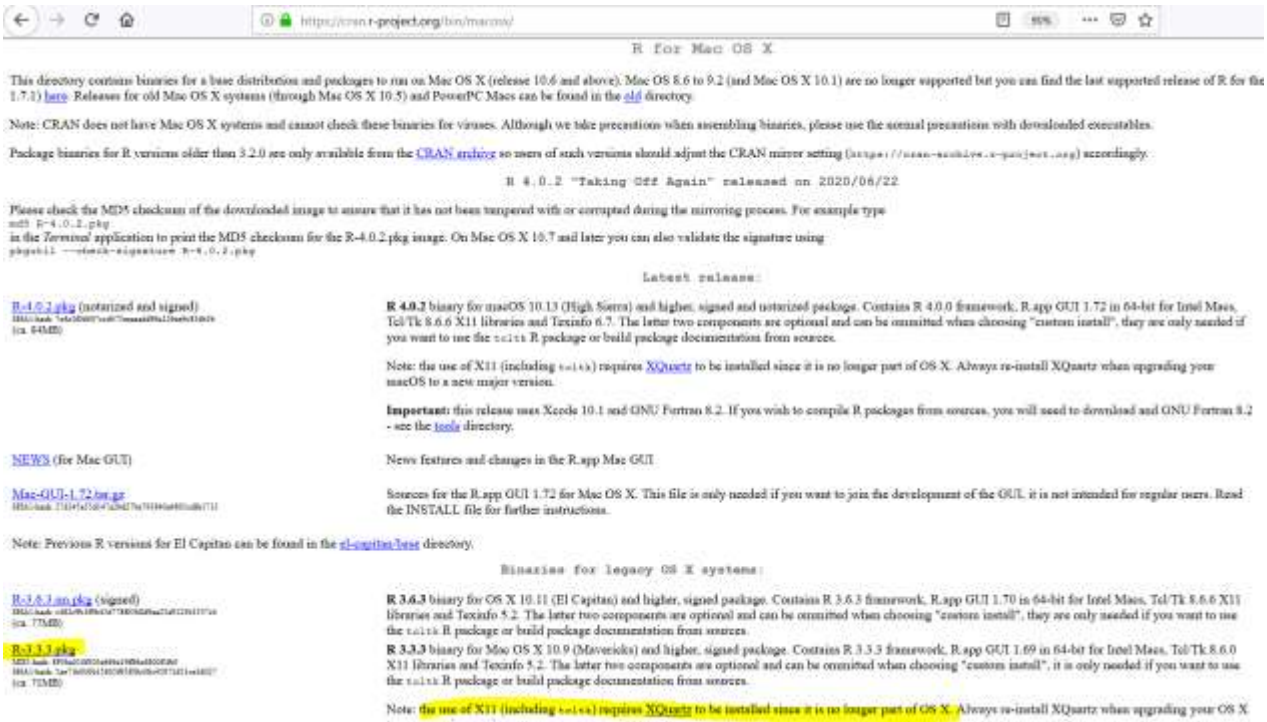
Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2020-02-29

b. For Mac users, click “R-3.3.3.pkg” to install version R 3.3.3. Make sure that you install XQuartz at https://www.xquartz.org/ as well.



R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS X 10.1) are no longer supported but you can find the last supported release of R for the 1.7.1) here. Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the old directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, please use the usual precautions with downloaded executables.

Package binaries for R versions older than 3.2.0 are only available from the CRAN archive so users of such versions should adjust the CRAN mirror setting (https://cran-archive.r-project.org) accordingly.

R 4.0.2 "Taking Off Again" released on 2020/06/22

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type
md5 R-4.0.2.pkg
in the *Terminal* application to print the MD5 checksum for the R-4.0.2.pkg image. On Mac OS X 10.7 and later you can also validate the signature using
pkgutil --check-signature R-4.0.2.pkg

Latest release:

R-4.0.2.pkg (notarized and signed) (ca. 84MB)

R 4.0.2 binary for macOS 10.13 (High Sierra) and higher, signed and notarized package. Contains R 4.0.0 framework, R.app GUI 1.72 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 6.7. The latter two components are optional and can be omitted when choosing "custom install", they are only needed if you want to use the tcltk R package or build package documentation from sources.

Note: the use of X11 (including tcltk) requires XQuartz to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your macOS to a new major version.

Important: this release uses Xcode 10.1 and GNU Fortran 8.2. If you wish to compile R packages from sources, you will need to download and GNU Fortran 8.2 - see the tools directory.

NEWS (for Mac GUI)

News features and changes in the R.app Mac GUI

Mac-GUI-1.72.tar.gz

Sources for the R.app GUI 1.72 for Mac OS X. This file is only needed if you want to join the development of the GUI, it is not intended for regular users. Read the INSTALL file for further instructions.

Note: Previous R versions for El Capitan can be found in the el-capitan/base directory.

Binaries for legacy OS X systems:

R-3.6.3.nn.pkg (signed) (ca. 77MB)

R 3.6.3 binary for OS X 10.11 (El Capitan) and higher, signed package. Contains R 3.6.3 framework, R.app GUI 1.70 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 5.2. The latter two components are optional and can be omitted when choosing "custom install", they are only needed if you want to use the tcltk R package or build package documentation from sources.

R-3.3.3.pkg (ca. 72MB)

R 3.3.3 binary for Mac OS X 10.9 (Mavericks) and higher, signed package. Contains R 3.3.3 framework, R.app GUI 1.69 in 64-bit for Intel Macs, Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. The latter two components are optional and can be omitted when choosing "custom install", it is only needed if you want to use the tcltk R package or build package documentation from sources.

Note: the use of X11 (including tcltk) requires XQuartz to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your OS X

1. 2 Install the R Commander Package

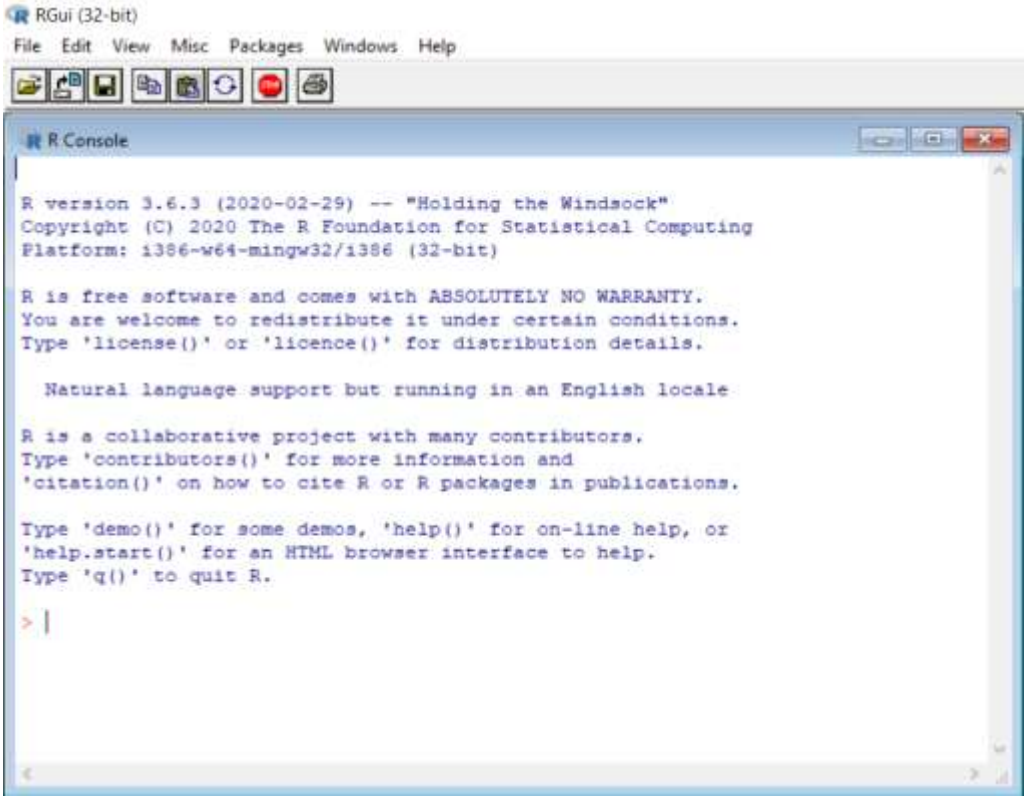
There are two ways to install the R Commander package.

The first way to install R Commander (an easier way):

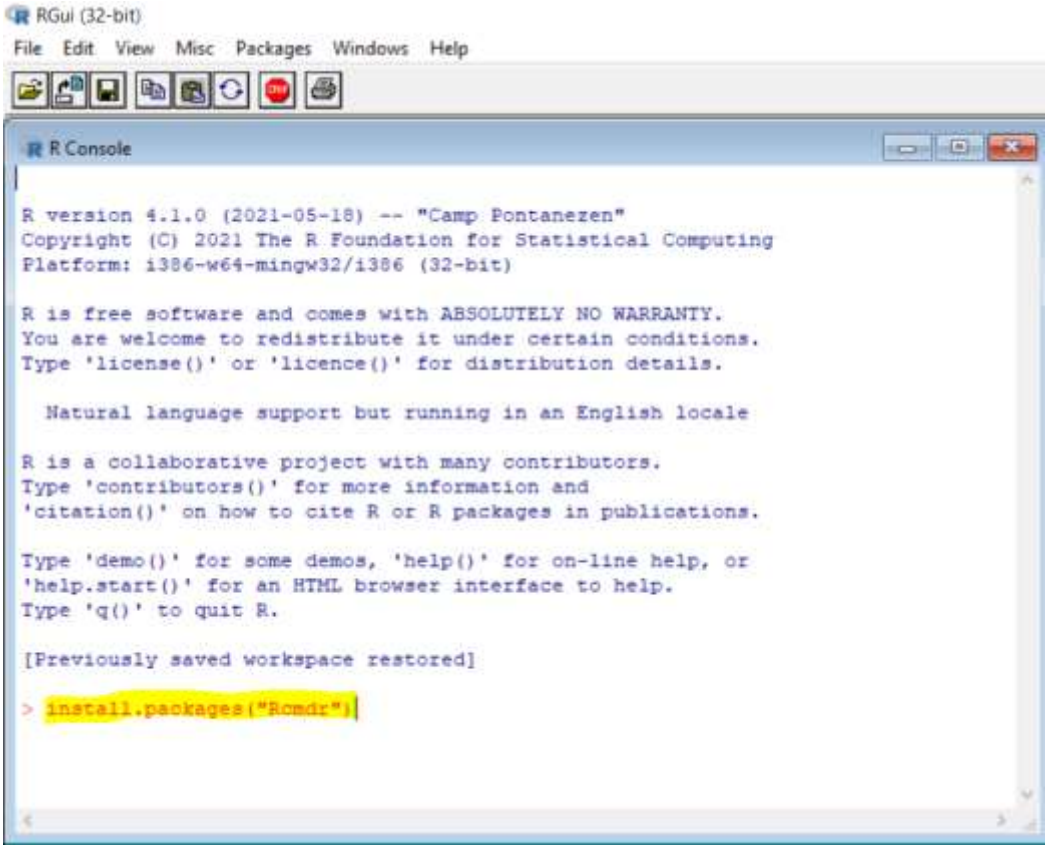
1. Once you have installed R, open it by double-clicking on the icon.

To launch R, double click the icon

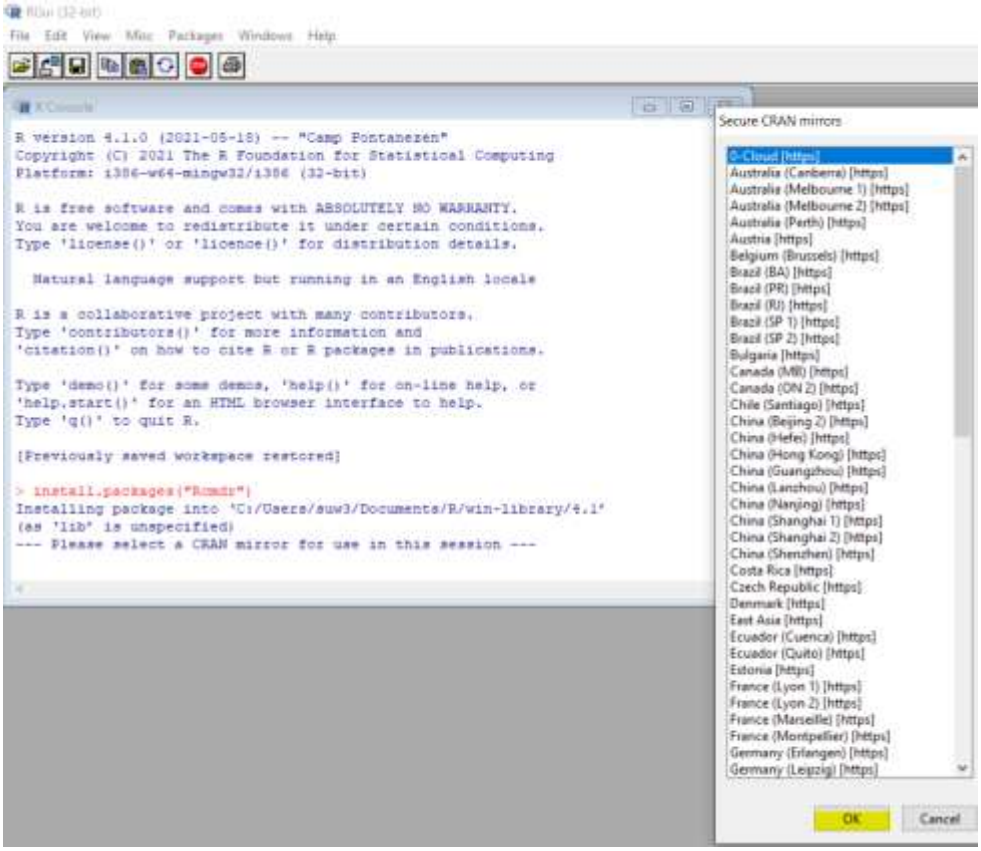
2. A window called “R Console” will open.



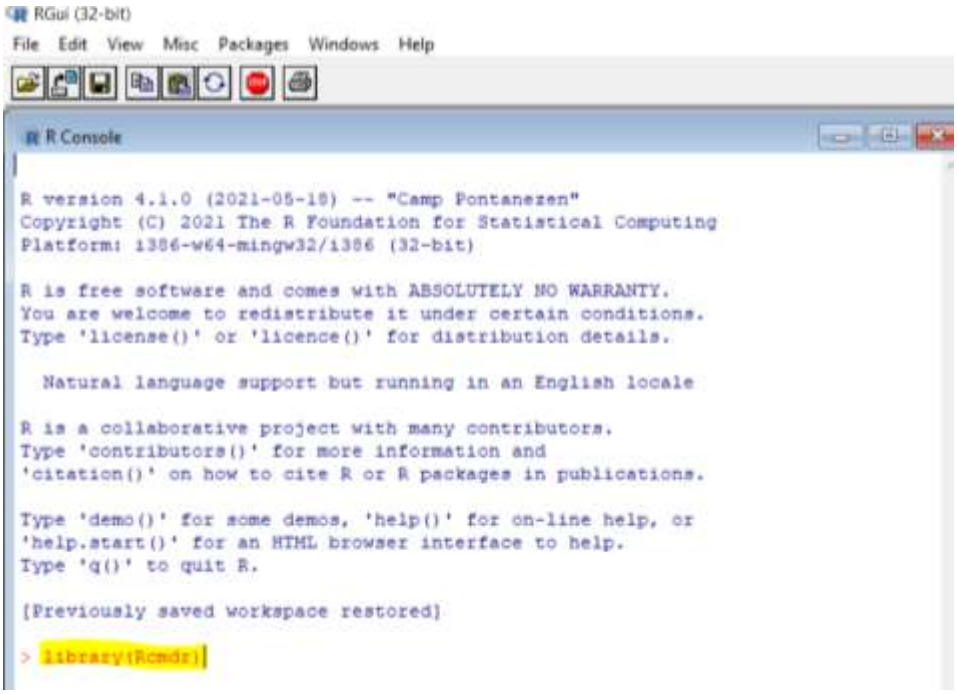
3. At the > command prompt, type the command **install.packages("Rcmdr")**, and click “enter”.



4. R will ask you to select a CRAN mirror; pick the first, "0-Cloud" mirror, or a mirror site near you.



5. Once the R commander package is installed, to load the **Rcmdr** package, just type the command library(Rcmdr) beside the > prompt and click “enter”. The name of the package is case sensitive.

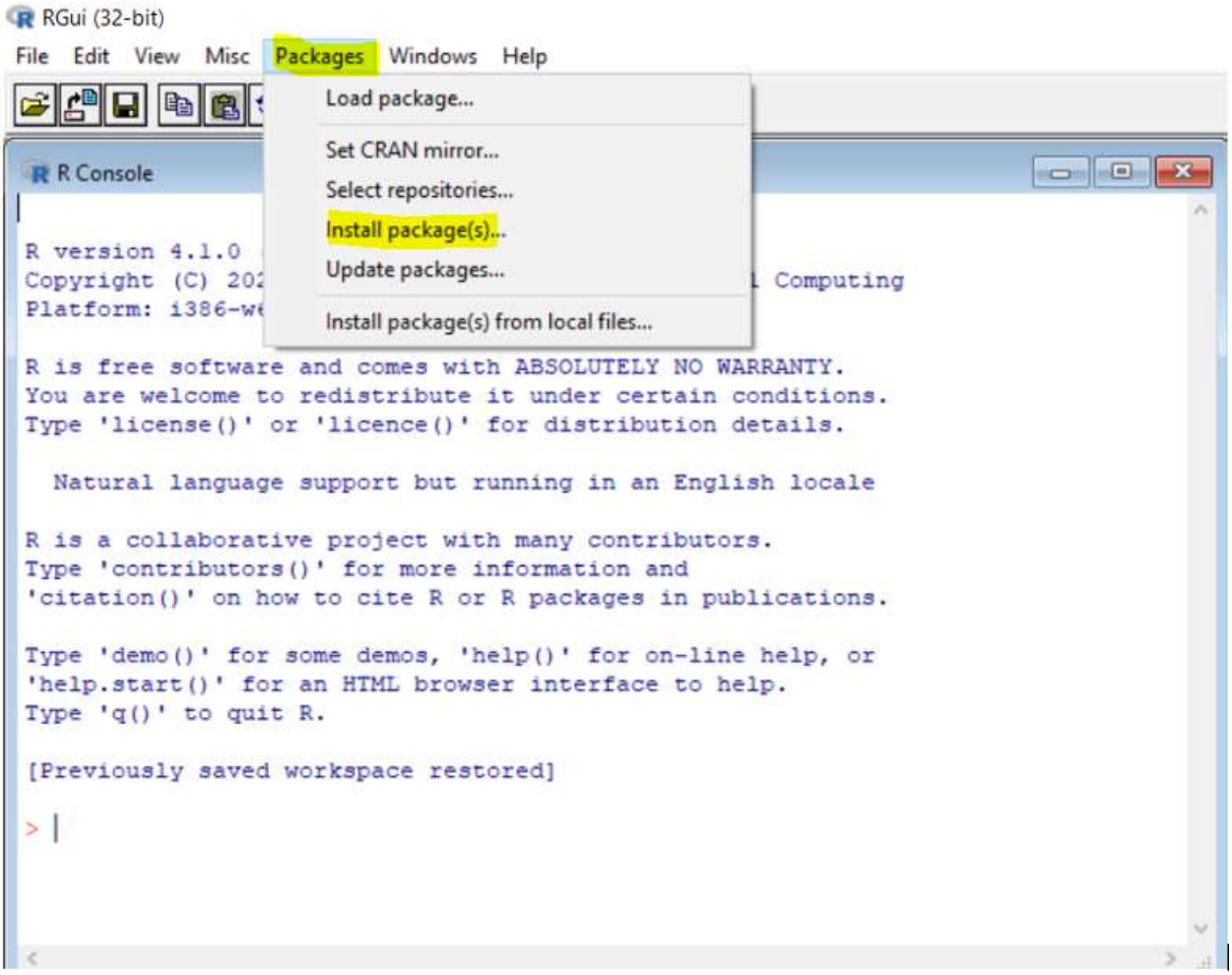


The second way to install the R commander package:

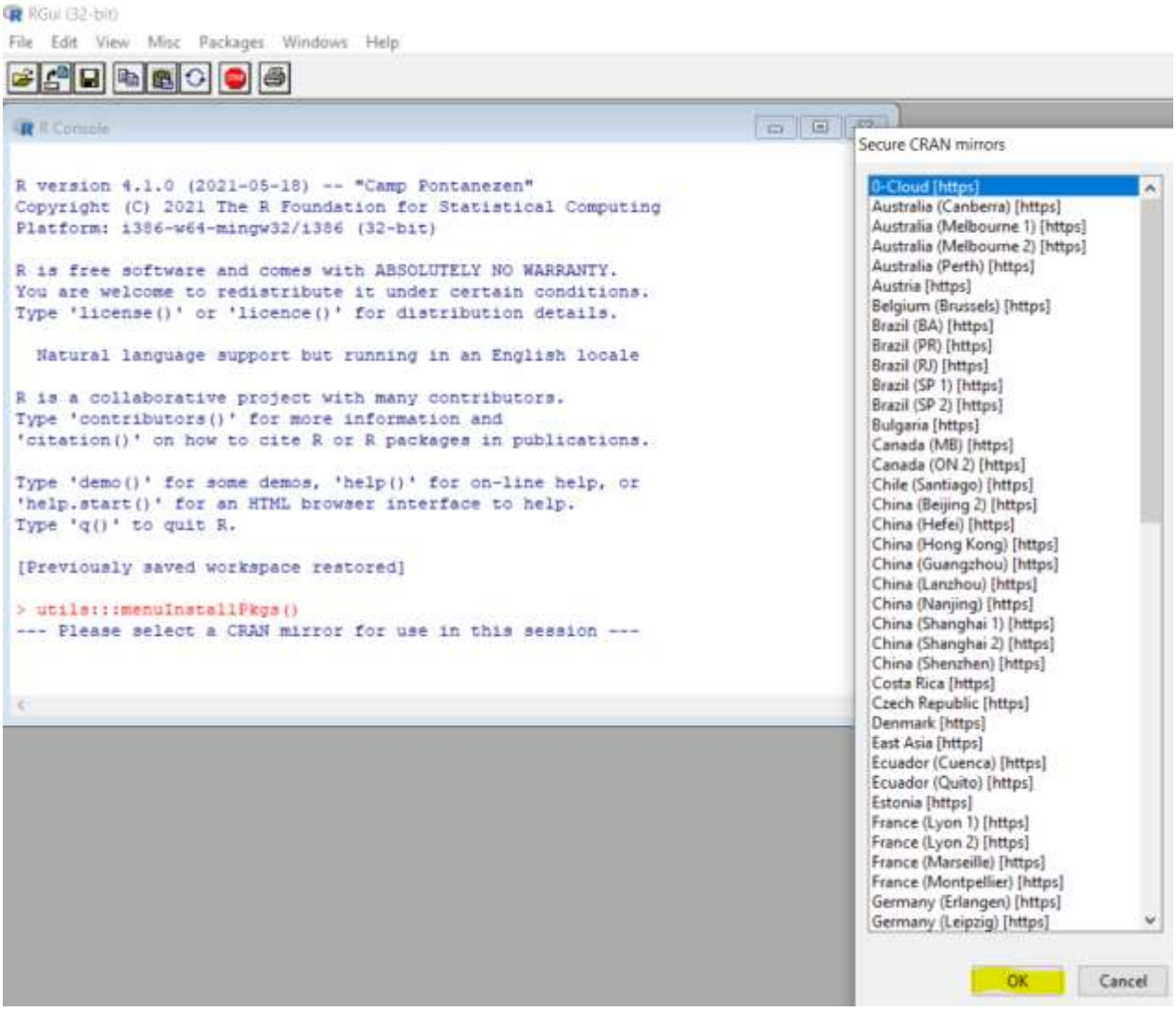
1. Once you have installed R, open it by double-clicking on the icon.

2. A window called “R Console” will open.

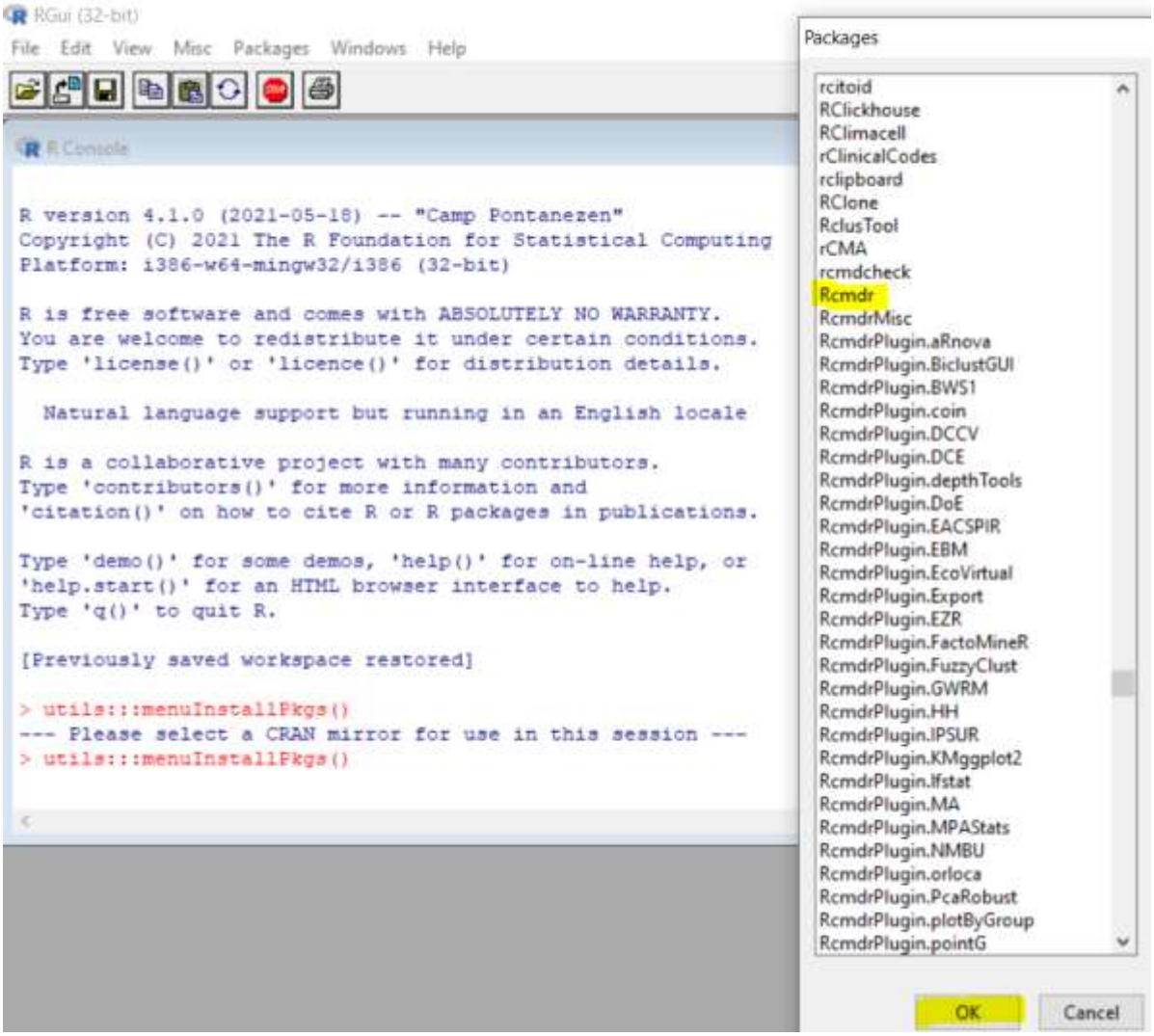
3. Click “Packages” on the menu bar, select “Install package(s)…” in the drop-down menu.



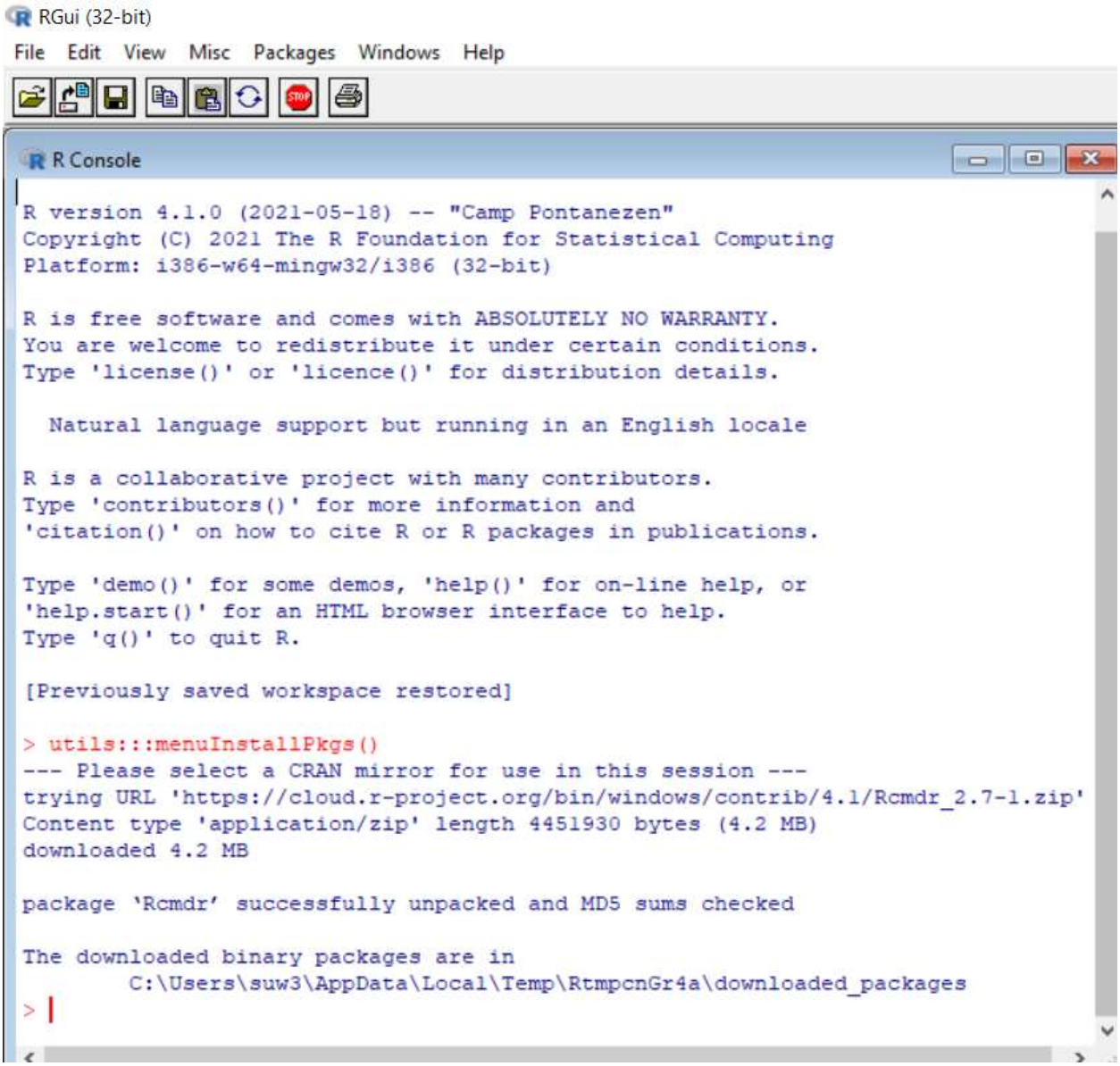
4. Click “OK” or select a location closest to you in the “HTTPS CRAN mirror” drop-down menu, and click “OK”.



5. Scroll down in the “Packages” drop-down menu, select the package “Rcmdr” and click “OK”.



6. Once the package is installed, the message “package ‘Rcmdr’ successfully unpacked and MD5 sums checked” should be shown in the R Console window.

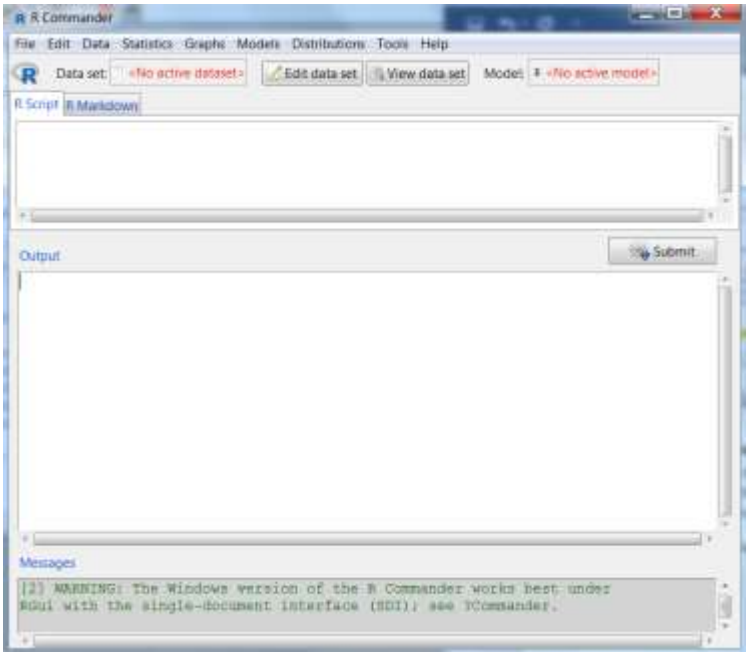


7. Once the R commander package is installed, to load the **Rcmdr** package, just type the command library(Rcmdr) and click "enter".

1.3 Starting R Commander

If R is not already open, open it by clicking on its icon. To open R Commander, at the > prompt type **library("Rcmdr")** and press Enter. If an error message says "lack of some packages, would you like to install those packages", click "Yes" and select "download from CRAN".

You should see a large new window pop up, labeled R Commander.



You are now ready to analyze your data with R Commander. If you close this window while R is still open, you can start R Commander again by entering the command "**Commander()**" in R Console. Entering "library(Rcmdr)" in this situation will not work unless you close R and open it again.

1.4 Trouble Shooting

One possible way to fix the problem is to copy the error message to Google and you might find a remedy. Here are some common problems when installing R commander, the "Rcmdr" package.

1. Error messages say something like "Warning in install.packages("Rcmdr") : 'lib = "C:/Program Files/R/R-3.6.3/library"' is not writable".
   a. Run R with Administrator privileges by right-clicking on the R shortcut and selecting 'Run as Administrator'.
   b. Double check whether you have any anti-virus program or security setting blocking installing software from so-called unknown developers. If yes, you might need to set your default secure cran mirror as trustable site.
2. Any error related to the **tcltk** package:
   a. You might have installed the most current version of R, but your system has not been updated. Try installing a previous version, say R 3.6.3 for Windows users and R 3.3.3 for Mac users.
   b. For Mac users, make sure that XQuartz has been installed.

3. Something like .zip file is not writable. Change the path before installing Rcmdr:

.libPaths("C:\\Program Files\\R\\R-3.6.3\\library")

4. Make sure that you run XQuartz before running R. Restart your computer if opening XQuartz behand does not work.

```
Attaching package: 'carData'

The following objects are masked from 'package:car':

    Guyer, UN, Vocab

lattice theme set by effectsTheme()
See ?effectsTheme for details.
xcode-select: note: no developer tools were found at '/Applications/Xcode.ap
requesting install. Choose an option in the dialog to download the command l
developer tools.
Error : .onAttach failed in attachNamespace() for 'Rcmdr', details:
  call: structure(.External(.C_dotTclObjv, objv), class = "tclObj")
  error: [tcl] invalid command name "image".

In addition: Warning messages:
1: running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
Resources/library/tcltk/libs//tcltk.so'' had status 1
2: In fun(libname, pkgname) : couldn't connect to display ":0"
Error: package or namespace load failed for 'Rcmdr'
> |
```

LAB 2 FIRST TASTE OF R AND R COMMANDER

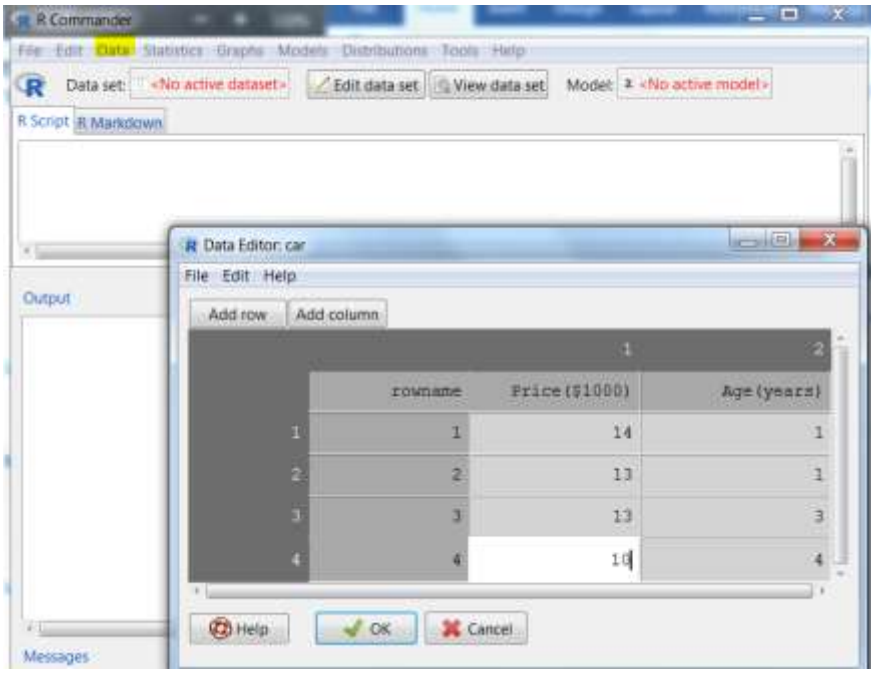
This lab introduces how to enter data into R and explore the data using figures and numerical summaries.

