

Using R in an Introductory Statistics Course*

Alan T. Arnholt

Department of Mathematical Sciences

Appalachian State University

arnholt@math.appstate.edu

18th Annual International Conference on Technology in
Collegiate Mathematics, March, 2006, Orlando, Florida

What is R?

R U Ready?

Audience and Class

My Audience

Course Objectives

Examples

R and Tinn-R

My History

R Features

Tinn-R

Installing R

Installing R and Using Packages

Resources

Resources for the introductory course

R U Ready?

R U Ready?

- R is a statistical programming language not unlike the non-GUI commands in S-PLUS. R is a derivative of the original S System, an innovative software program that helps users to manage and extract useful information from data.

R U Ready?

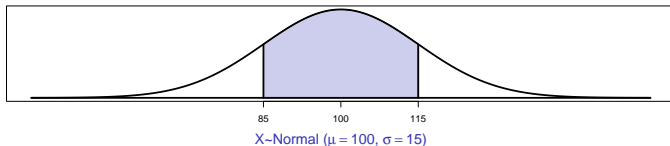
- R is a statistical programming language not unlike the non-GUI commands in S-PLUS. R is a derivative of the original S System, an innovative software program that helps users to manage and extract useful information from data.
- John Chambers, the developer of the S System, is now a core member of the R development team.

R U Ready?

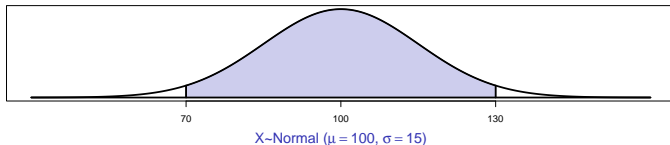
- R is a statistical programming language not unlike the non-GUI commands in S-PLUS. R is a derivative of the original S System, an innovative software program that helps users to manage and extract useful information from data.
- John Chambers, the developer of the S System, is now a core member of the R development team.
- In fact, in the February, 2005, issue of the *Journal of Statistical Software*, Jan de Leeuw, said “It is obvious now, and it was obvious then, that S was rapidly becoming the **lingua franca** of statistics.”

Graphical Illustration of Empirical Rule

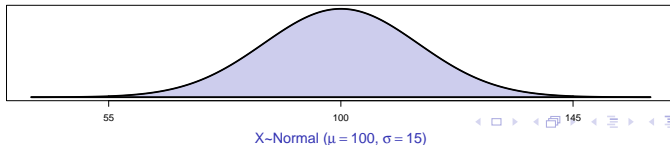
The area between 85 and 115 is 0.6827



The area between 70 and 130 is 0.9545

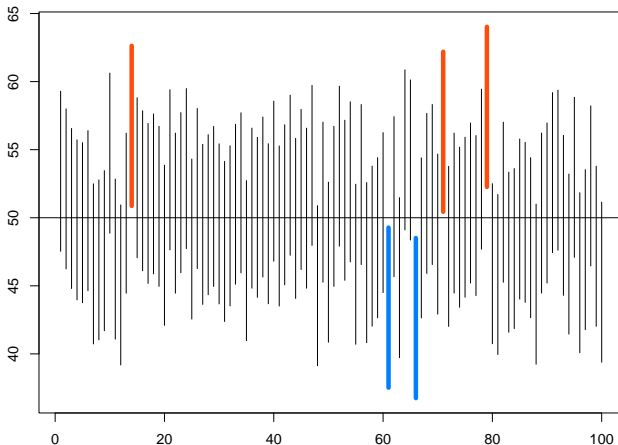


The area between 55 and 145 is 0.9973



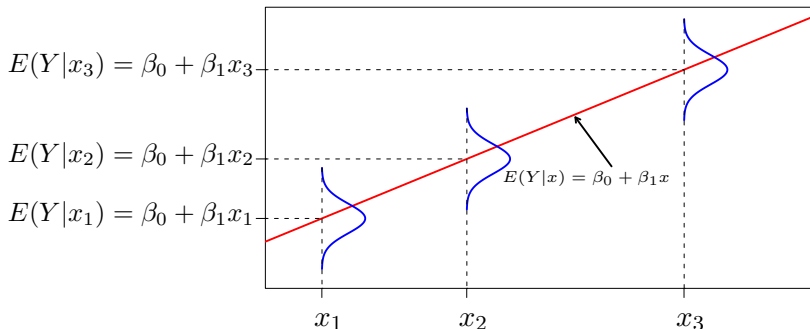
Confidence Interval Simulation

100 random 95% confidence intervals where $\mu = 50$



Note: 5% of the random confidence intervals do not contain $\mu=50$

Graphical Representation of Simple Linear Regression



Statistical Problems

Many statistical problems can be broken down into three components:

