GMBA 7098: Statistics and Data Analysis (Fall 2014)

Introduction

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 15, 2014

Road map

- ▸ **What is statistics**?
- ▸ Syllabus.
- ▸ Basic statistical concepts.
- ▸ The R programming language.

Coffee pricing

- How to set the price p of a cup of coffee?
- Suppose the problem is like this:
  - Supply: unit production cost is c.
  - Demand: $D(p) = a - bp$.
  - What is the optimal price that maximizes the coffee shop's profit?

Coffee pricing

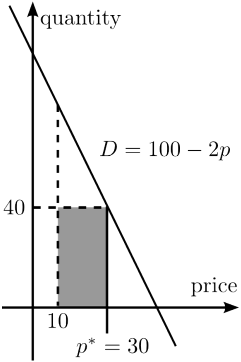
- Econ 101:

$$\max_p \, (p - c)(a - bp).$$

- First order condition:

$$\frac{\partial}{\partial p}\Big[(p - c)(a - bp)\Big] = a - 2bp + bc.$$

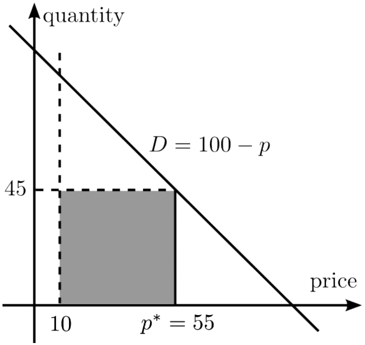
- $p^* = \dfrac{a + bc}{2b} > 0$ is the optimal price.



$(a = 100,\ b = 2,$
$c = 10)$

Coffee pricing

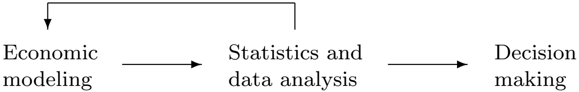
- For different demand function, we have different optimal prices.
- But what is **the** demand function?
  - How to measure a and b?
  - Is $D(p)$ really a and b?
  - If not, what factors also affect D?



$(a = 100,\ b = 1,\ c = 10)$

Measuring unknowns in the world

- It is always challenging to **measure unknowns** in the world.
- To help us measure unknowns, people develop **statistics**.
- Statistics is the **science** of gathering, analyzing, interpreting, and presenting **numerical** data.
  - For texts: text mining, natural language processing, etc.
  - For images: image recognition, digital image processing, etc.
- Mathematics (particularly probability) is required.
- Goal: to achieve better decision making.



Economic modeling

Statistics and data analysis

Decision making

What is Statistics?

- Many things are unknown...
  - Consumers' tastes.
  - Quality of a product.
  - Stock prices.
  - Employees' preferences.
- The study of Statistics includes:
  - Descriptive Statistics.
  - Probability.
  - Inferential Statistics: Estimation.
  - Inferential Statistics: Hypothesis testing.
  - Inferential Statistics: Prediction.
- In summary: To estimate, test, and predict those unknowns.

Road map

- What is statistics?
- **Syllabus**.
- Basic concepts.
- The R programming language.

The instructing team

- Instructor:
  - Ling-Chieh Kung.
  - Third-year assistant professor.
  - Department of Information Management.
  - Office: Room 413, Management Building II.
  - Office hour: **9:00am-10:30am, Thursday** or by appointment.
  - E-mail: lckung@ntu.edu.tw.
- Teaching assistant:
  - Ho Ho and Ian Zhong.
  - First-year master students.
  - Office: Room 320C, Management Teaching and Research Building.
  - Ho's E-mail: r02725041@ntu.edu.tw.
  - Ian's E-mail: r02725040@ntu.edu.tw.

Language and references

- Language: **"All" English**.
  - All materials (including course videos) are in English.
  - Students are encouraged (but not required) to speak English in class.
  - The instructor speak Chinese or English in office hour.
  - The instructor will speak Chinese in lectures when it helps.
- References:
  - *Business Statistics: For Contemporary Decision Making* by Ken Black.
  - *Freakonomics* by Steven Levitt and Stephen Dubner.
  - *Learn R in a Day* by Steven Murray (Amazon Kindle e-books only).
  - *Big Data* by Viktor Mayer-Schnberger and Kenneth Cukier.