2.1 DATA ENTRY

There are several ways to enter data into R: manually enter, import from an existing data file, export from a built-in R package.

2.1.1 Manually Enter

1. Start a new data set through **Data →New data set…**
2. Enter a new name for the data set, say "usedcar" → OK
   Note: the name cannot have space and special symbols such as $
   Note: R is case-sensitive hence usedcar≠Usedcar
3. A data editor window where you can type in your data using a typical spreadsheet format. You can type rowname (say car), variable names (say price and age). Each row corresponds to one independent observation. For example, the spreadsheet below shows the price (in $1000) and the age (in year) of four used cars. The first car is 1 year old and its price is 14 ($1000).
4. Press Enter or click "Add row" if you need more rows.
5. Click "Add column" if you need more variables.
6. Click "OK".



2.1.2 Import From an Existing Data File

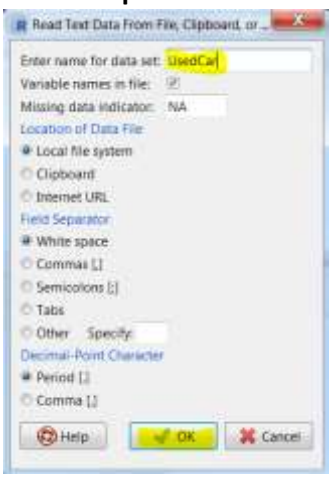
The existing data can be SPSS, Minitab, text, excel, SAS, and STATA data sets. We demonstrate with text, SPSS and Excel files. Data files used in this manual will be available in Blackboard (or another location specified by your instructor) and students can download them there.

Import from a text file

The data file needs to be organized as a classic data frame. Each column represents a single variable, e.g. price. Each row represents one individual. Header information needs to be contained to a single row.

For this example, please download the file called car.txt from online.

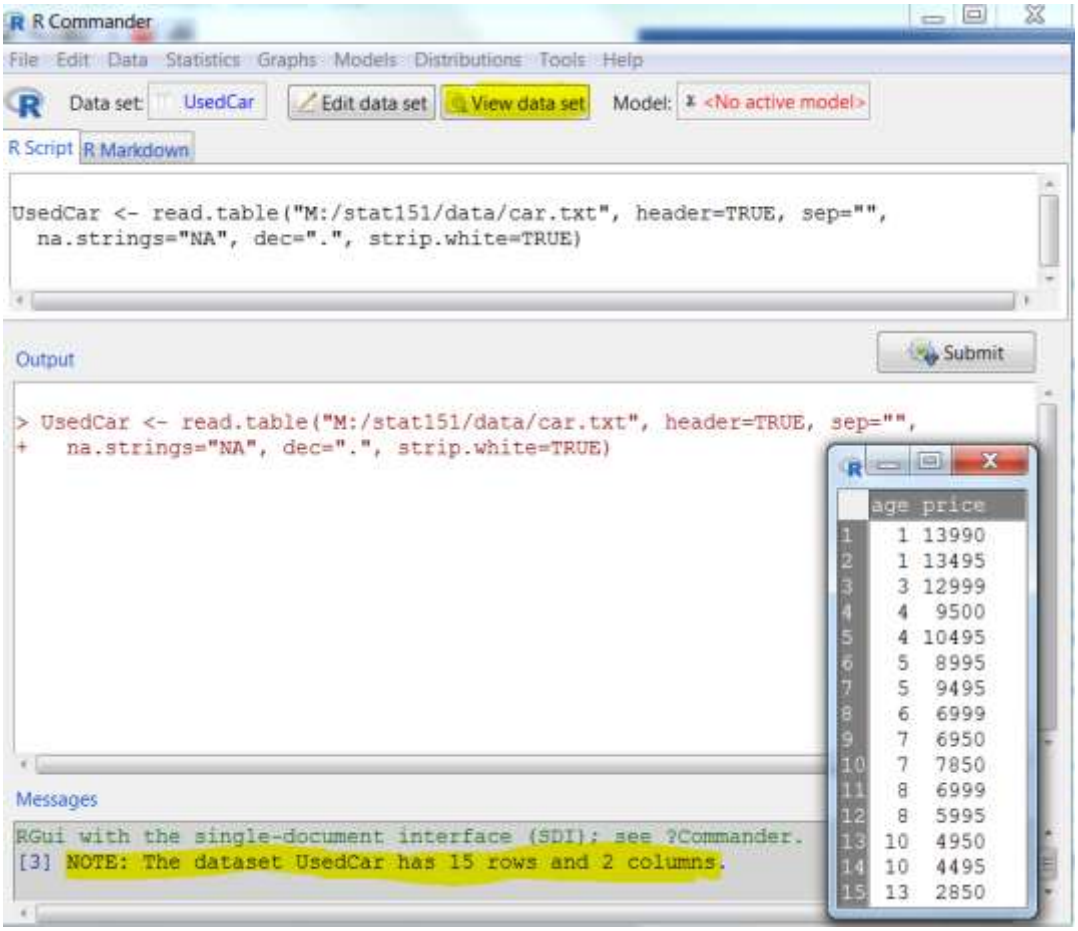
1. **Date→Import data→from text file, clipboard or URL…**



2. Enter the name (say car) for the data set and click “OK”.

3. Follow the path to where you stored the text file named car.txt is stored, and click “open”.

4. The imported data set “car” is now an active data set. Click “View data set” to view data.



Note: R commander was developed as an easy to use graphical user interface (GUI) for R language. The task can be also carried out by typing the commands directly in the R Console window. The corresponding commands are shown in the R Script sub-window. And the corresponding computer output is shown in the Output sub-window. In the Messages sub-window, it tells us that the data set has 15 rows and 2 columns.

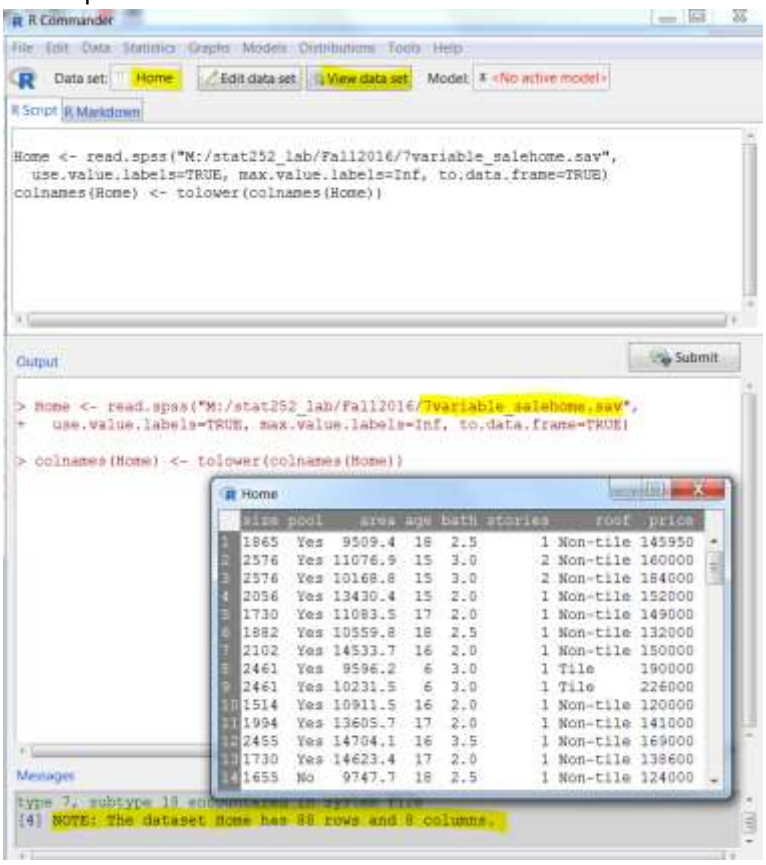
Import from an SPSS file

For this example, please download the file called 8variable_salehome.sav from online. This data set gives the price of 88 sale homes and had columns that detail eight features of the homes.

1. **Date→Import data→from SPSS data set…**

2. Enter the name (say Home) that you want to call the data set and click “OK”.
3. Go to the path where the sav file 8variable_salehome.sav is stored and click “open”.

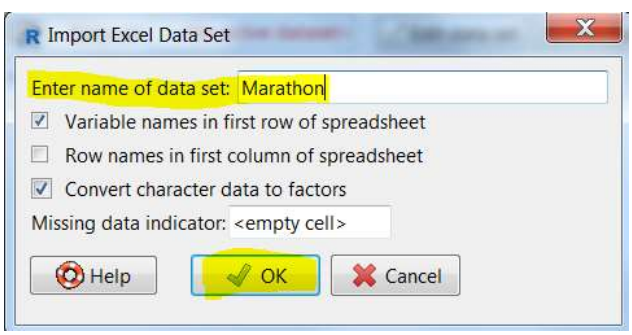
4. The imported data set “Home” is now an active data set. Click “View data set” to view the data.



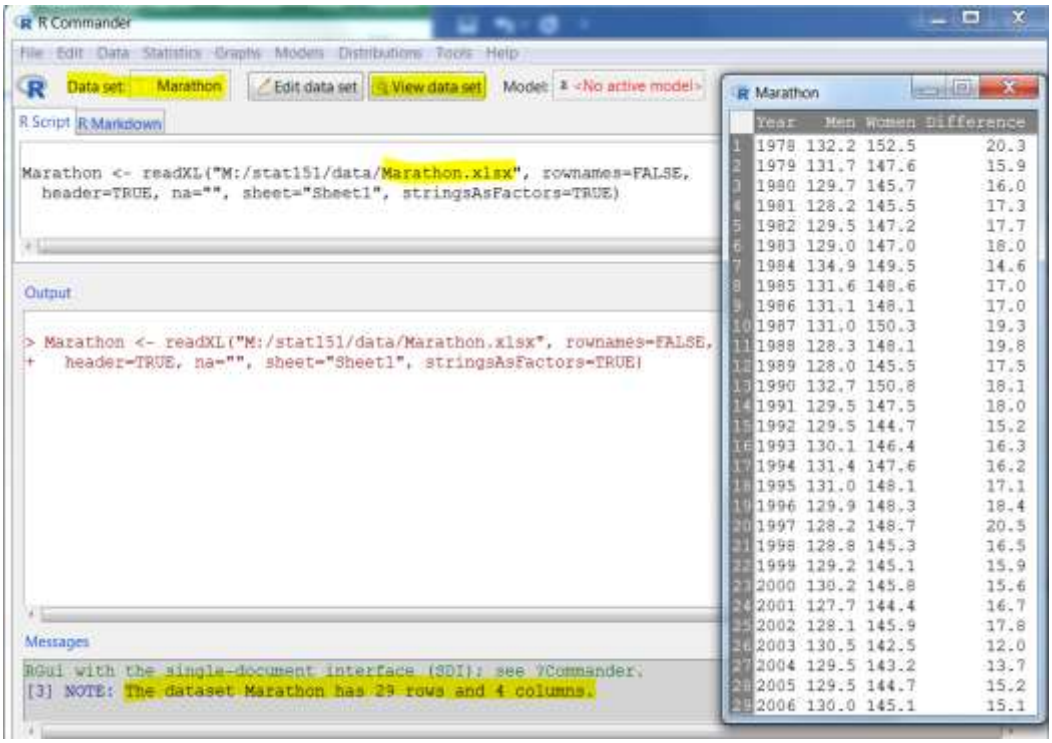
Import from an Excel file

For this example, please download the file called marathon.xlsx from online.

1. **Data→Import data→from Excel data set…**



2. Enter the name (say Marathon) for the data set and click “OK”.
3. Go to the path where the Excel data file is stored and select the file ***marathon.xlsx*** and click “open”.
4. The imported data set “Marathon” is now an active data set. Click “View data set” to view the data.



This data set gives the winning times (in minutes) for men and women in the New York City Marathon between 1978 and 2006 (www.nycmarathon.org). The last column gives the difference in winning time between female and male.

2.2 Explore Data Using R Commander

Basically, there are two types of statistics: descriptive statistics and inferential statistics.

- **Descriptive statistics** consists of numerical and graphical methods for organizing and summarizing the sampled data. One only focuses on the sampled data.
- **Inferential statistics** consists of methods for drawing conclusions about the population based on information obtained from the sampled data. It uses the sampled data to make estimates, decisions, predictions, or other generalizations about the population. For inferential study, look for the key words "estimate for all" or "prediction for all".

This lab session focuses on how to obtain descriptive statistics using R commander. Statistics is all about data. And data are information about a group of individuals organized in variables. There are two types of variables: qualitative/categorical and quantitative variables. The quantitative variable can be further classified as either continuous or discrete.

- **Qualitative variable**: A non-numerically valued variable that classifies subjects into different categories, such as "Name" and "Sex". The values of qualitative variables are not numbers. A qualitative variable is also called a **categorical variable**.
- **Quantitative variable**: A numerically valued variable (e.g., "Number of hours/day on internet"). There are two types of quantitative variable --- continuous and discrete.
    - **Continuous variable**: A quantitative variable whose possible values form some interval of numbers (e.g., height, length of feet, salary, age). Technically speaking, continuous

variables have an arbitrary number of decimal places. For any two possible values, we can always find some value in between.

- **Discrete variable**: A quantitative variable whose possible values can be listed (e.g., number of siblings, number of phone calls within an hour.)

The following table summarizes the descriptive methods for some standard statistical tasks.

	Descriptive Statistics	
Task	Graphical	Numerical
Display one qualitative variable	pie chart bar chart	frequency table relative frequency table
Display two qualitative variables	side-by-side pie or bar chart	contingency table
Display one quantitative variable	histogram dot plot boxplot stem-leaf plot	5-number summary (mean, SD) (median, IQR)
Display two quantitative variables	scatter plot	correlation coefficient r and r^2 (covered in Chapter 14)
Display one qualitative and one quantitative variables	side-by-side histogram boxplot stem-leaf plot	5-number summary (mean, SD) (median, IQR) by groups

The 8variable_salehome.xlsx price dataset that you can find and download from online will be used as a demo in this section. There are eight variables of different data types. Size, area, age, and price can be treated as quantitative continuous; bath (# of bathrooms) and stories (# of stories) can be treated as quantitative discrete, and pool and roof are qualitative (categorical). We first import the data set into R commander.

1. **Data→Import data→from Excel data set…**
2. Enter the name (say Home) for the data set and click "OK".
3. Go to the path where the Excel data file is stored and select the file ***8variable_salehome.xlsx*** and click "open".
4. The imported data set "8variable_salehome.xlsx" is now an active data set named Home in R. Click "View data set" to view the data.

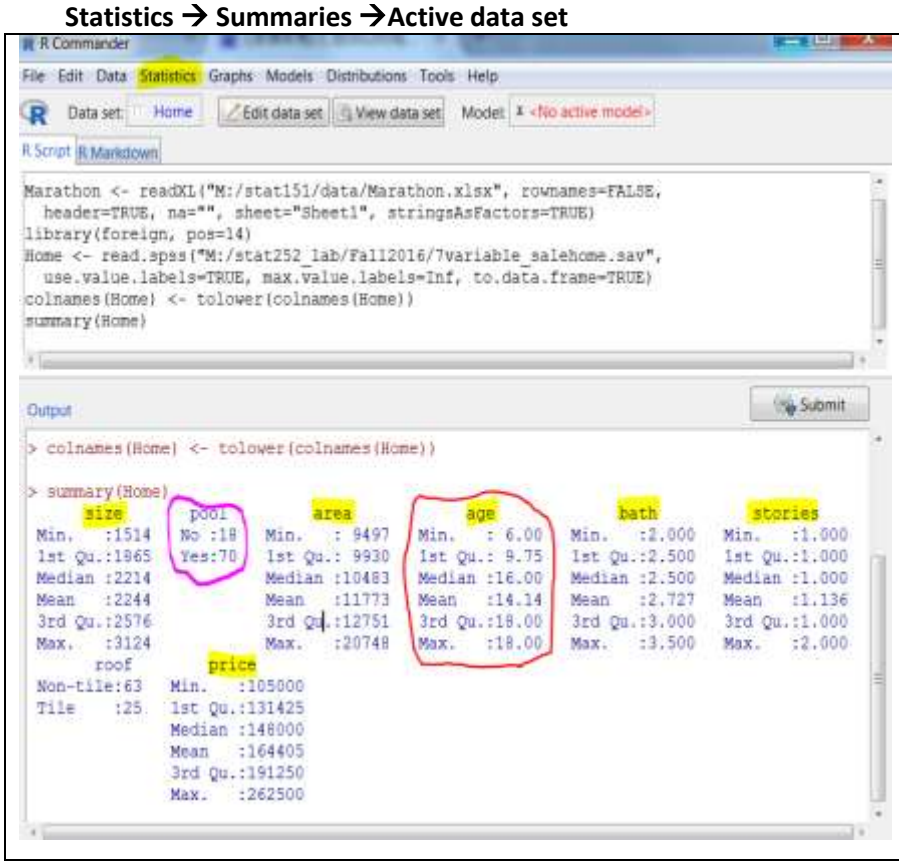
Home

	size	pool	area	age	bath	stories	roof	price
1	1865	Yes	9509.4	18	2.5	1	Non-tile	145950
2	2576	Yes	11076.9	15	3.0	2	Non-tile	160000
3	2576	Yes	10168.8	15	3.0	2	Non-tile	184000
4	2056	Yes	13430.4	15	2.0	1	Non-tile	152000
5	1730	Yes	11083.5	17	2.0	1	Non-tile	149000
6	1882	Yes	10559.8	18	2.5	1	Non-tile	132000
7	2102	Yes	14533.7	16	2.0	1	Non-tile	150000
8	2461	Yes	9596.2	6	3.0	1	Tile	190000
9	2461	Yes	10231.5	6	3.0	1	Tile	226000
10	1514	Yes	10911.5	16	2.0	1	Non-tile	120000
11	1994	Yes	13605.7	17	2.0	1	Non-tile	141000
12	2455	Yes	14704.1	16	3.5	1	Non-tile	169000
13	1730	Yes	14623.4	17	2.0	1	Non-tile	138600
14	1655	No	9747.7	18	2.5	1	Non-tile	124000
15	1865	Yes	9932.9	18	2.5	1	Non-tile	130000
16	1882	Yes	10274.4	18	2.5	1	Tile	150000
17	2718	Yes	9675.3	6	3.5	1	Tile	243000
18	1882	Yes	11825.1	18	2.5	1	Non-tile	137900
19	1882	No	14831.5	18	2.5	1	Non-tile	111500
20	1994	Yes	16122.5	17	2.0	1	Non-tile	152000
21	2214	Yes	12358.3	18	2.5	1	Non-tile	147000
22	2718	Yes	16214.1	6	3.5	1	Tile	245000
23	2576	Yes	12055.5	15	3.0	2	Non-tile	175000
24	3124	No	9497.6	6	3.5	1	Tile	242500
25	2128	Yes	9823.7	15	2.5	1	Non-tile	152000
26	1655	Yes	10520.5	18	2.5	1	Non-tile	137000
27	2214	No	10739.0	18	2.5	1	Non-tile	148000
28	2576	Yes	11087.7	15	3.0	2	Non-tile	175000
29	2928	Yes	16458.6	10	3.5	1	Tile	210000
30	2576	Yes	10368.5	15	3.0	2	Non-tile	169900

2. 2.1 Obtain Numerical Summaries

We can obtain the numerical summaries for each variable of the active data set:

Statistics → Summaries →Active data set

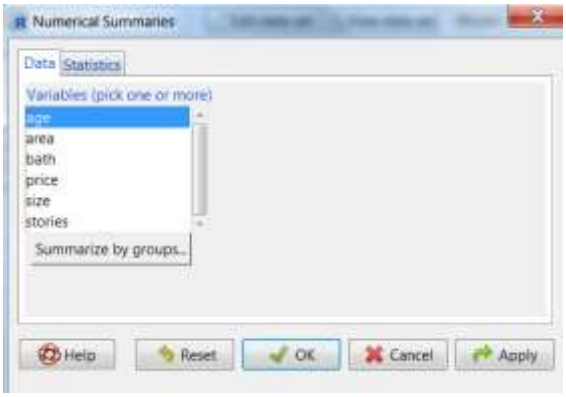


1. For quantitative variables, it gives the mean and five number summaries, i.e., minimum, 1st quartile, median (2nd quartile), 3rd, and maximum. Take age for example: the average age of those 88 sale homes is 14.14 years with a median 16 years. The newest 25% of homes are between 6 to 9.75 years old; another 25% are between 9.75 and 16; another 25% are between 16 and 18; the oldest 25% are 18 years old.

2. For qualitative (categorical) variables, it gives the frequencies (number of times) for which values occur. Take Pool for example: 18 out of 88 homes do not have a swimming pool and 70 have.

We can obtain the numerical summaries for a single **quantitative** variable.

1. **Statistics → Summaries →Numerical summaries…**
   Note: numerical summaries are only for quantitative variables. For categorical variables, we use frequency distributions to summarize counts of the variable values (see below).
2. Select the variable of interest, say age, from the list and click OK.



Output:

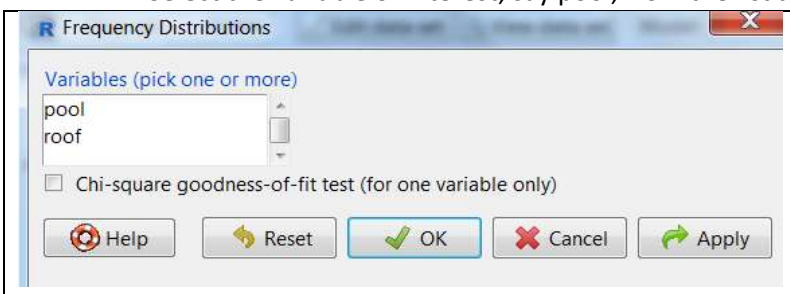
```
    mean       sd  IQR 0%  25% 50% 75% 100%  n
14.13636 4.823748 8.25  6 9.75  16  18   18 88
```

Understand the output:

mean	Sample mean, measure of central tendency
sd	Sample standard deviation, measure of spread (variation)
IQR	Inter-quartile range=3rd quartile-1st quartile, the middle 50% of the observations are within IQR
0%	Minimum value, 0th percentile
25%	1st quartile. The value below which 25 percent of the observations may be found.
50%	2nd quartile, the median. The value below which 50 percent of the observations may be found.
75%	3rd quartile. The value below which 75 percent of the observations may be found.
100%	Maximum value
n	Sample size, number of individuals in the sample

We can obtain the numerical summaries for a single **qualitative (categorical)** variable.

1. **Statistics → Summaries → Frequency distributions…**
2. Select the variable of interest, say pool, from the list and click OK.



Output:

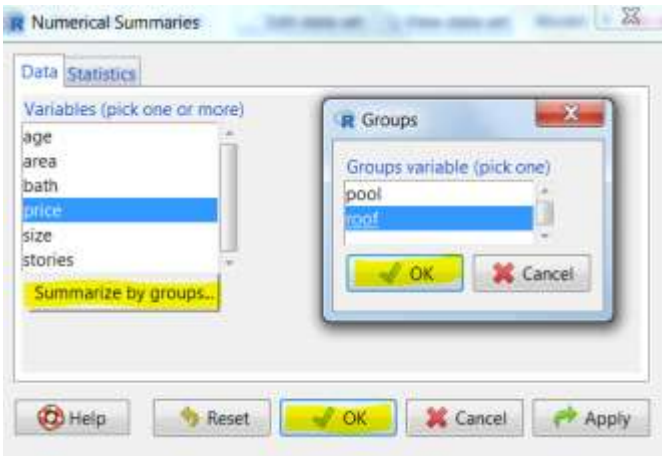
```
counts:
pool
No  Yes
 18  70

percentages:
pool
  No    Yes
20.45 79.55
```

1. Counts are the frequencies.
2. Percentages are the relative frequencies multiplied by 100 $=\frac{counts}{n}\times 100$.

We can obtain the numerical summaries of a single quantitative variable among different sub-groups.

1. **Statistics → Summaries → Numerical summaries…**
2. Select the variable of interest from the list, e.g., price
3. Click "**Summarize by groups…**"
4. In the pop-up window "Groups", select the categorical variable defining the sub-groups (say the roof type indicating the whether the home has a tile roof or non-tile roof) and click OK.
5. Click OK in the pop-up window Numerical Summaries.



Output:

```
             mean       sd   IQR     0%    25%    50%    75%   100% data:n
Non-tile 139225.7 20080.80 27500 105000 123500 137000 151000 185500     63
Tile     227856.0 29833.54 35000 150000 210000 237000 245000 262500     25
```

Interpretation of the computer output:

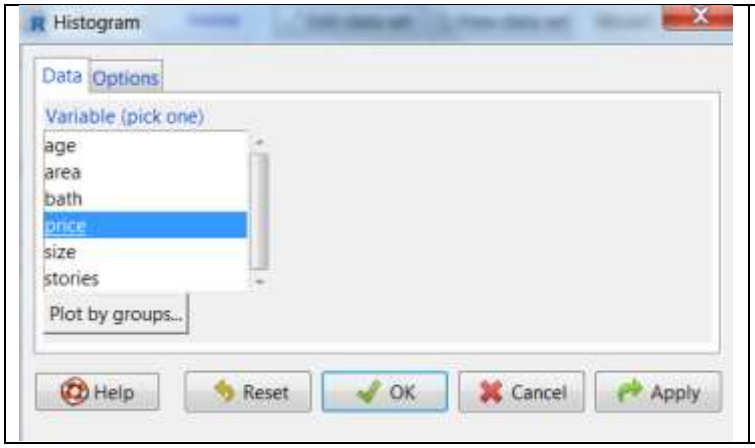
1. Out of those 88 sale homes, 63 homes have a non-tile roof and 25 have a tile roof.
2. The average price of homes with a tile roof is $227856.0 and the average price of homes with a non-tile roof is $139225.7, which means on average homes with a tile roof are more expensive than homes with a non-tile roof.
3. The price of homes with a tile roof has a larger variation than the price of homes with a non-tile roof, because it has a larger sample standard deviation ($29833.54 versus $20080.80) and a larger IQR ($35000 versus $27500).
4. The price of homes with a tile roof also has a larger minimum, quartiles, and maximum, respectively.

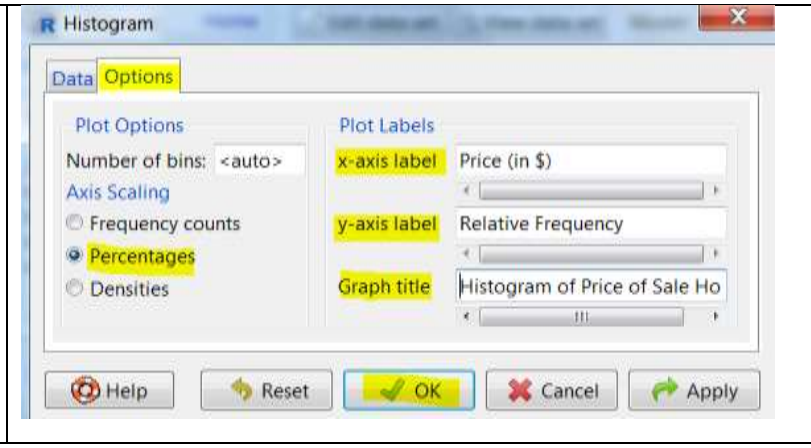
2.2.2 Obtain Graphs

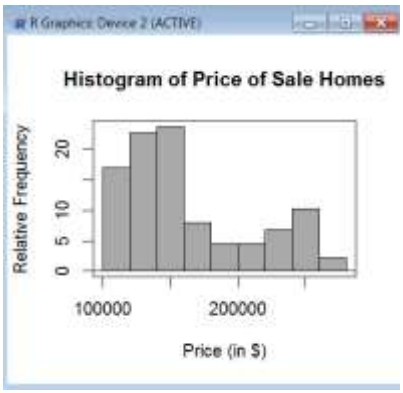
Almost all graphs can be found under **Graphs** in the menu bar. In general, the bar chart and pie chart are for qualitative (categorical) variables, while the histogram, boxplot, dot plot, and stem-and-leaf display are for quantitative variables. The scatter plot is for two quantitative variables. The quantile-comparison (QQ) plot is used to check whether the data follow a certain distribution. We can use it to check whether the data follow a normal distribution; this is called the normal probability plot in the textbook.

Histogram for a single quantitative variable:

1. **Graphs → Histogram…**
2. Select the variable of interest from the list, e.g., price
3. Click **Options** to specify the Axis scaling; use Frequency counts for frequency and Percentages for relative frequency. Specify the labels and the title of the histogram if you want.
4. Click OK







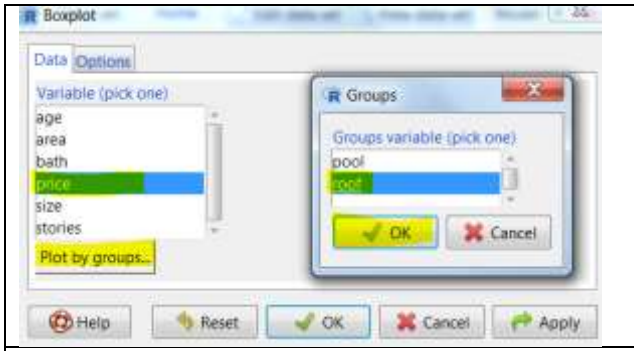
Side-by-side histogram to compare a single quantitative variable among different sub-groups

Graphs → Histogram…

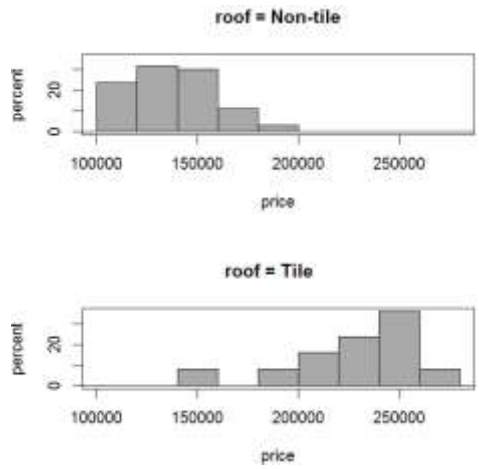
1. Select the variable of interest from the list, e.g., price
2. Click **Plot by groups…**, select the categorical variable defining the sub-groups (say roof), click OK
3. Click **Options** to specify the Axis scaling, making sure to use **Percentage** for a side-by-side plot
4. Click OK

We can also draw the side-by-side boxplots to compare the price of homes with a tile and non-tile roof (see the boxplots output below)

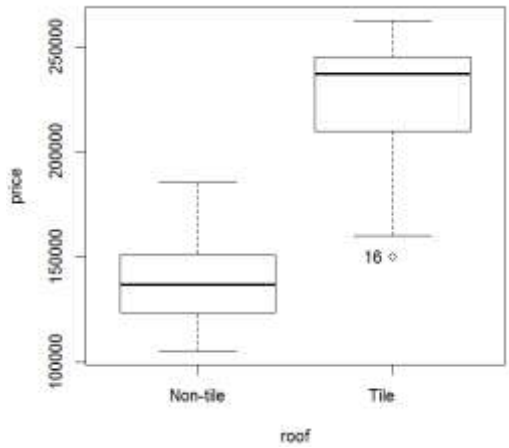
1. **Graphs → Boxplot…**
2. Select the variable of interest from the list, e.g., price
3. Click **Plot by groups…**, select the categorical variable defining the sub-groups (say roof), click OK
4. Click OK



Side-by-side Histogram of Price of Sale Homes

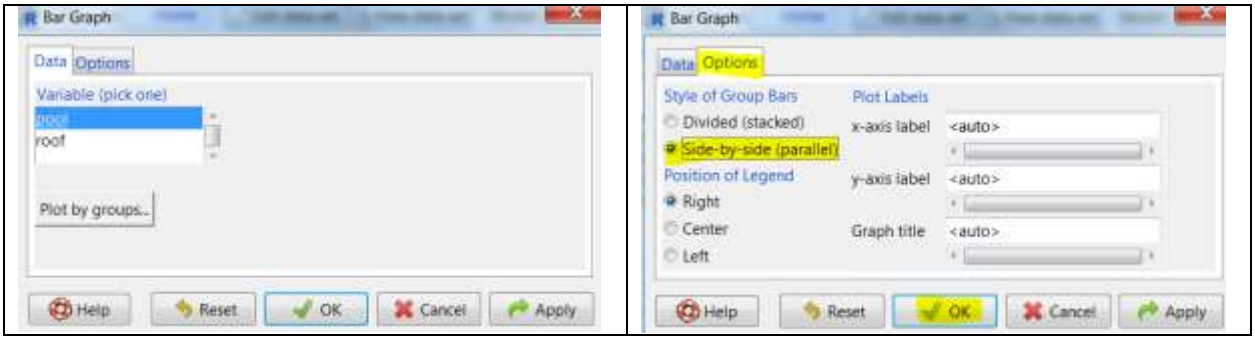


Side-by-side Boxplot of Price of Sale Homes



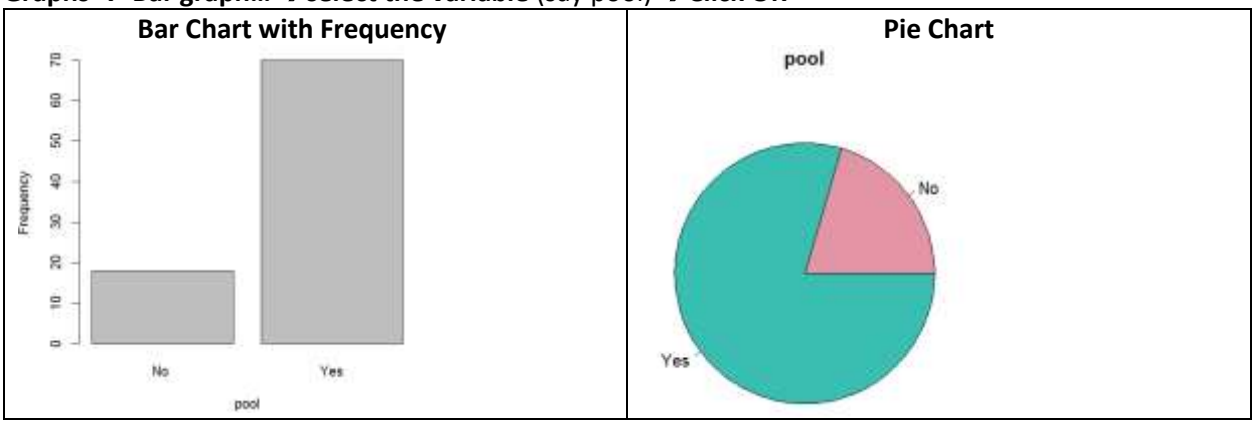
Bar Chart for a single qualitative (categorical) variable)

1. **Graphs → Bar graph…**
2. Select the variable of interest from the list, e.g., pool (whether the home has a swimming pool)
3. Click **Options** to specify style of the bars. Click OK
4. Click OK



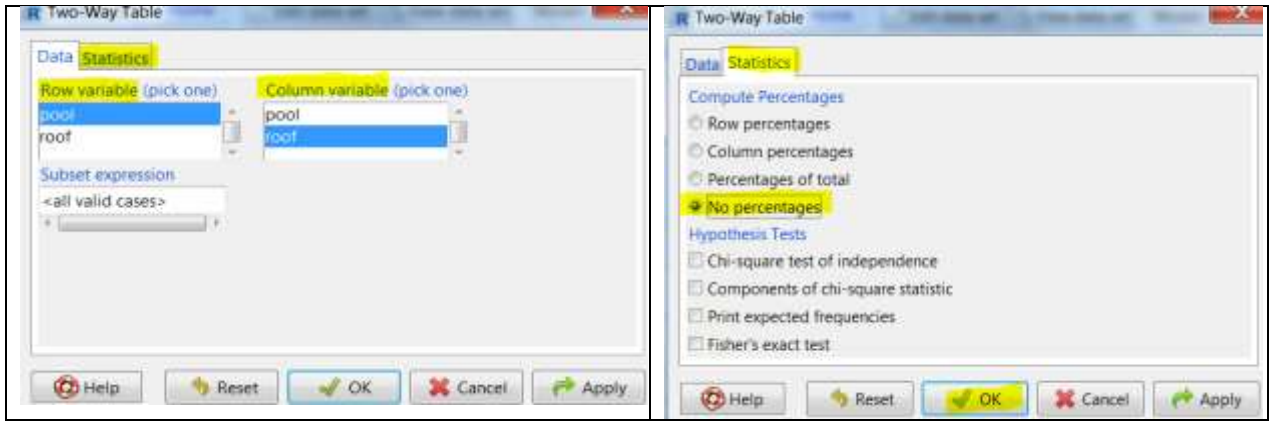
Pie Chart for a single qualitative (categorical) variable

Graphs → Bar graph… →select the variable (say pool) **→Click OK**



Contingency (two-way) table for two categorical variables

1. **Statistics→Contingency table→Two-way table…**
2. Specify the row variable and column variable (say pool and roof, respectively)
3. Click **Statistics→No percentage** (only gives the counts in each cell)
4. Click OK



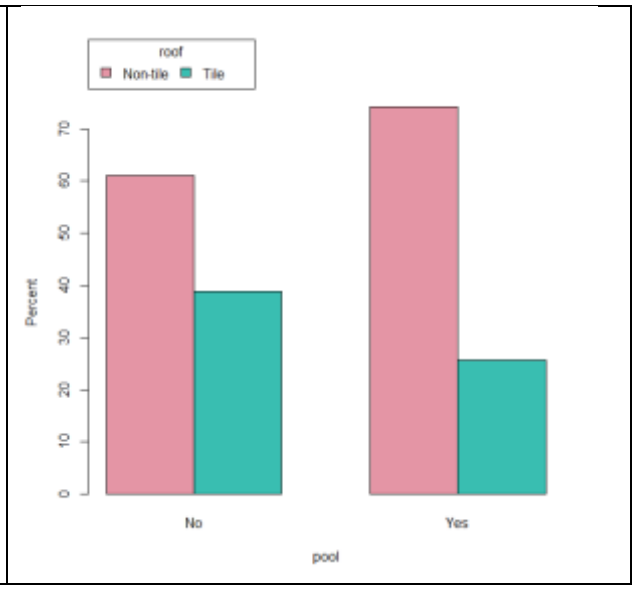
Output:	
Frequency table: roof pool Non-tile Tile No 11 7 Yes 52 18	1. The row variable is pool, and the column variable is roof 2. 11 out of 88 homes do not have a swimming pool and have a non-tile roof; 7 have no pool but a tile roof; 52 have a pool and non-tile roof; and 18 homes have a pool and a tile roof.

