Introduction, Sections 1.1 and 1.3, The R Statistical Package

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Grading, homework, exams

- Stat 205: Elementary Statistics for the Biological and Life Sciences, Tuesday/Thursday 1:15pm-2:30pm in Leconte College 201A.
- Grade is 40% exams: 20% each for exams I and II.
- 50% of your grade is homework, 10% attendance.
- Homework: there will be 8 or 9 homeworks graded over the course; most use R.
- Exam I will be in October, Exam II Dec. 12.
- No late homework.

Homework and exams

- 8 or 9 homeworks you will turn in for credit. Most will be a statistical analysis in R, with pertinent output included and a short write-up. First homework next week.
- Homework problems from each section will be assigned *but not collected or graded*. These will form the basis of exam questions.
- Two non-cumulative exams, each covering about half of the course. Each exam is worth 20% of your grade.
- Attendance is 10% of your grade; 2.5 hours a week is a small investment for the wealth of knowledge you will gain!
- Strong, positive correlation between attendance and your final grade.

Topics we'll cover

- Graphical displays and summary statistics
- Probability, random variables (normal and binomial)
- Confidence intervals for μ and p
- Two-sample testing and CI
- 2×2 tables: relative risk & odds ratios
- Analysis of variance
- Linear regression
- Logistic regression, survival analysis, diagnostic screening
- Use of the statistical package R to analyze real data

Motivation: why analyze data?

- Clinical trials/drug development compare existing treatments with new methods to cure disease.
- Agriculture enhance crop yields, improve pest resistance.
- Ecology study how ecosystems develop/respond to environmental impacts.
- Lab studies learn more about biological tissue/cellular activity.

1.1 Statistics and the life sciences

- Statistics is the science of
  - collecting,
  - summarizing,
  - analyzing, and
  - interpreting

  data.
- Goal: to understand the underlying biological phenomena that generate the data.
- Statistics separates *signal* from *noise*.
- Are there *associations* or *relationships* among variables in the data?

Example 1.1.2: liver tumors in mice

Table 1.1.2 Incidence of liver tumors in mice

Response	Treatment	
	E. coli	Germ free
Liver tumors	8	19
No liver tumors	5	30
Total	13	49
Percent with liver tumors	62%	39%

- Is there an association between germ environment (germ-free vs. *E. coli*) and whether liver tumors develop?
- Is the association perfect?
- Statistics can help answer *whether* there's a difference and further *quantify* the effect of germ exposure (Chapter 10).

Example 1.1.4: MOA and schizophrenia

Table 1.1.4 MAO activity in schizophrenic patients

Diagnosis	MAO activity				
I:	6.8	4.1	7.3	14.2	18.8
Chronic undifferentiated	9.9	7.4	11.9	5.2	7.8
schizophrenic	7.8	8.7	12.7	14.5	10.7
(18 patients)	8.4	9.7	10.6		
II:	7.8	4.4	11.4	3.1	4.3
Undifferentiated with	10.1	1.5	7.4	5.2	10.0
paranoid features	3.7	5.5	8.5	7.7	6.8
(16 patients)	3.1				
III:	6.4	10.8	1.1	2.9	4.5
Paranoid schizophrenic (8 patients)	5.8	9.4	6.8		

- Monoamine oxidase (MOA) enzyme thought to regulate behavior.
- Blood from $n = 42$ schizophrenia patients collected, *stratified* by diagnosis (I, II, III).
- Is there an association between MOA and diagnosis?