"Flipped classroom"

- Lectures in **videos**, then discussions in classes.
- Before each Monday, the instructor uploads a video of lectures.
  - Ideally, the video will be no longer than one and a half hour.
  - Students must watch the video by themselves before that Monday.
- During the lecture, we do three things:
  - Discussing the lecture materials (0.5 to 1 hour).
  - Doing on-site exercises (1 to 2 hours).
  - Further discussions (0.5 to 1 hour).
  - Solving **lecture problems** to earn points.
- Teams:
  - Students form teams to work on class problems and case studies.
  - Students will be **randomly** grouped into teams with about three people.
  - For different modules, one may have different teammates.

Homework, office hour, project, and exam

- No homework!
- Office hour:
  - 9:00am-10:30am, Thursday or by appointment.
- Case studies:
  - Three case studies about real stories or real data.
  - One for each module.
- Midterm exam:
  - In-class and open whatever you have (including all electronic devices).
  - No information is allowed to be transferred among students.
  - There is no final exam.
- Final project:
  - Students form teams to apply the techniques learned in this course to a **self-selected** problem.
  - Each team does an oral presentation in one of the last two weeks.
  - All team members must be in class for the team to present.

Grading

- Class participation: 10%.
- Lecture problems: 20%.
- Three case reports: 15% (5% each).
- One case presentation: 10%.
- Midterm exam: 15%.
- Final project: 30%.
- The final letter grades will be given according to the following conversion rule:

Letter	Range	Letter	Range	Letter	Range
A+	$[90, 100]$	B+	$[77, 80)$	C+	$[67, 70)$
A	$[85, 90)$	B	$[73, 77)$	C	$[63, 67)$
A−	$[80, 85)$	B−	$[70, 73)$	C−	$[60, 63)$

Important dates and tentative plan

- Important dates:
  - Week 4 (2014/10/6): TA session because the instructor is in the military.
  - Week 9 (2014/11/10): Midterm exam.
  - Weeks 17 and 18 (2015/1/5 and 2015/1/12): Project presentations.
- Tentative plan:
  - Foundation (five weeks).
  - Inferential Statistics (four weeks).
  - Advanced Techniques (five weeks).
  - Applications (four weeks).

Online resources

- CEIBA.
  - Viewing your grades.
  - Receiving announcements.
- `http://www.ntu.edu.tw/~lckung/courses/SDA-Fa14/`.
  - Downloading course materials.
- The bulletin board “NTUIM-lckung” on PTT.
  - Discussions.
- YouTube:
  - Watching lecture videos.

Road map

- What is statistics?
- Syllabus.
- **Basic concepts**.
- The R programming language.

Populations vs. samples

- A **population** is a collection of persons, objects, or items.
  - A **census** is to investigate the whole population.
- A **sample** is a portion of the population.
  - **Sampling** is to investigate only a subset of the population.
  - We then use the information contained in the sample to **infer** ("guess") about the population.
- What are samples for the following populations?
  - All students in NTU.
  - All students in the business school.
  - All chips made in one factory.
  - All consumers who have bought iPhone 5.
- Two important questions:
  - **Why** sampling?
  - Is a sample **representative**?

Descriptive vs. inferential statistics

- **Descriptive statistics**:
  - Graphical or numerical summaries of data.
  - Describing (visualizing or summarizing) a set of data.
- **Inferential statistics**:
  - Making a "scientific guess" on unknowns.
  - Trying to say something about the population.
- Which is descriptive and which is inferential?
  - Calculating the average height of 1000 randomly selected NTU students.
  - Using this number to estimate the average height of all NTU students.
- Another example (pharmaceutical research):
  - All the potential patients form the population.
  - A group of randomly selected patients is a sample.
  - Use the result on the sample to infer the result on the population.