Side-by-side bar charts (conditional on sample size of sub-groups) for two categorical variables

Based on the contingency table, we can draw a side-by-side bar chart to check whether those homes with a swimming pool and without a swimming pool share the same pattern regarding to the roof type.

1. **Graphs → Bar graph…**
2. Select the variable for X-axis, e.g., pool
3. Click **Plot by groups**, select the variable whose pattern is of interest (say roof type here). Click ok.
4. Click **Options** to specify style of the group bars. Under **Axis Scaling**, choose **Percentages**. Under **Style of Group Bars**, choose **Side-by-side (parallel)**. Under **Percentages for the Group Bars**, choose **Conditional.** This will account for the sample size in each sub-group, and the provided percentage in each cluster of bars will be the percentage from each subgroup. Click OK.
5. Click OK

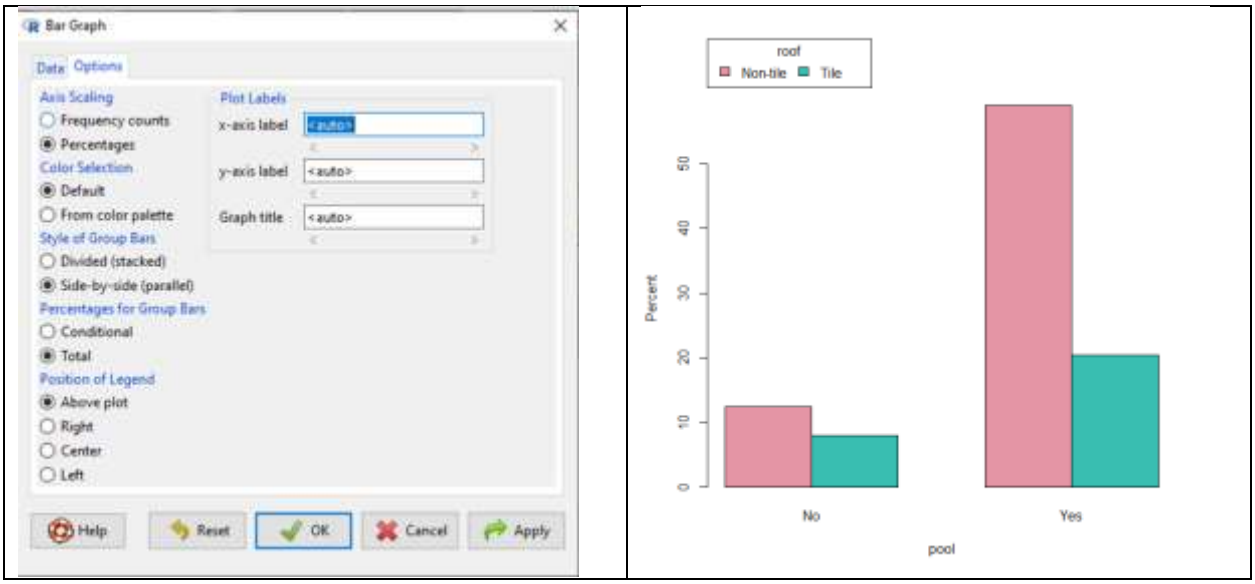
   Below, you can see:
   1) the bar for no pool and non-tile roof is at a height of 11/18 = 61.1% and the bar for no pool and no-tile is at 7/18 = 38.9%. Percentages add to 100% for the no pool group.
   2) the bar for yes pool and non-tile roof is at a height of 52/70 = 74.3% and the bar for yes pool and tile roof is at 18/70 = 25.7%. Percentages add to 100% for the yes pool group.



Side-by-side bar charts (using overall sample sizes) for two categorical variables

If at step 4 above, you had chosen **Total** under **"Percentages for Group Bars",** the bars did not consider the sample size of each subgroup, you would divide the total of each bar by the overall total number of observations in the dataset. This is not useful or desirable when samples sizes are different, but the example is included here so you can see what happens.

Here you can see:
1) the bar for no pool and non-tile roof is at a height of 11/88 = 12.5% and the bar for no pool and no-tile is at 7/88 = 8.0%.
2) the bar for yes pool and non-tile roof is at a height of 52/88 = 59.0% and the bar for yes pool and tile roof is at 18/88 =20.5 %.
3)The total of all the percentages over all the four bars is 100%.



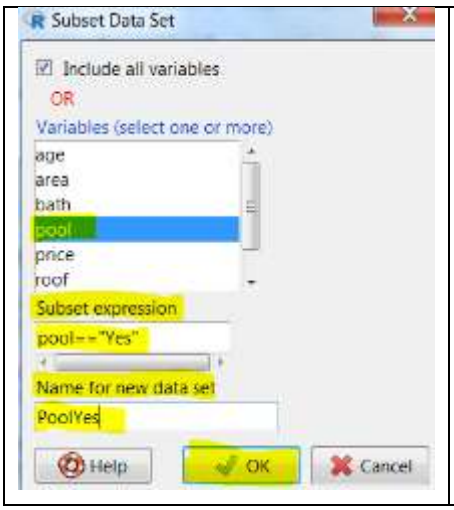
Side-by-side pie charts (with subset data sets) for two categorical variables

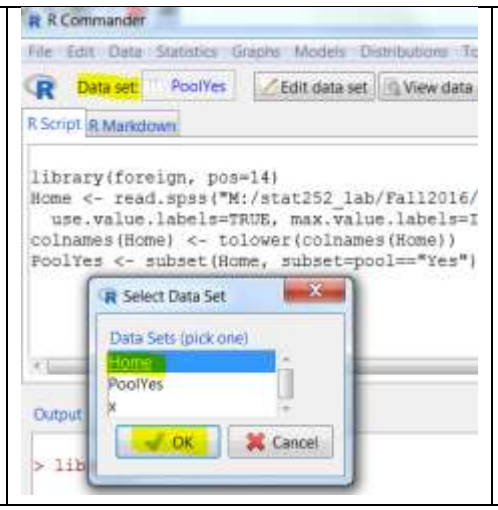
There is no easy way to draw a side-by-side pie chart; we need to select the subset of cases of interest and then draw an individual pie chart for each subset. For this example, we begin with the active data set you called Home in R (that was from the Excel file 8variable_salehome.xlsx that we have been using throughout section 2.2 of the manual) and then select homes with a swimming pool and save the data in a new data subset called PoolYes, and then we select homes without a swimming pool from the active data Home and save that data in a new data subset called PoolNo. And then we draw one pie chart on roof type for each subset dataset PoolYes and PoolNo.

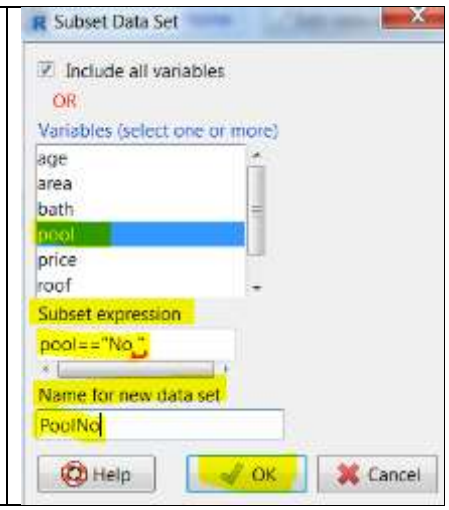
1. **Data→Active data set→Subset active data set…**
2. Select the variable to split the data (say pool here)
3. In **Subset expression**, type the selection condition. For example, **pool=="Yes''**
   Note: if the value is not numerical, you need to surround the value with double quotes. **Also, the variable name "pool" is case sensitive, the outcome "Yes" is case sensitive, and you must use two equal signs.**
4. In **Name of new data set**, type the name of the new data set. For example, PoolYes contains all homes with a swimming pool.

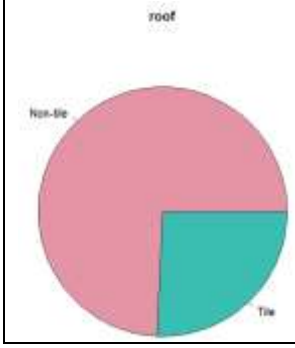
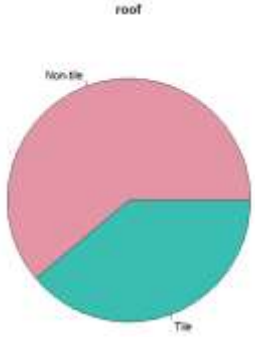
Note: Now the active data set is PoolYes. Make sure you switch the active data set back to Home before selecting homes without a swimming pool.

5. Click **Data set**, select the whole data set (Home) and click OK
6. Repeat for homes without a swimming pool. Use the **Subset expression pool=="No"** and use the name PoolNo as your **Name of New data set.**
   Note: if the value is not numerical, you need to surround the value with double quotes. **Also, the variable name "pool" is case sensitive, the outcome "No" is case sensitive, and you must use two equal signs. Also, there is a space after No!**
7. Click **Data set**, select PoolYes as the active data set and click OK
8. **Graphs → Pie Chart…**, select roof and click OK
9. Click **Data set**, select PoolNo as the active data set and click OK
10. **Graphs → Pie Chart…**, select roof and click OK







With Swimming Pool	Without Swimming Pool	Comments:
		1. The patterns regarding to roof type are different for homes with and without a swimming pool, since the two pie charts are not similar. 2. Percentage of homes with a tile roof is higher for homes without a swimming pool.

For your reference, the following table summarizes selection operators in R.

Symbol/code	Name	Use
==	equality	used to indicate the variable should equal
!=	Inequality	used to indicate the variable should not equal
&	And	used to combine multiple expressions
|	Or	used to combine multiple expressions
is.na(varname)		Include the missing values of a variable
!is.na(varname)		Exclude the missing values of a variable
>	Greater than	
<	Less than	
>=		More than or equal to
<=		Less than or equal to

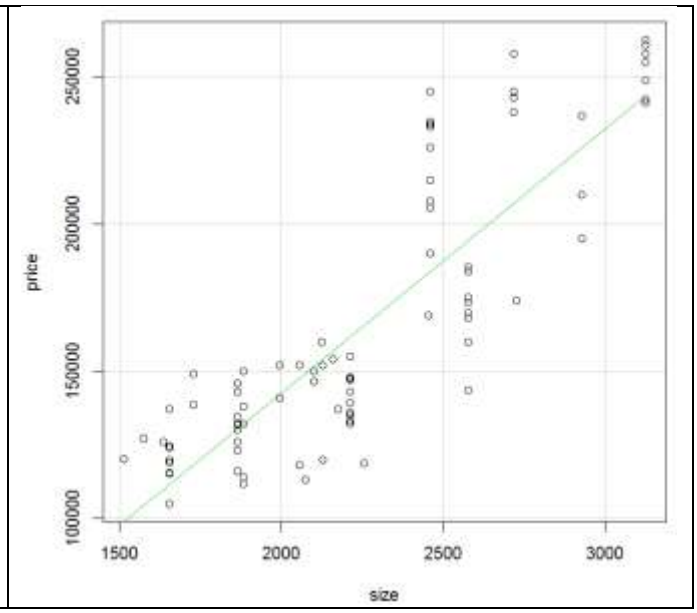
Scatterplot and Pearson correlation coefficient for two numerical quantitative variables.

Here we will investigate the relationship between two quantitative variables.

We again use the Home data.

Draw a scatter plot of price (Y-axis) versus size (X-axis). Could we model their relationship using a straight line? How does price change when size increases?

1. Click **Data set**, select Home as the active data set and click OK
2. **Graphs → Scatterplot…**
3. Choose **size** as the **y-variable** and **price** as the **x variable.**
4. Click **Options**, select **Least-squares line** under **Plot Options.** Click OK.

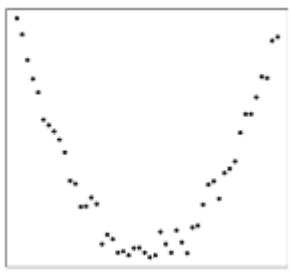
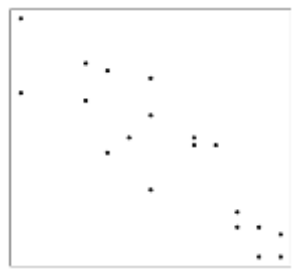
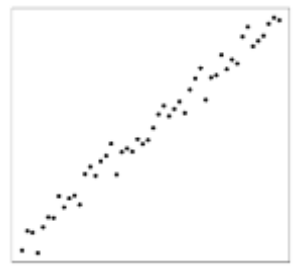


Comments: It might be okay to model the relationship between price and size using a straight line. When size increases the price increases. This means price and size have a positive association.

Like the five-number summary is the numerical summary of a boxplot, the numerical summary for a scatter plot is the Pearson correlation coefficient r; it captures the association between the response variable y (e.g., price) and the predictor variable x (e.g., size) in three aspects:

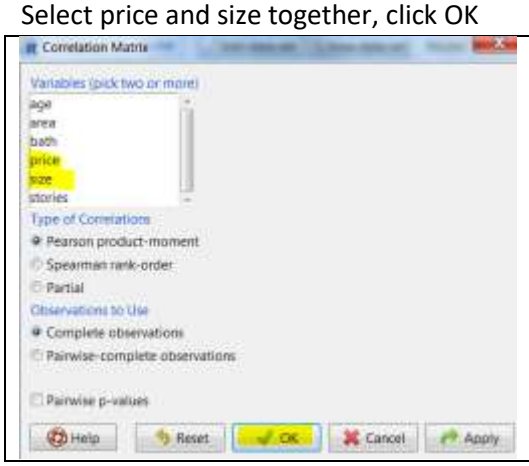
- Pattern: it captures only the linear association. Do not use the correlation coefficient r to describe non-linear association.
- Strength: the closer r is to either +1 or -1, the stronger the linear association. $r \approx 0$ indicates no or weak linear association.
- Direction: positive or negative. Positive association ($r > 0$) means that y and x change in the same direction. That is, y increases (decreases) if x increases (decreases). Negative association ($r < 0$) means that y and x change in the opposite direction. That is, y increases (decreases) if x decreases (increases).

The following figure gives four scatter plots and their corresponding correlation coefficients.

$r = -0.020$	$r = 0.697$	$r = -0.887$	$r = 0.989$
			

Calculate the Pearson correlation coefficient between price and size.

1. **Statistics→Summaries→Correlation Matrix**
2. Select price and size together, click OK



Output

```
            price      size
price 1.0000000 0.8571687
size  0.8571687 1.0000000
```

The correlation coefficient between price and size is r =0.857. The value is quite close to +1. There is **fairly strong**, **positive**, **linear** association between price and size.

We can also calculate the correlation coefficient for each pair of the quantitative variables. To do this, select all the variables when you run the correlation matrix commands above.

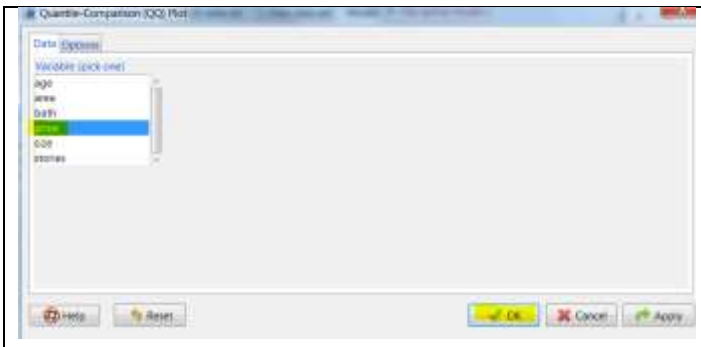
	age	area	bath	price	size	stories
age	1.00000000	0.11507397	-0.7494382	-0.92896864	-0.7536454	0.04393443
area	0.11507397	1.00000000	0.0111107	0.03110192	0.2010353	0.19604329
bath	-0.74943817	0.01111070	1.0000000	0.79871946	0.8234954	0.26569831
price	-0.92896864	0.03110192	0.7987195	1.00000000	0.8571687	0.03983124
size	-0.75364538	0.20103532	0.8234954	0.85716868	1.0000000	0.27797829
stories	0.04393443	0.19604329	0.2656983	0.03983124	0.2779783	1.00000000

Price and age have a strong, negative, linear association. Size and bath have a moderately strong, positive association.

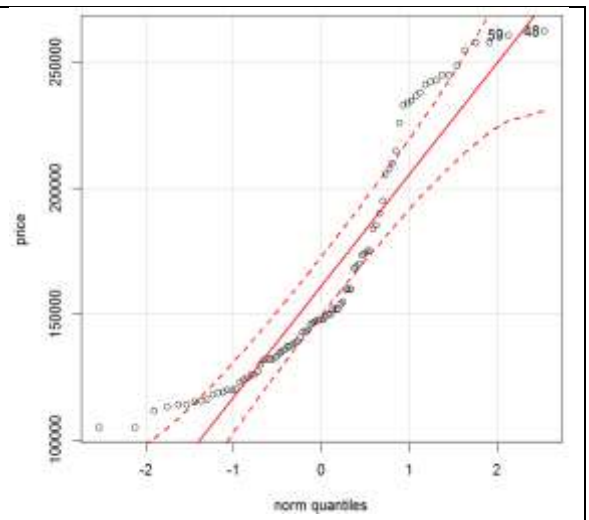
In statistics, it is important to check whether the data are taken from a normal population. The graphical tool used is called the normal probability plot. The normal probability plot is also called the normal Q-Q (Quantile-Quantile) plot since it is a scatter plot with the observed and theoretical quantiles as the axis. It does not matter whether we put the observed quantile on the x-axis or on the y-axis. If the data are taken from a normal population, the points roughly fall on a straight line. By default, R commander plots the theoretical quantile on the x-axis and the observed quantile on the y-axis.

Using the home data set, check whether the price of sale homes follows a normal distribution.

1. Click **Data set**, select Home as the active data set and click OK
2. **Graphs → Quantile-comparison plot…**
3. Select price and click OK



Since the points are not roughly on a straight line, we can conclude that price of the sale home does not follow a normal distribution.



LAB 3 PROBABILITY DISTRIBUTIONS (BINOMIAL AND NORMAL)

This chapter introduces how to use R commander to calculate probabilities related to Binomial distributions (a discrete distribution) and normal distributions (a continuous distribution).

3.1 BINOMIAL DISTRIBUTION

A Bernoulli trial is a chance experiment with only two possible outcomes: success or failure. Let p be the probability of success. Repeat the Bernoulli trial n times and let X=number of successes out of these n outcomes. X follows a Binomial distribution with parameters n (# of trials) and p (probability of success).

3.1.1 Steps to Apply the Binomial Formula

- Identify n (# of trials) and p (probability of success);
- Write down the event using the variable X;
- Re-write the event in the form of $P(X = a)$ or $P(X \leq a)$ or $P(X > a)$ if necessary.

3.1.2 Example: Application of Binomial Distribution

A quiz consists of 10 multiple choices questions with four choices A, B, C and D. I did not study and randomly picked one answer for each question.

(a) Find the probability that I got six correct answers.
(b) Find the probability that I got at least one correct answer.
(c) Find the probability that I got at least nine correct answers.
(d) Find the probability that I got eight to ten correct answers.

Solutions: For each question, I either got the correct answer or not. Each question is one Bernoulli trial. Since I randomly picked one answer, each of the four choices has the same chance to be chosen. There is only one correct answer and the probability of obtaining the correct answer is ¼. Whether I obtain the correct answer for the current question will not affect the chance of getting the correct answer for the next question, so the trials are independent with the same probability of success. Let X =# of correct answers. X follows a binomial distribution. Its probability distribution is

$$P(X = x) = {}_nC_x p^x (1-p)^{n-x} = {}_{10}C_x (\tfrac{1}{4})^x (1-\tfrac{1}{4})^{10-x} = {}_{10}C_x (0.25)^x (1-0.25)^{10-x}, \quad x=0, 1, \ldots, 10.$$

Re-write the events in the form of $P(X = a)$ (binomial probabilities) or $P(X \leq a)$ (binomial **lower tail** probabilities) or $P(X > a)$ (binomial **upper tail** probabilities).

(a) Find the probability that I got six correct answers. $P(X = 6)$
   1. **Distributions→Discrete distributions→Binomial distribution→Binomial probabilities**
   2. In "**Binomial Probability**" window, put n in **Binomial trials** and p in **Probability of success**

Output gives the probability distribution, i.e., all possible values x in the first column and their corresponding probabilities $P(X = x)$ in the second column.

Output:

	Probability
0	5.631351e-02
1	1.877117e-01
2	2.815676e-01
3	2.502823e-01
4	1.459980e-01
5	5.839920e-02
6	1.622200e-02
7	3.089905e-03
8	3.862381e-04
9	2.861023e-05
10	9.536743e-07

	Probability
0	0.0563135147095
1	0.1877117156982
2	0.2815675735474
3	0.2502822875977
4	0.1459980010986
5	0.0583992004395
6	0.0162220001221
7	0.0030899047852
8	0.0003862380981
9	0.0000286102295
10	0.0000009536743

$$P(X = 6) = 0.016222$$

Note: Your computer output may use notation with e- in it, depending on your version of R. In computer outputs of R, $1.622200e - 02 = 1.622200 \times 10^{-2} = 0.016222$, $1.622e + 02 = 1.622 \times 10^{2} = 162.2$, $2.861023e - 05 = 2.861023 \times 10^{-5} = 0.00002861023$.

(b) Find the probability that I got at least one correct answer. $P(X \geq 1)$
Note that $P(X \geq 1) = \boldsymbol{P(X > 0)} = P(X = 1) + P(X = 2) + \cdots + P(X = 10) = \boldsymbol{1 - P(X = 0)}$

Therefore, there are two ways to calculate the answer:

- Based on the output of probability distribution, we find
$$\boldsymbol{1 - P(X = 0)} = 1 - 0.05631351 = 0.9436865$$

- We can use the upper tail probability $\boldsymbol{P(X > x)}$**.** In this question, we want $\boldsymbol{P(X > 0)}$**.**
  1. **Distributions→Discrete distributions→Binomial distribution→Binomial tail probabilities…**
  2. In "**Binomial Probability**" window, put x in **Variable value(s)**, n in **Binomial trials**, and p in **Probability of success**. In this example, $x = 0, n = 10, p = 0.25$
  3. Select **Upper tail**, since we want the upper tail probability (greater than)
  4. Click OK

Output:

```
[1] 0.9436865
```

The result is the same as the one obtained using the first method.

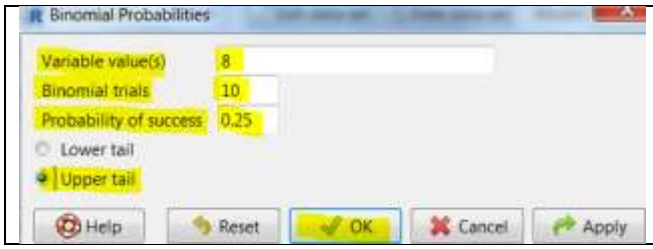
(c) Find the probability that I got at least nine correct answers. $P(X \geq 9)$
Note that $P(X \geq 9) = \boldsymbol{P(X > 8)} = \boldsymbol{P(X = 9)} + \boldsymbol{P(X = 10)}$

Therefore, there are two ways to calculate:

- Based on the output of probability distribution,
$$\boldsymbol{P(X = 9)} + \boldsymbol{P(X = 10)} = 0.00002861023 + 0.0000009536743 = 0.0000295639$$

- Use the upper tail probability $\boldsymbol{P(X > x)}$. In this question, we want $\boldsymbol{P(X > 8)}$.
  1. **Distributions→Discrete distributions→Binomial distribution→Binomial tail probabilities…**

2. In "**Binomial Probability**" window, put x in **Variable value(s)**, n in **Binomial trials**, and p in **Probability of success**. In this question, $x = 8, n = 10, p = 0.25$
3. Select **Upper tail**, since we want the upper tail probability (greater than)
4. Click OK



Output:

```
[1] 2.95639e-05
```

$$\boldsymbol{P(X > 8) = 0.0000295639}$$

(d) Find the probability that I got eight to ten correct answers, inclusively. $P(8 \le X \le 10)$
Note that

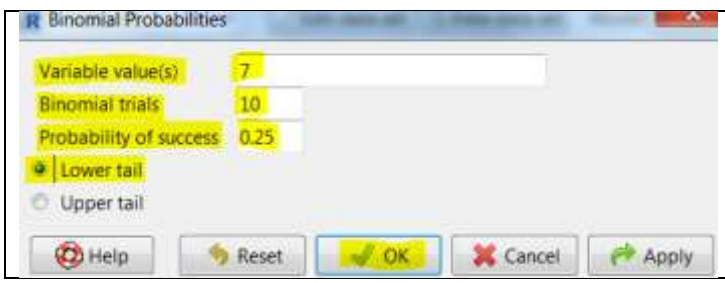
$$\boldsymbol{P(8 \le X \le 10) = P(X = 8) + P(X = 9) + P(X = 10) = P(X \le 10) - P(X \le 7) = 1 - P(X \le 7)}$$

Therefore, there are two ways to calculate:

- Based on the output of probability distribution,

$$\boldsymbol{P(X = 8)} + \boldsymbol{P(X = 9)} + \boldsymbol{P(X = 10)} = 0.0003862381 + 0.00002861023 + 0.0000009536743 = 0.000415802$$

- Use the lower tail probability $\boldsymbol{P(X \le x)}$. In this question, we want $\boldsymbol{P(X \le 7)}$.
  1. **Distributions→Discrete distributions→Binomial distribution→Binomial tail probabilities…**
  2. In "**Binomial Probability**" window, put x in **Variable value(s)**, n in **Binomial trials**, and p in **Probability of success**. In this question, $x = 7, n = 10, p = 0.25$
  3. Select **Lower tail**, since we want the lower tail probability (less than or equal to)
  4. Click OK



Output:

```
[1] 0.9995842
```

$$\boldsymbol{P(8 \le X \le 10) = 1 - P(X \le 7)} = 1 - 0.9995842 = 0.0004158$$

3.2 Normal Distribution

We use the density curve to describe the distribution of a continuous variable. The total area under a density curve is one, and the area under the curve is related to the probability of a certain event. The most widely used continuous distribution is the normal distribution, which is well known as the bell-shaped and symmetric curve. The normal density function has two parameters: the mean μ and the standard deviation σ. The parameter μ controls the center (location) of the distribution and σ controls the shape of the distribution. When σ is larger, the curve appears to be shorter and fatter; when σ is smaller, the curve appears to be taller and slimmer. If a random variable X follows a normal distribution with mean μ and standard deviation σ, we write $X \sim N(\mu, \sigma)$. Its probability density function $f(x)$ is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty \ < x < \infty, \text{ with } \pi \approx 3.142, \ e \approx 2.718.$$

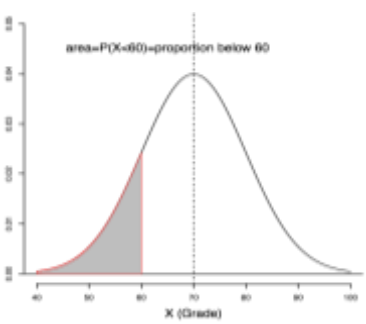
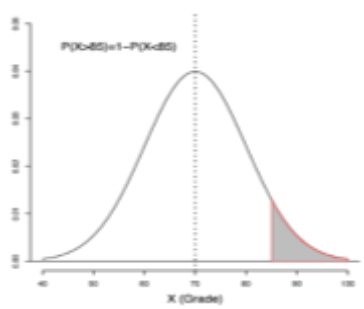
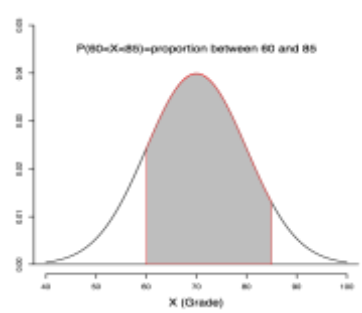
Recall that for a binomial distribution, $P(X \leq x) = P(X = 0) + P(X = 1) + \cdots + P(X = x)$. For a normal distribution which is continuous, $P(X = x) = 0$ and therefore, $P(X \leq x) = P(X < x)$. There are two main applications of normal distributions: to find the probabilities given the x values (tail probabilities) and to find the x values given the probabilities (quantiles).

3.2.1 Find the Probabilities Related to Normal Distributions

Suppose grade X follows a normal distribution with a mean 70 and a standard deviation 10. That is $X \sim N(70, 10)$. We are interested in the probabilities of the following events.

1. Find the probability that a student has a grade below 60. $P(X < 60)$
2. Find the probability that a student has a grade above 85. $P(X > 85)$
3. Find the probability that a student has a grade between 60 and 85. $P(60 < X < 85)$

The following graphs show their corresponding probabilities:

$P(X < 60)$=area to the **left** of 60	$P(X > 85)$=area to the **right** of 85	$P(60 < X < 85)$=area between 60 and 85
		

(a) Find the probability that a student has a grade below 60.

We want $P(X < 60)$, which is a lower tail probability.

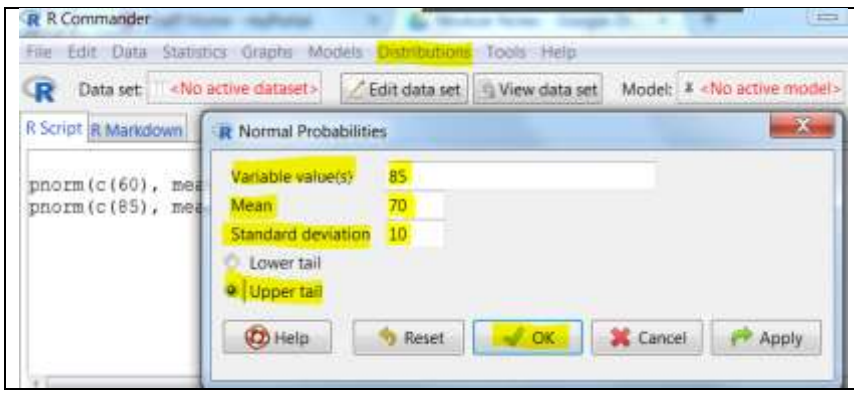
1. **Distributions→Continuous distributions→Normal distribution→Normal probabilities**
2. In "**Normal Probability**" window, put x in **Variable value(s)**, μ in **Mean**, and σ in **Standard deviation**. In this question, $x = 60, \mu = 70, \sigma = 10$
3. Select **Lower tail**, since we want the lower tail probability (less than)
4. Click OK

Output:

```
[1] 0.1586553
```

If $X \sim N(70, 10)$, $P(X < 60)$=0.1587. If we randomly pick one student, the probability that the student obtains a grade below 60 is 0.1587. Or 15.87% of the students obtain a grade below 60.

(b) Find the probability that a student has a grade above 85.
We want $P(X > 85)$ which is an **upper tail** probability.
1. **Distributions→Continuous distributions→Normal distribution→Normal probabilities**
2. In "**Normal Probability**" window, put x in **Variable value(s)**, μ in **Mean**, and σ in **Standard deviation**. In this question, $x = 85, \mu = 70, \sigma = 10$
3. Select **Upper tail**, since we want the upper tail probability (greater than)
4. Click OK

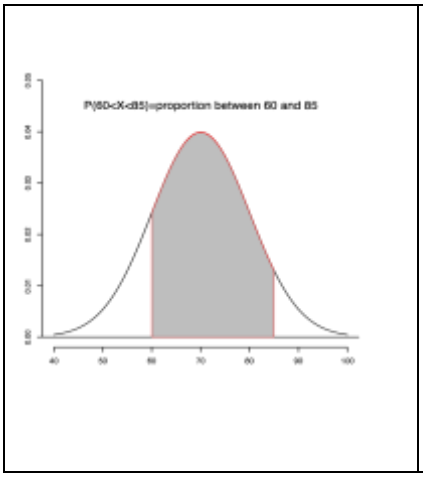


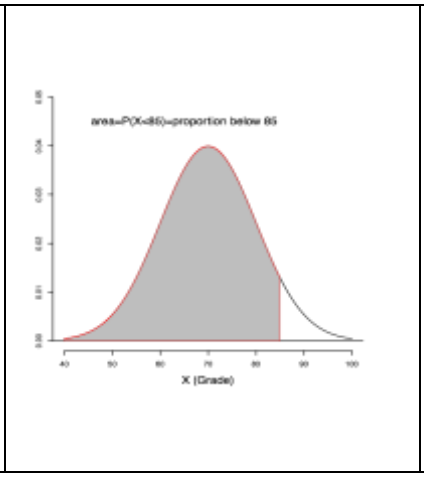
Output:

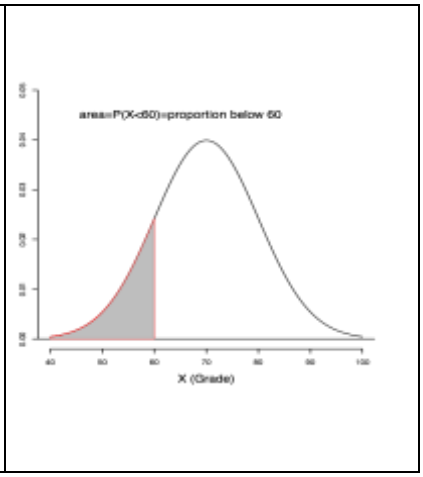
```
[1] 0.0668072
```

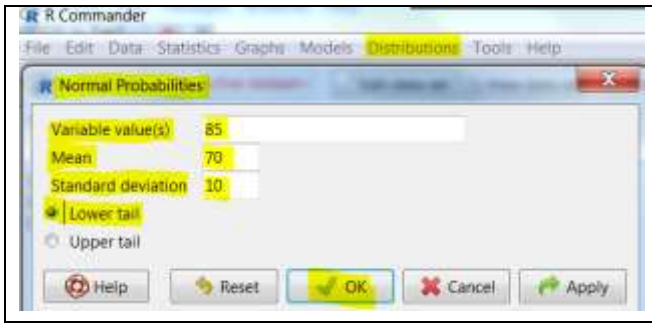
If $X \sim N(70, 10)$, $P(X > 85)$=0.0668. If we randomly pick one student, the probability that the student obtains a grade above 85 is 0.0668. Or 6.68% of the students obtain a grade above 85.