- Collecting data

Statistical Problems

Many statistical problems can be broken down into three components:

- Collecting data
- Analyzing / summarizing the collected data

Statistical Problems

Many statistical problems can be broken down into three components:

- Collecting data
- Analyzing / summarizing the collected data
- Interpreting the analyzed data

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

Personal Position

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

Personal Position

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)
- Computing is integrally related to statistics.

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

Personal Position

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)
- Computing is integrally related to statistics.
- Any programming language or software program is simply a tool to implement a concept.

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

Personal Position

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)
- Computing is integrally related to statistics.
- Any programming language or software program is simply a tool to implement a concept.
- Working with large data sets by hand is not time effective.

The course I am teaching assumes data is collected correctly. This talk will focus on the analyzing / summarizing of collected data, for which my class uses R.

Personal Position

- If you cannot implement a concept, you do not understand the concept. (Example: Given some data, find the percent of values that fall within plus or minus two standard deviations of the mean.)
- Computing is integrally related to statistics.
- Any programming language or software program is simply a tool to implement a concept.
- Working with large data sets by hand is not time effective.
- Simulations are even more effective when the students can code them versus when the students use the instructor's code or an applet.

To whom am I teaching?

To whom am I teaching?

- Appalachian State University is a comprehensive state university.

To whom am I teaching?

- Appalachian State University is a comprehensive state university.
- Students come from a variety of majors (50% psychology, 25% biology, and 25% many disciplines).

To whom am I teaching?

- Appalachian State University is a comprehensive state university.
- Students come from a variety of majors (50% psychology, 25% biology, and 25% many disciplines).
- Incoming freshman average 1100 on the SAT.

To whom am I teaching?

- Appalachian State University is a comprehensive state university.
- Students come from a variety of majors (50% psychology, 25% biology, and 25% many disciplines).
- Incoming freshman average 1100 on the SAT.
- The majority of students are sophomores.

How does my class work?

How does my class work?

- Students install R and Tinn-R (a text editor) on machines in class — first class period.

How does my class work?

- Students install R and Tinn-R (a text editor) on machines in class — first class period.
- Students have a pen drive on which they install R and Tinn-R — second class period.

How does my class work?

- Students install R and Tinn-R (a text editor) on machines in class — first class period.
- Students have a pen drive on which they install R and Tinn-R — second class period.
- Students use R!

How does my class work?

- Students install R and Tinn-R (a text editor) on machines in class — first class period.
- Students have a pen drive on which they install R and Tinn-R — second class period.
- Students use R!
- Students use scripts that mirror R on slides — this allows students to take notes (Tinn-R) during class.

How does my class work?

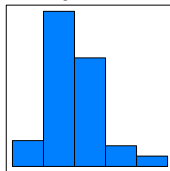
- Students install R and Tinn-R (a text editor) on machines in class — first class period.
- Students have a pen drive on which they install R and Tinn-R — second class period.
- Students use R!
- Students use scripts that mirror R on slides — this allows students to take notes (Tinn-R) during class.
- Students are in front of computers **every class**.

Topics and course objectives

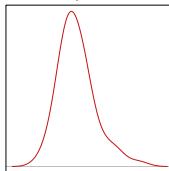
Students should be able to organize and summarize univariate data.