Example 1.1.4: MOA and schizophrenia

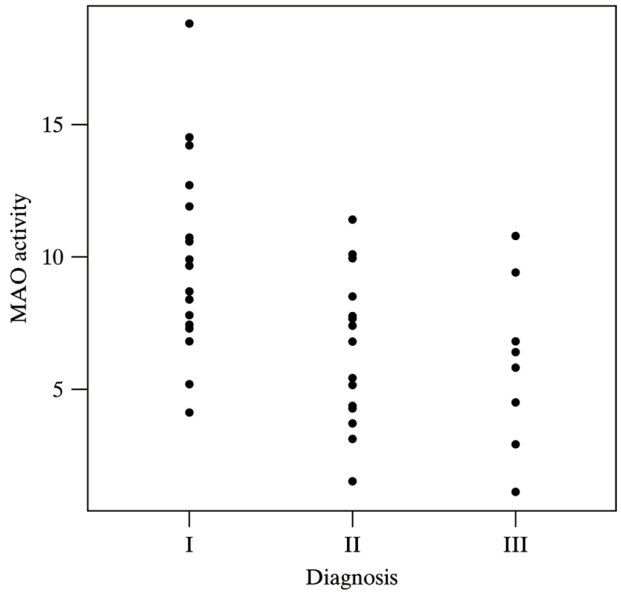


Figure 1.1.2 MAO activity in schizophrenic patients

- What happens to MOA as severity of diagnosis increases? Is the relationship perfect?
- These are side-by-side dotplots, described in Sec. 2.2. Formal approach in Chapter 11.

Example 1.1.6: Body size & energy expenditure

Table 1.1.6 Fat-free mass and energy expenditure

Subject	Fat-free mass (kg)	24-hour energy expenditure (kcal)	
1	49.3	1,851	1,936
2	59.3	2,209	1,891
3	68.3	2,283	2,423
4	48.1	1,885	1,791
5	57.6	1,929	1,967
6	78.1	2,490	2,567
7	76.1	2,484	2,653

- Fat-free body mass (kg) & 24-hour sedentary energy expenditure (kcal) measured *twice* for each of $n = 7$ men.
- Question: is there an association between body mass and energy expenditure? How can we formally assess this?
- We can informally assess association via a scatterplot of the data; formally in Chapter 12.

Example 1.1.6: Body size & energy expenditure

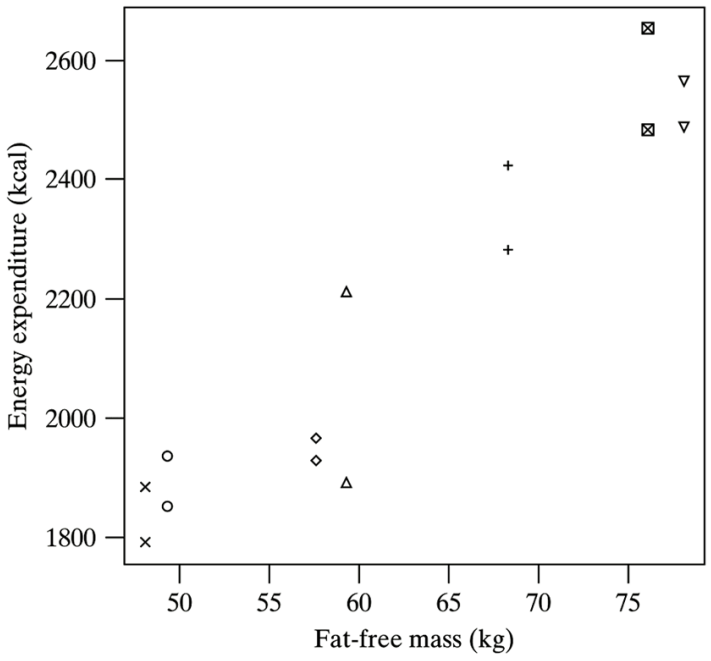


Figure 1.1.4 Fat-free mass and energy expenditure in seven men. Each man is represented by a different symbol.