(c) Find the probability that a student has a grade between 60 and 85.
We want $P(60 < X < 85)$, the area between 60 and 80, which is equal to the area to the left of 85 minus the area to the left of 60.









Output:

```
[1] 0.9331928
```

If $X \sim N(70, 10)$, $P(X < 85)$=0.9332.

$$P(60 < X < 85) = P(X < 85) - P(X < 60)$$
$$= 0.9331928 - 0.1586553$$
$$= 0.7745375$$

77.45% of the students obtain a between 60 and 85.

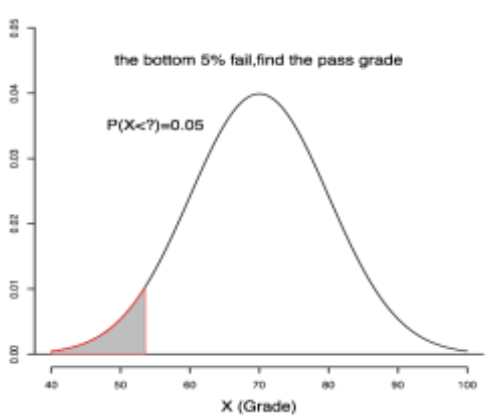
3.2.2 Find the Quantiles of Normal Distribution

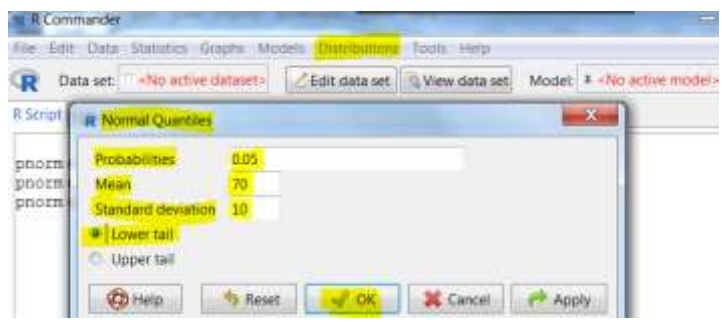
That is given the percentage or probability q, find the x value such that $q = P(X < x)$. The x value is called the quantile of the distribution corresponding to q.

Suppose grade X follows a normal distribution with a mean 70 and a standard deviation 10. That is $X \sim N(70, 10)$.

(a) If the bottom 5% of students will fail, find the passing grade.
We want to find the x value such that $P(X < x) = 0.05$, i.e., 5% of grades below what value.

1. **Distributions→Continuous distributions→Normal distribution→Normal quantiles**
2. In "**Normal Quantiles**" window, put q in **Probabilities**, μ in **Mean**, and σ in **Standard deviation**. In this question, $q = 0.05, \mu = 70, \sigma = 10$
3. Select **Lower tail**, since we want the x value corresponding to a lower tail probability (less than)
4. Click OK





Output:

```
[1] 53.55146
```

The passing grade is 53.55, since $P(X < 53.55) = 0.05$.

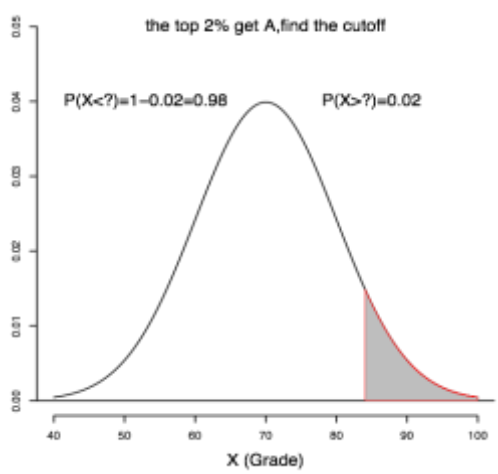
(b) If the top 2% of students will get an A, find the cutoff of getting an A.
We want to find the x value such that $P(X > x) = 0.02$, i.e., 2% of grades above what value or 98% of grades below what value.

Approach 1: upper tail probability, find the x value such that $P(X > x) = 0.02$.

1. **Distributions→Continuous distributions→Normal distribution→Normal quantiles**
2. In "**Normal Quantiles**" window, put q in **Probabilities**, μ in **Mean**, and σ in **Standard deviation**. In this question, $q = 0.02, \mu = 70, \sigma = 10$
3. Select **Upper tail**, since we want the x value corresponding to an upper tail probability (greater than)
4. Click OK

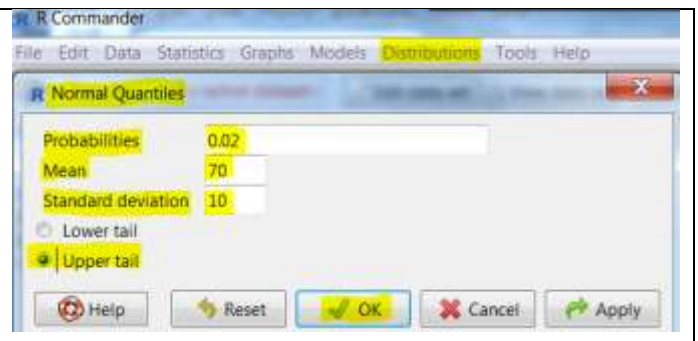
Approach 2: lower tail probability, find the x value such that $P(X < x) = 0.98$.
Note that **2%** of grades **above** what value=**98%** of grade **below** what value. That is
$P(X > x) = 0.02$ is that same as $1 - P(X > x) = 1 - 0.02 \Longrightarrow \boldsymbol{P(X < x) = 0.98}$

1. **Distributions→Continuous distributions→Normal distribution→Normal quantiles**
2. In "**Normal Quantiles**" window, put q in **Probabilities**, μ in **Mean**, and σ in **Standard deviation**. In this question, $q = 0.98, \mu = 70, \sigma = 10$
3. Select **Lower tail**, since we want the x value corresponding to a lower tail probability (less than)
4. Click OK




```
[1] 90.53749
```

The cutoff of A is 90.54. Students get a grade above 90.54 will get an A.



3.3 Generate Simple Random Samples from a Certain Distribution

3.3.1 Setting a Seed

Although you can let the software choose a random seed prior to generating simple random samples, **examples that require the generation of simple random samples in the manual will require you to set a given seed that is provided for you**. This allows the output in the manual examples to match what you get as you work through them. Setting a seed retires meticulous input to R.

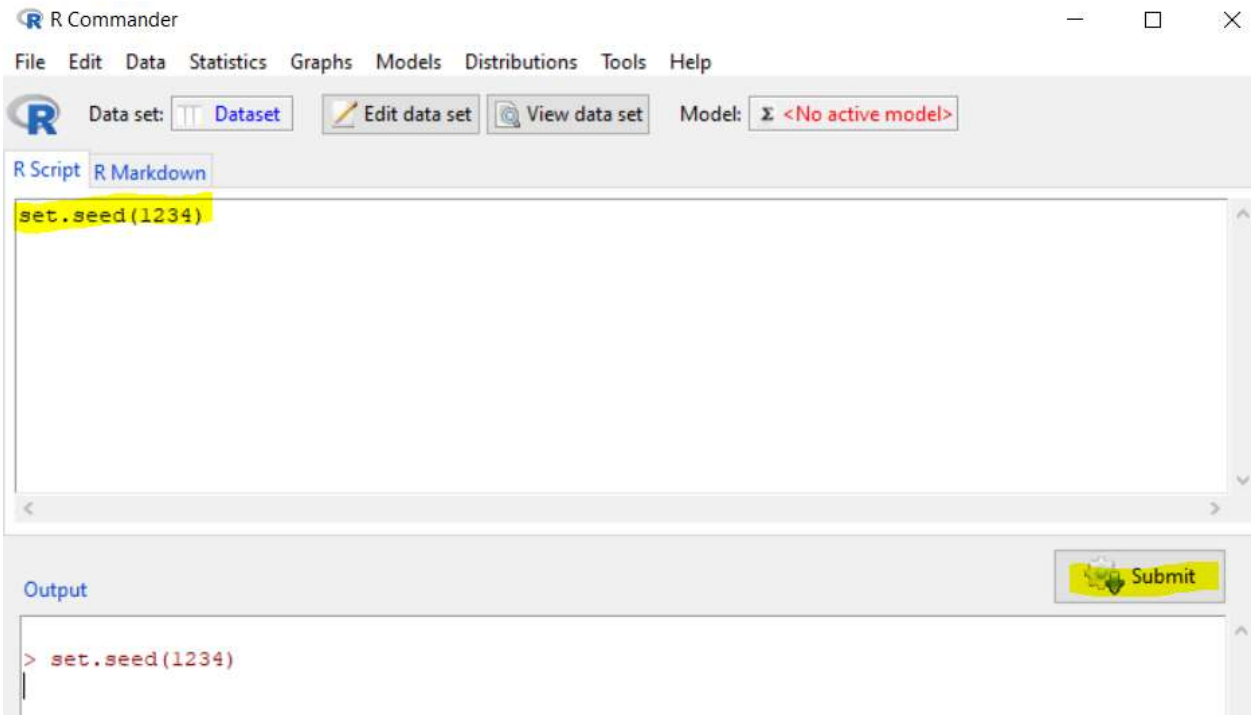
You must set your desired seed every time you do a new problem!

Instructions are below for setting a seed of 1234.

Approach 1 (fastest, but you must be meticulously accurate, and it is easy to mess up):

Type the command **set.seed(1234) in the R Script box** (not in the Output box!!!), then click "**Submit**" (do not hit "Enter", it will not work). The command line will be executed and appear in the Output box.

Make sure there are no characters in front of your set.seed(1234) command and that the command is typed flush against the left side of the R Script box in a new line all by itself. See below.



Approach 2: tedious and kind of mission impossible.

Drop down **Distributions→Set random number generator seed**. A box appears with a suggested seed. Your box may have a different suggested seed.

Move the two boxes together to get as close to 1234 as you can. The closest I can get is 1191. The closest number you can get may be different.

Click carefully, as many times as necessary, in the grey bar directly beside the boxes to move the seed number you have there to 1234. This is very tedious. Again, you must set the seed to the given seed each time you do a problem.

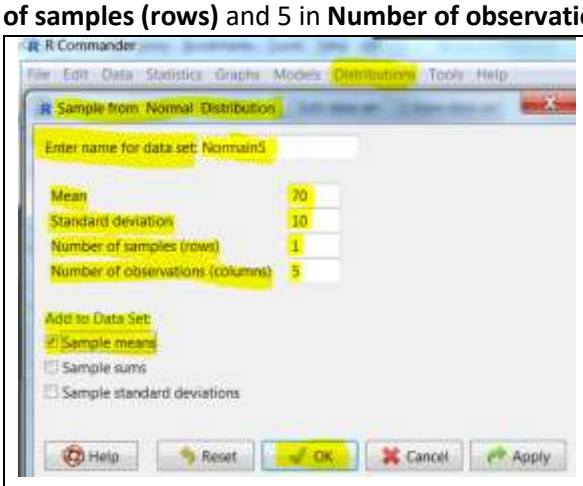
Once succeed, click OK.

3.3.2 Generate Simple Random Sample from a Normal Distribution

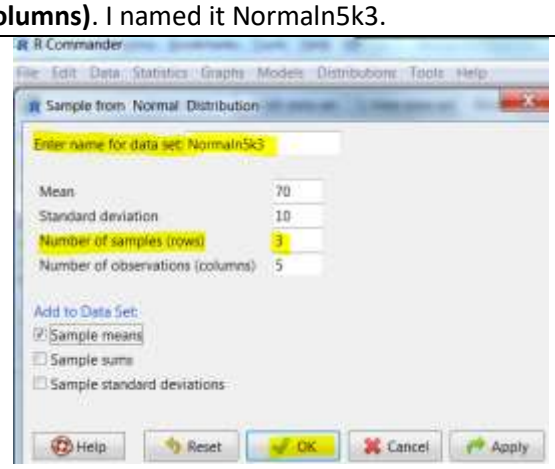
Suppose we want to generate n=5 observations from a normal distribution with mean μ=70 and standard deviation σ=10. We set a seed of 1234. We call the one dataset Normaln5 since the sample size is 5 and we are doing only 1 set (of samples of size 5).

1. Type **set.seed(1234) in the R Script box** (on its own line and flush against the left side of the box). Click Submit.
2. **Distributions→Continuous distributions→Normal distribution→Sample from normal distribution…**
3. In the "**Sample from Normal Distribution**" window, perform the following. **Enter name of data set** (say Normaln5), put μ in **Mean**, and σ in **Standard deviation**, number of samples in **Number of samples (rows)**, and the sample size n in **Number of observations (columns)**. In this question, $\mu = 70, \sigma = 10$, we only want one simple random sample, with sample size $n = 5$.
4. Select **Sample means** under **Add to Data Set.** It will store the sample mean of the sample in the last column. Click OK.
5. Select Normaln5 under **Data set** to make it as active data set
6. Click **View data set** to view the sampled data

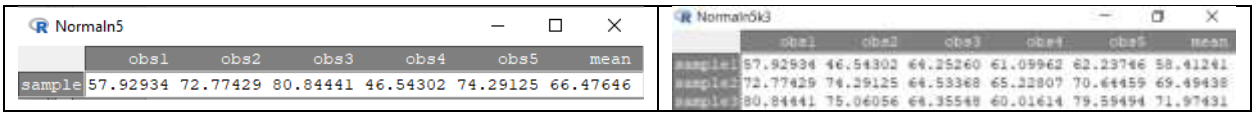
We can also generate K sets of simple random samples of size n by setting the value of **Number of observations (columns)** to be K. For example, if we want to generate three simple random samples of size 5, we would follow the steps 1 to 6 above (including setting the seed to 1234) and put 3 in **Number of samples (rows)** and 5 in **Number of observations (columns)**. I named it Normaln5k3.



Output



Output



	obs1	obs2	obs3	obs4	obs5	mean
sample	57.92934	72.77429	80.84441	46.54302	74.29125	66.47646

	obs1	obs2	obs3	obs4	obs5	mean
sample1	57.92934	46.54302	64.25260	61.09962	62.23746	58.41241
sample2	72.77429	74.29125	64.53368	65.22807	70.64459	69.49438
sample3	80.84441	75.06056	64.35548	60.01614	79.59494	71.97431

3.3.3 Generate Simple Random Sample from an Exponential Distribution

An exponential distribution is an extremely right skewed continuous distribution which is widely used to model the lifetime of products. The density function of exponential distribution is given by:

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}},\ x > 0,\ \text{with } e \approx 2.718$$

denoted as $X \sim Exp(\lambda)$ where λ is the mean (expected value) of X. One property of an exponential distribution is the mean and standard deviation both equal λ, i.e., $\mu = \sigma = \lambda$.

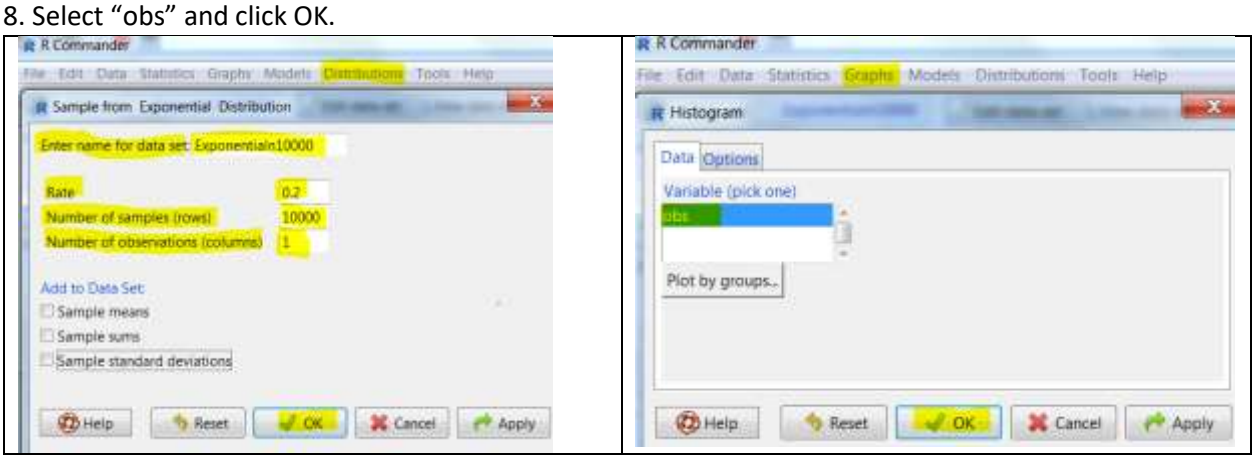
Suppose the survival time of liver cancer patients, X, follows an exponential distribution with mean and standard deviation 5 years, i.e., $\mu = \lambda = 5, \sigma = \lambda = 5$.

(a) Generate 10000 observations from an exponential population distribution with mean $\lambda = 5$ or rate $\frac{1}{\lambda} = \frac{1}{5} = 0.2$. Use the seed 1235 and save the data in the file "Exponentialn1000".

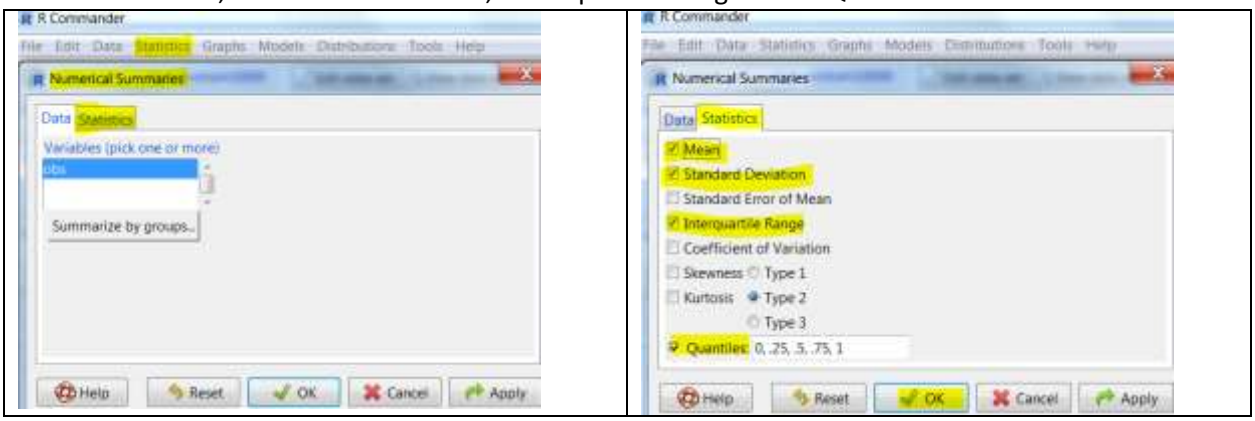
(b) Draw a histogram using those 10000 observations. With 10000 observations, this sample histogram provides an excellent approximation of an exponential population distribution with mean 5.

(c) Calculate the sample mean and sample standard deviation and compare them with the population mean and standard deviation.

1. Type **set.seed(1235) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.
2. **Distributions→Continuous distributions→Exponential distribution→Sample from exponential distribution…**
3. In the "**Sample from Exponential Distribution"** window, type "Exponentialn10000" in **Enter name of data set**, put 0.2 in **Rate**, 10000 in **Number of samples (rows)**, and 1 in **Number of observations (columns)**.
4. Click OK
5. Select Exponentialn10000 under **Data set** to make it as active data set
6. Click **View data set** to view the sampled data. The samples are stored in the column "obs", the data set has one column and 10000 rows.
7. **Graphs→Histogram**
8. Select "obs" and click OK.

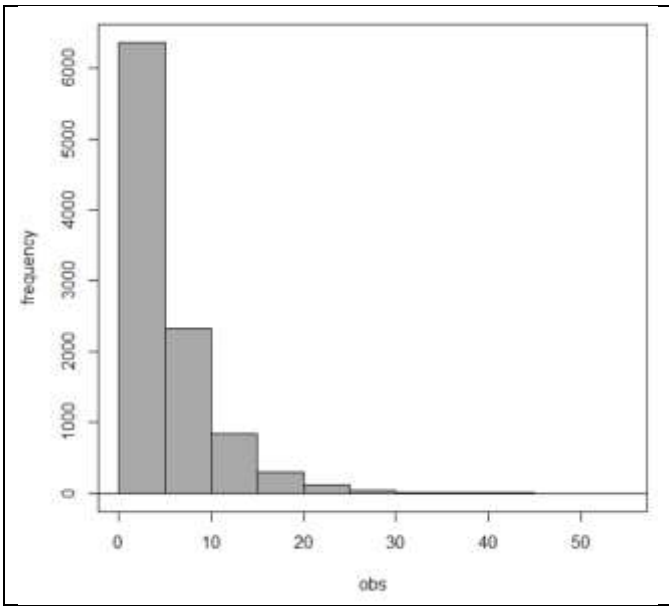


8. **Statistics→Summaries→Numerical Summaries**
9. In the **Numerical Summaries** window, select “obs” and click **Statistics**
10. Check “Mean”, “Standard Deviation”, “Interquartile Range” and “Quantiles”.



Outputs

mean	sd	IQR	0%	25%	50%	75%	100%	n
4.92683	4.913667	5.328878	0.0007534213	1.447858	3.431467	6.776736	51.99725	10000



- The histogram shows that the distribution of exponential with mean $\mu = \lambda = 5$ or rate $\frac{1}{\lambda} = \frac{1}{5} = 0.2$ is extremely right skewed.
- The sample mean based on $n = 10000$ observations is $\bar{x} = 4.927$ which is very close to the population mean $\mu = 5$ due to the large sample size.
- The sample standard deviation $s = 4.914$ which is also very close to the population standard deviation $\sigma = \lambda = 5$. Note that for an exponential distribution, the population mean and standard deviation are equal. That is $\mu = \sigma = \lambda$.

LAB 4 DISTRIBUTION OF THE SAMPLE MEAN & CENTRAL LIMIT THEOREM

In this lab, we are going to investigate the distribution of the sample mean $\bar{X}$ by generating samples with different sample sizes from different population distributions. The central limit theorem states that when the sample size n is large enough (rule of thumb: $n \geq 30$), the sample mean $\bar{X}$ is approximately normally distributed regardless of the population distribution. We can understand the central limit theorem by simulation.

4.1 OBTAIN THE DISTRIBUTION OF THE SAMPLE MEAN FROM A CERTAIN DISTRIBUTION

1. Take a simple random sample of size n from a certain distribution.
2. Calculate the sample mean $\bar{x}$.
3. Suppose the population size is N (i.e., there are N individuals in the population), so there are NC_n (N choose n) distinct samples. One sample will generate one value of the sample mean $\bar{x}$.
4. Drawing a histogram on those NC_n $\bar{x}$-values gives the distribution of the sample mean $\bar{X}$ for sample size n.
5. If NC_n is too large for us to consider all possible samples, we can generate a sufficiently large number of samples, say 10000, to approximate the distribution of the sample mean $\bar{X}$.

For the distribution of the sample mean $\bar{X}$ with sample size n, we have the following conclusions:

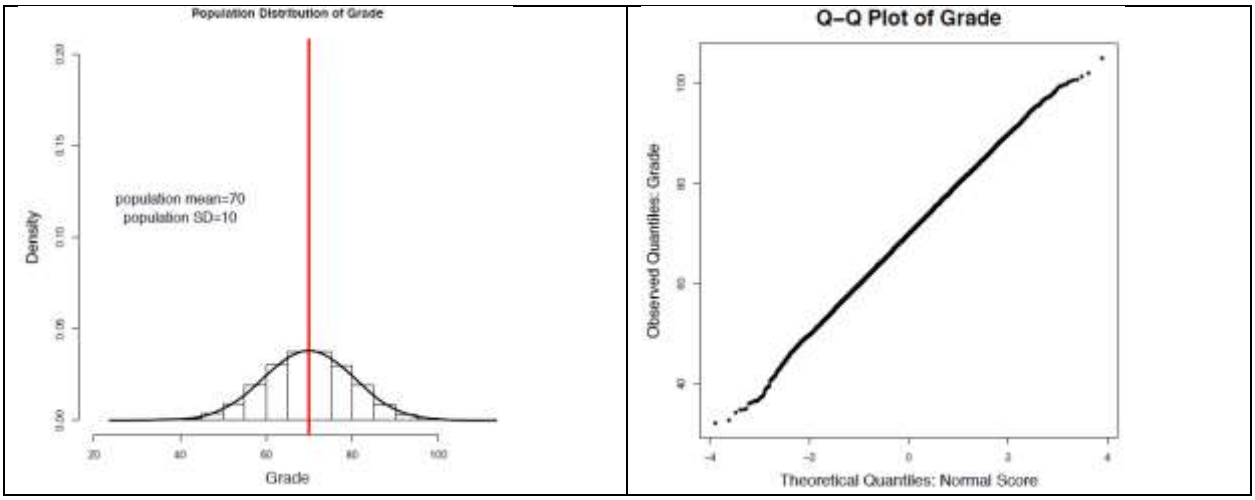
- The **mean of the sample mean** $\bar{X}$ equals the population mean μ; that is
$$\mu_{\bar{X}} = \mu.$$
- The **standard deviation of the sample mean** $\bar{X}$ equals the population standard deviation σ divided by the square root of the sample size; that is
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

These two conclusions are always true for any population distribution and for any sample size n.
We discuss the **shape of the distribution of the sample mean $\bar{X}$ in two cases**:

1. When the population distribution (the distribution of the variable under consideration X) is normal, the sample mean $\bar{X}$ is **exactly** normally distributed regardless of the sample size n.
2. When the population distribution is not normal, but the sample size n is large, the sample mean $\bar{X}$ is **approximately** normally distributed. This is guaranteed by the central limit theorem.

4.2 DISTRIBUTION OF THE SAMPLE MEAN WHEN THE POPULATION DISTRIBUTION IS NORMAL

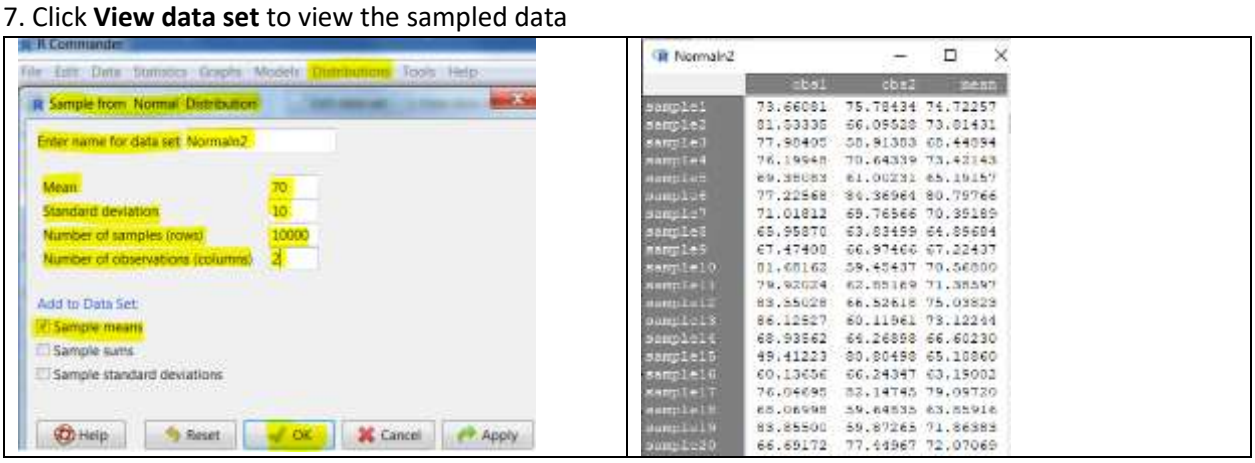
Suppose a population consists of $N = 100$ students and the variable of interest is the grade X. The histogram of the grades of these 100 students gives the population (or parent) distribution, the distribution of X. The mean and standard deviation of these 100 grades give the population mean and population standard deviation, respectively, as $\mu = 70$ and $\sigma = 10$. The normal QQ plot shows that the grade X follows a normal distribution, since all the data points roughly lie on a straight line.



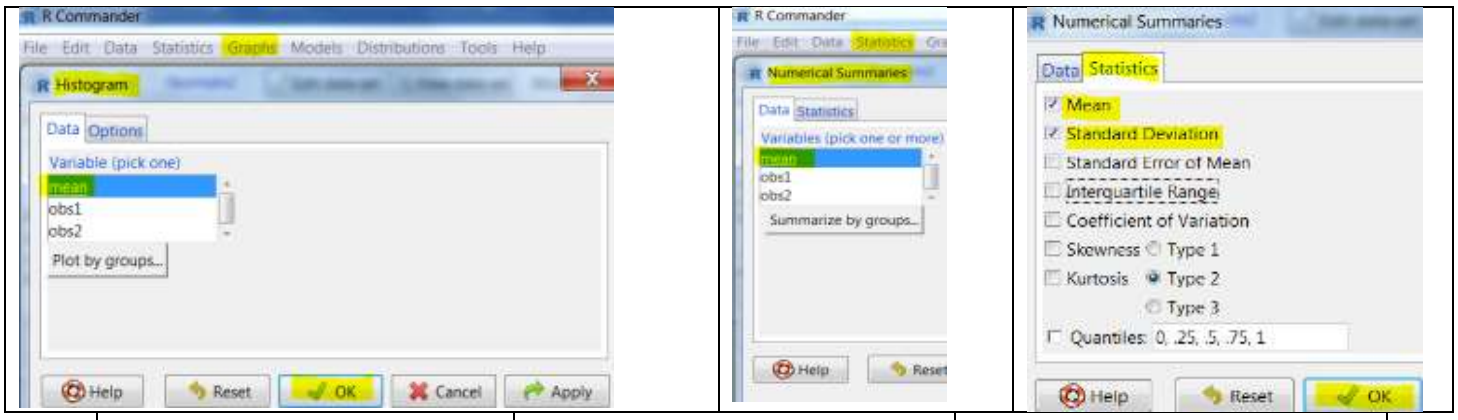
Let us examine the distribution of sample mean $\bar{X}$ with sample size $n = 2, 5, 30$ respectively.

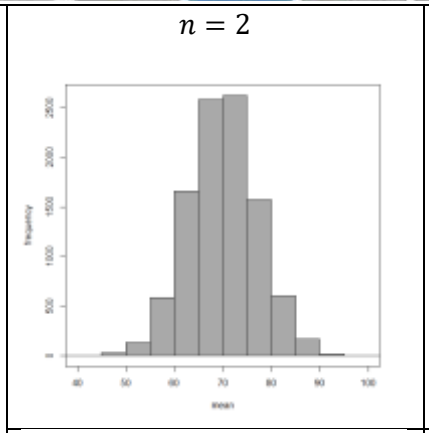
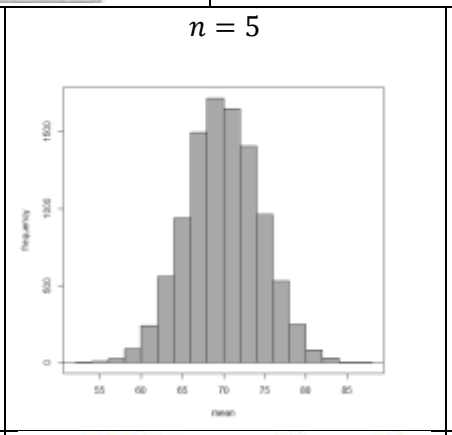
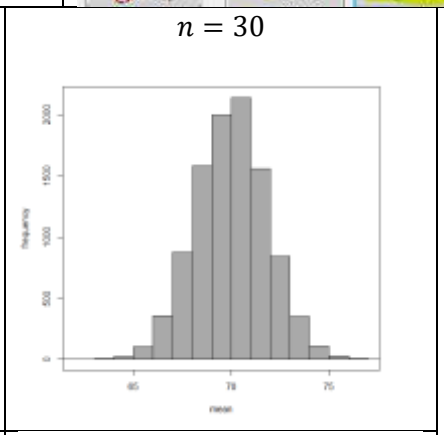
For each sample size n (# of observations (columns)), generate 10000 samples (# of rows). Use the seed 5942 for each n. Calculate the sample mean $\bar{x}$ for each sample by calculating the average of each row and store the value in the last column of the data file. Draw a histogram on the last column to obtain the distribution of the sample mean.

1. Type **set.seed(5942) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.
2. **Distributions→Continuous distributions→Normal distribution→Sample from normal distribution…**
3. In the "**Sample from Normal Distribution**" window, type Normaln2 in **Enter name of data set**, put 70 in **Mean**, and 10 in **Standard deviation**, 10000 in **Number of samples (rows)**, and 2 in **Number of observations (columns)**
4. Select **Sample means** under **Add to Data Set**, and the dataset will store the sample mean of the sample in the last column (labeled "mean").
5. Click OK
6. Select Normaln2 under **Data set** to make it as active data set
7. Click **View data set** to view the sampled data



8. **Graphs→Histogram**
9. Select "mean" and click OK.
10. **Statistics→Summaries**
11. In the "**Numerical Summaries**" window, select "mean" and click **Statistics**
12. Check "Mean", "Standard Deviation"
13. Repeat steps 1-12 for sample size $n = 5$ and $n = 30$ in "**Sample from Normal Distribution**" window). For each repetition of the steps, type set.seed(5942) in the R Script box and click submit. Use the file names Normaln5 and Normaln30. **Never write over a file.**

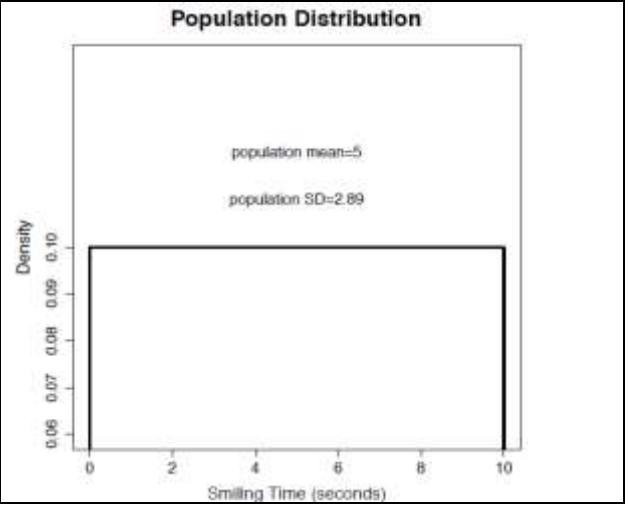


$n = 2$	$n = 5$	$n = 30$
		
mean sd n 70.03121 7.101047 10000	mean sd n 69.95981 4.497305 10000	mean sd n 69.99608 1.821243 10000

Findings:

- The mean of the sample mean is always very close to the population mean $\mu = 70$ regardless of the sample size n. The difference is because we did not consider all possible samples of size n, but only 10000 samples.
- The standard deviation of the sample mean is always close to theoretical value $\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{n}}$. When $n = 2, \frac{10}{\sqrt{n}} = \frac{10}{\sqrt{2}} = 7.071$; when $n = 5, \frac{10}{\sqrt{n}} = \frac{10}{\sqrt{5}} = 4.472$; when $n = 30, \frac{10}{\sqrt{n}} = \frac{10}{\sqrt{30}} = 1.826$.
- The histogram of the sample mean has a bell-shaped curve regardless of the sample size $n = 2, 5$, or 30.