Parameters vs. statistics

- A numerical summary of a population is a **parameter**.
  - The average height of all NTU students.
  - a and b in the demand function $D(p) = a - bp$.
- A numerical summary of a sample is a **statistic**.
  - The average height of all NTU male students.
  - The demand function generated by 1000 randomly selected people.
- Almost always people use a statistic to infer a parameter.
  - Some statistics are "good" while some are "bad."

Parameters vs. statistics: an example

- A laptop manufacturer wants to know the largest weight one can put on a type of laptop without destroying it.
  - Let's call this number x_i for the ith laptop produced.
  - x_is may be different for different laptops.
- Suppose 100000 laptops have been produced.
- The **parameter**: $\theta = \min_{i=1,...,100000} \{x_i\}$.
  - This will be the number announced to the public.
- Can the manufacturer conduct a census?

Parameters vs. statistics: an example

- Probably 50 laptops will be randomly chosen as a sample.
- For each laptop, we do an experiment (and destroy it) to get a number x_i, $i = 1, 2, ..., 50$.
- These x_is form a sample.
- What is a **statistic**?
  - $\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50}$ is a statistic.
  - $x_{\min} = \min_{i=1,\ldots,50} \{x_i\}$ is another statistic.
- Which statistic is "closer to" the parameter?

Parameters vs. statistics

- A parameter is a **fixed number**.
  - E.g., $\theta = \min_{i=1,\ldots,100000}\{x_i\}$.
  - E.g., the average height of all NTU students.
- A statistic is a **function** whose outcome is **random**.
  - Two different random samples typically generate two values of a statistic.
  - The sampling process matters.

Another example

- (Suppose) there is a new proposal of increasing the tuition of all students by 5% in NTU.
- We want to know the percentage of students supporting it.
  - What is the population?
  - What statistics would you choose?
  - Is it fine to sample by standing in front of Building I of the College of Management? How would you form a sample?

Levels of data measurement

- Most data we will play with are numerical.
- Numerical data may be categorized to three levels:
  - Nominal.
  - Ordinal.
  - Quantitative: interval or ratio.

Nominal level

- A **nominal** scale classifies data into categories with **no ranking**.
- Data are labels or names used to identify an attribute of the element.
- The label may be numeric or non-numeric label.
- Examples:

Categorical variables	Values (Categories)
Laptop ownership	Yes / No
Citizenship	Taiwan / Japan / ...
Country code	886 / 86 / 1 / ...

- Arithmetic operations **cannot** be applied on nominal data.

Ordinal level

- An **ordinal** scale classifies data into categories with **ranking**.
- The order or rank of the data is meaningful.
- However, differences between numerical labels **do not** imply distances.
- Examples:

Categorical variables	Values (Categories)
Product satisfaction	Satisfied, neutral, unsatisfied
Professor rank	Full, associate, assistant
Ranking of scores	1, 2, 3, 4, ...

- It is still not meaningful to do arithmetic on ordinal data.
  - Assistant + associate = full?!
  - The grade difference between no. 1 and no. 5 may not be equal to that between no. 11 and no. 15.

Quantitative (interval and ratio) levels

- An **interval** scale is an ordered scale in which the **difference** between measurements is a meaningful quantity but the measurements do not have a true zero point.
- A **ratio** scale is an ordered scale in which the difference between measurements is a meaningful quantity and the measurements have a true **zero point**.
- Ratio data appear more often in the world.
  - Heights, weights, income, prices.
- Interval data are actually rare.
  - Degrees in Celsius or Fahrenheit.
  - GRE or GMAT scores.
- How about degrees in Kelvin?

Some remarks

- Nominal and ordinal data are called **qualitative data**.
- Interval and ratio data are called **quantitative data**.
- Most statistical methods are for quantitative data; some are for qualitative data.
  - Distinguishing nominal and ordinal scales is important.
  - Distinguishing interval and ratio scales is not.
- Sometimes quantitative data are called **numeric** data.

A short summary

- Understand these terms:
  - Populations vs. samples.
  - Parameters vs. statistics.
  - Inferential statistics vs. descriptive statistics.
- For each scale of measurement, is it meaningful to calculate ...