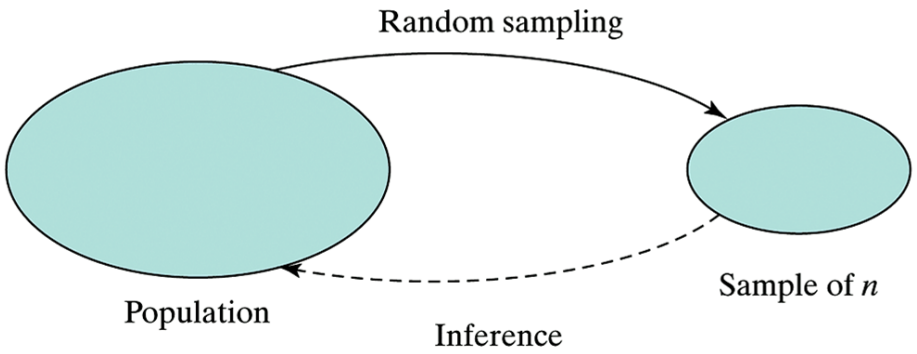
- Is there an association? Is it positive or negative? Is this what you would expect?
- Two sources of variability: *within* each man, and *among* men.

1.3 Random sampling

- Data can come from observational studies, planned experiments, clinical trials, etc.
- Data are *random*. Formally, a piece of data is a random variable (Chapter 3).
- The underlying mathematics that drives the methodology in this course relies on assuming data are a **random sample** from their population.
- A **random sample** is one it which each subject has the same probability of being measured, and subjects are chosen independently of each other.
- This provides a representative set of observations from the **population**, the data $Y_1, Y_2, \ldots, Y_n$.

Random sampling

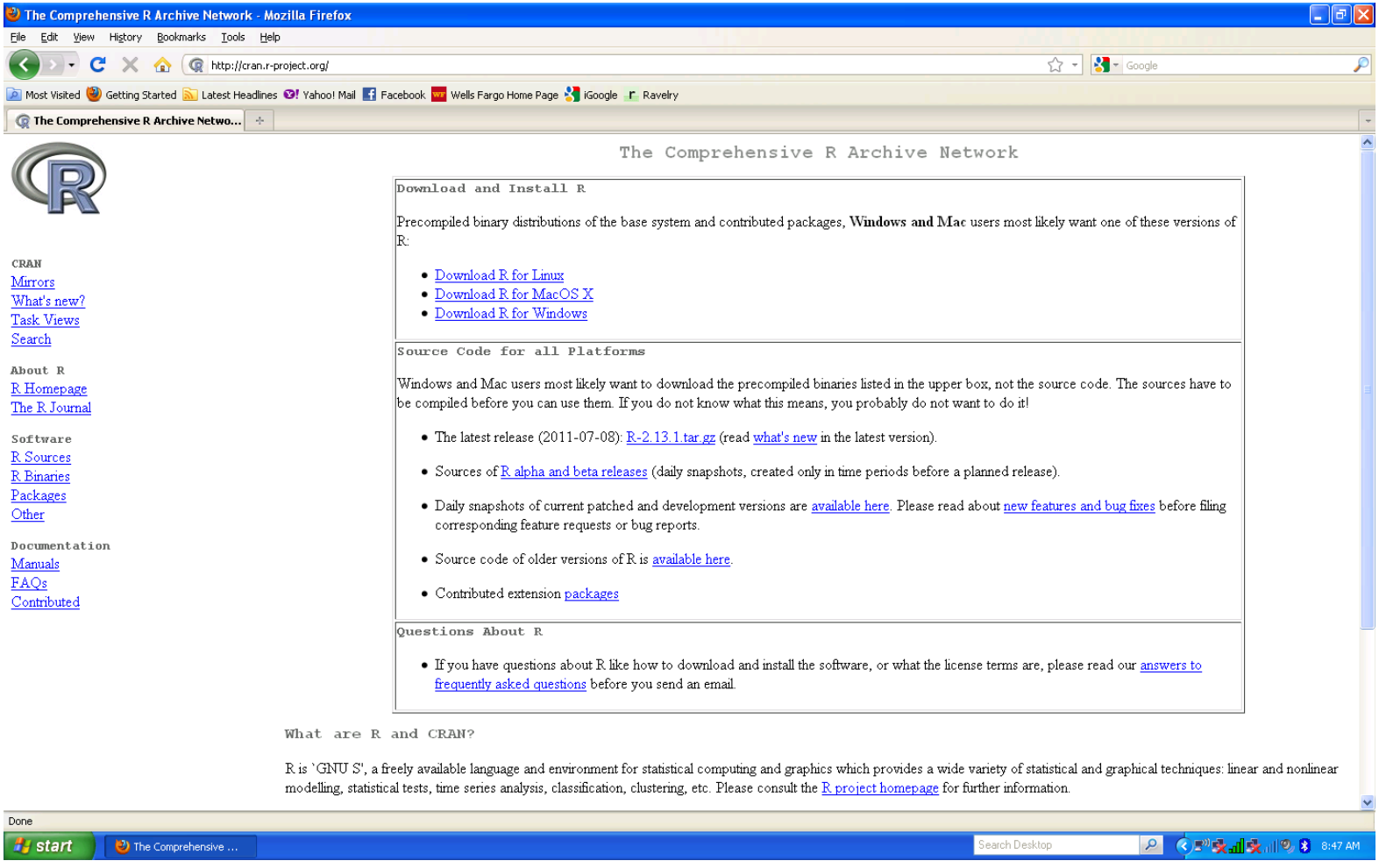
- The **population** is *all* the subjects/animals/specimens/etc. of interest.
- Since we can't measure the entire population (usually) we take a small sample of size n and use the data collected to *infer* about the population.