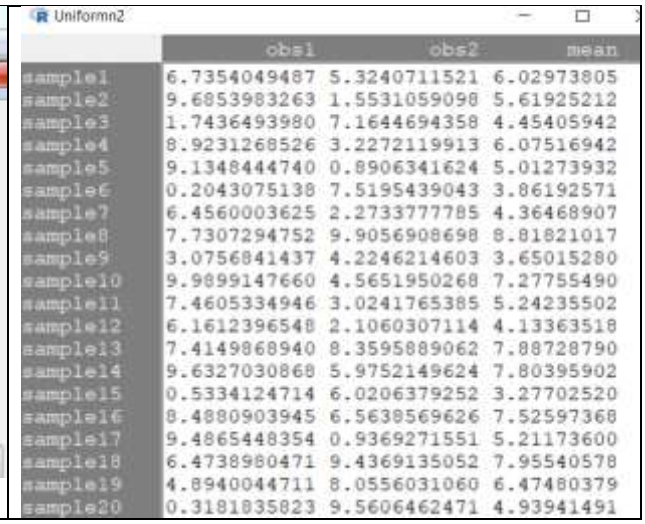
4.3 Distribution of the Sample Mean When the Population Distribution is Uniform

Suppose X, the smiling time of eight-week-old babies, follows a uniform distribution between 0 and 10 seconds. The density curve is shown in the right panel. The density curve forms a rectangle and hence not a normal curve for sure. The population mean $\mu = 5$ second and the population standard deviation $\sigma = 2.89$ second. Let's examine the distribution of sample mean $\bar{X}$ with sample size $n = 2, 5, 30$ respectively. That is the distribution of the average smiling time of n randomly selected babies.	

For each sample size n (# of columns), generate 10000 samples (# of rows). Use the seed 3921 for each n. Calculate the sample mean $\bar{x}$ for each sample by calculating the average of each row and store the value in the last column of the data file. Draw a histogram on the last column to obtain the distribution of the sample mean.

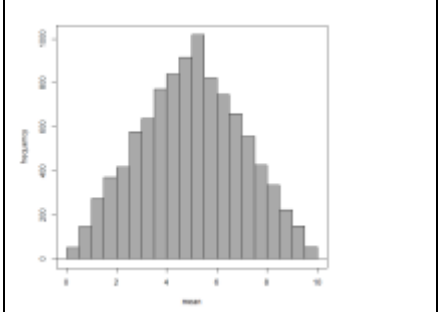
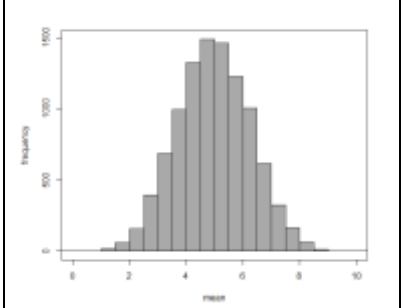
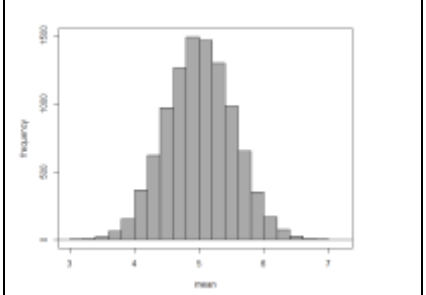
1. Type **set.seed(3921) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.
2. **Distributions→Continuous distributions→Uniform distribution→Sample from uniform distribution…**
3. In the "**Sample from Uniform Distribution**" window, type Uniformn2 in **Enter name of data set**, put 0 in **Minimum** and 10 in **Maximum**, 10000 in **Number of samples (rows)**, and 2 in **Number of observations (columns)**
4. Select **Sample means** under **Add to Data Set**, it will store the sample mean of the sample in the last column.
5. Click OK
6. Select Uniformn2 under Data set to make it as active data set

7. Click View data set to view the sampled data



	obs1	obs2	mean
sample1	6.7354049487	5.3240711521	6.02973805
sample2	9.6853983263	1.5531059098	5.61925212
sample3	1.7436493980	7.1644694358	4.45405942
sample4	8.9231268526	3.2272119913	6.07516942
sample5	9.1348444740	0.8906341624	5.01273932
sample6	0.2043075138	7.5195439043	3.86192571
sample7	6.4560003625	2.2733777785	4.36468907
sample8	7.7307294752	9.9056908698	8.81821017
sample9	3.0756841437	4.2246214603	3.65015280
sample10	9.9899147660	4.5651950268	7.27755490
sample11	7.4605334946	3.0241765385	5.24235502
sample12	6.1612396548	2.1060307114	4.13363518
sample13	7.4149868940	8.3595889062	7.88728790
sample14	9.6327030868	5.9752149624	7.80395902
sample15	0.5334124714	6.0206379252	3.27702520
sample16	8.4880903945	6.5638569626	7.52597368
sample17	9.4865448354	0.9369271551	5.21173600
sample18	6.4738980471	9.4369135052	7.95540578
sample19	4.8940044711	8.0556031060	6.47480379
sample20	0.3181835823	9.5606462471	4.93941491

8. **Graphs→Histogram**
9. Select "mean" and click OK.
10. **Statistics→Summaries**
11. In the "**Numerical Summaries**" window, select "mean" and click **Statistics**
12. Check "Mean", "Standard Deviation"
13. Repeat steps 1-12 for sample size $n = 5$, and $n = 30$ (number of columns in "**Sample from Uniform Distribution**" window). Type set.seed(3921) each time before sampling data from the uniform distribution. Use the file names Uniformn5 and Uniformn30. Never write over a file.

$n = 2$	$n = 5$	$n = 30$
		
mean sd n 4.970168 2.039542 10000	mean sd n 4.968316 1.281207 10000	mean sd n 5.006232 0.5238191 10000

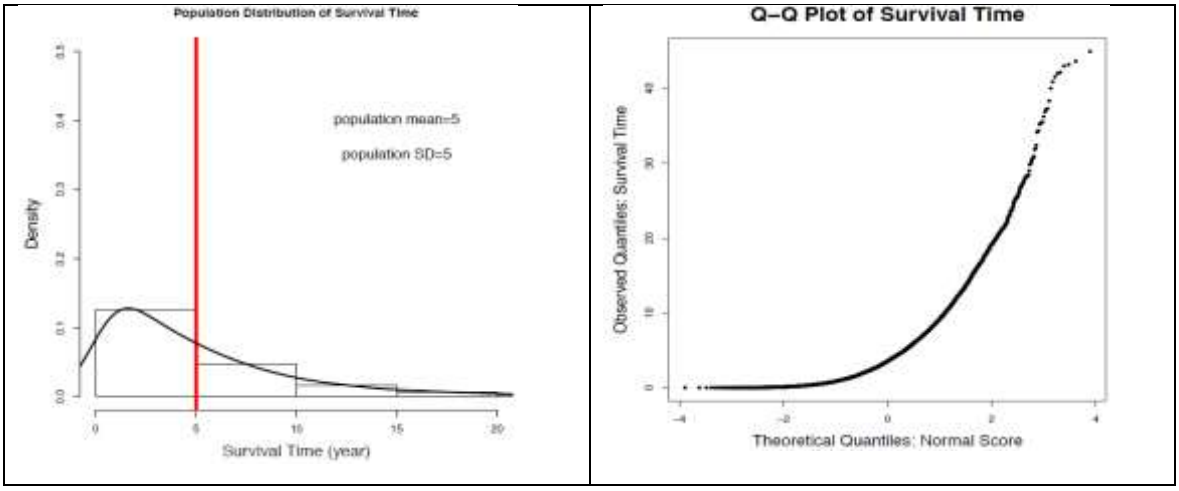
Findings:

- The mean of the sample mean is always very close to the population mean $\mu = 5$ regardless of the sample size n. The difference is because we did not consider all possible samples of size n, but only 10000 samples.
- The standard deviation of the sample mean is always close to the theoretical value $\frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{n}}$. When $n = 2, \frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{2}} = 2.044$; when $n = 5, \frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{5}} = 1.292$; when $n = 30, \frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{30}} = 0.528$.

- The population is symmetric, and the distribution of the sample mean is triangular when $n = 2$. The distribution of the sample mean appears to be normal for $n = 5$ and $n = 30$.

4.4 Distribution of the Sample Mean When the Population Distribution is Exponential

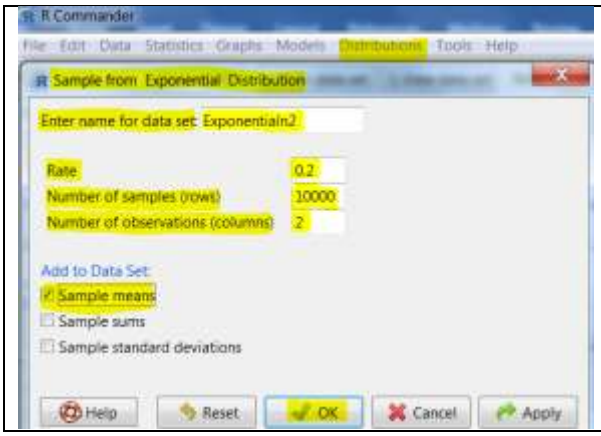
Suppose the survival time of liver cancer patients, X, follows an exponential distribution with mean and standard deviation 5 years, which is an extremely right skewed distribution.

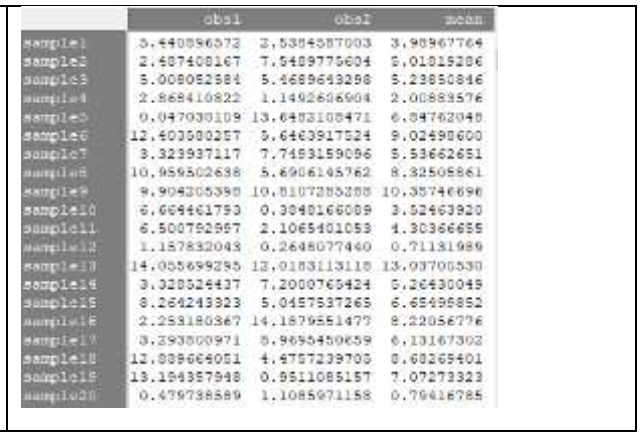


Let's examine the distribution of sample mean $\bar{X}$ with sample size $n = 2, 5, 30$ respectively. That is the distribution of the average of survival time of n randomly selected patients.

For each sample size n (# of columns), generate 10000 samples (# of rows). Use the seed 4518 for each n. Calculate the sample mean $\bar{x}$ for each sample by calculating the average of each row and store the value in the last column of the data file. Draw a histogram on the last column to obtain the distribution of the sample mean.

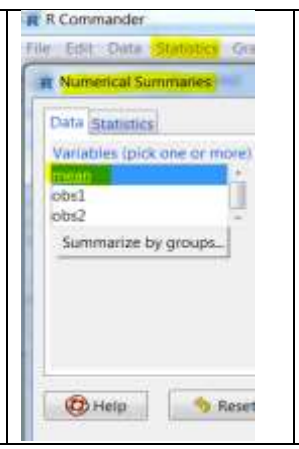
1. Type **set.seed(4518) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.
2. **Distributions→Continuous distributions→Exponential distribution→Sample from exponential distribution…**
3. In the "**Sample from Exponential Distribution**" window, type the name of the data file you would like to store the sampled data in **Enter name of data set** (say Exponentialn2), put 0.2 in **Rate**, 10000 in **Number of samples (rows)**, and 2 in **Number of observations (columns)**
4. Select **Sample means** under **Add to Data Set**, it will store the sample mean of the sample in the last column.
5. Click OK
6. Select Exponentialn2 under **Data set** to make it as active data set
7. Click **View data set** to view the sampled data

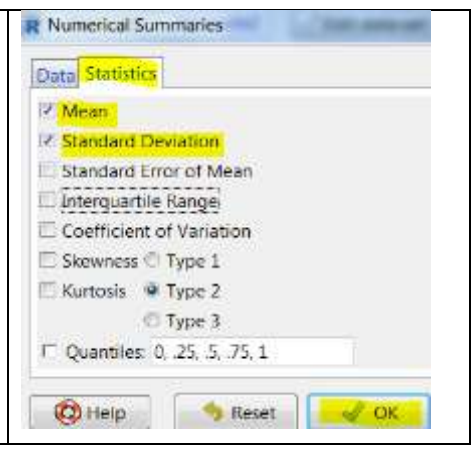


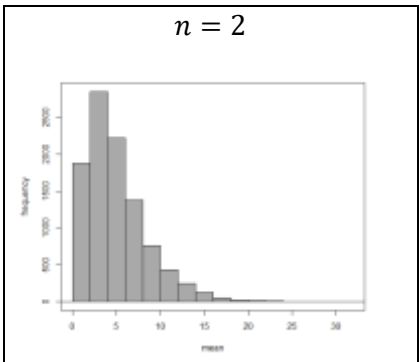
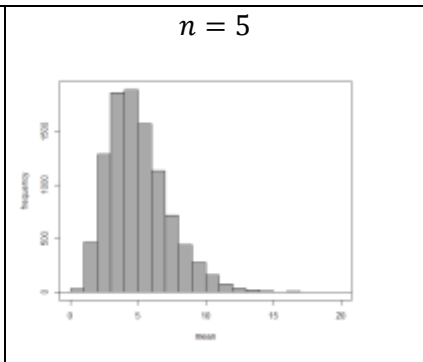
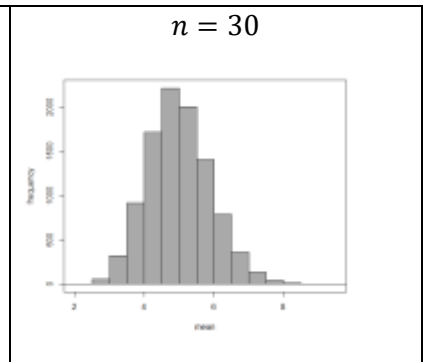


	obs1	obs2	mean
sample1	5.440896572	2.5384587003	3.98967764
sample2	2.487408167	7.5489779604	5.01815286
sample3	5.008052584	5.4689643298	5.23850846
sample4	2.868410822	1.1492606904	2.00883576
sample5	0.047030109	13.6483108471	6.84762048
sample6	12.403500257	5.6463917524	9.02490600
sample7	3.323937117	7.7453159096	5.53662651
sample8	10.959502638	5.6906145762	8.32505861
sample9	9.904205398	10.8107285288	10.35746696
sample10	6.664461793	0.3848166009	3.52463920
sample11	6.500792967	2.1065401053	4.30366655
sample12	1.157832043	0.2648077440	0.71131989
sample13	14.055699295	13.0183113118	13.03700530
sample14	3.320524437	7.2000765424	5.26430045
sample15	8.264243323	5.0457537265	6.65495852
sample16	2.253180367	14.1879551477	8.22056776
sample17	3.393500971	5.9695450659	4.13167302
sample18	12.039664051	4.4757239705	8.60265401
sample19	13.194357948	0.9511085157	7.07273323
sample20	0.479738589	1.1085971158	0.79416785

8. **Graphs→Histogram**
9. Select "mean" and click OK.
10. **Statistics→Summaries**
11. In the "**Numerical Summaries**" window, select "mean" and click **Statistics**
12. Check "Mean", "Standard Deviation"
13. Repeat steps 1-12 for sample sizes $n = 5$ and $n = 30$ (number of columns in "**Sample from Exponential Distribution**" window). Type set.seed(4518) and click submit each time before sampling data from the exponential distribution. Use the file names Exponentialn5 and Exponentialn30. Never write over a file.





$n = 2$	$n = 5$	$n = 30$
		
mean sd n 5.018273 3.562459 10000	mean sd n 5.028775 2.244574 10000	mean sd n 4.999087 0.9033595 10000

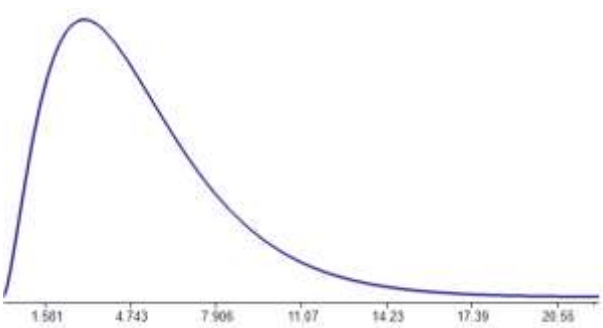
Findings:

- The mean of the sample mean is always very close to the population mean $\mu = 5$ regardless of the sample size n. The difference is because we did not consider all possible samples of size n, but only 10000 samples.
- The standard deviation of the sample mean is always close to the theoretical value $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{n}}$. When $n = 2, \frac{5}{\sqrt{n}} = \frac{5}{\sqrt{2}} = 3.536$; when $n = 5, \frac{5}{\sqrt{n}} = \frac{5}{\sqrt{5}} = 2.236$; when $n = 30, \frac{5}{\sqrt{n}} = \frac{5}{\sqrt{30}} = 0.913$.
- The population is extremely right skewed, and the distribution of the sample mean is still right skewed for the relatively small sample sizes of $n = 2$ and 5. But it is roughly normal when sample size $n \geq 30$.

4.5 Distribution of the Sample Mean When the Population Distribution is Chi-Square

The Chi-square distributions form a family of right skewed distributions where a parameter called "degrees of freedom" determines where the peak of the distribution is and how skewed the distribution is. The mean of the Chi-square distribution is equal to its number of degrees of freedom. The variance of a Chi-square distribution is equal to two times the number of its degrees of freedom. This distribution is used in Goodness of Fit Tests and in Tests of Independence (both of which we will work with later in the course) and is a distribution that can characterize magnetic resonance imaging data.

Suppose the random variable X, follows a chi-square distribution with 5 degrees of freedom. So, it has a mean μ = 5 and standard deviation σ = $\sqrt{\sigma^2}$ = $\sqrt{2 \times 5} = \sqrt{10}$ = 3.162278 (to 6 decimals). The density curve of the distribution is shown below.

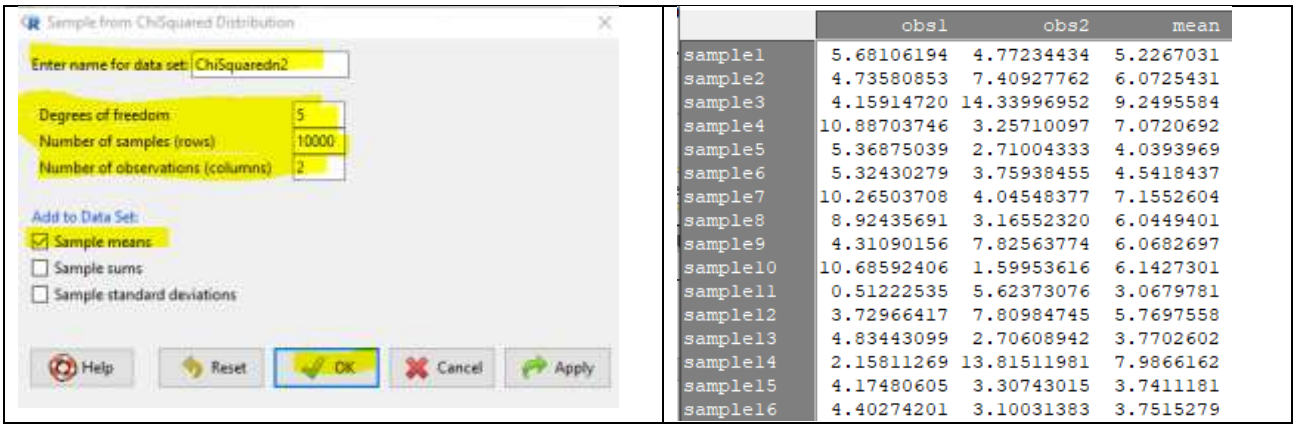


Let's examine the distribution of sample mean $\bar{X}$ with sample size $n = 2, 5, 30$ respectively.

For each sample size n (# of columns), generate 10000 samples (# of rows). Use the seed 6292 for each n. Calculate the sample mean $\bar{x}$ for each sample by calculating the average of each row and store the value in the last column of the data file. Draw a histogram on the last column to obtain the distribution of the sample mean.

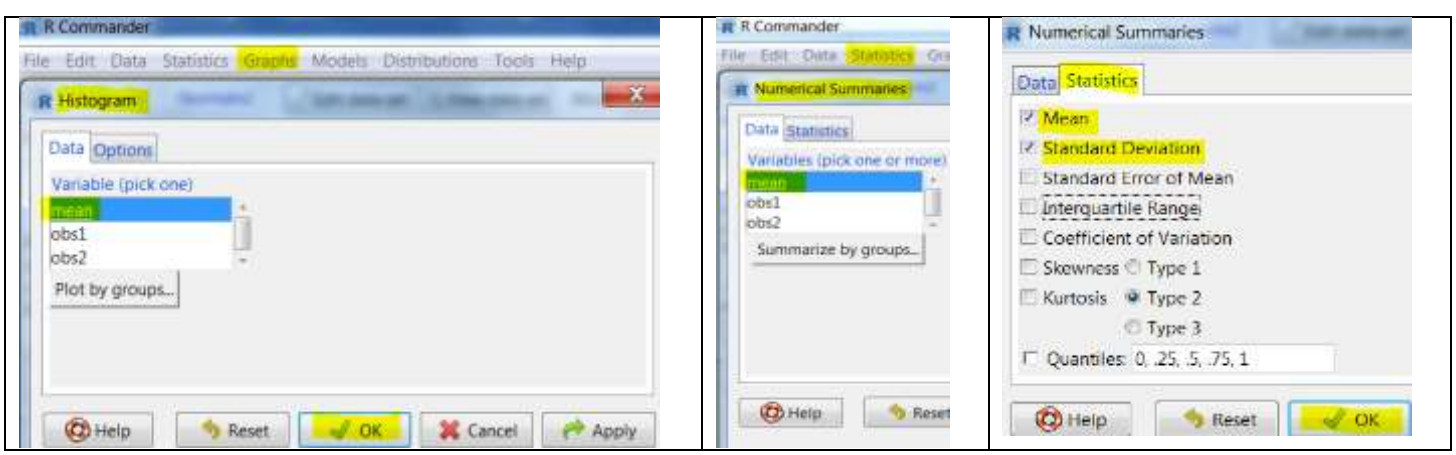
1. Type **set.seed(6292) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.

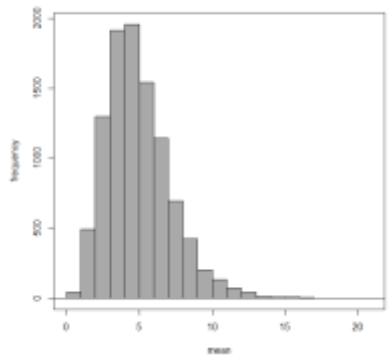
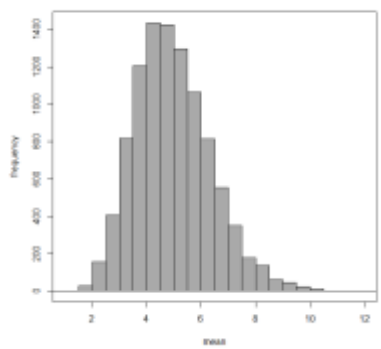
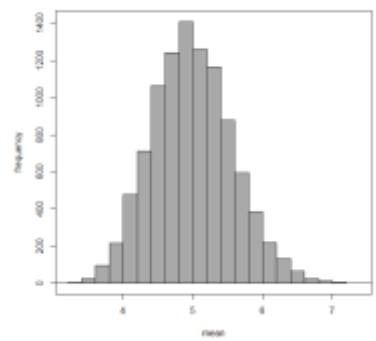
2. **Distributions→Continuous distributions→Chi-squared distribution→Sample from chi-squared distribution…**
3. In the "**Sample from ChiSquared Distribution**" window, type the name of the data file you would like to store the sampled data in **Enter name of data set** (say ChiSquaredn2), put 5 in **Degrees of Freedom**, 10000 in **Number of samples (rows)**, and 2 in **Number of observations (columns)**
4. Select **Sample means** under **Add to Data Set**, it will store the sample mean of the sample in the last column.
5. Click OK
6. Select ChiSquareln2 under **Data set** to make it as active data set
7. Click **View data set** to view the sampled data



	obs1	obs2	mean
sample1	5.68106194	4.77234434	5.2267031
sample2	4.73580853	7.40927762	6.0725431
sample3	4.15914720	14.33996952	9.2495584
sample4	10.88703746	3.25710097	7.0720692
sample5	5.36875039	2.71004333	4.0393969
sample6	5.32430279	3.75938455	4.5418437
sample7	10.26503708	4.04548377	7.1552604
sample8	8.92435691	3.16552320	6.0449401
sample9	4.31090156	7.82563774	6.0682697
sample10	10.68592406	1.59953616	6.1427301
sample11	0.51222535	5.62373076	3.0679781
sample12	3.72966417	7.80984745	5.7697558
sample13	4.83443099	2.70608942	3.7702602
sample14	2.15811269	13.81511981	7.9866162
sample15	4.17480605	3.30743015	3.7411181
sample16	4.40274201	3.10031383	3.7515279

8. **Graphs→Histogram**
9. Select "mean" and click OK.
10. **Statistics→Summaries**
11. In the "**Numerical Summaries**" window, select "mean" and click **Statistics**
12. Check "Mean", "Standard Deviation"
13. Repeat steps 1-12 for sample sizes $n = 5$ and $n = 30$ (number of columns in "**Sample from ChiSquared Distribution**" window). Type set.seed(6292) and click submit each time before sampling from the chi-square distribution. Use the file names Chisquaredn5 and Chisquaredn30. Never write over a file.



$n = 2$	$n = 5$	$n = 30$
		
mean sd n 4.948138 2.213075 10000	mean sd n 4.962986 1.390001 10000	mean sd n 4.99041 0.577546 10000

Findings:

- The mean of the sample mean is always very close to the population mean $\mu = 5$ regardless of the sample size n. The difference is because we did not consider all possible samples of size n, but only 10000 samples.
- The standard deviation of the sample mean is always close to the theoretical value $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2\times 5}}{\sqrt{n}}$. When $n = 2, \frac{\sqrt{10}}{\sqrt{n}} = \frac{3.162278}{\sqrt{2}} = 2.236$; when $n = 5, \frac{\sqrt{10}}{\sqrt{n}} = \frac{3.162278}{\sqrt{5}} = 1.414$; when $n = 30, \frac{\sqrt{10}}{\sqrt{n}} = \frac{3.162278}{\sqrt{30}} = 0.577$.
- The population is quite right skewed, and the distribution of the sample mean is still right skewed for the relatively small sample sizes of $n = 2$ and 5. But it is roughly normal when sample size $n \geq 30$.

4.6 Central Limit Theorem For the Sample Mean

The Central Limit Theorem (CLT) states that when the sample size n is large enough, the sample mean $\bar{X}$ is approximately normally distributed, regardless of the distribution of the variable under consideration (the population distribution).

Note that:

- The central limit theorem is about the **shape of the sample mean $\bar{X}$**. It is the random variable $\bar{X}$ that will be normally distributed if the sample size n is large enough.
- What constitutes a large enough value of n is depends on the shape of the population distribution. If the population distribution, the distribution of X, is symmetric, $n \geq 5$ might be large enough to claim that the sample mean $\bar{X}$ is normally distributed; if the distribution of X is not too extremely skewed, $n \geq 30$ should be enough; if the population is very skewed, we might need $n \geq 100$ (see the central limit theorem for proportion in the next section).

4.7 Central Limit Theorem For the Sample Proportion

Recall that the population mean $\mu = \frac{\sum x_i}{N}$, where N is the population size (number of individuals in the population), is a population parameter used to describe the population. The population proportion

$$p = \frac{\text{\# of individuals having a certain attribute}}{\text{population size}} = \frac{\text{\# of successes}}{N}$$

is another parameter used to describe the population.

For example, the proportion of female students at MacEwan is defined as

$$p = \frac{\text{\# of female students at MacEwan}}{\text{total number of students at MacEwan}} = \frac{\text{\# of successes}}{N},$$

where picking a female student is regarded as a success event.
Just as the sample mean $\bar{x} = \frac{\sum x_i}{n}$ is used to estimate the population mean μ, the sample proportion $\hat{p}$ which is defined as:

$$\hat{p} = \frac{\text{\# of individuals having a certain attribute in the sample}}{\text{sample size}} = \frac{\text{\# of successes in the sample}}{n}$$

is used to estimate the population proportion p.

Inference on the population mean μ is based on the distribution of the sample mean $\bar{X}$. Similarly, inference on the population proportion p is based on the distribution of the sample proportion $\hat{p}$.

Population proportion is defined as:

$$p = \frac{\text{\# of individuals having a certain attribute}}{\text{\# of individuals in thepopulation}} = \frac{\text{\# of successes}}{N}.$$

Population proportion can be regarded as a special population mean if we let the variable of interest be an indicator variable as follows:

$$x_i = \begin{cases} 1 & \text{if the } i\text{th individual has the attribute (a success)} \\ 0 & \text{if the ith individual does not have the attribute.} \end{cases}$$

Then the population proportion can be rewritten as:

$$p = \frac{\text{\# of individuals having a certain attribute}}{\text{\# of individuals in thepopulation}} = \frac{\text{\# of successes}}{N} = \frac{\sum X_i}{N}$$

The variable of interest X has only two possible values: 1 if the individual has the attribute and 0 if not. If we randomly select one individual, the probability that this individual has the attribute is p.

As a result, the probability distribution of X is:

x	1	0
$P(X = x)$	p	$1 - p$

with a population mean and population standard deviation:

$$\mu = \sum xP(X = x) = 1 \times p + 0 \times (1 - p) = p$$

$$\sigma = \sqrt{\sum x^2 P(X = x) - \mu^2} = \sqrt{1^2 \times p + 0^2 \times (1 - p) - \mu^2} = \sqrt{p - p^2} = \sqrt{p(1 - p)}.$$

When we take a simple random sample of size n, the proportion of individuals in the sample who have the specific attribute is the sample proportion (which can be regarded as a special sample mean $\bar{x}$).

$$\hat{p} = \frac{\text{\# of individuals having a certain attribute in the sample}}{\text{sample size}} = \frac{\text{\# of successes in the sample}}{n} = \frac{\sum x_i}{n} = \bar{x}$$

with $x_i = 1$ if the individual has the attribute and 0 if not.

Therefore, the sampling distribution of the sample proportion $\hat{p}$ has the following properties:

- **Center**: the mean of the sample proportion $\hat{p}$ equals the population mean μ; that is
$$\mu_{\hat{p}} = \mu = p.$$

- **Spread**: the standard deviation of the sample proportion $\hat{p}$ equals the population standard deviation σ divided by the square root of the sample size; that is
$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1 - p)}}{\sqrt{n}} = \sqrt{\frac{p(1 - p)}{n}}.$$

  These two results above are always true for any sample size n.

- **Shape**: The population distribution is non-normal. By the central limit theorem (CLT), however, $\hat{p}$ is approximately normal if n is large enough. The thumb of rule is to guarantee both $np \geq 5$ and $n(1 - p) \geq 5$, i.e., $n = \max\{\frac{5}{p}, \frac{5}{1-p}\}$, the larger value of $\frac{5}{p}$ and $\frac{5}{1-p}$. Some textbooks require both $np \geq 10$ and $n(1 - p) \geq 10$.

Central limit theorem for the sample proportion:

If the sample size n (rule of thumb: $np \geq 5$ and $n(1 - p) \geq 5$) is large enough, the **sample proportion** $\hat{p}$ is approximately normally distributed.

Suppose the population proportion is $p = 0.05$. By the rule of thumb, a sample size of at least

$$n = \max\left\{\frac{5}{p}, \frac{5}{1 - p}\right\} = \max\left\{\frac{5}{0.05}, \frac{5}{1 - 0.05}\right\} = \max\{100, 5.26\} = 100$$

is required to make the sample proportion $\hat{p}$ be normally distributed. A larger sample size n is required to make the sample proportion $\hat{p}$ to be approximately normally distributed when the population proportion is either closer to 0 or 1.

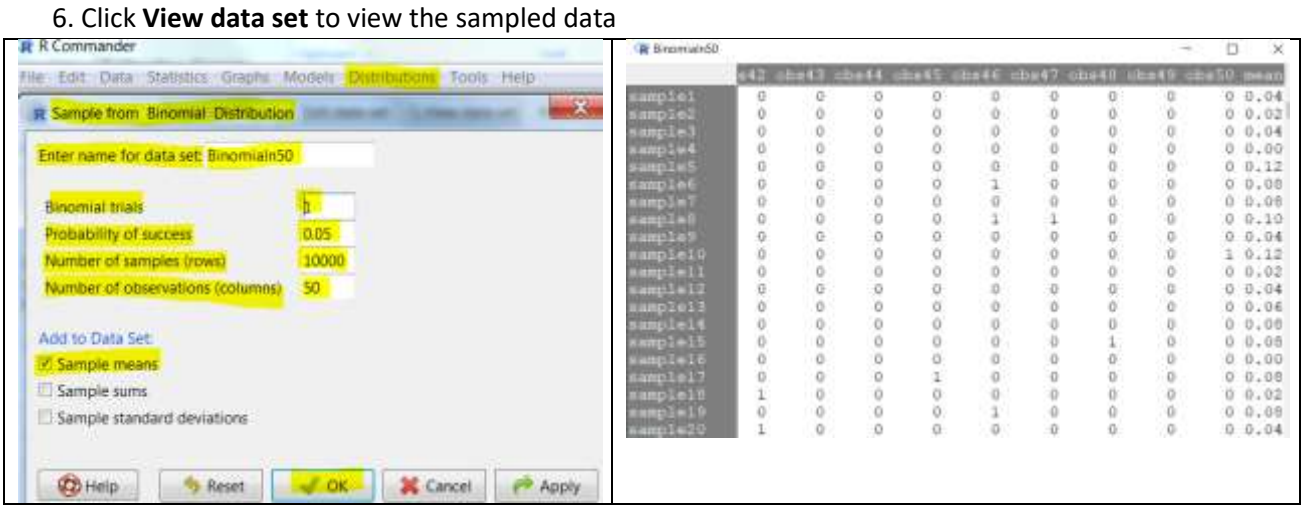
We can generate data from the population distribution $X = \begin{cases} 1 & \text{with probability } p = 0.05 \\ 0 & \text{with probability } 1 - p = 0.95, \end{cases}$

which is a special binomial distribution with number of trials $n = 1$ and probability of success $p = 0.05$.

For this population proportion distribution (where the attribute occurs with a probability of 0.05), we will investigate the sampling distribution of the sample proportion $\hat{p}$ with a sample size of $n = 50, 100, 200, 1000$ respectively. That is the distribution of the average number of individuals out of n randomly selected individuals who have a certain attribute.

For each sample size n (# of columns), generate 10000 samples (# of rows) sequence of 0s and 1s. Set the seed 59744 in each case. Calculate the sample mean for each sample by calculating the average of each row and store the value in the last column of the data file. Draw a histogram on the last column to obtain the distribution of the sample proportion.

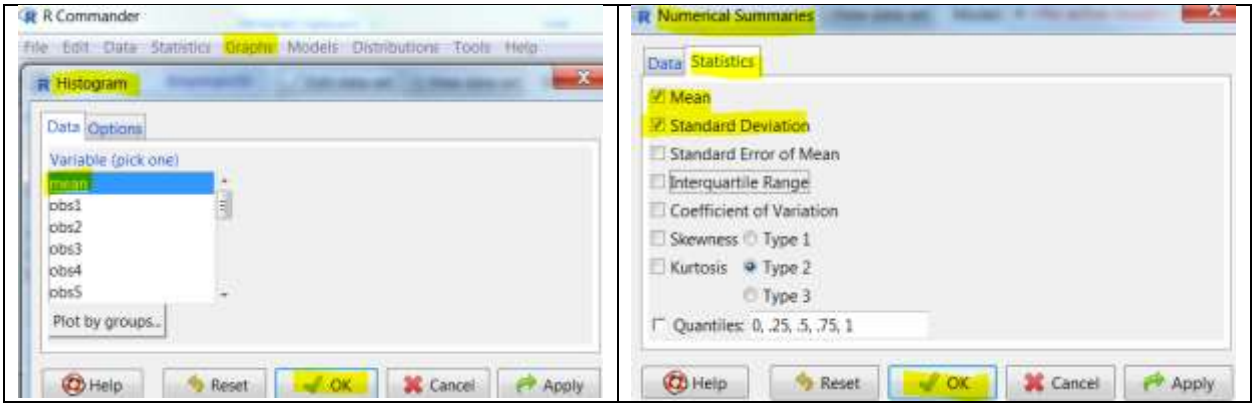
1. Type **set.seed(59744) in the R Script box (on its own line and flush against the left side of the box)**. Click Submit.
2. **Distributions→Discrete distributions→Binomial distribution→Sample from binomial distribution…**
3. In the "**Sample from Binomial Distribution**" window, type the name of the data file you would like to store the sampled data in **Enter name of data set** (say Binomialn50), put 1 in **Binomial trials**, 0.05 in **Probability of success**, 10000 in **Number of samples (rows)**, and 50 in **Number of observations (columns)**
4. Select **Sample means** under **Add to Data Set**, it will store the sample proportion of the sample in the last column. Click OK.
5. Select Binomialn50 under **Data set** to make it as active data set
6. Click **View data set** to view the sampled data



8. **Graphs→Histogram**
9. Select "mean" and click OK.
10. **Statistics→Summaries**
11. In the "**Numerical Summaries**" window, select "mean" and click **Statistics**

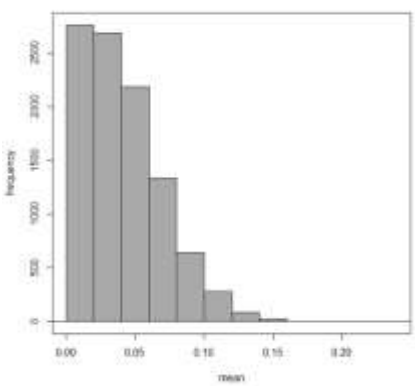
12. Check “Mean”, “Standard Deviation”

13. Repeat steps 1-11 for sample size $n = 100, 200, 1000$ (number of columns in “**Sample from Binomial Distribution**” window)



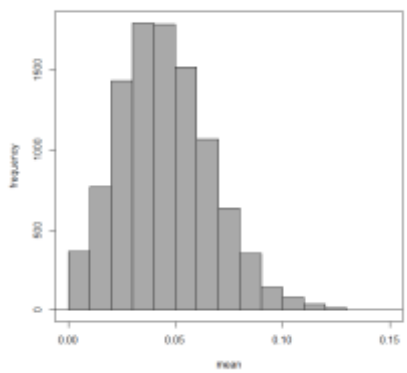
The following figures shows the sampling distribution of the sample proportion with different sample size n=50, 100, 200 and 1000.

