Lecture 02: Frequency Distributions

Kunlin Wei

Notation

- The individual measurements or scores obtained for a research participant will be identified by the letter X (or X and Y if there are multiple scores for each individual).
- The number of scores in a data set will be identified by N for a population or n for a sample.
- Summing a set of values is a common operation in statistics and has its own notation. The Greek letter sigma, Σ, will be used to stand for "the sum of." For example, ΣX identifies the sum of the scores.

Frequency Distributions

- After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.
- This is the goal of descriptive statistical techniques.
- One method for simplifying and organizing data is to construct a **frequency distribution**.

Frequency Distributions (cont.)

- A **frequency distribution** is an organized tabulation showing exactly how many individuals are located in each category on the scale of measurement. A frequency distribution presents an organized picture of the entire set of scores, and it shows where each individual is located relative to others in the distribution.

Frequency Distribution Tables

- A **frequency distribution table** consists of at least two columns - one listing categories on the scale of measurement (X) and another for frequency (f).
- In the **X** column, values are listed from the highest to lowest, without skipping any.
- For the **frequency** column, tallies are determined for each value (how often each X value occurs in the data set). These tallies are the frequencies for each X value. The sum of the frequencies should equal N.
- A third column can be used for the **proportion** (p) for each category: p = f/N. The sum of the p column should equal 1.00.
- A fourth column can **display the percentage of the distribution** corresponding to each X value. The percentage is found by multiplying p by 100. The sum of the percentage column is 100%.

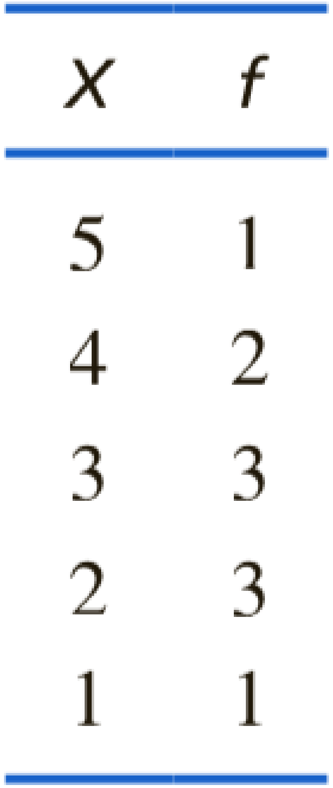
Frequency Distribution Table Example

The following set of $N = 20$ scores was obtained from a 10-point statistics quiz. We will organize these scores by constructing a frequency distribution table. Scores:

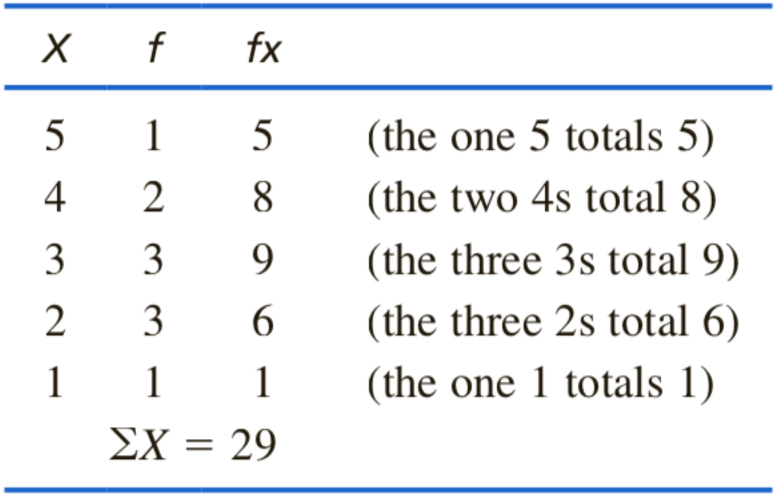
8	9	8	7	10	9	6	4	9	8
7	8	10	9	8	6	9	7	8	8

X	f
10	2
9	5
8	7
7	3
6	2
5	0
4	1

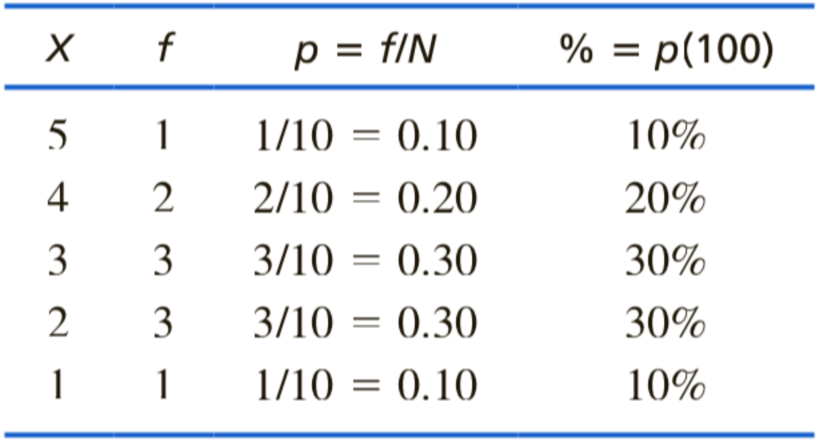
Sum, Proportions, and Percentages



X	f
5	1
4	2
3	3
2	3
1	1



X	f	fx	
5	1	5	(the one 5 totals 5)
4	2	8	(the two 4s total 8)
3	3	9	(the three 3s total 9)
2	3	6	(the three 2s total 6)
1	1	1	(the one 1 totals 1)
	$\Sigma X = 29$		