Level	Ranking	Distance
Nominal		
Ordinal		
Quantitative		

Road map

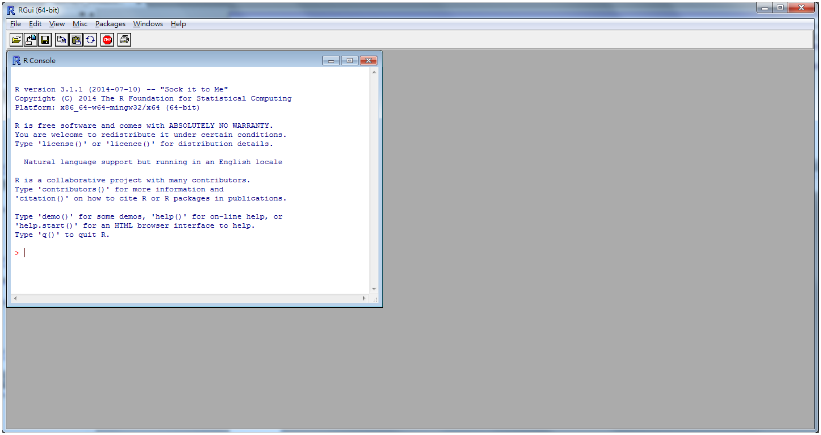
- What is statistics?
- Basic concepts.
- Syllabus.
- **The R programming language**.

The R programming language

- **R** is a programming language for statistical computing and graphics.
- R is open source.
- R is powerful and flexible.
  - It is fast.
  - Most statistical methods have been implemented as packages.
  - One may write her own R programs to complete her own task.
- http://www.r-project.org/.
- To download, go to http://cran.csie.ntu.edu.tw/, choose your platform, then choose the suggested one (the current version is 3.1.1).

The programming environment

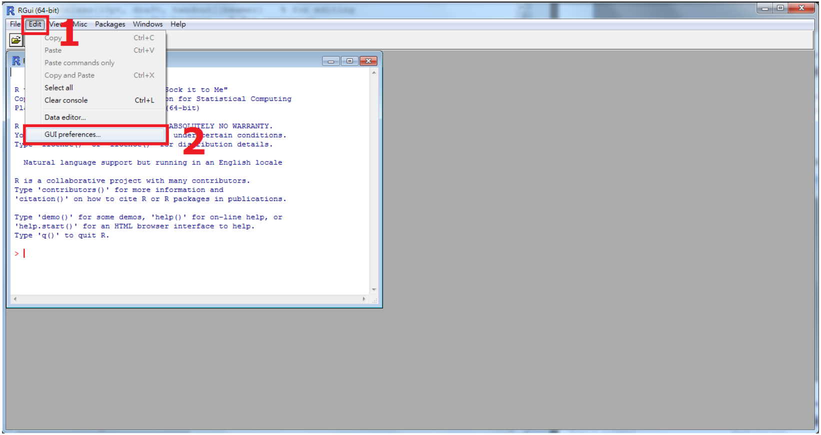
- When you run R, you should see this:



```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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.

> |
```

Interface setting

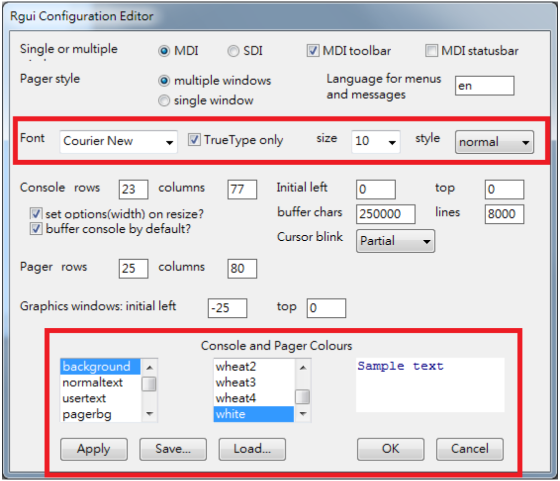
- You may start right away or (like me) change the interface setting:



```
R                                "Sock it to Me"
Co                               on for Statistical Computing
Pl                               (64-bit)

R                                ABSOLUTELY NO WARRANTY.
Yo                                under certain conditions.
Ty                                    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.

> 
```

Interface setting

- You may change font, font size, background color, text color, etc.