EXPLORATORY DATA ANALYSIS

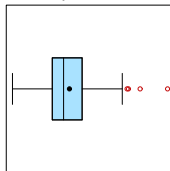
Histogram of cholest



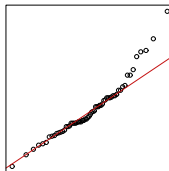
Density of cholest



Boxplot of cholest

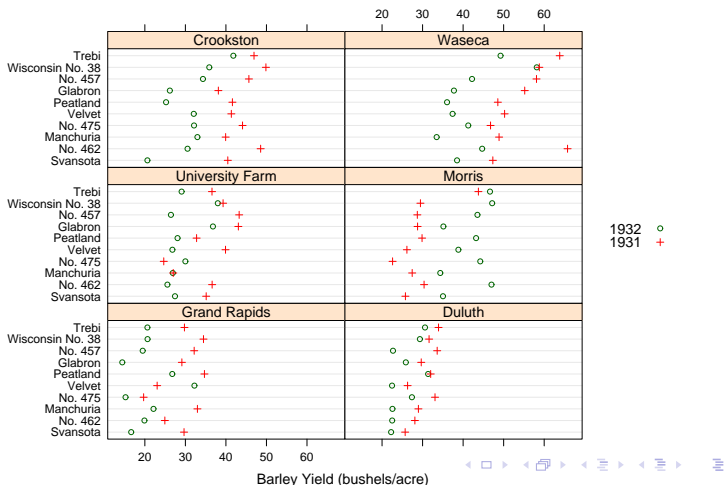


Q-Q Plot of cholest



Topics and course objectives

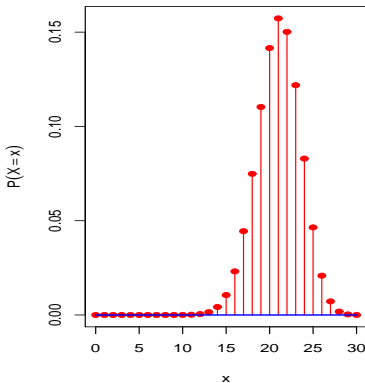
Students should be able to organize and summarize multivariate data.



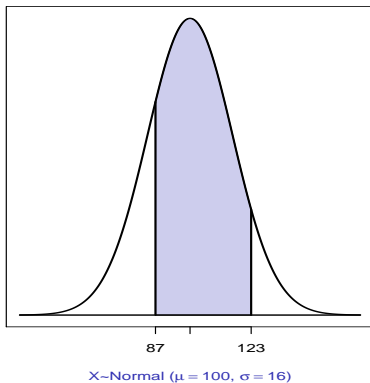
Topics and course objectives

Students should be able to solve problems involving the binomial and normal distributions.

X~Bin(30 , 0.7)

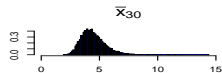
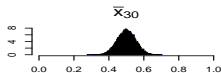
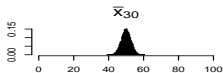
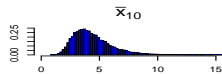
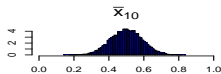
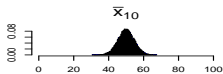
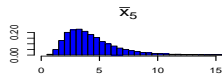
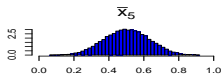
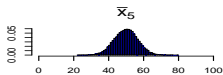
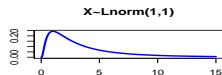
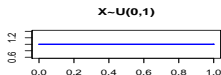
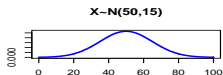


The area between 87 and 123 is 0.7165



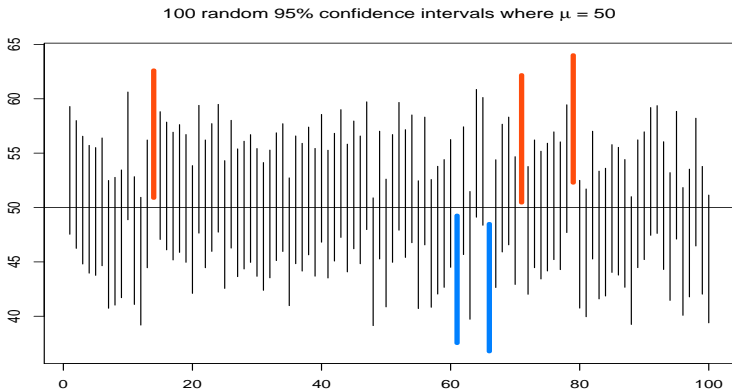
Topics and course objectives

Students should understand the ideas behind a sampling distribution.



Topics and course objectives

Students should understand the logic behind the creation as well as be able to compute and interpret confidence intervals for unknown parameters.



Note: 5% of the random confidence intervals do not contain $\mu=50$

Topics and course objectives

Students should understand the logic behind hypothesis testing and be able to implement that logic with practical scenarios.