R computing & graphics package

- R is a powerful, free statistical computing and graphics package.
- Popular with many researchers due to contributed packages: R functions to do specialized, advanced, & often complex statistical analyses.
- R can also do many important, routine calculations, analyses, and provide common graphical displays used in this course.
- Installed in several of the computing labs across campus, e.g. Sloan 108 & 109, Gambrell 003.
- You can download it and install it from CRAN: `https://cran.r-project.org/`

The Comprehensive R Archive Network

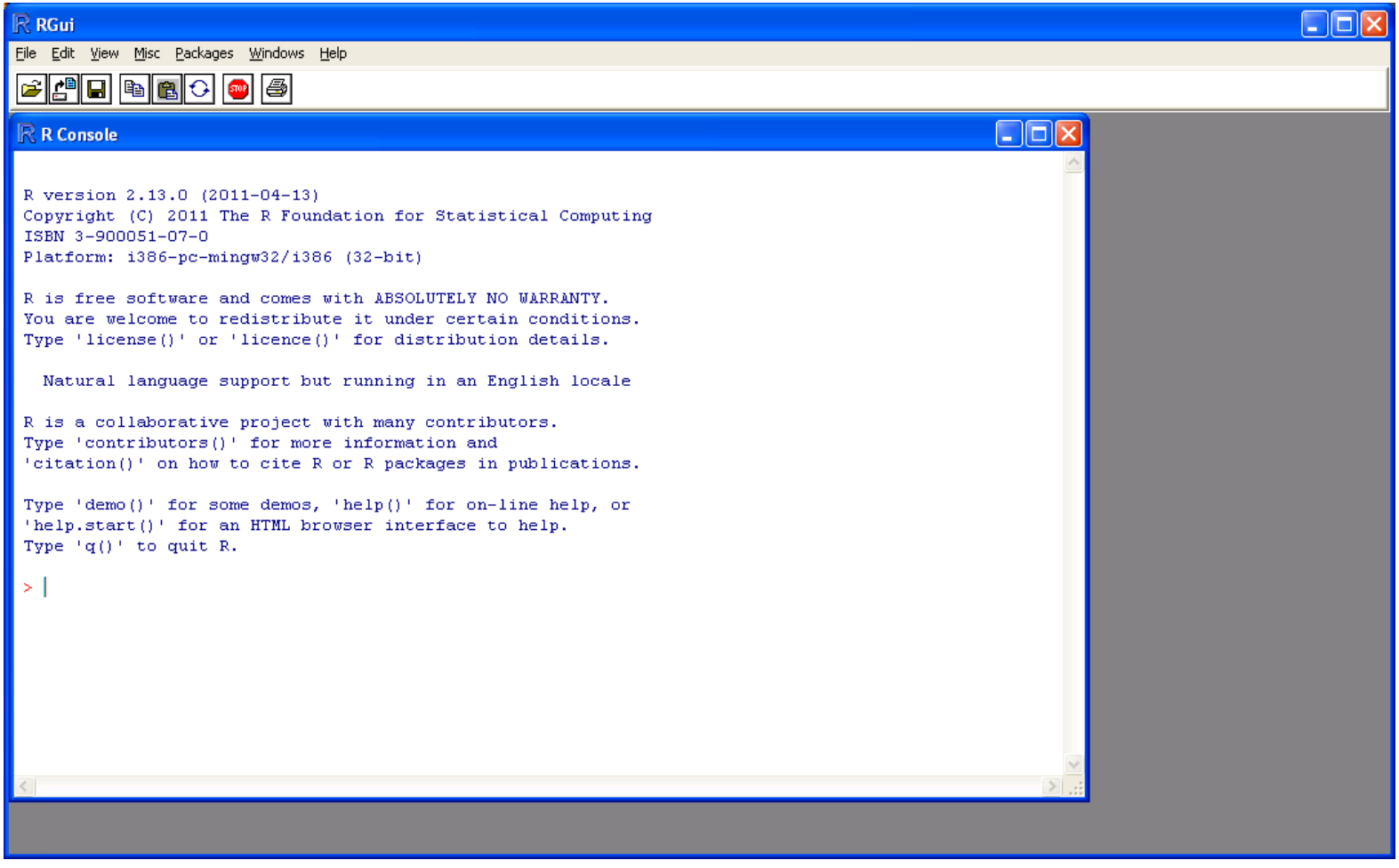


Here is where you download R.

Installing R

- From `https://cran.r-project.org/`, under `Download and Install R` click on your platform (Linux, MacOS X, or Windows).
- for Windows click on `base` and on the next page click on `Download R 3.4.1 for Windows` (this is the latest release as of August 2017).
- Click Save File and when it's done downloading run the executable by clicking on it – alternatively you can choose to Run Program directly after downloading from the web.
- The installation program will ask you a series of questions; choose the defaults. (e.g. English language, the suggested installation folder, the checked selected components to install, not to customize startup options, shortcut in the Start Menu, and additional tasks).
- When it's done, click on the new R desktop icon. Click on the console. This is where you will type commands to R.

The R interface



Initially, there is only the console window open. If you make plots, other windows will open too.

Some code to try

Note that the # sign is a "comment" – R ignores anything after #.

```
# generate some random normal data
data=rnorm(100)
# look at a histogram and a boxplot
hist(data)
boxplot(data)
# compute the sample mean, median, variance, standard deviation
mean(data)
median(data)
var(data)
sd(data)
# if you have a question about a command, preface it with ?
?hist
```

MOA data: R code

```
# read data from web, take 1st & 2nd columns as moa and group indicators, plot
stuff=read.table("http://people.stat.sc.edu/hansont/stat205/moa.txt",header=FALSE)
stuff
moa=stuff[,1]
moa
group=stuff[,2]
group
plot(group,moa)
# you can also read data from a file on your computer (text, Excel, etc.)
```