Rgui Configuration Editor

Single or multiple ● MDI ○ SDI ☑ MDI toolbar □ MDI statusbar

Pager style ● multiple windows ○ single window

Language for menus and messages: en

Font: Courier New ☑ TrueType only size: 10 style: normal

Console rows: 23 columns: 77

Initial left: 0 top: 0

buffer chars: 250000 lines: 8000

☑ set options(width) on resize?

☑ buffer console by default?

Cursor blink: Partial

Pager rows: 25 columns: 80

Graphics windows: initial left: -25 top: 0

Console and Pager Colours

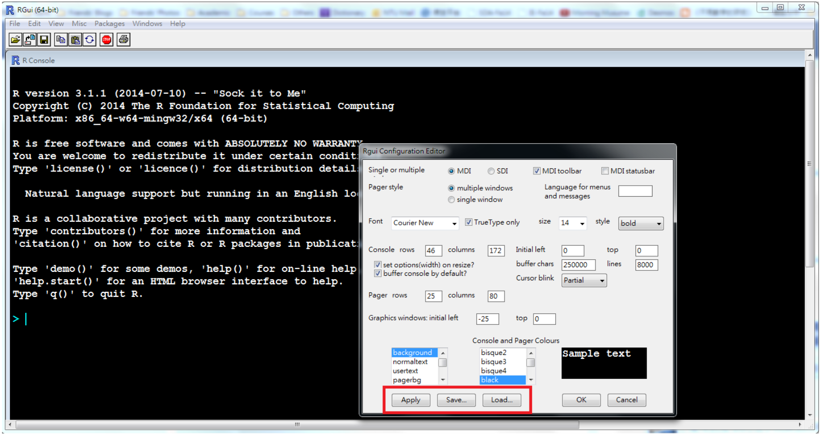
background
normaltext
usertext
pagerbg

wheat2
wheat3
wheat4
white

Sample text

Applying your setting

- **Apply** it. **Save** it for future use (by loading it).



```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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

  Natural language support but running in an English lo

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

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

> |
```

Single or multiple ● MDI ○ SDI ☑ MDI toolbar □ MDI statusbar

Pager style ● multiple windows ○ single window

Language for menus and messages

Font Courier New ☑ TrueType only size 14 style bold

Console rows 46 columns 172 Initial left 0 top 0

☑ set options(width) on resize?

☑ buffer console by default?

buffer chars 250000 lines 8000

Cursor blink Partial

Pager rows 25 columns 80

Graphics windows: initial left -25 top 0

Console and Pager Colours

background
normaltext
usertext
pagerbg

bisque2
bisque3
bisque4
black

Sample text

Try it!

- Type some mathematical expressions!

```
> 1 + 2
[1] 3
> 6 * 9
[1] 54
> 3 * (2 + 3) / 4
[1] 3.75
```

- Or if you prefer:

```
> log(2.718)
[1] 0.9998963
> 10 ^ 3
[1] 1000
> sin(3.1416)
[1] -7.34641e-06
```

Let's do statistics

- A wholesaler has 440 customers in Portugal:
  - 298 are "horeca"s (hotel/restaurant/café).
  - 142 are retails.
- These customers locate at different regions:
  - Lisbon: 77.
  - Oporto: 47.
  - Others: 316.
- http://archive.ics.uci.edu/ml/datasets/Wholesale+customers.