X	f	$p = f/N$	$\% = p(100)$
5	1	1/10 = 0.10	10%
4	2	2/10 = 0.20	20%
3	3	3/10 = 0.30	30%
2	3	3/10 = 0.30	30%
1	1	1/10 = 0.10	10%

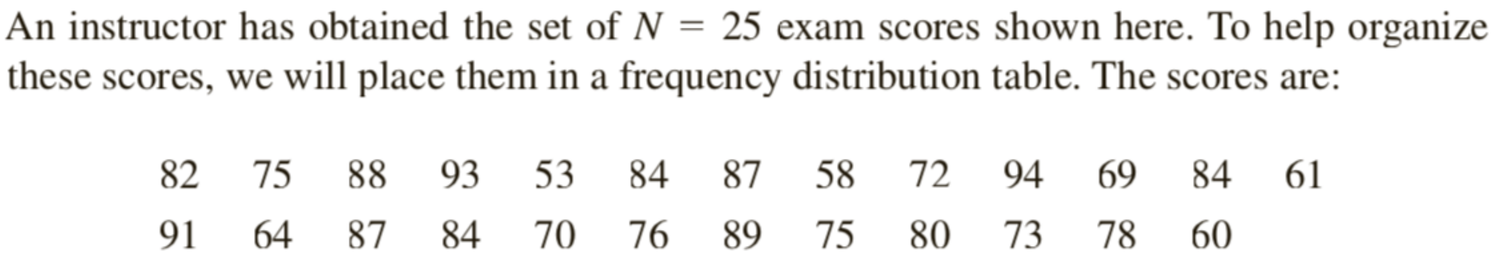
Frequency Distribution： Regular vs grouped

- When a frequency distribution table lists all of the individual categories (X values) it is called a **regular frequency distribution**.
- Sometimes, however, a set of scores covers a wide range of values. In these situations, a list of all the X values would be quite long - too long to be a “simple” presentation of the data.
- To remedy this situation, a **grouped frequency distribution** table is used

Grouped Frequency Distribution (cont.)

- In a grouped table, the X column lists groups of scores, called **class intervals**, rather than individual values.
- **These intervals all have the same width**, usually a simple number such as 2, 5, 10, and so on.
- Each interval begins with a value that is a multiple of the interval width. The interval width is selected so that the table will have approximately ten intervals.

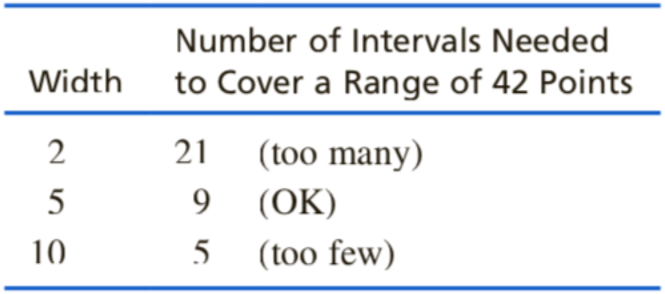
How to construct a grouped distribution table



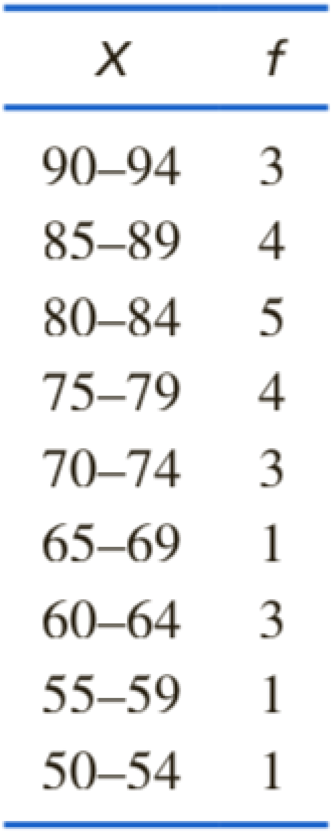
An instructor has obtained the set of $N = 25$ exam scores shown here. To help organize these scores, we will place them in a frequency distribution table. The scores are:

82	75	88	93	53	84	87	58	72	94	69	84	61
91	64	87	84	70	76	89	75	80	73	78	60	

1. Determine the range of scores



Width	Number of Intervals Needed to Cover a Range of 42 Points	
2	21	(too many)
5	9	(OK)
10	5	(too few)



X	f
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

2. Identify the intervals

3. Count the frequencies

(4. Interpret the table)