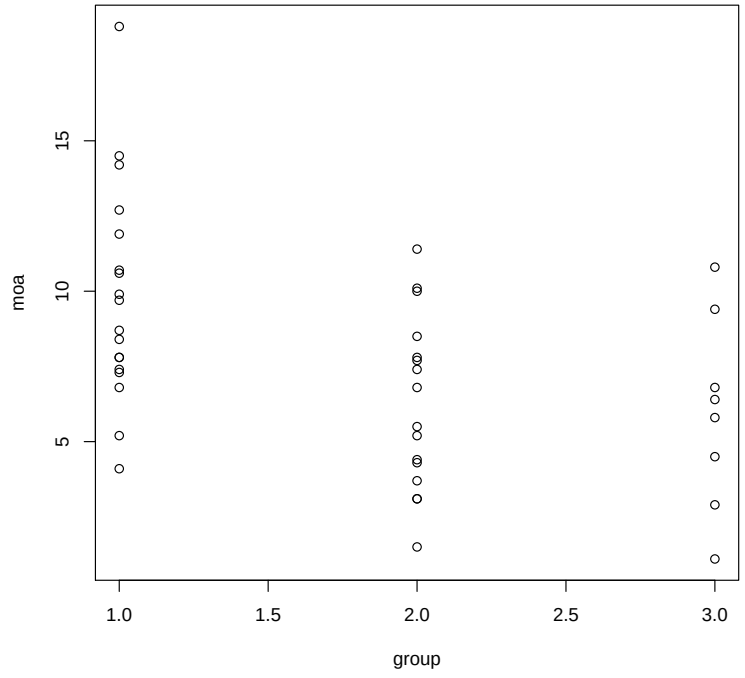
MOA data: output

```
> stuff=read.table("http://people.stat.sc.edu/hansont/stat205/moa.txt",header=FALSE)
> stuff
     V1 V2
1   6.8  1
2   4.1  1
3   7.3  1
4  14.2  1
5  18.8  1
6   9.9  1
7   7.4  1
8  11.9  1
9   5.2  1
10  7.8  1
11  7.8  1
12  8.7  1
13 12.7  1
14 14.5  1
15 10.7  1
16  8.4  1
17  9.7  1
18 10.6  1
19  7.8  2
20  4.4  2
21 11.4  2
22  3.1  2
23  4.3  2
24 10.1  2
25  1.5  2
```