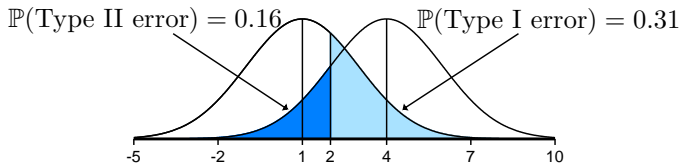


Figure: Graphical representation of type I and type II errors when $H_0: \mu = 1$ versus $H_1: \mu = 4$.

How did I get to R?

How did I get to R?

- Used Minitab™ from 1993-2002.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.
 - Student frustration went down.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.
 - Student frustration went down.
 - Students learning went down as well.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.
 - Student frustration went down.
 - Students learning went down as well.
- Fall 2004 started using R in service course.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.
 - Student frustration went down.
 - Students learning went down as well.
- Fall 2004 started using R in service course.
- Because R is free and so similar in functionality to S-PLUS, I could not justify the cost of renewing the S-PLUS license.

How did I get to R?

- Used Minitab™ from 1993-2002.
- Used S-PLUS 2002/2003.
- Fall 2002 students used S-PLUS by typing commands (encountered some frustration).
- Wrote GUI front end for everything students used in course.
 - Student frustration went down.
 - Students learning went down as well.
- Fall 2004 started using R in service course.
- Because R is free and so similar in functionality to S-PLUS, I could not justify the cost of renewing the S-PLUS license.
- Wanted to create robust course materials (GUI materials hard to keep current)

What is great about R?

What is great about R?

Advantages of using R include:

What is great about R?

Advantages of using R include:

- It is free.

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.
 - Extremely portable

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.
 - Extremely portable
 - Students use it more.

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.
 - Extremely portable
 - Students use it more.
 - Students don't have to go to a lab if they have a home computer.

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.
 - Extremely portable
 - Students use it more.
 - Students don't have to go to a lab if they have a home computer.
- An incredible amount of documentation

What is great about R?

Advantages of using R include:

- It is free.
- Students have no excuse not to get it.
- Can be installed and run from a pen drive.
 - Extremely portable
 - Students use it more.
 - Students don't have to go to a lab if they have a home computer.
- An incredible amount of documentation
- Extremely powerful and difficult to misuse

What is great about R? (continued)

- Commands do not change as GUI menus do.

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.
- Easily extensible

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.
- Easily extensible
- Open source

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.
- Easily extensible
- Open source
- Analyses are reproducible.

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.
- Easily extensible
- Open source
- Analyses are reproducible.
- Seamless integration into reports/slides using [Sweave](#)

What is great about R? (continued)

- Commands do not change as GUI menus do.
- Simulations are easy.
- Easily extensible
- Open source
- Analyses are reproducible.
- Seamless integration into reports/slides using [Sweave](#)
- Programming language of choice for research statisticians

Power of a programming language

The cholesterol levels of 62 subjects in the Framingham Heart Study are stored in the variable `chol` of the data frame `Framingh` found in the BSDA package.

Power of a programming language

The cholesterol levels of 62 subjects in the Framingham Heart Study are stored in the variable `cholest` of the data frame `Framingh` found in the BSDA package.

- What percent of the actual data falls within one standard deviation of the mean?

Power of a programming language

The cholesterol levels of 62 subjects in the Framingham Heart Study are stored in the variable `chol` of the data frame `Framingh` found in the BSDA package.

- What percent of the actual data falls within one standard deviation of the mean?
- What percent of the actual data falls within two standard deviations of the mean?

Power of a programming language

The cholesterol levels of 62 subjects in the Framingham Heart Study are stored in the variable `chol` of the data frame

Framingh found in the BSDA package.

- What percent of the actual data falls within one standard deviation of the mean?
- What percent of the actual data falls within two standard deviations of the mean?
- Create a histogram of the cholesterol values and depict the range of values for the mean \pm two standard deviations with a double arrow.