See the guidelines on the book

Apparent limits, real limits and interval width

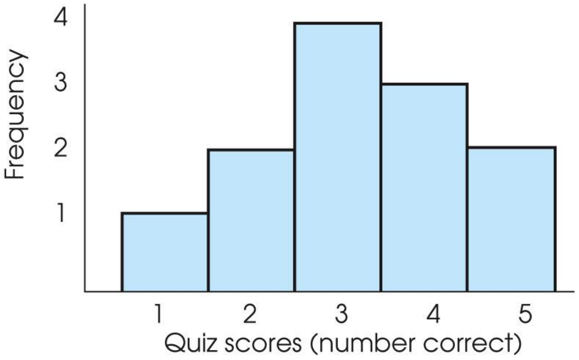
- Suppose a class interval of 40–49 contains scores from X = 40 to X = 49.
- The **apparent limits** of the interval are the values that appear to form the upper and lower boundaries for the class interval (40 and 49 in this case).
- If X is a continuous variable, X = 40 is actually an interval from 39.5 to 40.5. **The real limits** of the class interval are thus 39.5 (lower real limit) and 49.5 (upper real limit). The distance between these two real limits (10 points) is the width of the interval.
- If X is a discrete variable that takes integer values, the width of the **class interval** is 49-40+1 = 10.

Frequency Distribution Graphs

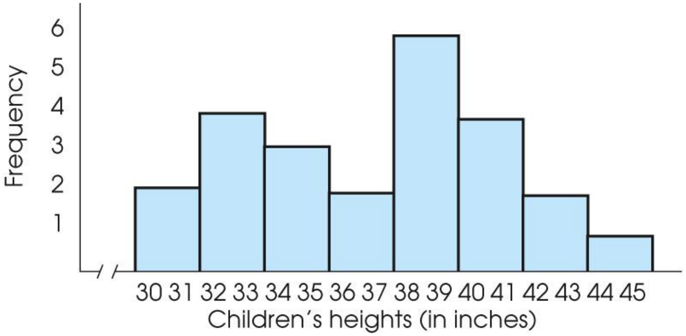
- In a **frequency distribution graph**, the score categories (X values) are listed on the X axis and the frequencies are listed on the Y axis.
- When the score categories consist of numerical scores from an interval or ratio scale, the graph should be either a histogram or a polygon.

Histograms

- In a **histogram**, a bar is centered above **each score** (or **class interval**) so that the height of the bar corresponds to the frequency and the width extends to the **real limits**, so that adjacent bars touch.



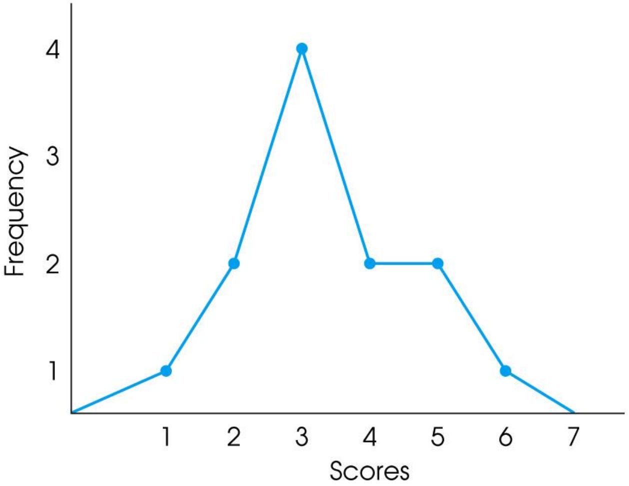
X	f
5	2
4	3
3	4
2	2
1	1



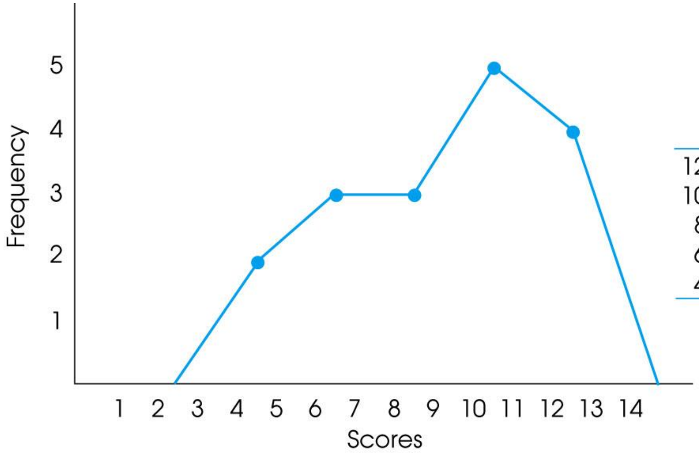
X	f
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

Polygons

- In a **polygon**, a dot is centered above each score so that the height of the dot corresponds to the **frequency**. The dots are then connected by straight lines. An additional line is drawn at each end to bring the graph back to a zero frequency.