MOA data: output continued

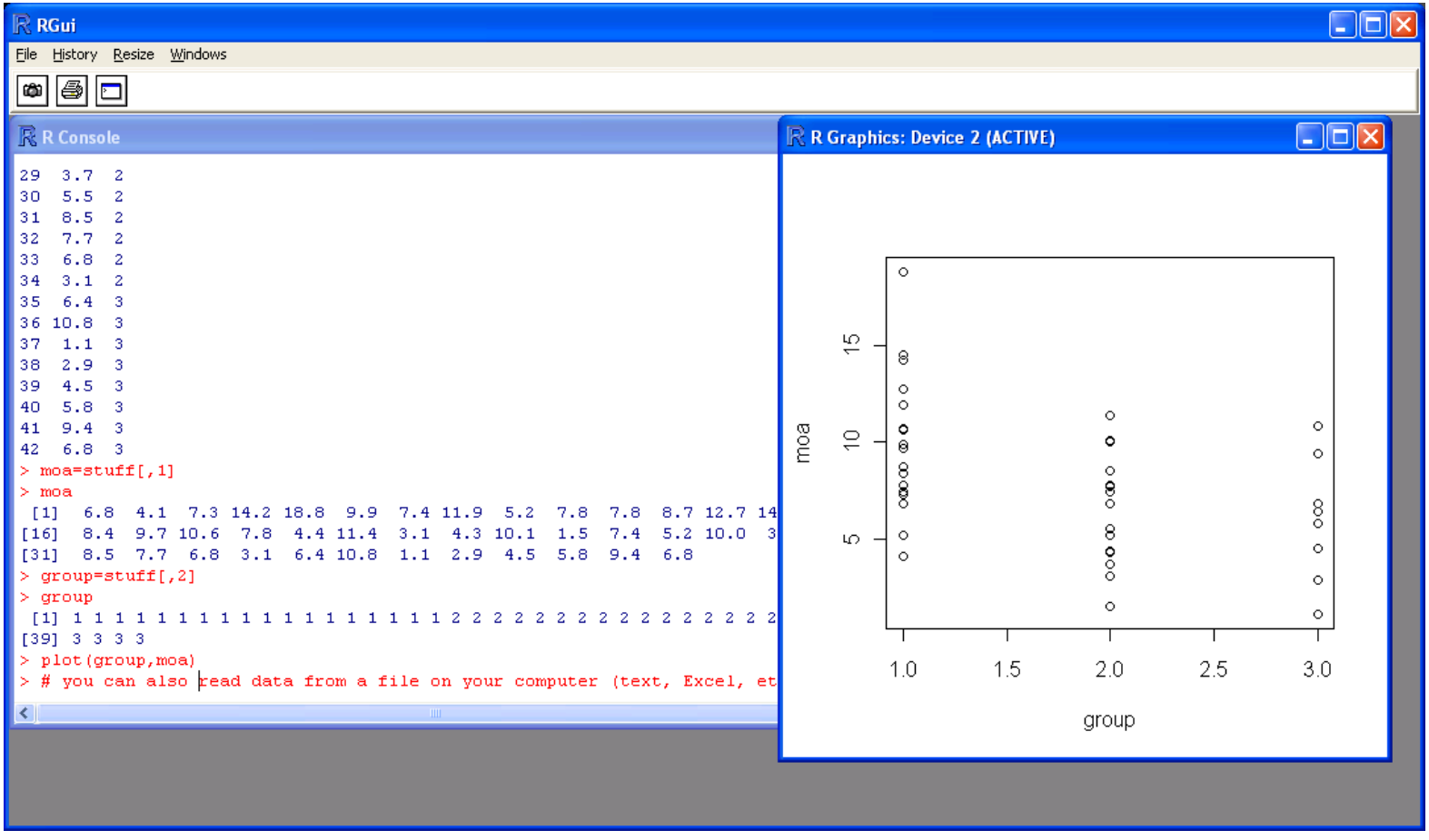
```
26  7.4  2
27  5.2  2
28 10.0  2
29  3.7  2
30  5.5  2
31  8.5  2
32  7.7  2
33  6.8  2
34  3.1  2
35  6.4  3
36 10.8  3
37  1.1  3
38  2.9  3
39  4.5  3
40  5.8  3
41  9.4  3
42  6.8  3
> moa=stuff[,1]
> moa
 [1]  6.8  4.1  7.3 14.2 18.8  9.9  7.4 11.9  5.2  7.8  7.8  8.7 12.7 14.5 10.7  8.4  9.7
[18] 10.6  7.8  4.4 11.4  3.1  4.3 10.1  1.5  7.4  5.2 10.0  3.7  5.5  8.5  7.7  6.8  3.1
[35]  6.4 10.8  1.1  2.9  4.5  5.8  9.4  6.8
> group=stuff[,2]
> group
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3
> plot(group,moa)
```

Plot of MOA data from R



You can right click on an R plot to save it to the clipboard as a metafile or bitmap. These can be saved into Microsoft applications such as Word. You can also leftclick on the plot then under `Save` choose `Save as` and save the plot, e.g. PDF.

Plot of MOA data from R



R window after cutting and pasting the commands a few slides ago.

More on R

- R will allows you to do all analyses covered in this course, and beyond.
- There are some tutorials, both installed in R and on the web. Under `Help` choose `Manuals (in PDF)` and choose `An introduction to R`. This can get you started.
- For homework, I'll give you a skeleton set of commands to get the basic job done with no frills.
- R's error messages can be cryptic and therefore R is not as "user friendly" as some other packages such as Minitab.
- However it is free; now being used by hundreds of thousands of people.