$n = 50$



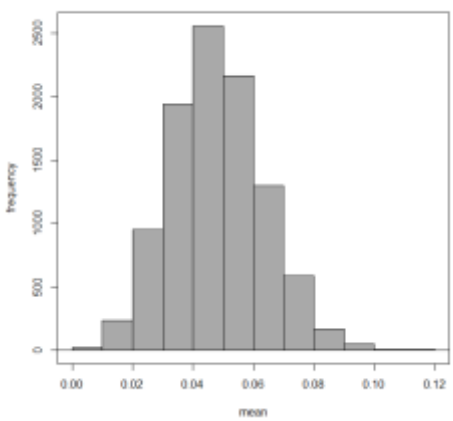

```
      mean          sd
  0.049926  0.03072212
```

$n = 100$



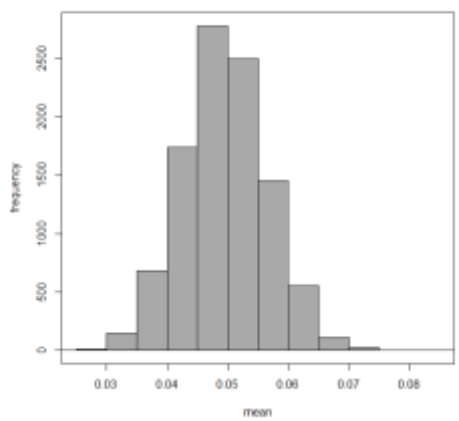

```
      mean          sd
   0.05008   0.0217518
```

$n = 200$




```
      mean          sd
 0.0503595  0.01547731
```

$n = 1000$




```
      mean          sd
 0.0500404  0.006898415
```

Findings:

- The mean of the sample proportion is always very close to the population proportion $p = 0.05$ regardless of the sample size n. The difference is because we did not consider all possible samples of size n, but only 10000 samples.
- The standard deviation of the sample proportion is always close to the theoretical value $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05(1-0.05)}{n}}$. When $n = 50$, $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05(1-0.05)}{50}} = 0.0308$; when $n = 100$, $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05(1-0.05)}{100}} = 0.0218$; when $n = 200$, $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05(1-0.05)}{200}} = 0.0154$; when $n = 1000$, $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05(1-0.05)}{1000}} = 0.0069$
- The population is extremely right skewed, and the distribution of the sample proportion is still right skewed for relatively large sample sizes $n = 50$. It is still slightly right skewed when $n = 100$, even though $n = 100$ should large enough according to the rule of thumb. But it is roughly normal when sample size $n = 200$ and 1000.

Recall that the central limit theorem tells us that the sample mean $\bar{X}$ will be approximately normally distributed when the sample size n is large enough. The rule of thumb is $n \geq 30$. However, how large n is large enough to make the sample mean $\bar{X}$ be normally distributed depends on how far the population distribution departs from a normal distribution; the further the population distribution is away from a normal distribution, the larger the sample size n is required. If the population distribution is continuous and not extremely skewed, n=30 should be large enough; however, if the population distribution is discrete (like the Bernoulli distribution for sample proportion), a much larger n is required, say n=200 or more.

LAB 5 CONFIDENCE INTERVAL AND HYPOTHESIS TESTS FOR ONE MEAN

There are two types of statistics: descriptive and inferential statistics. We will focus on inferential statistics hereafter. Inferential statistics include estimation and hypothesis testing. Estimation is to estimate the value of a population parameter; hypothesis testing is to test whether a statement about the value of a population parameter is true or false. This lab illustrates how to obtain a confidence interval and conduct a hypothesis test for the population mean μ based on one simple random sample.

A general form for a confidence interval for a population parameter is

$$point\ estimate \pm error = point\ estimate \pm mulitplier \times Standard\ Error\ of\ the\ estimator.$$

General steps to set up the hypotheses:

1. Look for the key words, write down what we want to claim under the alternative H_a.
2. Take the opposite of the alternative H_a to obtain the null H_0.
   Depending on the purpose of the hypothesis test, there are three choices for H_a:

Two tailed	**Right (upper) tailed**	**Left (lower) tailed**
$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
"differ", "change"	"more than", "increase"	"less than", "decrease"

Depending on whether the population standard deviation σ is known or not, we can use the one-sample z test and interval or the one-sample t test and interval.

5.1 ONE-SAMPLE z TEST AND INTERVAL WHEN THE POPULATION STANDARD DEVIATION IS KNOWN

Use the one-sample z test and z interval when the population standard deviation σ is known. The assumptions and steps to conduct a one-sample z test and z interval are as follows.

Assumptions:

1. A simple random sample (SRS)
2. Normal population or large sample size ($n \geq 30$)
3. The population standard deviation σ is known

Steps:

1. Set up the hypotheses:

$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$

2. State the significance level α.
3. Compute the value of the test statistic: $z_o = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
4. Find the P-value or rejection region:

	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
P-value	$2P(Z \geq \lvert z_o \rvert)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$	$Z \geq z_\alpha$	$Z \leq -z_\alpha$

$z_{\alpha/2}$ is the z value for which $\mathrm{P}(Z > z_{\alpha/2}) = \alpha/2$

5. Reject the null H_0 if P-value$\leq \alpha$ or z_o falls in the rejection region.
6. Conclusions.

A corresponding $(1-\alpha)\times 100\%$ one-sample z confidence interval is given by

	Two-sided Interval	**Upper Tailed Interval**	**Lower Tailed Interval**
	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
$(1-\alpha)\times 100\%$ **CI**	$(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$	$(\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$	$(-\infty, \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}})$
Decision	**Reject H_0 if μ_0 is outside the interval**		

Interpretation of the confidence interval: we can be $(1-\alpha)\times 100\%$ confident that the population mean μ is within the interval.

5.2 One-Sample t Test and Interval When the Population Standard Deviation is Unknown

Given that the population is normal **OR** the sample size n is large enough, the sample mean $\bar{X}$ can be regarded to be normally distributed, i.e., $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

The population standard deviation σ is usually unknown and can be estimated by the sample standard deviation s.

When the population distribution is normal, the standardized variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.

When the population distribution is normal, the studentized variable

$$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t \text{ distribution with } df = n-1.$$

The assumptions and steps to conduct a one-sample t test and t interval for one population mean μ are as follows.

Assumptions:

1. A simple random sample (SRS)
2. Normal population or large sample size ($n \geq 30$)
3. The population standard deviation σ is unknown

Steps:

1. Set up the hypotheses:

$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$

2. State the significance level α.
3. Compute the value of the test statistic: $t_o = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with $df = n-1$.
4. Find the P-value or rejection region:

	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
P-value	$2P(t \geq \lvert t_o \rvert)$	$P(t \geq t_o)$	$P(t \leq z_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_\alpha$	$t \leq -t_\alpha$

$t_{\alpha/2}$ is the t value for which $\mathrm{P}(t > t_{\alpha/2}) = \alpha/2$.

5. Reject the null H_0 if P-value$\leq \alpha$ or t_o falls in the rejection region.
6. Conclusions.

	Two-sided Interval for Two-sided Test	**Upper Tailed Interval for Right Tailed Test**	**Lower Tailed Interval for Left Tailed Test**
	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$
$(\mathbf{1-\alpha}) \times \mathbf{100\%}$ **CI**	$(\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$	$(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}, \infty)$	$(-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}})$
Decision	**Reject H_0 if μ_0 is outside the interval**		

Interpretation of the confidence interval: we can be $(1-\alpha) \times 100\%$ confident that the population mean μ is within the interval.

NUANCE: Students should note that although the Central Limit Theorem tells us that for any unknown population distribution shape with large n, the sampling distribution of Z= $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately normal, it actually does not tell us that for any unknown population distribution shape with large n, t=$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ is approximately normal or approximately a t distribution.

However, it is sensible to think that s will be close to sigma (a good estimate) when n is large, and therefore that t values calculated will be close to z values when n is large. So it is not untoward to think that $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ values will be approximately $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ values for large n and thus the sampling distribution will indeed have a normal shape (regardless of the parent population shape).
We also note that a t distribution with n – 1 degrees of freedom is approximately normal for large n.

Some textbooks suggest that students doing problems that entail finding the test statistic t = $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ proceed to calculate p-values and rejection region critical values using the standard normal distribution, Z. This is useful because t tables are not comprehensive when n >=30.

Other textbooks suggest that students doing problems that entail finding the test statistic t = $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ use the t distribution with n – 1 df to calculate p-values and rejection region critical values when n is large. This is generally just fine because these values can be readily calculated online.

The software R Commander finds p values and critical values for a t distribution with n – 1 degrees of freedom when you use it to do a single sample t test.

Example: A machine fills beer into bottles whose volume is supposed to be 341 ml, but the exact amount varies from bottle to bottle. We randomly pick 50 bottles and actual volume of each bottle is given in the data file. The sample mean volume is 338.428 ml and sample standard deviation $s = 5.238$ ml.

343.8	339.8	347.3	348.4	338.1	333.1	345.8	342.7	341.0	336.5
338.6	337.8	339.2	341.7	339.0	343.0	333.4	332.8	337.1	338.0
338.8	331.3	343.6	331.8	338.4	345.3	333.7	344.4	337.0	347.0
336.0	341.4	330.5	328.7	340.8	337.4	336.9	326.4	344.3	329.2
334.4	339.6	341.5	334.2	333.0	337.8	343.3	337.4	346.4	333.8

Note: the data were generated from a normal distribution with mean 339 and standard deviation 5 with random number generator seed 4067, rounded to one decimal place.

For this problem, please download the dataset beer.xlsx from online. Then import it into R commander and called it beer, say.

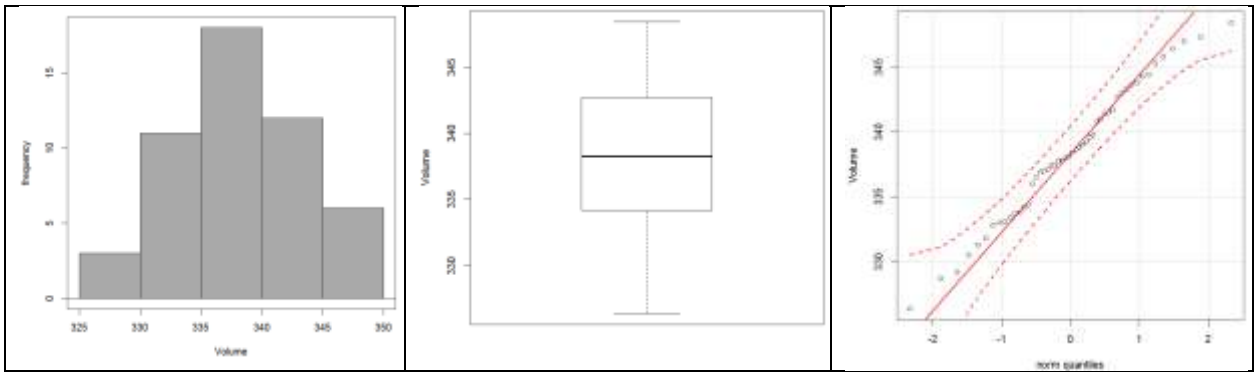
(a) Test at the 5% significance level whether the machine is NOT working properly.
(b) Obtain a 95% confidence interval for the population mean volume. Interpret the interval.
(c) Does the confidence interval obtained in part (b) support the conclusion of the test in part (a)?
(d) Test at the 1% significance level whether the mean volume is below 341 ml.

Check the assumptions:

- We have a simple random sample.
- We have a large sample with sample size $n = 50 > 30$; therefore, it does not matter whether the population is normal or not. However, we can draw a normal probability (Q-Q) plot, a histogram, and a boxplot to check the normality of the sample data. For your imported dataset called beer, use the **Graphs→Histogram, Graphs→boxplot, and Graphs→Quantile-comparison plot** commands. All the graphs of the summarized sample data shown below suggest (or do not contradict) that the sample data was taken from a normal population.

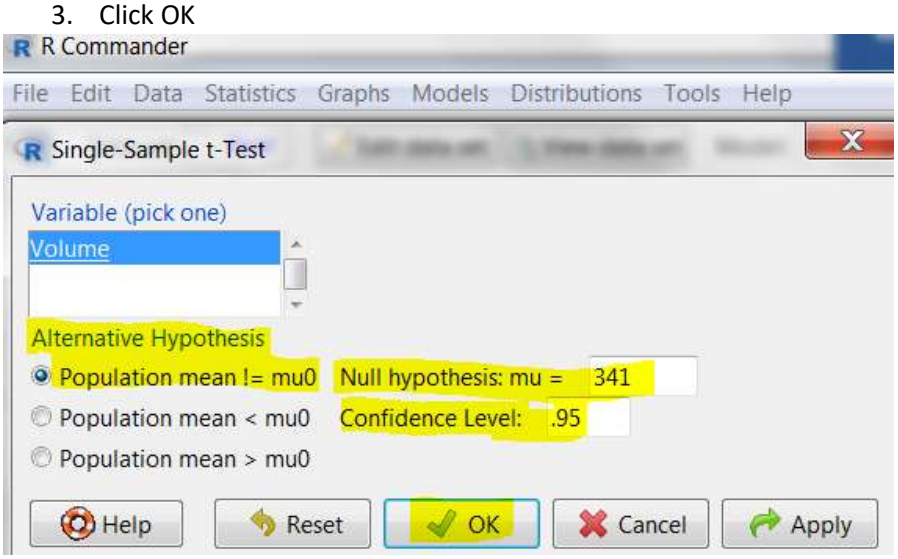
  Please note that the best way to check the normality assumption is a normal Q-Q plot, especially when the sample size is not very large. In general, a boxplot cannot show whether the data are from a normal population. A histogram can be misleading and cannot show whether the data have a bell-shaped distribution when the sample size is not large enough.
- The population standard deviation σ is unknown.

The assumptions for a one-sample t test are met.



To run a one-sample t test in R Commander:

1. **Statistics→Means→Single-sample t test**
2. In the "**Single-Sample t-Test**" window, pick "**Volume**" as the variable. Choose the alternative hypothesis: two-tailed (!=mu0). Specify the hypothesized value "**mu=341**", i.e., $\mu_0 = 341$. Specify the "**Confidence Level: 0.95**", i.e., the significance level $\alpha = 0.05$.
3. Click OK




```
        One Sample t-test

data:  Volume
t = -3.4718, df = 49, p-value = 0.001089
alternative hypothesis: true mean is not equal to 341
95 percent confidence interval:
 336.9392 339.9168
sample estimates:
mean of x
  338.428
```

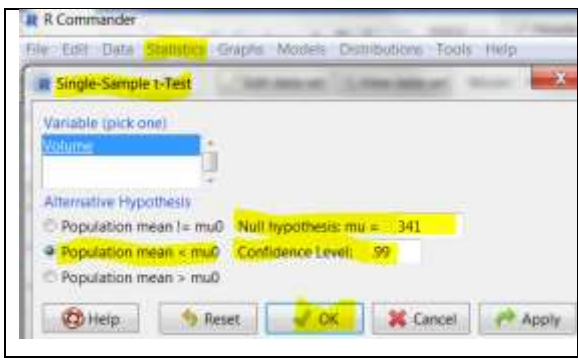
(a) Test at the 5% significance level whether the machine is NOT working properly.
If the machine is working properly, $\mu = 341$ ml; if the machine is not working properly, $\mu \neq 341$ ml.
The steps for a one-sample t test are:

- Hypotheses. $H_0: \mu = 341$ ml versus $H_a: \mu \neq 341$ ml
- The significance level is $\alpha = 0.05$.
- Compute the value of the test statistic: $t_o = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = -3.4718$, with $df = n - 1 = 49$
- The P-value= $2P(t \geq |t_o|) = 2P(t \geq 3.4718) = 0.001089$
- Since P-value=$0.001089 < 0.05$ (α), reject H_0.
- Conclusion: At the 5% significance level, the data provide sufficient evidence that the machine is NOT working properly.

(b) Obtain a 95% confidence interval for the population mean volume. Interpret the interval.
A 95% confidence interval for the population mean volume is (336.9392, 339.9168) ml.
Interpretation: we can be 95% confident that the population mean volume μ is somewhere between 336.9392 ml and 339.9168 ml.

(c) Does the confidence interval obtained in part (b) support the conclusion of the test in part (a)?
Yes. In part (a), we reject H_0 and claim that the machine is not working properly, i.e., $\mu \neq 341$ ml. In part (b), the interval does not contain 341; therefore, we can be 95% confident that $\mu \neq 341$ ml and it supports the conclusion of the hypothesis test in part (a).

(d) Test at the 1% significance level whether the mean volume is **below** 341 ml.



Output

```
        One Sample t-test

data:  Volume
t = -3.4718, df = 49, p-value = 0.0005447
alternative hypothesis: true mean is less than 341
99 percent confidence interval:
     -Inf 340.2096
sample estimates:
mean of x
  338.428
```

- Hypotheses. $H_0: \mu \geq 341$ ml versus $H_a: \mu < 341$ ml
- The significance level is $\alpha = 0.01$.
- Compute the value of the test statistic: $t_o = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = -3.4718$, with $df = n - 1 = 49$
- The P-value= $P(t < t_o) = P(t < -3.3718) = P(t \geq 3.4718) = 0.0005447$
- Since P-value=$0.0005447 < 0.01$ (α), reject H_0.
- Conclusion: At the 1% significance level, the data provide sufficient evidence that the mean volume is below 341 ml.

(e) Obtain a confidence interval corresponding to the test in part (d). Does the interval support the conclusion of the test in part (d)?

A left-tailed test at the 1% significant level corresponds to a 99% lower-tailed confidence interval. A 99% lower-tailed confidence interval for the population mean volume is $(-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}}) = (-\infty, 340.2096)$. Interpretation: we can be 99% confident that the population mean volume μ is somewhere below 340.2096 ml. Since the entire interval is below 341, we can claim that $\mu < 341$ ml. This supports the conclusion of the hypothesis test in part (d).

5.3 Relation Between Confidence Interval and Hypothesis Tests

Recall:

Two-sided confidence intervals correspond to two-tailed tests, upper-tailed confidence intervals correspond to right-tailed tests, and lower-tailed confidence intervals correspond to left-tailed tests.

A $(1-\alpha) \times 100\%$ two-sided t confidence interval is given in the form $(\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$.

A $(1-\alpha) \times 100\%$ upper-tailed t confidence interval is given by $(\bar{x} - t_\alpha \frac{s}{\sqrt{n}}\ , \infty)$ and the number $\bar{x} - t_\alpha \frac{s}{\sqrt{n}}$ is called the lower bound of the interval.

A $(1-\alpha) \times 100\%$ lower-tailed t confidence interval is given by $(-\infty, \bar{x} + t_\alpha \frac{s}{\sqrt{n}})$ and the number $(\bar{x} + t_\alpha \frac{s}{\sqrt{n}})$ is called the upper bound of the interval.

Remember:

We can use the confidence intervals to make conclusions about hypothesis tests: reject the null hypothesis H_0 at the significance level α if the corresponding $(1-\alpha) \times 100\%$ confidence interval does not contain the hypothesized value μ_0.

Confidence interval (CI) and hypothesis testing (HT) should give consistent results: we should not reject H_0 at the significance level α if the corresponding $(1-\alpha) \times 100\%$ confidence interval contains the hypothesized value μ_0.

LAB 6 CONFIDENCE INTERVAL & HYPOTHESIS TESTS FOR TWO MEANS

Suppose we have two populations with means μ_1 and μ_2 respectively. This lab covers how to obtain a confidence interval and conduct a hypothesis test for the difference between the two population means, i.e., $\mu_1 - \mu_2$, using R commander. Depending on whether the two samples are independent or paired, we have a two-sample t test or a paired t test, respectively.

6.1 TWO-SAMPLE t TEST AND t INTERVAL BASED ON TWO INDEPENDENT SAMPLES

The two-sample t test can be used to test hypotheses on the difference between two population means. Depending on whether the two population standard deviations (σ_1 and σ_2) are equal or not, we use the non-pooled and pooled two sample t test and t interval, respectively. Minor advantages of the pooled t test are that it provided a slightly narrower confidence interval, a slightly more powerful test, and a simpler formula for the degrees of freedom. However, a pooled t test is valid only when the two population standard deviations are identical; otherwise, it gives invalid results. Therefore, we recommend using the non-pooled t test unless we are very confident that $\sigma_1 = \sigma_2$ (which is very difficult to verify).

6.1.1 Non-pooled Two-Sample t Test and t Interval

Assumptions:

1. Simple random samples;
2. Two samples are independent;
3. Normal populations or large sample sizes (rule of thumb: $n_1 \geq 30, n_2 \geq 30$).

Steps:

1. Set up the hypotheses:

Two tailed test	Right (upper) tailed test	Left (lower) tailed test
$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_0: \mu_1 - \mu_2 \geq \Delta_0$
$H_a: \mu_1 - \mu_2 \neq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$

Note that Δ_0 can be zero or any value you would like to test.

2. State the significance level α.
3. Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ with $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$ rounded **down** to the nearest integer, i.e., take the integer part.
4. Find the P-value or rejection region:

	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_0: \mu_1 - \mu_2 \geq \Delta_0$
	$H_a: \mu_1 - \mu_2 \neq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$
P-value	$2P(t \geq \lvert t_o \rvert)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_\alpha$	$t \leq -t_\alpha$

5. Decision: reject the null H_0 if P-value$\leq \alpha$ or if t_o falls in the rejection region.
6. Conclusions.

A $(1-\alpha)\times 100\%$ two-sample t confidence interval for $\mu_1-\mu_2$ is:

	Two-sided Interval for Two-sided Test	**Upper Tailed Interval for Right Tailed Test**	**Lower Tailed Interval for Left Tailed Test**
	$H_0: \mu_1-\mu_2=\Delta_0$	$H_0: \mu_1-\mu_2\leq\Delta_0$	$H_0: \mu_1-\mu_2\geq\Delta_0$
	$H_a: \mu_1-\mu_2\neq\Delta_0$	$H_a: \mu_1-\mu_2>\Delta_0$	$H_a: \mu_1-\mu_2<\Delta_0$
Interval	$(\bar{x}_1-\bar{x}_2)\pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$((\bar{x}_1-\bar{x}_2)-t_{\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\infty)$	$(-\infty,(\bar{x}_1-\bar{x}_2)+t_{\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}})$
Decision	**Reject H_0 if Δ_0 is outside the interval**		

Example: Two-sample t Test and t Interval Assuming Standard Deviations Not Equal

Some students attend class regularly, but some do not. An instructor wants to compare the class average for those who attend lectures regularly (μ_1) with those who do not (μ_2). Simple random samples are randomly selected from attendees and non-attendees. Their attending status (Attend/Non-Attend) and final grade (in %) are given in the following table (grades to 2 decimals places). Data are stored in "example_twosample_grade.xlsx", which can be found online, and has 13 decimal places for grades. Note that practising students should download the online file and use it, as typing or copying/pasting the data shown below to their own Excel file and using it (with grades to 2 decimal places) will not yield the answers found in the descriptions and inference done below.

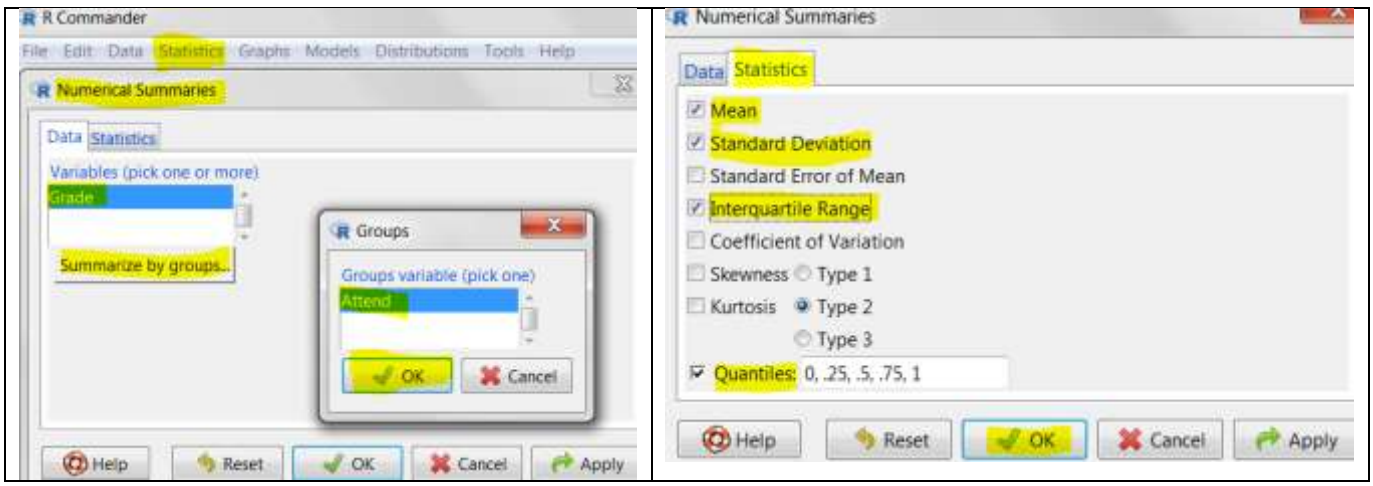
Attend	69.68	Attend	77.56	Attend	65.03	Attend	89.30	Attend	87.75
Non-Attend	61.21	Non-Attend	64.76	Non-Attend	65.04	Attend	68.54	Non-Attend	35.62
Attend	80.43	Attend	66.01	Attend	57.08	Attend	71.24	Attend	96.51
Attend	80.97	Attend	78.10	Attend	95.86	Attend	49.19	Non-Attend	65.81
Non-Attend	60.74	Attend	95.54	Attend	83.32	Non-Attend	39.30	Attend	82.82
Attend	83.34	Attend	67.50	Attend	74.53	Non-Attend	78.46	Attend	83.00
Attend	72.03	Attend	93.30	Attend	55.24	Attend	81.23	Non-Attend	42.94
Non-Attend	77.11	Attend	85.03	Attend	76.27	Non-Attend	80.67	Attend	80.14
Attend	75.49	Non-Attend	82.50	Attend	74.76	Non-Attend	64.32	Attend	79.47
Attend	75.03	Non-Attend	54.10	Attend	61.58	Attend	47.77	Attend	72.49
Attend	90.86	Non-Attend	78.71	Attend	62.91	Attend	93.26	Non-Attend	85.07
Attend	86.87	Non-Attend	47.67	Non-Attend	51.30	Non-Attend	70.68	Non-Attend	55.65
Attend	96.32	Attend	76.51	Attend	77.06	Attend	68.40	Attend	72.66
Attend	50.62	Attend	85.97	Attend	80.24	Non-Attend	39.39	Attend	90.25
Attend	83.13	Attend	85.19	Attend	70.16	Attend	68.12	Attend	87.75
Non-Attend	72.80	Attend	78.40	Attend	66.06	Attend	86.51	Non-Attend	55.38
Attend	71.22	Non-Attend	67.34	Non-Attend	42.39	Non-Attend	87.30	Non-Attend	80.88

(a) Use the proper descriptive statistics tools (figures and numerical summaries) to summarize the data.
(b) Test at the 1% significance level whether those who attend lectures have a **higher average**, i.e., $\mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$.
(c) Obtain a confidence interval for the difference between the class average for attendees and non-attendees, $\mu_1 - \mu_2$, that corresponds to the test in part (b).
(d) Based on the interval obtained in part (c), can we claim that the class average of attendees is at least 5% higher than that of the non-attendees? How about 10% higher?

Solutions:

(a) Use the proper descriptive statistics tools (figures and numerical summaries) to summarize the data. We want to compare the grade between attendants and non-attendants. Note that grade is a quantitative continuous variable. Hence, to compare the two groups numerically, we use the five-number summary (min, Q_1, median, Q_3, max) , mean and standard deviation for each group, while graphically, we use a side-by-side histogram and/or a side-by-side boxplot.

1. **Statistics→Summaries→Numerical Summaries...**
2. In the "**Numerical Summaries**" window, select "**Grade**" as the variable.
3. Click "**Summarize by groups…**", in the "**Groups**" window, choose "**Attend**" as the grouping variable. Click OK
4. Click "**Statistics**", check "**Mean**", "**Standard Deviation**", "**Interquartile Range**", and "**Quantiles**", click OK.




```
              mean       sd      IQR       0%      25%      50%      75%     100% data:n
Attend     76.92475 11.82723 15.35319 47.76920 69.79938 77.83021 85.15257 96.51480     58
Non-Attend 63.22769 15.48007 25.08418 35.62145 52.69786 64.76214 77.78204 87.29546     27
```

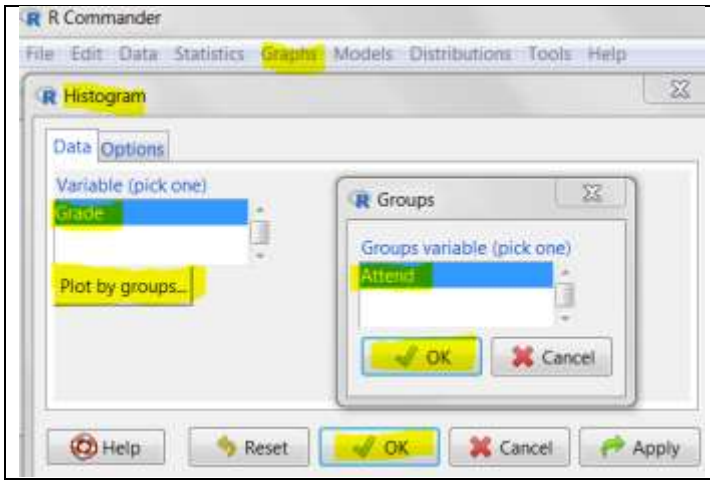
Here are the findings from the numerical summaries:

1. There are $n_1 = 58$ attendees and $n_2 = 27$ non-attendees.
2. The sample mean for the attendees is $\bar{x}_1 = 76.925\%$. The sample mean for the non-attendees is $\bar{x}_2 = 63.228\%$, which is 13.697% lower than the mean of the attendees. The attendees also have a larger median (50% quantile) than their non-attendees counterpart, 77.830% versus

64.762% (that is; the median for the non-attendees is 13.068% lower than the median for the attendees).

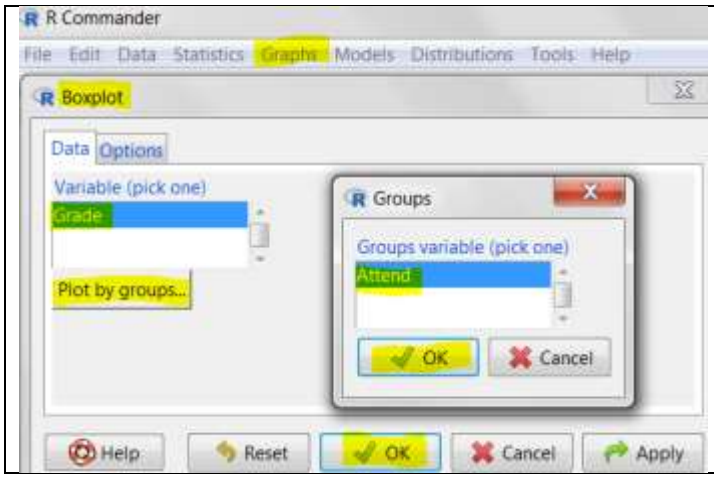
3. The sample standard deviation for the attendees is $s_1 = 11.827\%$ and the sample standard deviation for the non-attendees is $s_2 = 15.480\%$. There is a larger variation in grade among non-attendees. This can be also found through the IQR. The IQR is 15.353% for attendees and 25.084% for non-attendees.
4. The attendees have a larger maximum grade than non-attendees, 96.514% versus 87.295%; the attendees also have a higher minimum grade, 47.769% versus 35.621%.

All the findings above can be also seen from the plots created below.



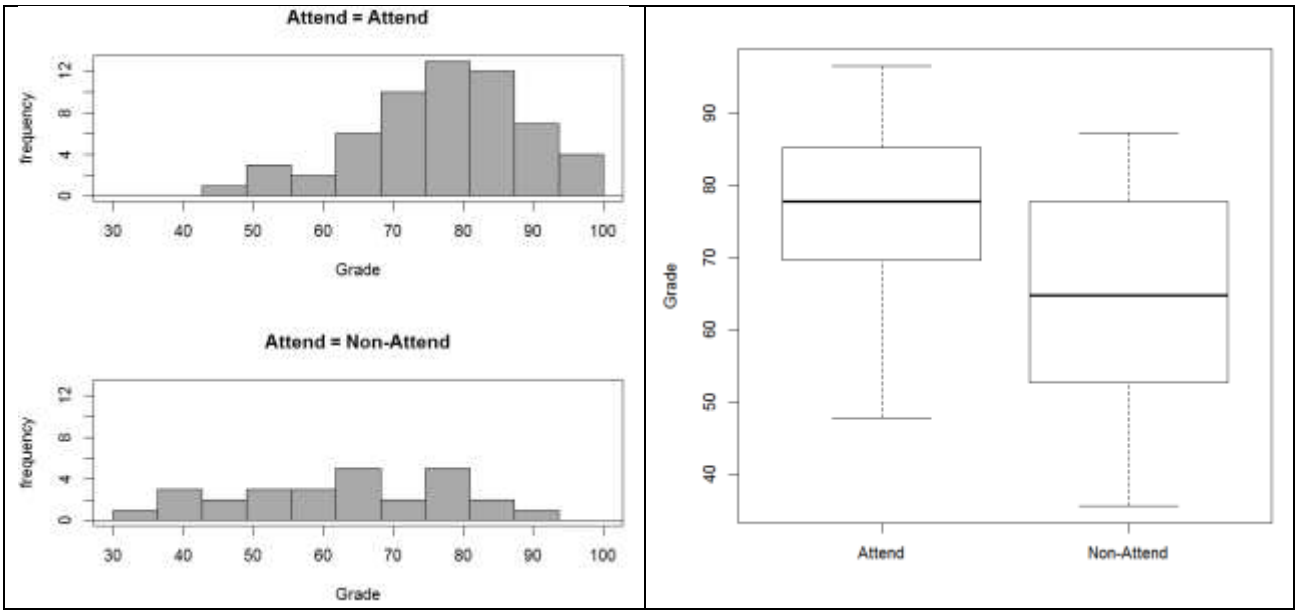
Steps for side-by-side histogram

1. **Graphs→Histogram...**
2. In the "**Histogram**" window, select "**Grade**" as the variable
3. Click "**Plot by groups…**", in the "**Groups**" window, choose "**Attend**" as the grouping variable. Click OK
4. Click OK



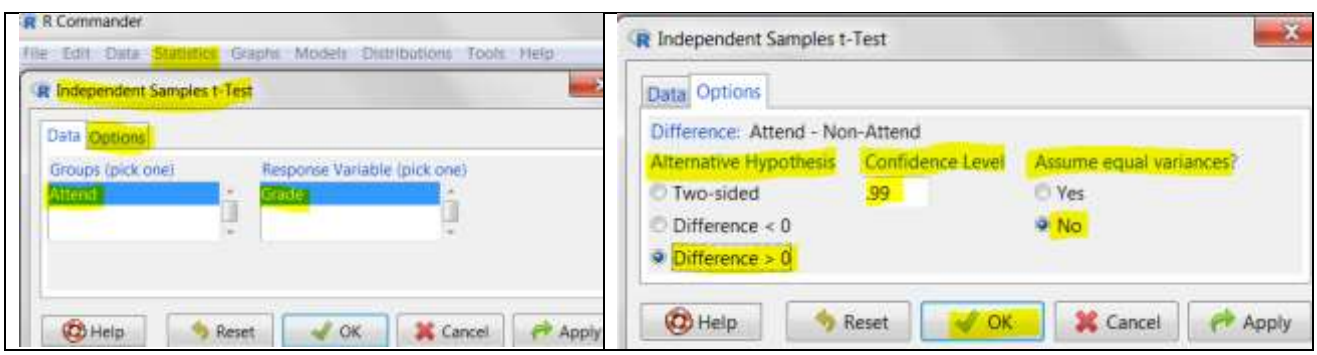
Steps for side-by-side boxplot

1. **Graphs→Boxplot...**
2. In the "**Boxplot**" window, select "**Grade**" as the variable
3. Click "**Plot by groups…**", in the "**Groups**" window, choose "**Attend**" as the grouping variable. Click OK
4. Click OK



(b) **Test at the 1% significance level whether** those who attend lectures have a **higher average**, i.e., $\mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$.
Use a two-sample t test since the samples (the attendees and the non-attendees) are independent.