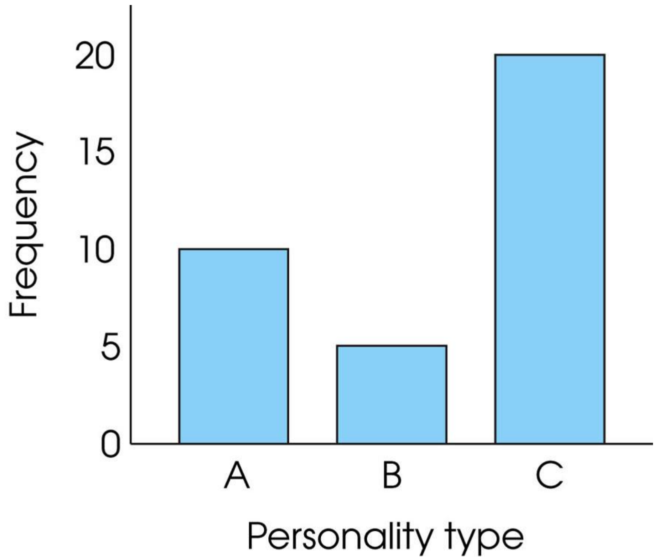
X	*f*
6	1
5	2
4	2
3	4
2	2
1	1



X	*f*
12–13	4
10–11	5
8–9	3
6–7	3
4–5	2

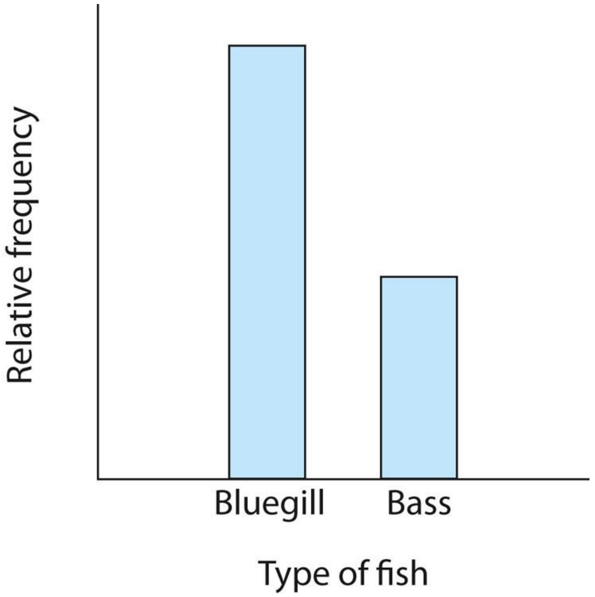
Bar graphs

- When the score categories (X values) are measurements from a nominal or an ordinal scale, the graph should be a bar graph.
- A **bar graph** is just like a histogram except that gaps or spaces are left between adjacent bars.



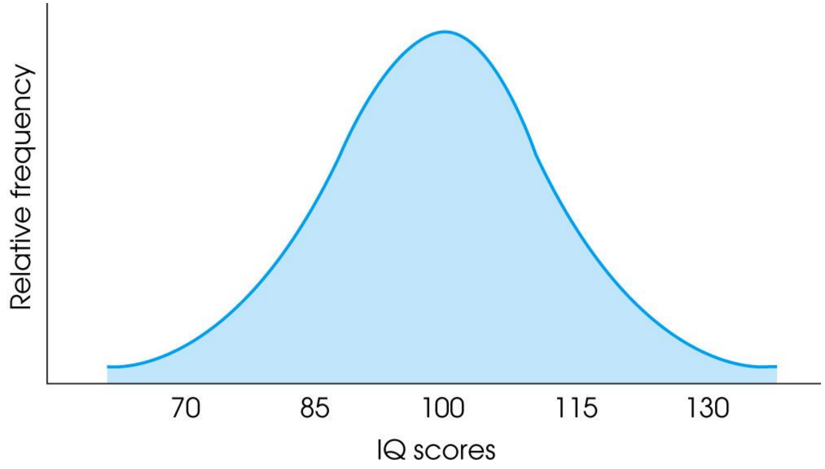
Relative frequency

- Many populations are so large that it is impossible to know the exact number of individuals (frequency) for any specific category.
- In these situations, population distributions can be shown using **relative frequency** instead of the absolute number of individuals for each category.



Smooth curve

- If the scores in the population are measured on an interval or ratio scale, it is customary to present the distribution as a **smooth curve** rather than a jagged histogram or polygon.
- The smooth curve emphasizes the fact that the distribution is not showing the exact frequency for each category.



What to see in frequency distribution graphs

- Frequency distribution graphs are useful because they show the entire set of scores.
- At a glance, you can determine the **highest** score, the **lowest** score, and where the scores are **centered**.
- The graph also shows whether the scores are clustered together or **scattered** over a wide range.