Tinn-R Code

The screenshot shows the Tinn-R editor interface. The title bar reads "Tinn-R". The menu bar includes "File", "Edit", "Format", "Projects", "Search", "Options", "Tools", "R", "View", "Window", "Web", and "Help". The toolbar contains various icons for file operations, editing, and execution. The main text area shows the following R code:

```

C:\texmf\atascrce\beamer\ICTCM2006\talk.r
library(BSDA)
attach(Framingh)
str(Framingh)
Altblue <- "#CDCDED"
mx <- mean(cholest)
sx <- sd(cholest)
c(mx, sx)
c(mx-sx, mx+sx)
c(mx-2*sx, mx+2*sx)
(sum(cholest>(mx-1*sx) &cholest<(mx+1*sx))/length(cholest))*100
(sum(cholest>(mx-2*sx) &cholest<(mx+2*sx))/length(cholest))*100

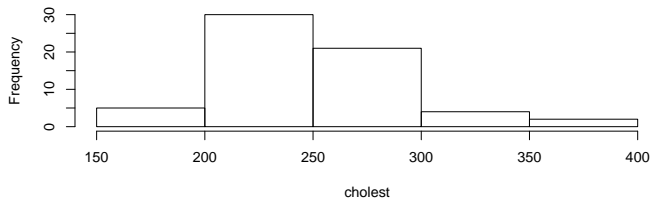
par(mfrow=c(2,1))
hist(cholest)
hist(cholest, col=Altblue, breaks=10, ylim=c(0, 30)
, xlim=c(150, 400), main="Fancy Histogram")
arrows(mx-2*sx, 20, mx+2*sx, 20, lwd=2, code=3, length=.1)
text(mx, 24, expression(hat(mu) %+-% 2*hat(sigma)))
text(mx, 29, expression(bar(x) %+-% 2*s))
par(mfrow=c(1,1))

```

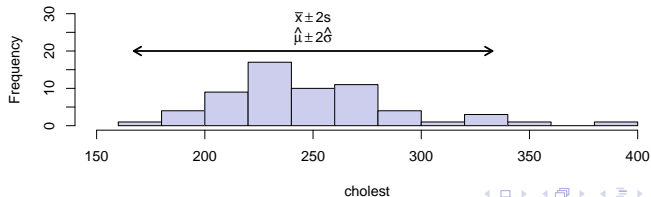
The status bar at the bottom left shows "talk.r".

R Graphs

Histogram of cholest



Fancy Histogram



Graphs and R Code

Histogram of cholest

Frequency

cholest

Fancy Histogram

Frequency

cholest

$\bar{x} \pm 2s$
 $\hat{\mu} \pm 2\hat{\sigma}$

```

Attaching package: 'BSDA'

The following object(s) are masked from package:dat:

  Orange

> attach(Framingh)
> str(Framingh)
'data.frame': 62 obs. of 1 variable:
 $ cholest: int 393 353 334 336 327 300 300 308 283 285 ...
> altblue <- "#CDCDED"
> mx <- mean(cholest)
> sx <- sd(cholest)
> c(mx,sx)
[1] 250.03226 41.44321
> c(mx-sx,mx+sx)
[1] 208.5890 291.4755
> c(mx-2*sx,mx+2*sx)
[1] 167.1458 332.9187
> (sum(cholest>[mx-1*sx] &cholest<[mx+1*sx])/length(cholest))$
[1] 77.41935
> (sum(cholest>[mx-2*sx] &cholest<[mx+2*sx])/length(cholest))$
[1] 91.93548
>
> par(mfrow=c(2,1))
> hist(cholest)
> hist(cholest,col=altblue,breaks=10,ylim=c(0,30)
+ ,xlim=c(150,400),main="Fancy Histogram")
> arrows(mx-2*sx,20,mx+2*sx,20,lwd=2,code=3,length=.1)
> text(mx,24,expression(hat(mu) %+-% 2*hat(sigma)))
> text(mx,29,expression(bar(x) %+-% 2*s))
> par(mfrow=c(1,1))
  
```

How can you get R working for you and your students?

How can you get R working for you and your students?

The directions in what follows apply to R under Windows. For other operating systems, please follow the directions provided in the [FAQ - \(2.5 How can R be installed?\)](#)

How can you get R working for you and your students?

The directions in what follows apply to R under Windows. For other operating systems, please follow the directions provided in the [FAQ - \(2.5 How can R be installed?\)](#)

- Go to your nearest CRAN site to download R
<http://cran.r-project.org/mirrors.html>.
- In the Precompiled Binary Distributions click on [Windows \(95 and later\)](#)
- Next click on [base](#)
- Download the current version of R by clicking on the file [R-2.2.1-win32.exe](#). When the file download prompt appears, select save. Make sure you note where you save the download!

