

MA331 Intermediate Statistics

Lecture 01 Data, Variables and Distribution ¹

Xiaohu Li

Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, New Jersey 07030

Week 01



¹Based on Chapter 1.

1. From sample to big data

✌ Statisticians find solution for scientific problems through **data-crunching**.

✌ Data size:

- Sample (about 100 obs) from scientific experiments – **OUR FOCUS**.
- Data set (hundreds of obs for several variables), databases from small business and industry.
- Big data (several data sets with millions obs of hundreds variables), databases from large business and industry.

✌ Statistics aim to extract useful information behind the data, and it gets the following two branches

- **Descriptive statistics** on the data.
- **Inferential statistics** on the information behind the data.

✌ This course focuses on ideas and methods, and the concerned computations have to be done with softwares such as R, SAS and SPSS.



2. One example – exercise on cholesterol levels

✌ Assess the effect of exercise on cholesterol levels

- One group exercises and the other does not.
Is cholesterol reduced in exercise group?
- People have naturally different cholesterol levels.
- Response to the same amount of exercise differs (e.g. genetics).
- The level may vary in adherence to exercise regimen.
- The diet may have an effect, and
exercise may affect other factors (e.g. appetite, energy, schedule).

✌ So, we have to collect

- observations of **cholesterol levels of the two groups** along with
- other **related covariates** such as starting level, amount of exercise, regimen of exercise, diet style etc.



3. Important points on data

- ✎ Example: Investigating the body weight of a certain group of students.
- ✎ The **randomness** in the data collecting gives rise to the variability in data.
- ✎ **Statistics** is the science of understanding data and making decisions in the context of variability.
- ✎ Methods to reduce the variability:
 - Better **experimental design** before collecting the data.
 - Employ a reasonable **statistic** to analyze the data.
 - Proper interpretation of the output of software.
 - An insightful discovering on what data is telling you.

HOMEWORK 1: (i) Download and install R on your computer. (ii) Download R-studio to get a convenient interface.



4. Basics of statistics

📖 **Individuals**: objects described by a set of data (patients, industrial systems/elements, animals, things).

📖 **Variables**: related descriptions of an individual, taking different values for different subjects.

📖 Three questions to ask before data collecting:

- Why: Purpose of study?
- Who: Members of the sample, how many?
- What: What variables should be measured?

📖 Example: A study on how the party-time-spent impact on GPA, variables like age, student's major, gender etc.


📖 We focus on statistical analysis and thus **always assume a sample** at hand.

📖 **Sampling theory** handles the sample design and data collecting.



5. Variable types

 **Categorical** variables have outcomes falling into **finite categories**.

 **Quantitative** variables have numerical outcomes.

- continuous: height, weight, distance etc. take any value within an interval.
- discrete: number of phone calls next week, number of students getting A this Fall etc. take all possible integers.

 **Example: Information on employees**

	A	B	C	D	E	F
1	Name	Job Type	Age	Gender	Race	Salary
2	Cedillo, Jose	Technical	27	Male	White	52,300
3	Chambers, Tonia	Management	42	Female	Black	112,800
4	Childers, Amanda	Clerical	39	Female	White	27,500
5	Chen, Huabang	Technical	51	Male	Asian	83,600
6						

Ready NUM



HOMEWORK 2: Input this table into R and save it as a data object 'Employees'.

6. Distribution of a variable

- ✎ Distribution comprises of all possible values a variable may take.
- ✎ For a categorical variable, just list count or percentage of individuals in each category.
- ✎ Methods to understand the distribution of a quantitative variable:
 - Graphical tools (bar graph, pie chart, histogram) visually display the distribution.
 - Numerical summaries (mean, variance) provide outlines of important characteristics.

HOMEWORK 3: Refer any R guidance for commands (`barplot`, `pie`, `table`) and practice them by using the built-in data.