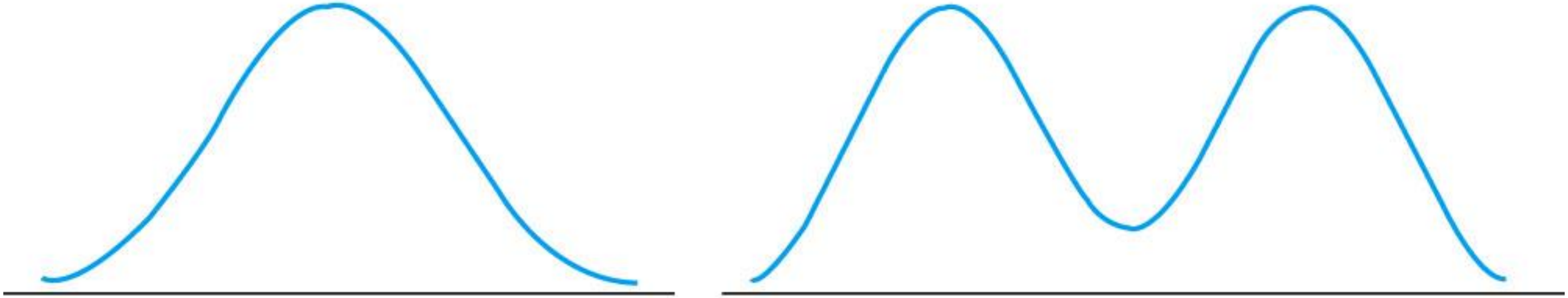
Shape

- A graph shows the **shape** of the distribution.
- A distribution is **symmetrical** if the left side of the graph is (roughly) a mirror image of the right side.
- One example of a symmetrical distribution is the bell-shaped normal distribution.
- On the other hand, distributions are **skewed** when scores pile up on one side of the distribution, leaving a "tail" of a few extreme values on the other side.

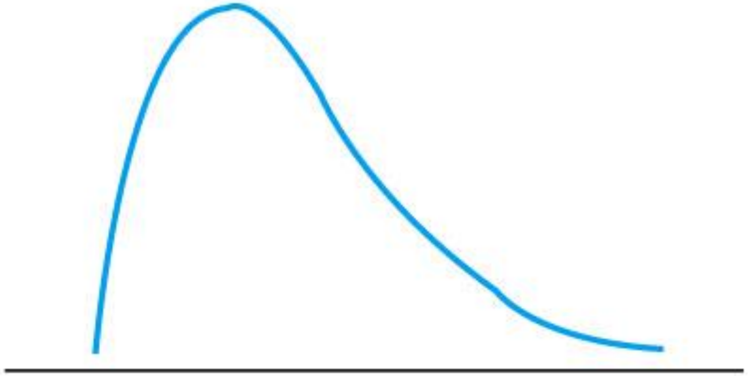
Positively and Negatively Skewed Distributions

- In a **positively skewed** distribution, the scores tend to pile up on the left side of the distribution with the tail tapering off to the right. a.k.a. right skewed
- In a **negatively skewed** distribution, the scores tend to pile up on the right side and the tail points to the left. a.k.a. left skewed

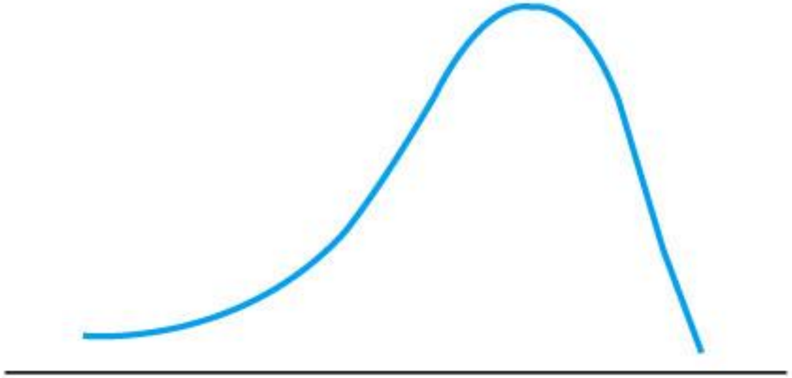
Symmetrical distributions



Skewed distributions



Positive skew



Negative skew

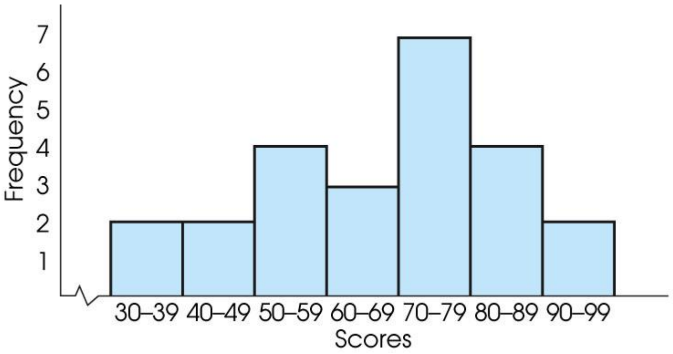
Stem-and-Leaf Displays

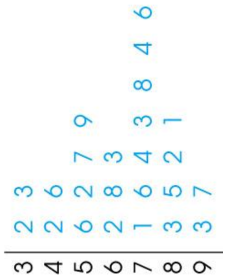
- A **stem-and-leaf display** provides a very efficient method for obtaining and displaying a frequency distribution.
- Each score is divided into a **stem** consisting of the first digit or digits, and a **leaf** consisting of the final digit.
- Finally, you go through the list of scores, one at a time, and write the leaf for each score beside its stem.
- The resulting display provides an organized picture of the entire distribution. The number of leafs beside each stem corresponds to the frequency, and the individual leafs identify the individual scores.