Let's do statistics

- The data:

Channel	Label	Fresh	Milk	Grocery	Frozen	D. & P.	Deli.
1	1	30624	7209	4897	18711	763	2876
1	1	11686	2154	6824	3527	592	697
				⋮			
2	3	14531	15488	30243	437	14841	1867

- The wholesaler records the annual amount each customer spends on six product categories:
  - Fresh, milk, grocery, frozen, detergents and paper, and delicatessen.
  - Amounts have been scaled to be based on “monetary unit.”
- Channel: hotel/restaurant/café = 1, retailer = 2.
- Region: Lisbon = 1, Oporto = 2, others = 3.

The work directory

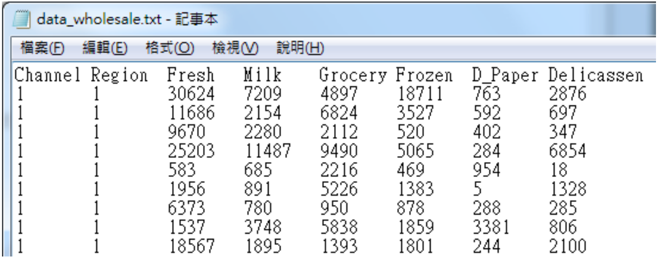
- The data are provided in a TXT file “data_wholesale.txt.”
- To start our analysis with R, first we set up our **work directory**.
- To set up the work directory:

```
> setwd("C:/Users/user/Documents/R")
> getwd()
[1] "C:/Users/user/Documents/R"
```

  - Create the directory before you use it!

Loading data from a TXT file

- Loading data from a TXT file with columns separated by tabs:


data_wholesale.txt - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

Channel	Region	Fresh	Milk	Grocery	Frozen	D_Paper	Delicassen
1	1	30624	7209	4897	18711	763	2876
1	1	11686	2154	6824	3527	592	697
1	1	9670	2280	2112	520	402	347
1	1	25203	11487	9490	5065	284	6854
1	1	583	685	2216	469	954	18
1	1	1956	891	5226	1383	5	1328
1	1	6373	780	950	878	288	285
1	1	1537	3748	5838	1859	3381	806
1	1	18567	1895	1393	1801	244	2100

```
> W <- read.table("data_wholesale.txt", header = TRUE)
```

- W is a **data frame** that stores the data.
- <- assigns the values at its right to the variable at its left.

Browsing data

- To browse the data stored in a data frame:

```
> W
> head(W)
> tail(W)
```

- To extract a row or a column:

```
> W[1, ]
> W$Channel
> W[, 1]
```

- What is this?

```
> W[1, 2]
```

Extracting more rows or columns

- To extract multiple rows or columns:

```
> W[1:6, ]
> W[, 1:3]
> head(W[, 1:3])
```

- How about nonconsecutive rows or columns?

```
> W[c(1, 4:6), ]
> head(W[, c(2, 5:6)])
```

- In general, `c()` does all kinds of concatenations and `i:j` produces a sequence of integers from `i` to `j`.
- How about these?

```
> head(cbind(W$Channel, W$Region))
> head(cbind(Channel = W$Channel, Region = W$Region))
```

Basic statistics

- The **mean** (average) expenditure on milk:

```
> mean(W$Milk)
```

- The **standard deviation** of expenditure on milk:

```
> sd(W$Milk)
```

- What is the mean expenditure on milk for those who
  - live in Lisbon (`Region` is 1) and
  - consume at hotel/restaurant/café (`Channel` is 1)?

```
> mean(W$Milk[1:59])
```

- There must be a better way!

Extracting rows by conditions

- Let's find those records for consumption at hotel/restaurant/café:

  ```
  > which(W$Channel == 1)
  ```

  - `which()` takes a vector and examine whether each element satisfies the given condition. If so, it returns that index.
  - `W$Channel[1]` is 1, `W$Channel[400]` is 2, etc.
- `=` is for **assignment** and `==` is for **comparison**!
  - To assign a value to a variable, use `=`.
  - To test whether two values are equal, use `==`.
- Now, we know what this is:

  ```
  > mean(W$Milk[which(W$Channel == 1)])
  ```

- What is next?