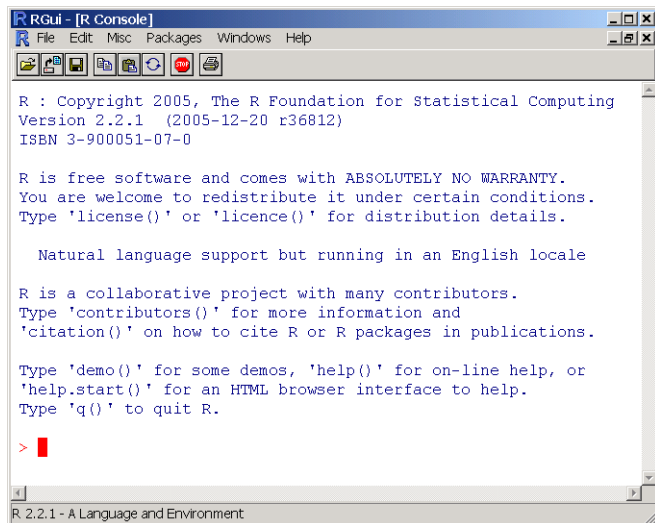
Installing R

- Navigate to the folder where the file `R-2.2.1-win32.exe` was saved.
- Double click on the file `R-2.2.1-win32.exe` and answer the Setup questions.
- Note: You may not have permission to install R on your Lab computers. However, you can always install R to a pen drive (provided your pen drive has at least 100 megs of free space - this may take 15-20 minutes) and subsequently run R from your pen drive. If you are installing R on a pen drive, make sure to specify the location where you would like R to be installed. For example, if your pen drive is in the F drive, you might specify `F:/Program Files/R/R-2.2.1` as your install folder.
- Use the default values for your installation unless you know what you are doing.

Launching R

- You should have a shortcut R icon appear on the machine where you downloaded R provided you choose the default installation values. However, if you installed R to a pen drive on a University/Lab computer, the shortcut icon will more than likely disappear when the machine is shut down.
- To launch R, either click on the R shortcut icon on the desktop or navigate to the bin folder (Program Files/R/R-2.2.1/bin) and click on the Rgui.exe file.

R Console



```
RGui - [R Console]
File Edit Misc Packages Windows Help
[Icons: Home, Open, Save, Print, Refresh, Stop, Copy]

R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.2.1 (2005-12-20 r36812)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █

R 2.2.1 - A Language and Environment
```

Downloading Packages (BSDA)

My class relies heavily on the BSDA package which needs to be both installed and loaded. To install BSDA,

Downloading Packages (BSDA)

My class relies heavily on the BSDA package which needs to be both installed and loaded. To install BSDA,

- Click on *Packages > Install Package(s)*.
 - Select an appropriate mirror.
 - Select the packages you want to install (BSDA).
 - Click on *OK* and BSDA and six additional packages required by BSDA will be downloaded and installed.
 - To load BSDA, click on *Packages > Load Package > BSDA*.
- Note:** You only install a package once. However, to use the package, you must load it each time you launch R.

Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is **strongly** recommended. There are several editors to choose from.

Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is **strongly** recommended. There are several editors to choose from.
- Tinn-R is an excellent choice for students who will only use the editor to interact with R. The most recent stable version of Tinn-R can be found at <http://www.sciviews.org/Tinn-R/index.html>.

Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is **strongly** recommended. There are several editors to choose from.
- Tinn-R is an excellent choice for students who will only use the editor to interact with R. The most recent stable version of Tinn-R can be found at <http://www.sciviews.org/Tinn-R/index.html>.
- If you have installed R on your pen drive, you will also want to install Tinn-R on your pen drive.