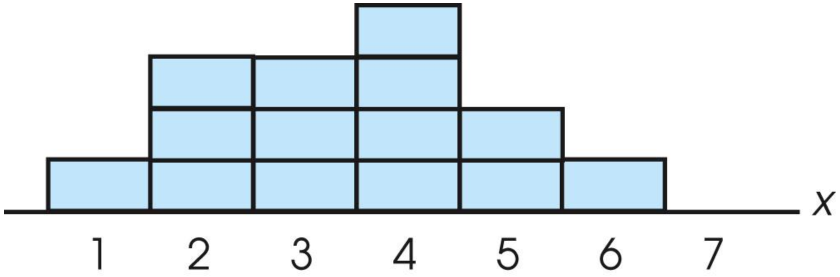
TABLE 2.3

A set of $N = 24$ scores presented as raw data and organized in a stem and leaf display.

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		







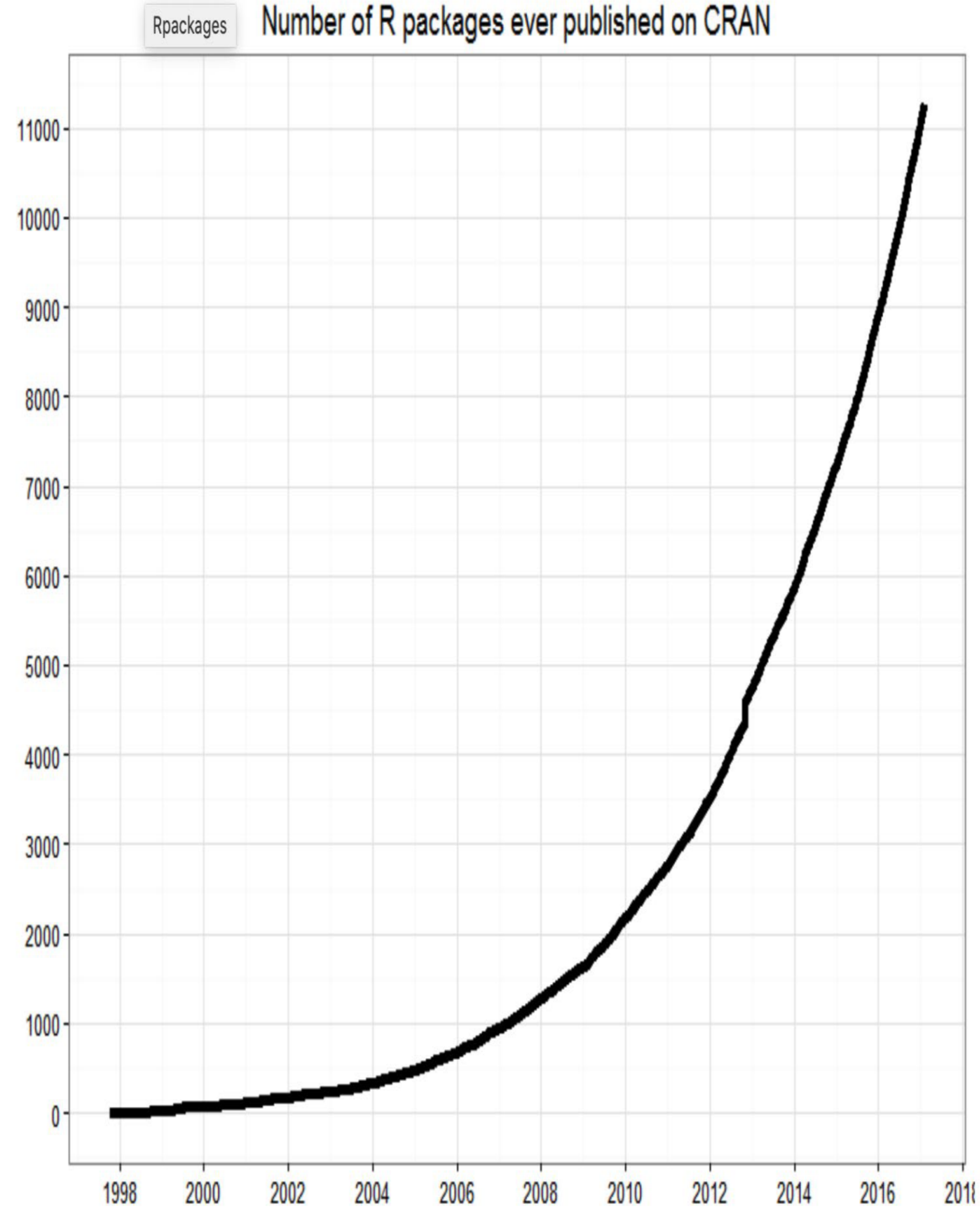
Review of Learning Goals

- Use and create frequency distribution tables and explain how they are related to the original set of scores.
- Choose when it is useful to set up a grouped frequency distribution table, and use and create this type of table for a set of scores.
- Describe how the three types of frequency distribution graphs—histograms, polygons, and bar graphs—are constructed and identify when each is used.
- Explain how frequency distribution graphs for populations differ from the graphs used for samples.
- Identify the shape of a distribution—symmetrical, and positively or negatively skewed—based on a set of scores or a frequency distribution graph.

R 软件初步介绍

Beware of both the convenience and the danger!

Growth of R packages through 2012



Number of R packages ever published on CRAN

R的介绍

1）下载R Studio

2）可能需要下载几个包：

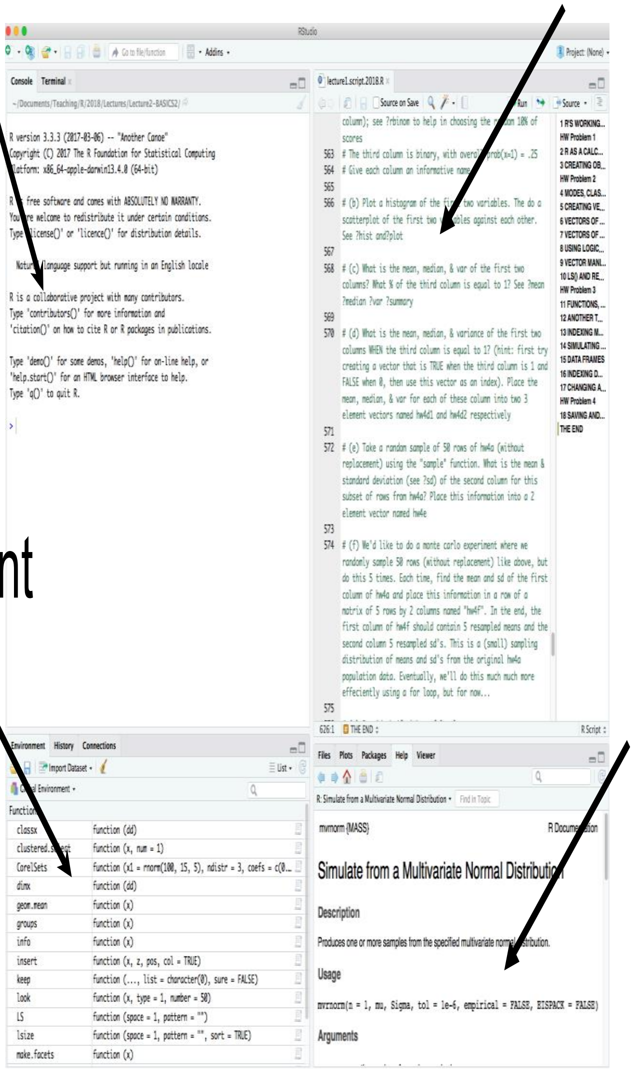
```
# 对付excel数据表格
install.packages("readxl")
install.packages(“cli”)  #command line interface
install.packages(“utf8”)   # unicode text processing
```

3）初级教程：https://www.runoob.com/r/r-scatterplots-charts.html

Typical Rstudio session

- Console – output & temporary input - usually unsaved

- Script – tells R what to do. Save this



```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

>
```

```
column); see ?rbinom to help in choosing the random 10% of scores