Combining conditions

- To specify an "and" operation, use & (ampersand).

```
> mean(W$Milk[which(W$Channel == 1 & W$Region == 1)])
```

- To specify an "or" operation, use | (bar).

```
> mean(W$Milk[which(W$Channel == 1 | W$Region == 1)])
```

- To specify a "not" operation, use ! (exclamation).

```
> mean(W$Milk[which(W$Channel == 1 | !(W$Region == 1))])
```

- This also works:

```
> index <- which(m$Channel == 1 & m$Region == 1)
> mean(m$Milk[index])
```

Exercises

- Fill in this table:

Channel	Region		
	1	2	3
1	3870.20		
2			

Mean expenditures on milk

- What is this?

```
> mean(which(W$Channel == 1 & W$Region == 1))
[1] 30
```

Some more basic statistics

- Counting:

```
> length(which(W$Channel == 1 & W$Region == 1))
```

- Median:

```
> median(W$Milk[which(W$Channel == 1 & W$Region == 1)])
```

- Maximum and minimum:

```
> max(W$Milk[which(W$Channel == 1 & W$Region == 1)])
> min(W$Milk[which(W$Channel == 1 & W$Region == 1)])
```

- Correlation coefficient:

```
> a <- W$Milk[which(W$Channel == 1 & W$Region == 1)]
> b <- W$Grocery[which(W$Channel == 1 & W$Region == 1)]
> cor(a, b)
[1] 0.654953
```

Some more basic statistics

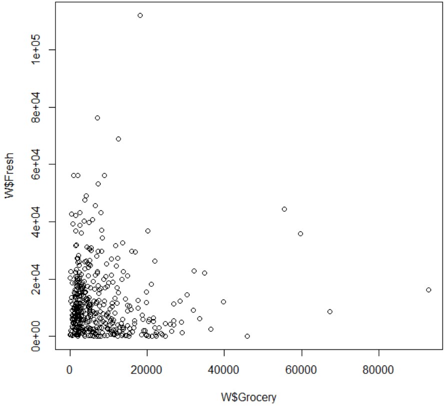
- In fact, you may simply do:

  ```
  > cor(W[, 3:8])
  ```

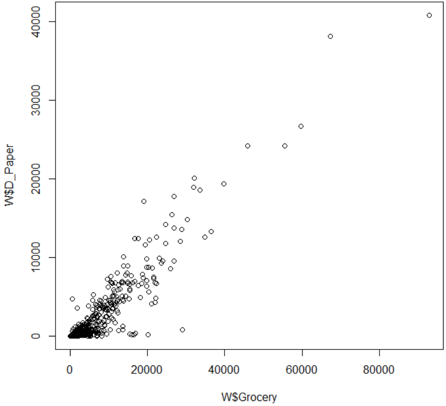
- How to find the correlation coefficients of `Grocery` and each of the other five variables?
  - Hint: Apply extractions with `c()` and `:` on the matrix produced by `cor(W[, 3:8])`.

Basic graphs: Scatter plots

```
> plot(W$Grocery, W$Fresh)
```

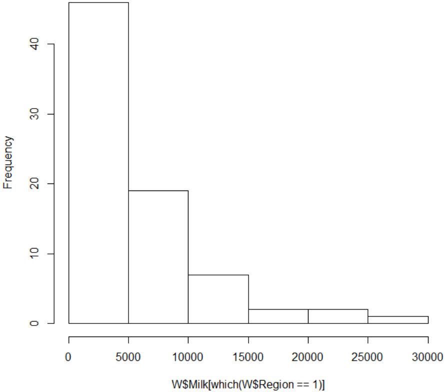


```
> plot(W$Grocery, W$D_Paper)
```



Basic graphs: histograms

```
> hist(W$Milk[which(W$Region == 1)])
```



Storing data to a TXT file

- To store the results of our calculation permanently:

```
> C <- cor(W[, 3:8])
> write.table(C, "cor_wholesale.txt")
> write.table(C, "cor_wholesale.txt", col.names = NA,
              row.names = TRUE, quote = FALSE, sep = "\t")
```

- Before you close your R environment:
  - Save the current work **image** to store all the variables and their values.