1. **Statistics→Means→Independent Sample t-Test...**
2. In the "**Independent Sample t-Test**" window, select "**Attend**" as the grouping variable and "**Grade**" as **the** response variable, since we want to compare the grades between attendants and non-attendants.
3. Click "**Options**", in the "**Options**" window, choose "**Difference>0**" as the **Alternative Hypothesis**, because we want to test whether $\mu_1 > \mu_2$ or the difference $\mu_1 - \mu_2 > 0$. Type **0.99** in the box under "**Confidence Level**", since the significance level $\alpha = 0.01$ which corresponds to a confidence level $1 - \alpha = 1 - 0.01 = 0.99$. Check "**No**" under "**Assume equal variances**" for a non-pooled two-sample t test. Click OK.
4. Click OK



```
        Welch Two Sample t-test

data:  Grade by Attend
t = 4.077, df = 40.68, p-value = 0.0001032
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 5.561483      Inf
sample estimates:
    mean in group Attend mean in group Non-Attend
                76.92475                63.22769
```

Steps:

- Hypotheses. $H_0: \mu_1 - \mu_2 \leq \mathbf{0}$ ml versus $H_a: \mu_1 - \mu_2 > 0$
- The significance level is $\alpha = 0.01$.
- Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mathbf{0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.077$, with degrees of freedom $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = 40.68$
- The P-value= $P(t > t_o) = P(t > 4.077) = 0.0001032$
- Since P-value=$0.0001032 < 0.01$ (α), reject H_0.
- Conclusion: At the 1% significance level, the data provide sufficient evidence that those who attend lectures have a higher average.

(c) Obtain a confidence interval for the difference between the class average for attendees and non-attendees $\mu_1 - \mu_2$ corresponding to the test in part (b).

For a **right-tailed** test at significance level $\alpha = 0.01$, the corresponding confidence interval is a $(1 - \alpha) \times 100\% = 99\%$ **upper-tailed** confidence interval. Based on the computer output above, a 99% confidence interval is

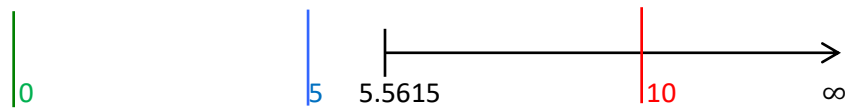
$$\left((\bar{x}_1 - \bar{x}_2) - t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \infty\right) = (5.5615, \infty).$$

Interpretation: we can be 99% confident that the difference between the class average for attendees and non-attendees $\mu_1 - \mu_2$ is at least 5.5615%, i.e., we can be 99% confident that the class average for attendees is at least 5.5615% higher than that of the non-attendees.

(d) Based on the interval obtained in pat (c), can we claim that the class average of attendees is at least 5% higher than that of the non-attendees? How about 10% higher?

We can claim that the class average of attendees is at least 5% higher than that of the non-attendees since the entire interval for $\mu_1 - \mu_2$ is above 5%, that is, $\mu_1 - \mu_2 > 5$ with $\Delta_0 = 5$.

We can not claim that the class average of attendees is at least 10% higher than that of the non-attendees since the entire interval contains 10. Therefore, we do not have sufficient evidence to claim $\mu_1 - \mu_2 > 10$ where $\Delta_0 = 10$.



$$\mu_1 - \mu_2 > 5.5615$$

6.1.2 Pooled Two-Sample t Test and t Interval

If the two population standard deviations are equal, i.e., $\sigma_1 = \sigma_2 = \sigma$, we can pool the two samples together to get a better estimate of the common standard deviation σ

$$\hat{\sigma} = s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

where the term $(n_1 - 1)s_1^2 = \sum_{sample\ 1}(x - \bar{x}_1)^2$ is the variation of the data within sample 1, and $(n_2 - 1)s_2^2 = \sum_{sample\ 2}(x - \bar{x}_2)^2$ is the variation of the data within sample 2. Recall that the standard deviation of $\bar{X}_1 - \bar{X}_2$ is $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. Thus, if $\sigma_1 = \sigma_2 = \sigma$, then $\sigma_{\bar{X}_1 - \bar{X}_2}$ reduces to $\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Estimating σ with s_p leads to the pooled test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\ distribution$$

with $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

The assumption $\sigma_1 = \sigma_2$ is very difficult to verify. Some textbooks suggest a rule of thumb: if the ratio of the larger to the smaller sample standard deviation is less than 2, then the assumption is considered to be met, i.e., $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} < 2$. The assumptions and steps for a two-sample pooled t test are as follows.

Assumptions:

1. Simple random samples;
2. Two samples are independent;
3. Normal populations or large samples ($n_1 \geq 30, n_2 \geq 30$);
4. Equal standard deviation $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} < 2$.

Steps:

1. Set up the hypotheses:

Two tailed test	Right (upper) tailed test	Left (lower) tailed test
$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_0: \mu_1 - \mu_2 \geq \Delta_0$
$H_a: \mu_1 - \mu_2 \neq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$

Note that Δ_0 can be zero or any value you would like to test.

2. State the significance level α.
3. Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1-\bar{x}_2)-\Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ with $df = n_1 + n_2 - 2$.
4. Find the P-value or rejection region:

	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_0: \mu_1 - \mu_2 \geq \Delta_0$
	$H_a: \mu_1 - \mu_2 \neq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$
P-value	$2P(t \geq \lvert t_o \rvert)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_\alpha$	$t \leq -t_\alpha$

5. Decision: reject the null H_0 if P-value$\leq \alpha$ or t_o falls in the rejection region.
6. Conclusions.

A $(1-\alpha) \times 100\%$ two-sample pooled t confidence interval for $\mu_1 - \mu_2$ is:

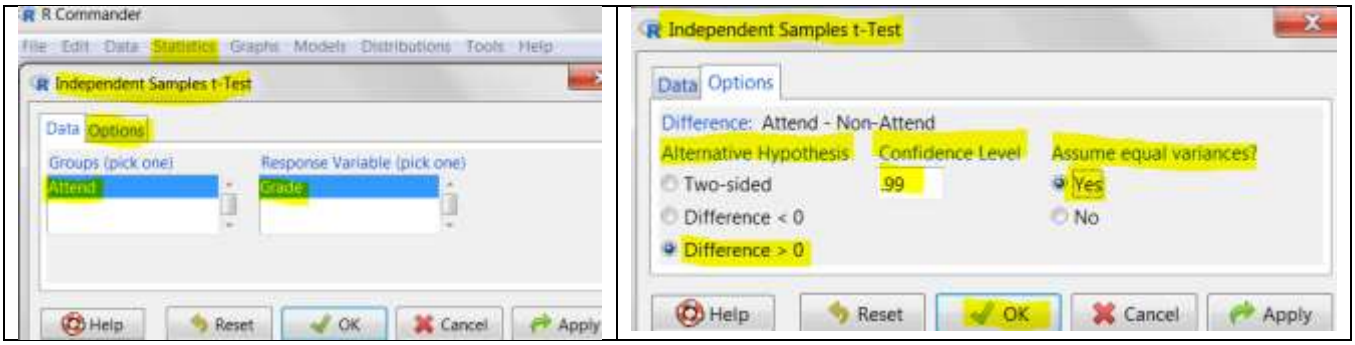
	Two-sided Interval for Two-sided Test	**Upper Tailed Interval for Right Tailed Test**	**Lower Tailed Interval for Left Tailed Test**
	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 \leq \Delta_0$	$H_0: \mu_1 - \mu_2 \geq \Delta_0$
	$H_a: \mu_1 - \mu_2 \neq \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 < \Delta_0$
Interval	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$((\bar{x}_1 - \bar{x}_2) - t_\alpha \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \infty)$	$(-\infty, (\bar{x}_1 - \bar{x}_2) + t_\alpha \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$
Decision	**Reject H_0 if Δ_0 is outside the interval**		

Example: **Pooled two-sample t Test and Interval**

Is it reasonable to conduct a pooled two-sample t test to test whether those who attend lectures have a higher average? If yes, conduct the test at the 1% significance level.

Since $\frac{\max\{s_1, s_2\}}{\min\{s_1, s_2\}} = \frac{\max\{11.827, 15.480\}}{\text{mix}\{11.827, 15.480\}} = \frac{15.480}{11.827} < 2$, it is reasonable to conduct a pooled two-sample t test.

1. **Statistics→Means→Independent Sample t-Test...**
2. In the "**Independent Sample t-Test**" window, select "**Attend**" as the grouping variable and "**Grade**" as the response variable, since we want to compare the grades between attendants and non-attendants.
3. Click "**Options**", in the "**Options**" window, choose "**Difference>0**" as the **Alternative Hypothesis**. Type **0.99** in the box under "**Confidence Level**". Check "**Yes**" under "**Assume equal variances**" for a pooled two-sample t test. Click OK.
4. Click OK




```
        Two Sample t-test

data:  Grade by Attend
t = 4.4942, df = 83, p-value = 1.121e-05
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 6.467476      Inf
sample estimates:
    mean in group Attend mean in group Non-Attend
                76.92475                63.22769
```

Steps:

- Hypotheses. $H_0: \mu_1 - \mu_2 \leq \mathbf{0}$ ml versus $H_a: \mu_1 - \mu_2 > 0$
- The significance level is $\alpha = 0.01$.
- Compute the value of the test statistic: $t_o = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mathbf{0}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = 4.4942$, with degrees of freedom $df = n_1 + n_2 - 2 = 58 + 27 - 2 = 83$.
- The P-value= $P(t > t_o) = P(t > 4.4942) = 0.00001121$.
- Since P-value=$0.00001121 < 0.01$ (α), reject H_0.
- Conclusion: At the 1% significance level, the data provide sufficient evidence that those who attend lectures have a higher average.

The corresponding 99% upper-tailed interval is $(6.4675, \infty)$. The result is very similar to that of a non-pooled two-sample t test.

6.1.3 Non-Pooled Versus Pooled Two-Sample t Test

Now, it comes to the question: shall we use pooled or non-pooled t-test?

The advantages of the pooled t test are:

- A much simpler formula to calculate the degrees of freedom;
- A slightly narrower confidence interval and a slightly more powerful test.

However, the pooled t test is valid only when the standard deviations of two groups are the same; otherwise, the pooled method gives misleading results.

It is even harder to test whether the two standard deviations are equal or not. Therefore, we recommend using the non-pooled two-sample t test by default; apply the pooled two-sample t test only if you are very confident that the two standard deviations are the same.

6.2 Paired t Test and t Interval Based on Paired Sample

Two samples are considered **paired** if each observation in the first sample is related to one and only one observation in the second sample. A paired t test and a paired t interval are exactly a one-sample t test and a one-sample t interval on the **paired differences** respectively.

Assumptions:

1. The paired difference $d_i, i = 1, \cdots, n$ is a simple random sample (SRS) from all possible pairs
2. The paired differences follow a normal distribution or large number of pairs ($n \geq 30$)

Steps:

1. Set up the hypotheses:

$H_0: \mu_1 - \mu_2 = \delta_0$	$H_0: \mu_1 - \mu_2 \leq \delta_0$	$H_0: \mu_1 - \mu_2 \geq \delta_0$
$H_a: \mu_1 - \mu_2 \neq \delta_0$	$H_a: \mu_1 - \mu_2 > \delta_0$	$H_a: \mu_1 - \mu_2 < \delta_0$

Note: δ_0 can be any value tested, in most cases $\delta_0 = 0$. Some textbooks state the hypotheses using $\mu_d = \mu_1 - \mu_2$.

2. State the significance level α.
3. Compute the value of the test statistic: $t_o = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}}$ with degree of freedom $df = n - 1$, where n is the number of pairs and

$$\bar{d} = \frac{\sum d_i}{n}, s_d = \sqrt{\frac{(\sum d_i^2) - \frac{(\sum d_i)^2}{n}}{n-1}}.$$

4. Find the P-value **or** rejection region:

	$H_0: \mu_1 - \mu_2 = \delta_0$	$H_0: \mu_1 - \mu_2 \leq \delta_0$	$H_0: \mu_1 - \mu_2 \geq \delta_0$
	$H_a: \mu_1 - \mu_2 \neq \delta_0$	$H_a: \mu_1 - \mu_2 > \delta_0$	$H_a: \mu_1 - \mu_2 < \delta_0$
P-value	$2P(t \geq \lvert t_o \rvert)$	$P(t \geq t_o)$	$P(t \leq t_o)$
Rejection region	$t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$	$t \geq t_\alpha$	$t \leq -t_\alpha$

5. Reject the null H_0 if P-value$\leq \alpha$ or t_o falls in the rejection region.
6. Conclusions.

A $(1-\alpha) \times 100\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ corresponding to a hypothesis test at the significance level α is:

	$H_0: \mu_1 - \mu_2 = \delta_0$	$H_0: \mu_1 - \mu_2 \leq \delta_0$	$H_0: \mu_1 - \mu_2 \geq \delta_0$
	$H_a: \mu_1 - \mu_2 \neq \delta_0$	$H_a: \mu_1 - \mu_2 > \delta_0$	$H_a: \mu_1 - \mu_2 < \delta_0$
$(1-\alpha) \times 100\%$ CI	$(\bar{d} - t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}})$	$(\bar{d} - t_\alpha \frac{s_d}{\sqrt{n}}, \infty)$	$(-\infty, \bar{d} + t_\alpha \frac{s_d}{\sqrt{n}})$
Decision	Reject H_0 if δ_0 is outside the interval		

Example: **Paired t Test and Paired t Interval**

Eleven people participate in a diet program, their weights in pounds before and after taking the program are listed below. Please download the file pair_diet.xlxs from online and import it into R commander.

Before (in lb)	After (in lb)	Paired Differences $d_i = \text{Before} - \text{After}$
130	100	30
140	115	25
160	140	20
110	115	-5
120	120	0
150	130	20
160	130	30
100	110	-10
180	140	40
200	150	50
130	120	10

(a) Test atthe 1% significance level whether the diet program is effective in reducing weight.

(b) Obtain a confidence interval corresponding to the test in part (a).

(c) Does the interval in part (b) support the conclusion in part (a)?

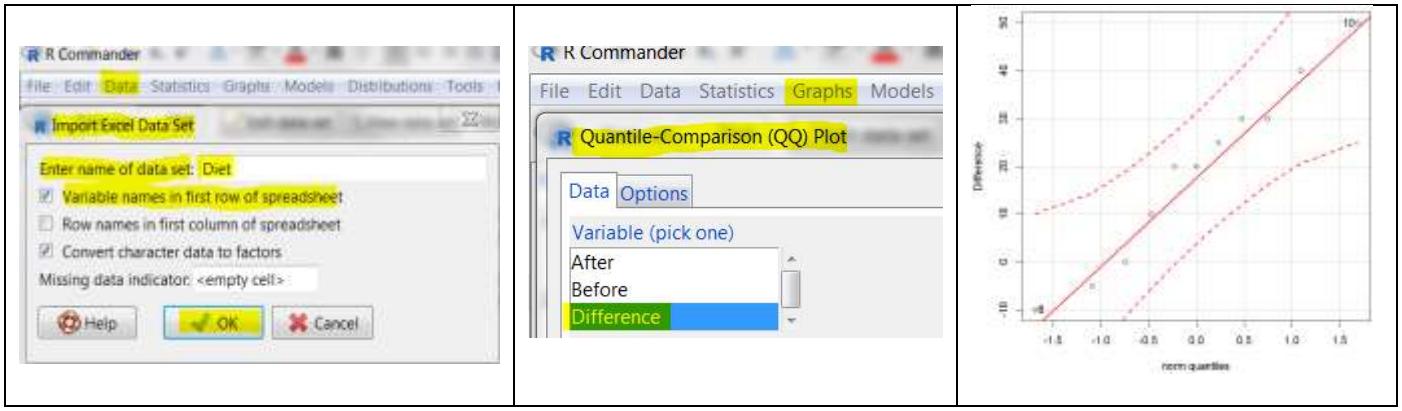
(d) Is it possible to claim that on average the diet program can reduce weight by more than 5 pounds? Explain why.

Check the assumptions:

1. We have a simple random sample in the paired differences.
2. We have eleven pairs, not a large number of pairs ($n < 30$). Therefore, we need to check whether the paired differences are taken from a normal population.

Draw a normal probability plot on the sample of paired differences and look for a straight line.

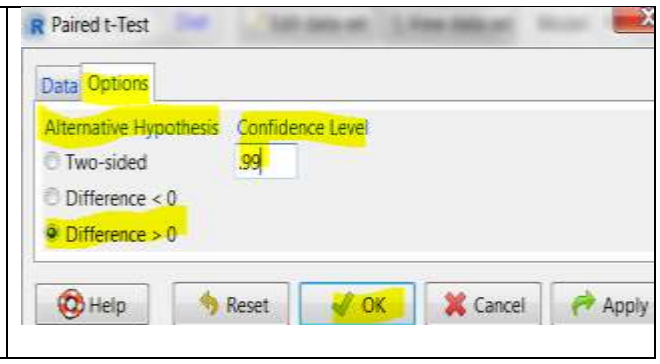
1. Import the data. **Data→Import data→from Excel file pair_diet.xlsx (name it "diet")**
2. **Graphs→Quantile-comparison plot…**
   In the "**Quantile-Comparison (QQ) Plot**" window, choose "**Difference**" as the variable to plot. Click OK.



Since all the points roughly lie on a straight line, we can assume that the paired differences are from a normal population. Therefore, the assumptions for a paired t test are satisfied.

(a) Test at the 1% significance level whether the diet program is effective in reducing weight.

1. Import the data. **Data→Import data→from Excel file pair_diet.xlxs (name it "diet")**
2. **Statistics→Means→Paired t-Test...**
3. In the "**Paired t-Test**" window, select "**Before**" as the First variable and "**After**" as the second variable, since we define the paired difference as Before-After.
4. Click "**Options**", in the "**Options**" window, choose "**Difference>0**" as the **Alternative Hypothesis**. Type **0.99** in the box under "**Confidence Level**". Click OK.
5. Click OK




```
        Paired t-test

data:  Before and After
t = 3.3648, df = 10, p-value = 0.003592
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 3.410302      Inf
sample estimates:
mean of the differences
               19.09091
```

Steps:

- Hypotheses. $H_0: \mu_B - \mu_A \leq \mathbf{0}$ versus $H_a: \mu_B - \mu_A > 0$.
- The significance level is $\alpha = 0.01$.
- Compute the value of the test statistic: $t_o = \frac{\bar{d}-\delta_0}{s_d/\sqrt{n}} = 3.3648$, with degrees of freedom $df = n - 1 = 11 - 1 = 10$.
- The P-value= $P(t \geq t_o) = P(t \geq 3.3648) = 0.003592$.
- Since P-value=$0.003592 < 0.01$ (α), reject H_0.
- Conclusion: At the 1% significance level, the data provide sufficient evidence that the diet program is effective in reducing weight.

(b) Obtain a confidence interval corresponding to the test in part (a).

For a right-tailed test at significance level $\alpha = 0.01$, the corresponding confidence interval should be a 99% upper-tailed interval, which is $(3.410302, \infty)$ from the computer output.

(c) Does the interval in part (b) support the conclusion in part (a)?
Yes. In part (a), we reject H_0 and claim that $\mu_B - \mu_A > 0$. In part (b), since the interval does not contain $\delta_0 = 0$ and the entire interval is above 0, we can claim that $\mu_B - \mu_A > 0$ with 99% confidence, which supports the results obtained in part (b).

(d) Is it possible to claim that on average the diet program can reduce more than 5 pounds? Explain why.
Here we will test $H_0: \mu_B - \mu_A \leq \underbrace{5}_{\delta_0=5}$ versus $H_a: \mu_B - \mu_A > 5$. Then $\delta_0 = 5$ in this question. The answer is "No", since $\delta_0 = 5$ is within the interval (3.410302, ∞). Therefore, we cannot reject $H_0: \mu_B - \mu_A \leq \underbrace{5}_{\delta_0=5}$ and claim that on average the diet program can reduce weight by more than 5 pounds.

LAB 7 INFERENCES FOR POPULATION PROPORTIONS

In this lab, we focus on inferences for another population parameter: the population proportion p. The population proportion is defined as the proportion (or percentage) of a population that have a specified attribute. For example, proportion of times that athletes wearing blue uniforms win the Judo games; proportion of customers who respond to the advertisement; proportion of women who suffer arthritis.

7.1 ONE-PROPORTION Z TEST & Z INTERVAL BASED ON ONE SAMPLE

Assumptions:

1. A simple random sample
2. Both np_0 and $n(1-p_0)$ are at least 5.

Steps:

1. Set up the hypotheses:

$H_0: p = p_0$	$H_0: p \leq p_0$	$H_0: p \geq p_0$
$H_a: p \neq p_0$	$H_a: p > p_0$	$H_a: p < p_0$

2. State the significance level α.
3. Compute the test statistic: $z_o = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ with $\hat{p} = \frac{x}{n}$ where x is the total successes in n observations.

4. Find the P-value **or** rejection region:

	$H_0: p = p_0$	$H_0: p \leq p_0$	$H_0: p \geq p_0$
	$H_a: p \neq p_0$	$H_a: p > p_0$	$H_a: p < p_0$
P-value	$2P(Z \geq \lvert z_o \rvert)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$	$Z \geq z_\alpha$	$Z \leq -z_\alpha$

5. Reject the null H_0 if P-value$\leq \alpha$ or z_o falls in the rejection region.
6. Conclusions.

A point estimate for the population proportion p is the sample proportion $\hat{p} = \frac{x}{n}$. A $(1-\alpha) \times 100\%$ confidence interval corresponding to a hypothesis test at the significance level α for the population proportion p are as shown in the table.

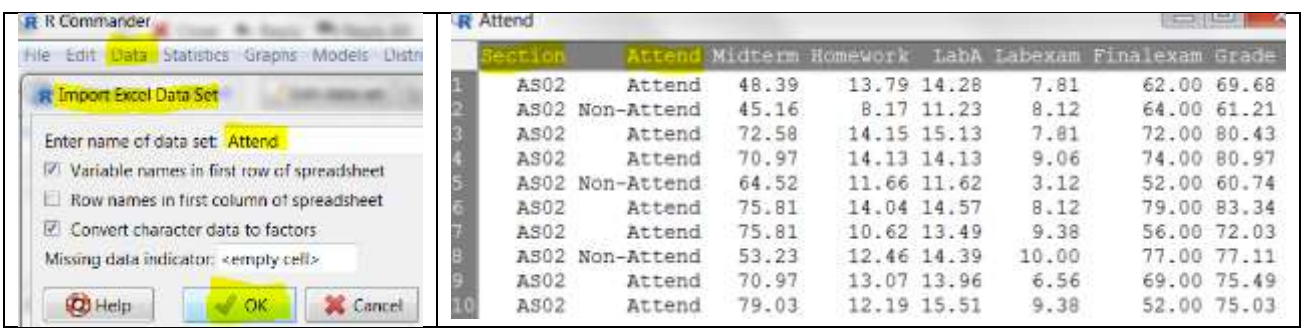
	$H_0: p = p_0$	$H_0: p \leq p_0$	$H_0: p \geq p_0$
	$H_a: p \neq p_0$	$H_a: p > p_0$	$H_a: p < p_0$
$(1-\alpha) \times 100\%$ CI	$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$(\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \infty)$	$(-\infty, \hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
Decision	Reject H_0 if p_0 is outside the interval		

Example: One-Proportion z Test and z Interval

Revisit the data set about the effect of attending lecture on grades. There are two sections: AS02 and AS04. Some students attend lectures regularly and some do not in both sections. We are interested in the attendance rate.

AS02	Attend	AS02	Attend	AS02	Attend	AS04	Attend	AS04	Attend
AS02	Non-Attend	AS02	Non-Attend	AS02	Non-Attend	AS04	Attend	AS04	Non-Attend
AS02	Attend	AS02	Attend	AS02	Attend	AS04	Attend	AS04	Attend
AS02	Attend	AS02	Attend	AS02	Attend	AS04	Attend	AS04	Non-Attend
AS02	Non-Attend	AS02	Attend	AS02	Attend	AS04	Non-Attend	AS04	Attend
AS02	Attend	AS02	Attend	AS02	Attend	AS04	Non-Attend	AS04	Attend
AS02	Attend	AS02	Attend	AS02	Attend	AS04	Attend	AS04	Non-Attend
AS02	Non-Attend	AS02	Attend	AS02	Attend	AS04	Non-Attend	AS04	Attend
AS02	Attend	AS02	Non-Attend	AS02	Attend	AS04	Non-Attend	AS04	Attend
AS02	Attend	AS02	Non-Attend	AS02	Attend	AS04	Attend	AS04	Attend
AS02	Attend	AS02	Non-Attend	AS02	Attend	AS04	Attend	AS04	Non-Attend
AS02	Attend	AS02	Non-Attend	AS02	Non-Attend	AS04	Non-Attend	AS04	Non-Attend
AS02	Attend	AS02	Attend	AS04	Attend	AS04	Attend	AS04	Attend
AS02	Attend	AS02	Attend	AS04	Attend	AS04	Non-Attend	AS04	Attend
AS02	Attend	AS02	Attend	AS04	Attend	AS04	Attend	AS04	Attend
AS02	Non-Attend	AS02	Attend	AS04	Attend	AS04	Attend	AS04	Non-Attend
AS02	Attend	AS02	Non-Attend	AS04	Non-Attend	AS04	Non-Attend	AS04	Non-Attend

Download attend_grade.xlsx from online. Import data ("attend_grade.xlsx") into R commander:
Data→Import data→from Excel file… (name it "Attend")
The data set consists of eight variables (columns) and 85 instances (rows). The variable "Section" indicating whether the student is in AS02 or AS04, "Attend" indicating whether the student attends lectures regularly or not; "Midterm", "Homework", "LabA", "Labexam", "Finalexam", "Grade" are the student's grades in midterm exam, homework assignments, lab assignments, lab exam, final exam, and the final grade.



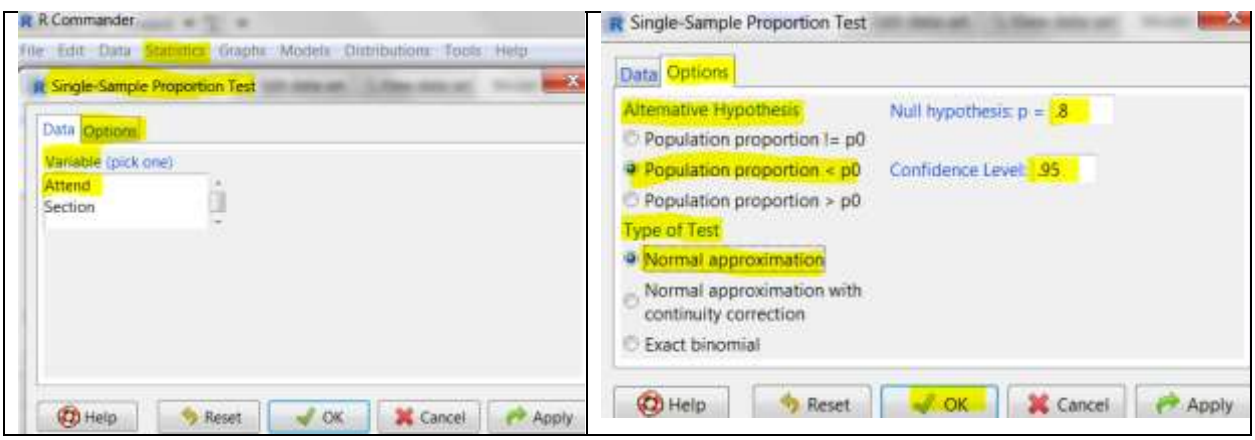
(a) What is the overall attendance rate in the two sections?

1. **Statistics→Summaries→Frequency Distributions**
2. In the "**Frequency Distributions"** window, choose "**Attend**" as the variable. Click OK.

There are 58+27=85 students altogether in both sections and 58 students attend lectures regularly. Therefore, the overall attendance rate is 58/85=0.6824 which is 68.24%.

(b) Test at the 5% significance level whether the overall attendance rate is **below** 80%.

1. **Statistics→Proportions→Single-sample proportion test…**
2. In the "**Single-Sample Proportion Test"** window, choose "**Attend**" as the variable.
3. Click "**Options**". In the "Options" window, choose "**Population proportion<p0**" as the **Alternative Hypothesis**. Specify the hypothesized value "p=0.8" under the "**Null hypothesis**". That is $p_0 = 0.8$. Type **0.95** in the box under "**Confidence Level**". Under "**Type of Test**", check "**Normal approximation**". Click OK.
4. Click OK.



```
        1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.8
X-squared = 7.3529, df = 1, p-value = 0.003348
alternative hypothesis: true p is less than 0.8
95 percent confidence interval:
 0.0000000 0.7586904
sample estimates:
        p
0.6823529
```

Steps:

- Hypotheses. $H_0: p \geq 0.8$ versus $H_a: p < 0.8$.
- The significance level is $\alpha = 0.05$.
- Compute the value of the test statistic: $z_o = -\sqrt{7.3529} = -2.71162$.
  Note: the computer output provides the chi-square score 7.3529 which is the square of the observed test statistic z_o.
  We can double check that the test statistic $z_o = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{58}{85}-0.8}{\sqrt{\frac{0.8(1-0.8)}{85}}} = -2.71163$.
  Note that ${z_o}^2 = (-2.71163)^2 = 7.3529$ which is the chi-square score.
- The P-value= $P(Z \leq z_o) = P(Z \leq -2.7116) = 0.003348$
- Since P-value=$0.003348 < 0.05$ (α), reject H_0.
- Conclusion: At 5% significance level, the data provide sufficient evidence that overall attendance rate is below 80%.

(c) Obtain a confidence interval corresponding to the test in part (b).
For a left-tailed test at the 5% significance level, the corresponding interval should be a 95% lower-tailed interval, which is (0, 0.7586904) obtained from the computer output.
Interpretation: we can be 95% confident that the overall attendance rate is below 0.75869, i.e., 75.869%.

(d) Does the interval in part (c) support the conclusion in part (b)?
Yes. In part (b), we reject H_0 and claim that $p < 0.8$. In part (c), since the interval does not contain $p_0 = 0.8$ and the entire interval is below 0.8, we can claim that $p < 0.8$ with 95% confidence, which supports the results obtained in part (c).

7.2 Two-Proportion z Test & z Interval Based on Two Independent Samples

For independent samples of size n_1 and n_2 from two populations, a point estimate for the difference between two population proportions $(p_1 - p_2)$ is the difference between the sample proportions $(\hat{p}_1 - \hat{p}_2)$ where $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$, and x1 and x2 are the number of successes in their samples.

7.2.1 Two-Proportion Z Interval

Assumptions:

1. Both samples are simple random samples from their own populations.
2. The two samples are independent.
3. Large samples, all the number of successes and the number of failures $x_1, n_1 - x_1, x_2$, and $n_2 - x_2$ are at least 5.

A $(1-\alpha) \times 100\%$ confidence interval for the difference between the population proportion $(p_1 - p_2)$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \quad \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$$

where $z_{\alpha/2}$ is the z score such that the area to its right is $\frac{\alpha}{2}$ under the standard normal curve. This is a two-tailed interval.

A $(1-\alpha) \times 100\%$ upper-tail confidence interval is $((\hat{p}_1 - \hat{p}_2) - z_{\boldsymbol{\alpha}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \infty)$.

And a $(1-\alpha) \times 100\%$ lower-tail confidence interval is $(-\infty, (\hat{p}_1 - \hat{p}_2) + z_{\boldsymbol{\alpha}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}})$.

7.2.2 Two-Proportion Z Test

Assumptions:
1. Both samples are simple random samples from their own populations.
2. The two samples are independent.
3. Large samples, all the number of successes and the number of failures $x_1, n_1 - x_1, x_2$, and $n_2 - x_2$ are at least 5.

Steps to perform a two-proportion z test:

1. Set up the hypotheses:

$H_0: p_1 = p_2$	$H_0: p_1 \leq p_2$	$H_0: p_1 \geq p_2$
$H_a: p_1 \neq p_2$	$H_a: p_1 > p_2$	$H_a: p_1 < p_2$

2. State the significance level α.
3. Compute the value of the test statistic:

$$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } \hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}, \hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$$

4. Find the P-value **or** rejection region:

	$H_0: p_1 = p_2$	$H_0: p_1 \leq p_2$	$H_0: p_1 \geq p_2$
	$H_a: p_1 \neq p_2$	$H_a: p_1 > p_2$	$H_a: p_1 < p_2$
P-value	$2P(Z \geq \lvert z_o \rvert)$	$P(Z \geq z_o)$	$P(Z \leq z_o)$
Rejection region	$Z \geq z_{\alpha/2}$ or $Z \leq -z_{\alpha/2}$	$Z \geq z_\alpha$	$Z \leq -z_\alpha$

5. Reject the null H_0 if P-value$\leq \alpha$ or z_o falls in the rejection region.
6. Conclusions.

Example: Two-Proportion Z Test and Z Interval

Revisit the data set attend_grade.xlsx (which you imported into R in the previous section) about the effect of attending lecture on grades. There are two sections: AS02 and AS04. Some students attend lectures regularly and some do not in both sections. We are interested in the attendance rate.

(a) What are the attendance rates in sections AS02 and AS04 respectively?

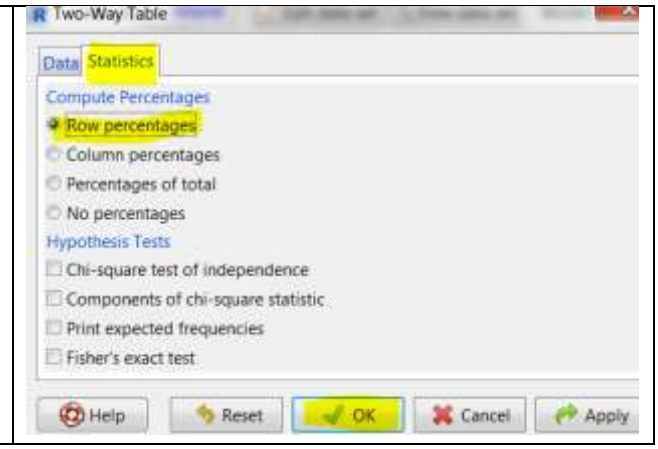
1**. Statistics→Contingency tables→Two-way table…**

2**.** In the "**Two-Way Table"** window, choose "**Section**" as the row variable and "**Attend**" as the column variable.

3. Check "**Statistics**". In the "Statistics" window, select "**Row percentage**" under "**Compute Percentages**". Click OK.