# The third column is binary, with overall prob(x=1) = .25
# Give each column an informative name

# (b) Plot a histogram of the first two variables. The do a scatterplot of the first two variables against each other. See ?hist and?plot

# (c) What is the mean, median, & var of the first two columns? What % of the third column is equal to 1? See ?mean ?median ?var ?summary

# (d) What is the mean, median, & variance of the first two columns WHEN the third column is equal to 1? (hint: first try creating a vector that is TRUE when the third column is 1 and FALSE when 0, then use this vector as an index). Place the mean, median, & var for each of these column into two 3 element vectors named hw4d1 and hw4d2 respectively

# (e) Take a random sample of 50 rows of hw4a (without replacement) using the "sample" function. What is the mean & standard deviation (see ?sd) of the second column for this subset of rows from hw4a? Place this information into a 2 element vector named hw4e

# (f) We'd like to do a monte carlo experiment where we randomly sample 50 rows (without replacement) like above, but do this 5 times. Each time, find the mean and sd of the first column of hw4a and place this information in a row of a matrix of 5 rows by 2 columns named "hw4f". In the end, the first column of hw4f should contain 5 resampled means and the second column 5 resampled sd's. This is a (small) sampling distribution of means and sd's from the original hw4a population data. Eventually, we'll do this much much more effeciently using a for loop, but for now...