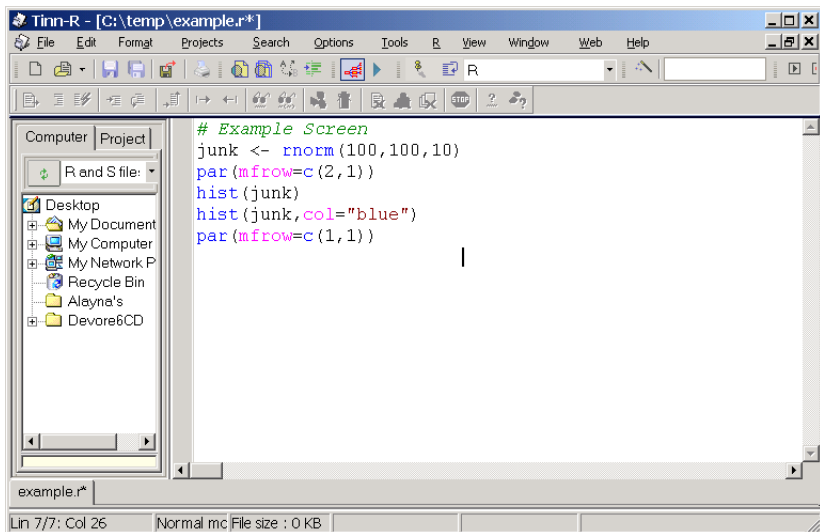
Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is **strongly** recommended. There are several editors to choose from.
- Tinn-R is an excellent choice for students who will only use the editor to interact with R. The most recent stable version of Tinn-R can be found at <http://www.sciviews.org/Tinn-R/index.html>.
- If you have installed R on your pen drive, you will also want to install Tinn-R on your pen drive.
- To launch Tinn-R, click on Tinn-R.exe which is in the Tinn-R/bin folder provided the default options were selected while installing Tinn-R.

Using an Editor (Optional)

- Although you can type commands directly in the R console, the use of an editor is **strongly** recommended. There are several editors to choose from.
- Tinn-R is an excellent choice for students who will only use the editor to interact with R. The most recent stable version of Tinn-R can be found at <http://www.sciviews.org/Tinn-R/index.html>.
- If you have installed R on your pen drive, you will also want to install Tinn-R on your pen drive.
- To launch Tinn-R, click on Tinn-R.exe which is in the Tinn-R/bin folder provided the default options were selected while installing Tinn-R.
- To use Tinn-R, type your commands in the Tinn-R window as shown in the next slide. Select *R > Send to R > All* to send all of the typed commands to R.

Tinn-R



The screenshot shows the Tinn-R editor window. The title bar reads "Tinn-R - [C:\temp\example.r*]". The menu bar includes File, Edit, Format, Projects, Search, Options, Tools, R, View, Window, Web, and Help. The toolbar contains various icons for file operations and execution. On the left, a file explorer shows the "R and S file:" view with a tree structure including Desktop, My Document, My Computer, My Network P, Recycle Bin, Alayna's, and Devore6CD. The main editor area contains the following R code:

```
# Example Screen
junk <- rnorm(100,100,10)
par(mfrow=c(2,1))
hist(junk)
hist(junk,col="blue")
par(mfrow=c(1,1))
```

The status bar at the bottom indicates "example.r*" and "Lin 7/7: Col 26 Normal mc File size : 0 KB".

R Script and Graph

The screenshot displays the RGui environment. On the left, two histograms titled "Histogram of junk" are shown. The top histogram has white bars, and the bottom one has blue bars. Both have a y-axis labeled "Frequency" ranging from 0 to 15 and an x-axis labeled "junk" ranging from 70 to 120. On the right, the R Console window shows the R version (2.2.1), ISBN (3-900051-07-0), and introductory text about R's license and features. Below the text, a series of R commands is entered, demonstrating the generation of a histogram with blue bars.

```

RGui
File History Resize Windows
R Graphics: Device 2 (ACTIVE)
R Console
Version 2.2.1 (2005-12-20 r36812)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO
You are welcome to redistribute it under certain
Type 'license()' or 'licence()' for distribution
information.

Natural language support but running in an
English locale.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in
publications.

Type 'demo()' for some demos, 'help()' for on-line
'help.start()' for an HTML browser interface to
help, and 'q()' to quit R.

> # Example Screen
> junk <- rnorm(100,100,10)
> par(mfrow=c(2,1))
> hist(junk)
> hist(junk,col="blue")
> par(mfrow=c(1,1))
> █
  
```

Resources for your Class

- [CRAN contributed Documentation](#) — Contributed materials not all in English.

Resources for your Class

- [CRAN contributed Documentation](#) — Contributed materials not all in English.
- Of particular interest to beginning students are the works “Using R for Data Analysis and Graphics — Introduction, Examples and Commentary” by John Maindonald and “Simple R” by John Verzani.

Resources for your Class

- [CRAN contributed Documentation](#) — Contributed materials not all in English.
- Of particular interest to beginning students are the works “Using R for Data Analysis and Graphics — Introduction, Examples and Commentary” by John Maindonald and “Simple R” by John Verzani.
- [Statistics and R](#) — A collection of slides and R scripts I have created.