4. Click OK.

Note that we chose "Section" as the row variable and we want the percentage of attendees within each section; therefore, we need to calculate the row percentages.




```
Frequency table:
       Attend
Section Attend Non-Attend
   AS02     34         12
   AS04     24         15

Row percentages:
       Attend
Section Attend Non-Attend Total Count
   AS02   73.9       26.1   100    46
   AS04   61.5       38.5   100    39
```

The attendance rate in AS02 is

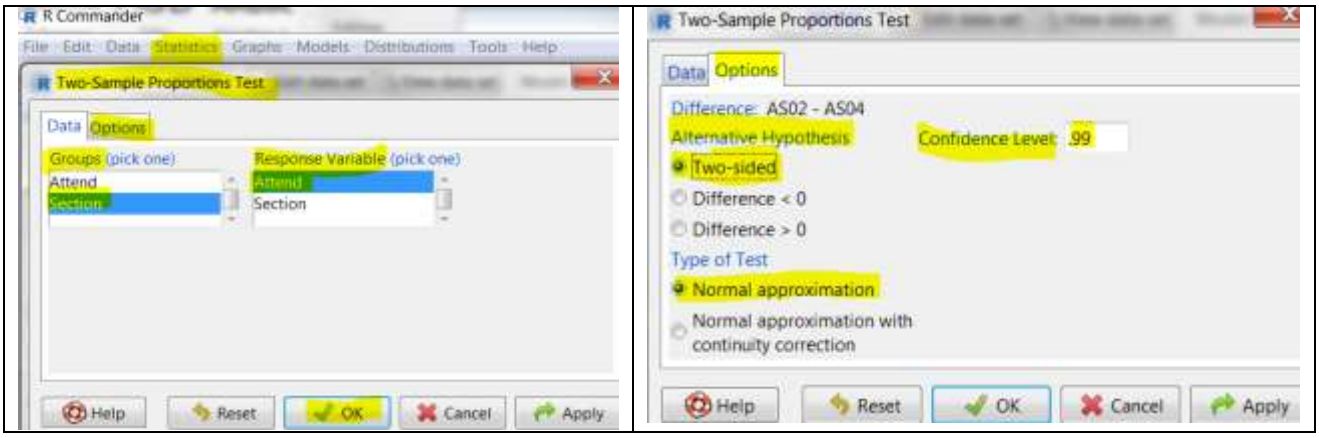
$\hat{p}_1 = \frac{x_1}{n_1} = \frac{34}{46} = 0.7391$ which is 73.91%.

The attendance rate in AS04 is

$\hat{p}_1 = \frac{x_2}{n_2} = \frac{24}{39} = 0.6154$ which is 61.54%.

(b) Test at the 1% significance level whether the attendance rates are **different** in both sections.

1. **Statistics→Proportions→Two-sample proportion test…**
2. In the "**Two-Sample Proportion Test"** window, choose "**Section**" as the row variable and "**Attend**" as the column variable. Click "**Options**". In the "Options" window, choose "**Two Sided**" as the **Alternative Hypothesis**. Type **0.99** in the box under "**Confidence Level**". Under "**Type of Test**", check "**Normal approximation**". Click OK.
3. Click OK.




```
        2-sample test for equality of proportions without continuity
        correction

data:  .Table
X-squared = 1.4911, df = 1, p-value = 0.222
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.1371712  0.3846628
sample estimates:
   prop 1    prop 2
0.7391304 0.6153846
```

Steps:

- Hypotheses. $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$, where p_1 is the attendance rate of section AS02 and p_2 is the attendance rate of section AS04.
- The significance level is $\alpha = 0.01$.
- Compute the value of the test statistic: $z_o = \sqrt{1.4911} = 1.2211$.
  Note: the computer output provides the chi-square score 1.4911 which is the square of the observed test statistic z_o.
  We can double check that the test statistic:
  $$z_o = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{0.7391304 - 0.6153846}{\sqrt{0.682353(1-0.682353)}\sqrt{\frac{1}{46}+\frac{1}{39}}} = 1.2211, \text{ with}$$
  $$\hat{p}_p = \frac{x_1+x_2}{n_1+n_2} = \frac{34+24}{46+39} = 0.682353, \hat{p}_1 = \frac{x_1}{n_1} = \frac{34}{46} = 0.7391304, \hat{p}_2 = \frac{x_2}{n_2} = \frac{24}{39} = 0.6153846.$$
- The P-value= $2P(Z \geq |z_o|) = 2P(Z \geq 1.2211) = 0.222$.
- Since P-value=$0.222 > 0.01$ (α), we cannot reject H_0.
- Conclusion: At the 1% significance level, the data do not provide sufficient evidence that the attendance rates are different in both sections.

(c) Obtain a confidence interval corresponding to the test in part (b).
For a two-tailed test at 1% significance level, the corresponding interval is a 99% two-sided interval for $p_1 - p_2$ which is (-0.1371712, 0.3846628) based on the computer output.

Interpretation: We can be 99% confident that $p_1 - p_2$ is somewhere between -0.1372 and 0.3847. That means, we can be 99% confident that the attendance rate of AS02 is between 13.72% lower to 38.47% higher than that of AS04.

We can double check that a 99% confidence for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (0.7391304 - 0.6153846) \pm$$

$$2.575\sqrt{\frac{0.7391304(1-0.7391304)}{46} + \frac{0.6153846(1-0.6153846)}{39}} = (-0.1370872, 0.3845788\ )$$

This is a little bit off due to rounding.

(d) Does the interval in part (c) support the conclusion in part (b)?

Yes. In part (b), we cannot reject H_0 and claim that the two attendance rates are significantly different. In part (c), since the interval contains 0, there is no significant difference between the attendance rates in both sections.

LAB 8 CHI-SQUARE TESTS

Lab 7 covers $\boldsymbol{z}$ test and $\boldsymbol{z}$ interval for one and two proportions. Chi-square tests should be used when more than two proportions are compared.

8.1 CHI-SQUARE GOODNESS-OF FIT TEST FOR ONE CATEGORICAL OR DISCRETE VARIABLE

The chi-square goodness-of-fit test can be applied to a categorical variable or a discrete quantitative variable that has only finitely possible values. The objective of a chi-square goodness-of-fit test is to test whether the variable follows the probability distribution specified in the null hypothesis H_0.

Assumptions:

1. All expected frequencies are at least 1.
2. At most 20% of the expected frequencies are less than 5.
3. Simple random sample (if you need to generalize the conclusion to a larger population)

Note: if the assumption 1 or 2 is violated, one can consider combining the cells to make the counts in those cells larger.

Before running a chi-square goodness-of-fit test, we should first check the assumptions. Calculate the expected frequency for each possible value of the variable using $E = np$, where n is the total number of observations and p is the relative frequency (or probability) specified in the null hypothesis. Check whether the expected frequencies satisfy assumptions 1 and 2. If not, consider combining some cells.

Steps to perform a chi-square goodness-of-fit test:

1. Set up the hypotheses:

$$H_0\text{: The variable has the specified distribution}$$
$$H_a\text{: The variable does not have the specified distribution}$$

2. State the significance level α.
3. Compute the value of the test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-\boldsymbol{E})^2}{\boldsymbol{E}}$ with $df = k - 1$.
4. Find the P-value **or** rejection region based on the χ^2curve with $df = k - 1$.

P-value	$P(\chi^2 \geq \chi_o^2)$ the area to the right of χ_o^2 under the curve
Rejection region	$\chi^2 \geq \chi_\alpha^2$ the region to the right of χ_α^2, the area is α

5. Reject the null H_0 if P-value$\leq \alpha$ or χ_o^2 falls in the rejection region.
6. Conclusions.

Example: Chi-square goodness-of-fit test

According to the results of the Federal election in 2015, 31.9% of votes supported Conservative, 39.5% supported Liberal, 19.7% supported New Democratic (NDP), 4.7% supported Bloc Québécois, and 3.4% supported Green (data from Wikipedia).

Federal proportions are summarized in this table.

Parties	Conservative	Green	Liberal	NDP	Bloc Québécois	Others
Proportion (p)	0.319	0.034	0.395	0.197	0.047	0.008

Thirty-seven students who voted in my Stat151 class responded to the online survey and their vote counts are summarized in the following table:

Parties	Conservative	Green	Liberal	NDP	Bloc Québécois	Others
Counts	9	2	17	6	0	3

Test at the 5% significance level whether the class has a different preference pattern from the whole nation (2015 election).

We check the assumptions. The expected frequencies (counts $E = np = 37 \times p$) for the outcome cells when n = 37 are:

Parties	Conservative	Green	Liberal	NDP	Bloc Québécois	Others
Proportion (p)	0.319	0.034	0.395	0.197	0.047	0.008
Expected Counts	11.803	1.258	14.615	7.289	1.739	0.296

Here we have one outcome cell with an expected count below 1, which violates an assumption. Furthermore, with $k = 6$ outcome cells, we wish to assume at most $6 \times 0.2 = 1.2$ cells with expected counts less than 5, and we have three cells less than 5. Also, our survey was taken in Alberta and no Bloc Québécois run in Alberta (although a student with a home riding of Quebec might have still voted that way).

We would like to do a test, so we need to combine some cells. Federally, we combine the cells "Green", "Bloc Québécois" and "Others" above and name the combined party "Others". In our sample data set, we also merge "Green" and "Others" and name the combined party "Others". This will lead us to have k = 4 outcomes for our federal population and k = 4 cell outcomes for our survey sample data, as follows.

As a result, the expected and observed frequencies are summarized as follows:

Parties	**Proportion** p	**Observed** (O)	**Expected** *(E)* $E = np = 37 \times p$
Conservative	0.319	9	$37 \times 0.319 = 11.803$
Liberal	0.395	17	$37 \times 0.395 = 14.615$
NDP	0.197	6	$37 \times 0.197 = 7.289$
Others	**0.089 = 0.034 + 0.047 + 0.008**	**2+3=5**	$37 \times 0.089 = 3.293$
	Sum=1	Sum=37	Sum=37

Now we have no cells with an expected count below 1, and 1 cell with an expected count below 5. So, we actually have 25% of our cells with an expected value below 5, which exceeds the assumption requirement that no more than 20% of our cells have an expected value below 5, but it is close, and we proceed for educational purposes.

The file "survey.xlxs" contains our sample data from the students. A column called "MergedParty" contains the data of interest (where the Green and Other cells have been renamed to "Others").

We import the data ("survey.xlsx") into R commander and perform the test:

Data→Import data→from Excel file… (name it "Survey")

Use R commander to run the chi-square goodness-of-fit test.

1. **Statistics→Summaries→Frequency distributions…**
2. In the "**Frequency Distributions"** window, choose "**Merged.Party**" as the variable. Check "**Chi-square goodness-of-fit test (for one variable only)**". Click OK.
3. In the "**Goodness-of-Fit Test**" window, specify the hypothesized proportions: 0.319 for Conservative, 0.395 for Liberal, 0.197 for NDP, and 0.089 for Others. Click OK.


```
counts:
MergedParties
Conservative      Liberal          NDP       Others
           9           17            6            5

percentages:
MergedParties
Conservative      Liberal          NDP       Others
       24.32        45.95        16.22        13.51

        Chi-squared test for given probabilities

data:  .Table
X-squared = 2.1677, df = 3, p-value = 0.5383
```

Steps to perform a chi-square goodness-of-fit test:

1. Set up the hypotheses:

$$H_0: p_C = 0.319,\ p_L = 0.395,\ p_{NDP} = 0.197,\ p_{Others} = 0.089$$
$$H_a: \text{At least one proportion is different the ones specified under } H_0$$

2. The significance level is $\alpha = 0.05$.
3. The test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-\boldsymbol{E})^2}{\boldsymbol{E}} = 2.1677$, with $df = k - 1 = 4 - 1 = 3$.

4. Find the P-value. Chi-square tests are always right tail.
   P-value= $P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 2.1677) = 0.5383$.
5. Decision: We do not reject the null H_0 since P-value= $0.5383 > 0.05(\alpha)$.
6. Conclusion: At the 5% significance level, we do not have sufficient evidence that the class has a different preference pattern from the whole nation (2015 election).

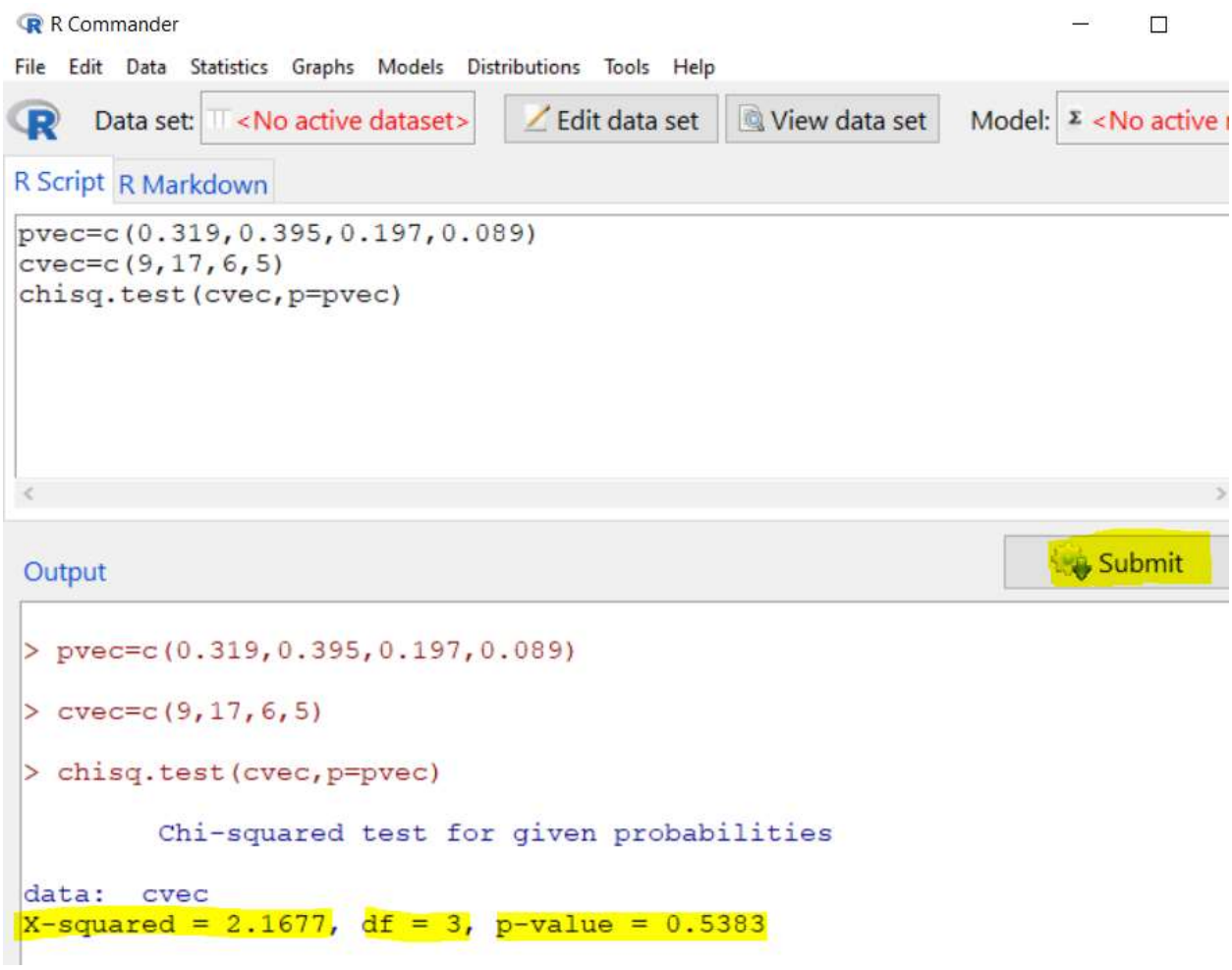
Another way to conduct a chi-square goodness-of-fit without the data is to type commands in the R Script window. We first need to let R commander know the proportions under the null and the observed counts.

1. Type **pvec=c(0.319,0.395,0.197,0.089)** in the R Script Window, click "**Submit**".

2. Type **cvec=c(9,17,6,5)** in the R Script Window, click "**Submit**".

3. Type **chisq.test(cvec,p=pvec)** in the R Script Window, click "**Submit**".

Note: for each line of the commands, put the mouse at the end of each line and click "**Submit**" to execute the command.



Computer output: we get the chi-square score is 2.1677, df=3, and p-value=0.5383, the same as the results obtained before in which we use the data.

8.2 Chi-square Independence Test

The chi-square independence test is used to test whether two categorical variables of a population are related (associated) or independent.

Assumptions:

1. All expected frequencies are at least 1.
2. At most 20% of the expected frequencies are less than 5.
3. Simple random sample (if you need to generalize the conclusion to a larger population)

Note: if the assumption 1 or 2 is violated, one can consider combining the cells to make the counts in those cells larger.

Before conducting a chi-square independence test, we first check the assumptions. Calculate the expected frequency for each possible value of the variable using $E = \frac{(r\text{th row total})\times(c\text{th column total})}{n}$, where n is the total number of observations. Check whether the expected frequencies satisfy assumptions 1 and 2. If not, consider combining some cells.

Steps to perform a chi-square independence test:

1. Set up the hypotheses:

$$H_0\text{: The two variables are independent}$$
$$H_a\text{: The two variables are associated}$$

2. State the significance level α.
3. Compute the value of the test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$ with $df = (r-1)\times(c-1)$, where $E = \frac{(r\text{th row total})\times(c\text{th column total})}{n}$, r is the number of rows and c is number of columns of the cells.
4. Find the P-value **or** rejection region based on the χ^2curve with $df = (r-1)\times(c-1)$.

P-value	$P(\chi^2 \geq \chi_o^2)$ the area to the right of χ_o^2 under the curve
Rejection region	$\chi^2 \geq \chi_\alpha^2$ the region to the right of χ_α^2, the area is α

5. Reject the null H_0 if P-value$\leq \alpha$ or χ_o^2 falls in the rejection region.
6. Conclusions.

Example: Chi-square Independence Test

Note: Data set is the Focus database described on Page 34, Introductory Statistics, 10th Edition (2016), by Neil A. Weiss, Pearson.

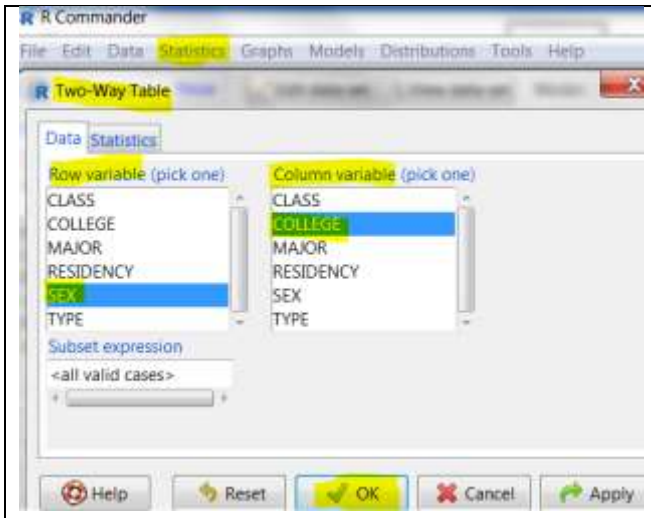
The Focus database contains information of a sample of 200 undergraduate students at the University of Wisconsin-Eau Claire. It has 13 variables including Sex, School/College, Classification (freshman, sophomore, junior, senior), ACT English Score, ACT math Score, ACT composite Score, and etc.

Test at the 5% significance level whether "Sex" and "College" are associated.

Download focus.xlsx from online and import the data into R commander:
Data→Import data→from Excel file… (name it "focus")

Use R-commander to run the chi-square goodness-of-fit test.

1. **Statistics→Contingency tables→Two-way table…**
2. In the "**Two-Way Table**" window, choose "**SEX**" as the row variable and "**COLLEGE**" as the column variable.
3. In the **Statistics** window of the "Two-Way Table" window, check **"Print Expected Frequencies"** and **"Components of chi-square statistic".**
4. Click OK.



Computer Output:

```
Frequency table:
   COLLEGE
SEX A&S Bus Educ Hss Nurs
  F  50  21   26  12    9
  M  47  25    6   4    0

        Pearson's Chi-squared test

data:  .Table
X-squared = 20.112, df = 4, p-value = 0.0004746


Expected counts:
   COLLEGE
SEX   A&S   Bus  Educ  Hss Nurs
  F 57.23 27.14 18.88 9.44 5.31
  M 39.77 18.86 13.12 6.56 3.69

Chi-square components:
   COLLEGE
SEX  A&S  Bus Educ  Hss Nurs
  F 0.91 1.39 2.69 0.69 2.56
  M 1.31 2.00 3.86 1.00 3.69
```

Messages

```
[26] WARNING:
1 expected frequencies are less than 5
```

Steps to perform a chi-square independence test:

1. Set up the hypotheses:

$$H_0\text{: The two varibles are independent}$$
$$H_a\text{: The two variables are associated}$$

2. The significance level is $\alpha = 0.05$.
3. The test statistic: $\chi_o^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = 20.112$,

with $df = (r-1) \times (c-1) = (2-1) \times (5-1) = 4$.

4. Find the P-value. Chi-square tests are always right tail.
   P-value= $P(\chi^2 \geq \chi_o^2) = P(\chi^2 \geq 20.112) = 0.0004746$.
5. Decision: We do not reject the null H_0 since P-value= $0.0004746 < 0.05(\alpha)$.
6. Conclusion: At the 5% significance level, we have sufficient evidence that "Sex" and "College" are associated, i.e., female and male students have significantly difference preference in choosing school/college.

Notice that all the expected cell values are above 1, so this assumption holds. Notice also the warning that the expected frequency for the Male and Nursing cell is below 5, but only 10% (one out of ten) of our expected cell frequencies are below 5, so this assumption holds.

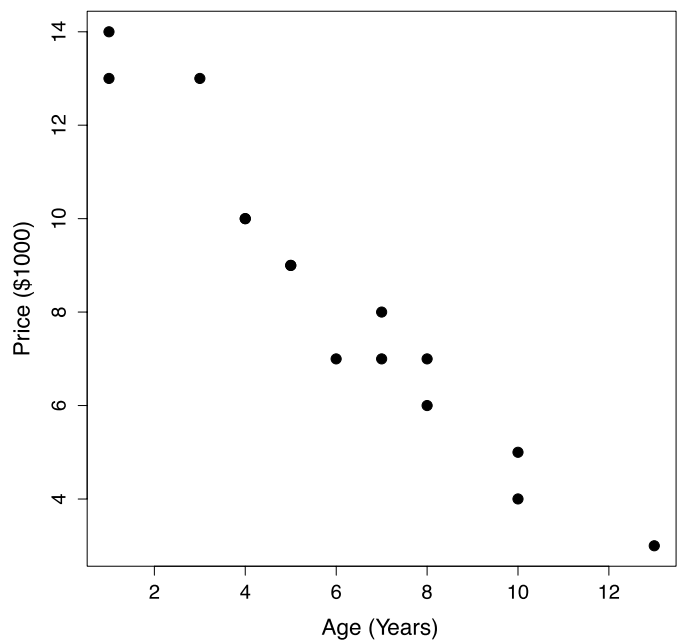
The fact that there are no observations in the Male and Nursing Cell is of note. An examination of the components of the chi-square test statistic does indicate that more females than expected were in nursing and less males than expected were in nursing. We also note that less females than expected were in education and more males than expected were in education. These four cells made the largest contributions towards obtaining a test statistic value that was large and led us to a significant result.

LAB 9 SIMPLE LINEAR REGRESSION

This lab covers when and how we could model the relationship between two quantitative variables using a straight line, which is called a simple linear regression model; and how to conduct a hypothesis test and obtain a confidence interval for the slope of the regression model.

The following table and scatter plot show the relationship between the price (in $1000) and the age (in years) of 15 used cars of a particular make and model. Download the dataset car.xls from online and then import it into R commander.

Age (x, in year)	Price (y, in $1000)
1	13.990
1	13.495
3	12.999
4	9.500
4	10.495
5	8.995
5	9.495
6	6.999
7	6.950
7	7.850
8	6.999
8	5.995
10	4.950
10	4.495
13	2.850



Example: Simple Linear Regression Model

(a) Import the data into R commander and re-produce the scatter plot. Could we use a straight line $\hat{y} = b_0 + b_1 x$ to model relationship between price and age of the used cars?
Data→Import data→from Excel file… (name it "car")
Draw the scatter plot:
1. **Graphs→Scatterplot…**
2. In the "**Scatterplot**" window, select "**age**" as **x-variable** and "**price**" as **y-variable**.
3. Click OK.

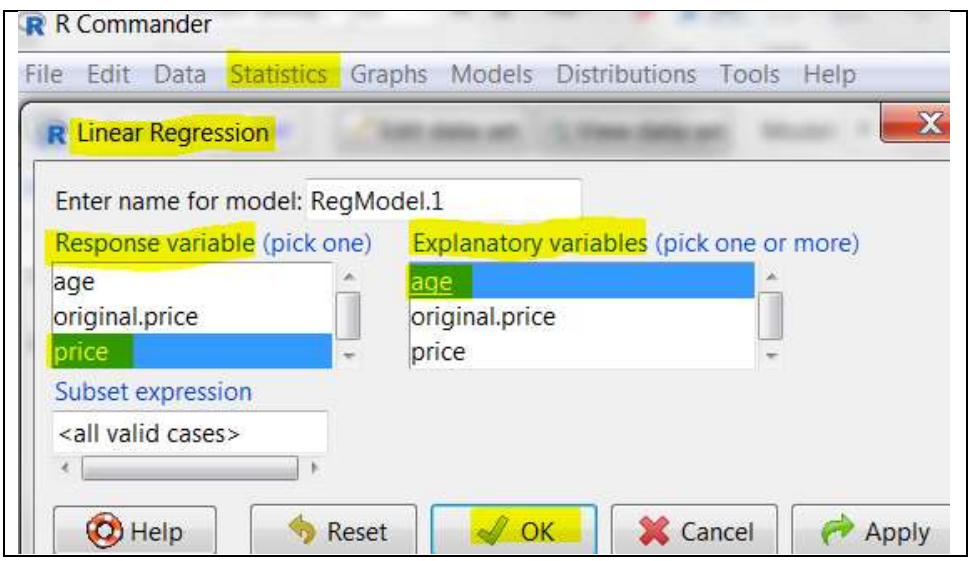
Note: The price is calculated as the original price divided by 1000.

Since all the data points are roughly on a straight line, we can use a straight line $\hat{y} = b_0 + b_1 x$ to model relationship between price and age of the used cars.

(b) Write down the least-squares regression equation.

Fit a regression model and obtain the least squares straight line:

1. **Statistics→Fit models→Linear regression…**
2. In the "**Linear Regression**", select "**price**" as the **Response variable** (dependent variable) and "**age**" as the **Explanatory variable** (independent variable).
3. Click OK.



The values of the intercept b_0 and the slope b_1 are given in the "**Estimate**" column. Based on the computer outputs, we have $b_0 = 14.28595$ and $b_1 = -0.95905$, and the fitted least-squares regression equation is

$$\hat{y} = b_0 + b_1 x \Longrightarrow \widehat{\text{price}} = 14.28595 + (-0.95905) \times \text{age} = 14.28595 - 0.95905 \times \text{age}$$

```
Call:
lm(formula = price ~ age, data = Car)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.53267 -0.55715  0.04524  0.33140  1.59019 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.28595    0.44867   31.84 1.01e-13 ***
age         -0.95905    0.06458  -14.85 1.56e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8162 on 13 degrees of freedom
Multiple R-squared:  0.9443, Adjusted R-squared:  0.9401 
F-statistic: 220.5 on 1 and 13 DF,  p-value: 1.562e-09
```

(c) Obtain and interpret the coefficient of determination r^2.
Based on the computer outputs, the coefficient of determination $r^2 = 0.9443$.
Interpretation: 94.43% of variation in the observed price of the used cars is due to the age of the used cars and can be explained by the fitted regression equation $\widehat{\text{price}} = 14.28595 - 0.95905 \times \text{age}$.

(d) Obtain and interpret the correlation coefficient r.
Since the correlation coefficient r and the slope b_1 have the same sign, and $b_1 = -0.95905$ which is negative, $r = -\sqrt{r^2} = -\sqrt{0.9443} = -0.9718$.
Interpretation: There is a strong, negative, linear association between price and age of the used cars.

(e) Test at the 5% significance level whether age is a **useful predictor** for the price of a used car.

Steps:

1. Set up the hypotheses. $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$.
2. The significance level is $\alpha = 0.05$.
3. Compute the value of the test statistic: $t_o = \frac{b_1}{\frac{s_e}{\sqrt{S_{xx}}}} = -14.85$ with $df = n - 2 = 13$.
4. Find the P-value. For a two tailed test with $df = 13$,
   P-value=$2P(t \geq |t_o|) = 2P(t \geq 14.158) = 1.56 \times 10^{-9}$.
5. Decision: reject the null H_0 since P-value= $1.56 \times 10^{-9} < 0.05(\alpha)$.
6. Conclusion: At the 5% significance level, we have sufficient evidence that age is a useful predictor for the price of a used car.

LAB 10 ONE-WAY ANOVA

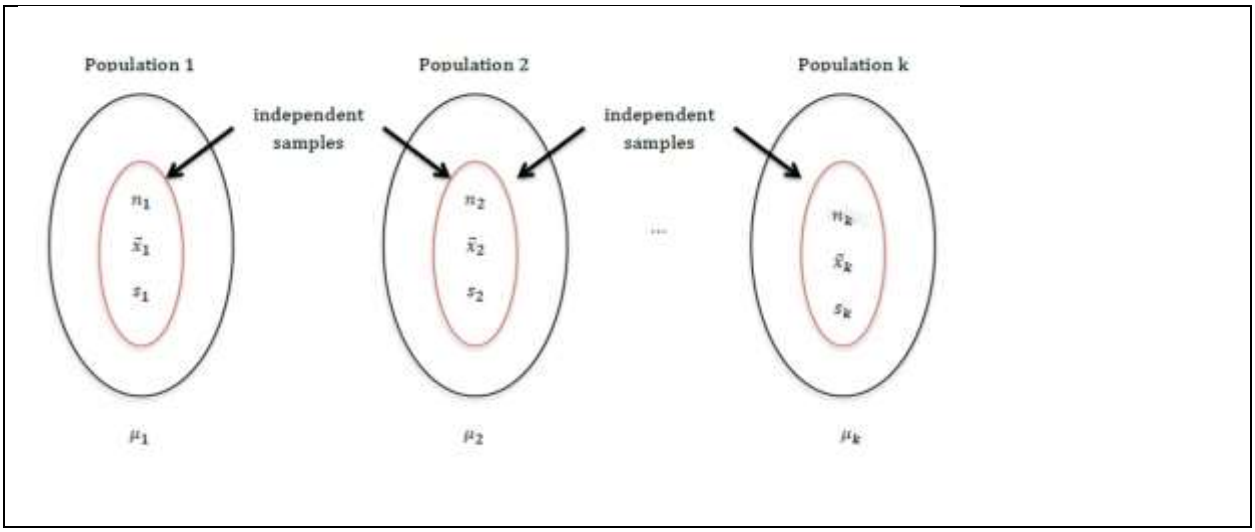
The two-sample t test can be used in comparing two population means based on two independent samples.

When comparing k $(k > 2)$ population means based on k independent samples, a one-way ANOVA can be used. ANOVA stands for **AN**alysis **O**f **VA**riance. This lab shows how to conduct a one-way ANOVA F test based on the computer output.

Let $\mu_1, \mu_2, \cdots, \mu_k$ be the population means of the k populations, respectively.

The hypotheses of one-way ANOVA are formulated as

- H_0: all means are equal, i.e., $\mu_1 = \mu_2 = \cdots = \mu_k$
- H_a: not all the means are equal.



In a two-sample t test, inference about the population means is based on two independent samples from two populations. In the ANOVA F test, inference about population means is based on k independent simple random samples from k populations.

If H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ is true, the sample means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ should be close to one another and hence the variation between sample means should be small. We should reject H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ if the sample means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ are very different.

Assumptions for One-way ANOVA F Test:

- Normal populations: for each population, the variable of interest is normally distributed.
- Equal variances: the variances of the variable of interest are the same for all populations.
- Independent samples: the samples from different populations are independent of one another.

- Simple random samples: the samples taken from the k populations should be simple random samples.

Steps:

1. Set up the hypotheses:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: \text{Not all means are equal}$$

2. State the significance level α.
3. Calculate the sums of squares SST, SSTR, SSE and the mean squares MSTR, MSE. Find the test statistic, F_o, and show the results in an ANOVA table:

Source	df	SS	$MS = \frac{SS}{df}$	F-statistic	p-value
Treatment	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F_o = \frac{MSTR}{MSE}$	$P(F \geq F_o)$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

4. Find the P-value **or** rejection region based on the F density curve with degrees of freedom $df_{numerator} = df_n = k-1, df_{denominator} = df_d = n-k$.

P-value	$P(F \geq F_o)$ the area to the right of F_o under the curve
Rejection region	$F \geq F_\alpha$ the region to the right of the critical value F_α

5. Reject the null H_0 if P-value$\leq \alpha$ or F_o falls in the rejection region.
6. Conclusions.

Example: One-way ANOVA F Test

A student performed an experiment to compare download speed at different times of the day. He placed a file on a remote server and then proceeded to download the file at three different time periods of the day: 7 a.m., 5 p.m., and 12 a.m. He downloaded the file 48 times, 16 times at each time period, and recorded the download time in seconds (De Veaux, Velleman, & Bock, 2008). Does the data below provide sufficient evidence that there is a difference between the mean download times at 7 a.m., 5 p,m, and 12 a.m.? Test at the 1 % significance level. The data can be found online in the Excel file downloading.xlsx.

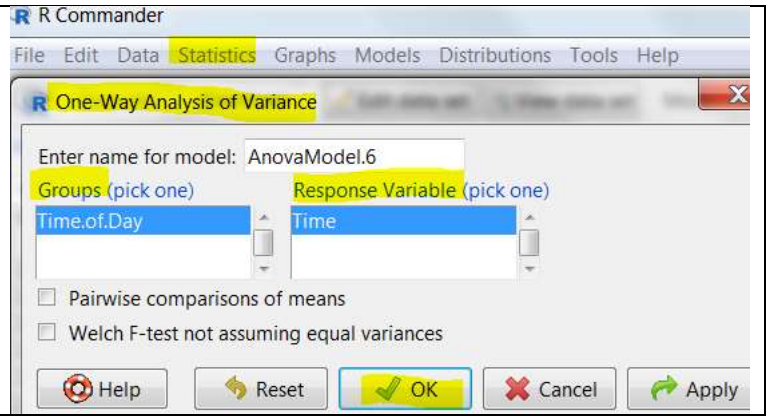
Time of Day	**Time (Sec)**	**Time of Day**	**Time (Sec)**	**Time of Day**	**Time (Sec)**
Early (7AM)	68	Evening (5 PM)	299	Late Night (12 AM)	216
Early (7AM)	138	Evening (5 PM)	367	Late Night (12 AM)	175
Early (7AM)	75	Evening (5 PM)	331	Late Night (12 AM)	274
Early (7AM)	186	Evening (5 PM)	257	Late Night (12 AM)	171
Early (7AM)	68	Evening (5 PM)	260	Late Night (12 AM)	187
Early (7AM)	217	Evening (5 PM)	269	Late Night (12 AM)	213
Early (7AM)	93	Evening (5 PM)	252	Late Night (12 AM)	221
Early (7AM)	90	Evening (5 PM)	200	Late Night (12 AM)	139

Early (7AM)	71	Evening (5 PM)	296	Late Night (12 AM)	226
Early (7AM)	154	Evening (5 PM)	204	Late Night (12 AM)	128
Early (7AM)	166	Evening (5 PM)	190	Late Night (12 AM)	236
Early (7AM)	130	Evening (5 PM)	240	Late Night (12 AM)	128
Early (7AM)	72	Evening (5 PM)	350	Late Night (12 AM)	217
Early (7AM)	81	Evening (5 PM)	256	Late Night (12 AM)	196
Early (7AM)	76	Evening (5 PM)	282	Late Night (12 AM)	201
Early (7AM)	129	Evening (5 PM)	320	Late Night (12 AM)	161

Import the data into R: **Data→Import data→from Excel file… (name it "downloading")**

Conduct the one-way ANOVA F test in R:

1. **Statistics→Means→One-way ANOVA…**
2. In the "**One-Way Analysis of Variance**" window, choose "**Time of Day**" as the Group variable and "**Time**" as the Response Variable.
3. Click OK.



Computer outputs

```
> summary(AnovaModel.6)
            Df Sum Sq Mean Sq F value   Pr(>F)
Time.of.Day  2 204641  102320   46.03 1.31e-11 ***
Residuals   45 100020    2223
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Steps to conduct a one-way ANOVA F-test:

1. Hypotheses

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_a: \text{Not all means are equal}$$

2. Significance level is $\alpha = 0.01$.
3. Test statistic $F_o = 46.03$ with $df_n = k - 1 = 3 - 1 = 2, df_d = n - k = 48 - 3 = 45$.
4. P-value= $P(F \geq F_o) = P(F \geq 46.03) = 1.31 \times 10^{-11}$ (given in the ANOVA table).
5. Reject H_0, since p-value=$1.31 \times 10^{-11} < 0.01\ (\alpha)$.
6. Conclusion: At the 1% significance level we have sufficient evidence that there is a significant difference between the mean downloading time at 7 a.m., 5 p.m., and 12 a.m.