```

Functions	
classx	function (dd)
clustered.select	function (x, num = 1)
CorelSets	function (x1 = rnorm(100, 15, 5), ndistr = 3, coefs = c(0...
dimx	function (dd)
geom.mean	function (x)
groups	function (x)
info	function (x)
insert	function (x, z, pos, col = TRUE)
keep	function (..., list = character(0), sure = FALSE)
look	function (x, type = 1, number = 50)
LS	function (space = 1, pattern = "")
lsize	function (space = 1, pattern = "", sort = TRUE)
make.facets	function (x)

mvrnorm {MASS} R Documentation

Simulate from a Multivariate Normal Distribution

Description

Produces one or more samples from the specified multivariate normal distribution.

Usage

```
mvrnorm(n = 1, mu, Sigma, tol = 1e-6, empirical = FALSE, EISPACK = FALSE)
```

Arguments

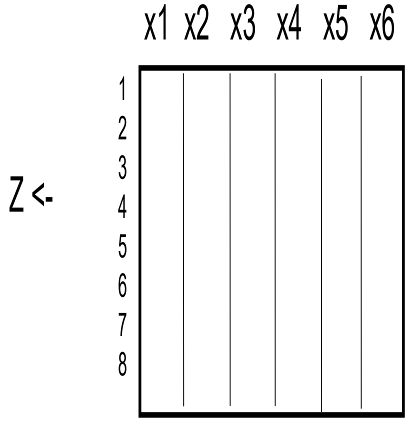
Environment

- Misc. windows, including help, files, etc.

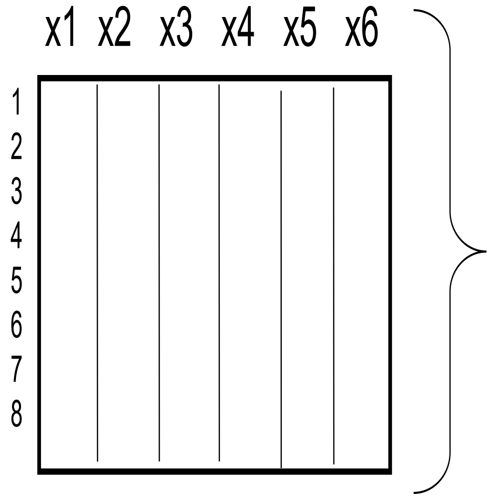
R Objects

- Almost all things in R – functions, datasets, results, etc. – are OBJECTS.
  - (graphics are written out and are not stored as objects)
- Script can be thought of as a way to make objects. Your goal is usually to write a script that, by its end, has created the objects (e.g., statistical results) and graphics you need.
- Objects are classified by two criteria:
  - MODE: how objects are stored in R - character, numeric, logical, list, & function
  - CLASS: how objects are treated by functions (important to know!) - [vector], matrix, array, factor, data.frame, & 1000s of special classes created by specific functions

R Objects



R Objects



The MODE of Z is determined automatically by the types of things stored in Z – numbers, characters, etc. Vectors & matrices must have their values all of the same mode. Lists can be a mix of modes.

R modes (to check, use mode() function):

numeric – numbers

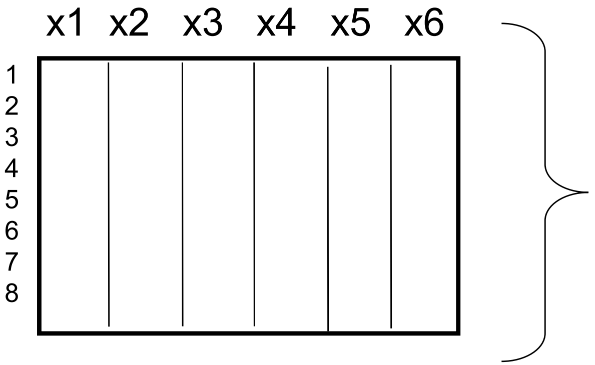
character

list – a concatenation of elements of different modes

logical – TRUE/FALSE

function

R Classes



The CLASS of Z is either set by default depending, on how it was created, or is explicitly set by user. You can check the objects' class and change it. It determines how functions deal with Z. If of class "lm", R searches for a function fun.lm

NOTE: If an object has two classes - c("first", "second") - R searches for a function called fun.first and, if it finds it, applies it to the object. If no such function is found, a function called fun.second is tried. If no class name produces a suitable function, the function fun.default is used.

R classes (to check, use class() function):

[for **vectors**, mode & class are same] - logical, numeric, character
[modes & class are same for these 2 as well] - function, list (when generic)
factor
matrix
array
data.frame

Learning R

- Read through the CRAN website & intro manual
- Know your objects' modes & classes: mode(x); class(x)
- Because R is interactive, errors are your friends!
- ?lm   gives you help on lm function. Reading help files can be very… helpful
- MOST IMPORTANT - the more time you spend using R, the more comfortable you become with it. After doing your first real project in R, you won't look back. I promise.