7. Examples of graph tools

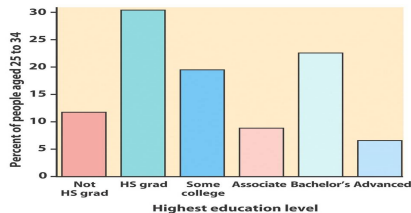


Figure 1-1a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

(a) Bar graph

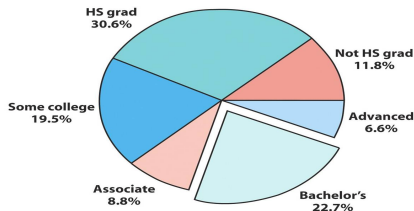
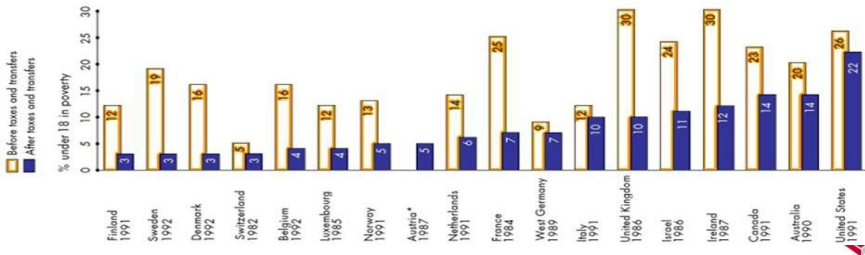


Figure 1-1b
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

(b) Pie chart



(c) Frequencies

8. Histogram through an example

TABLE 1.2 Percent of Hispanics in the adult population, by state (2000)

State	Percent	State	Percent	State	Percent
Alabama	1.5	Louisiana	2.4	Ohio	1.6
Alaska	3.6	Maine	0.6	Oklahoma	4.3
Arizona	21.3	Maryland	4.0	Oregon	6.5
Arkansas	2.8	Massachusetts	5.6	Pennsylvania	2.6
California	28.1	Michigan	2.7	Rhode Island	7.0
Colorado	14.9	Minnesota	2.4	South Carolina	2.2
Connecticut	8.0	Mississippi	1.3	South Dakota	1.2
Delaware	4.0	Missouri	1.8	Tennessee	2.0
Florida	16.1	Montana	1.6	Texas	28.6
Georgia	5.0	Nebraska	4.5	Utah	8.1
Hawaii	5.7	Nevada	16.7	Vermont	0.8
Idaho	6.4	New Hampshire	1.4	Virginia	4.2
Illinois	10.7	New Jersey	12.3	Washington	6.0
Indiana	3.1	New Mexico	38.7	West Virginia	0.6
Iowa	2.3	New York	13.8	Wisconsin	2.9
Kansas	5.8	North Carolina	4.3	Wyoming	5.5
Kentucky	1.3	North Dakota	1.0		

8. Histogram through an example

👉 Steps to construct a histogram:

- Arrange the data in the ascending order and determine as $\text{Range} = \text{Maximum} - \text{Minimum}$.
- Choose the interval width so as to divide data into 5 to 9 subintervals (classes) of equal width.
- Count the number of observations in each interval (class) and then plot the frequencies as their heights.

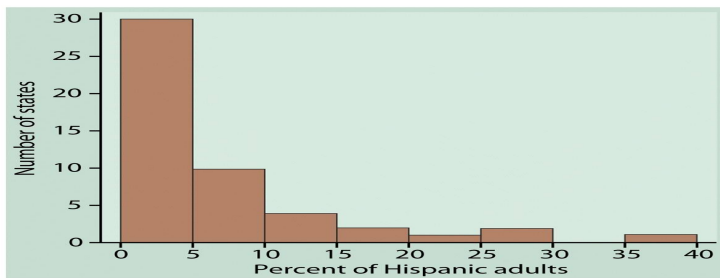
Class	Count	Percent	Class	Count	Percent
0.1-5.0	30	60	20.1-25	1	2
5.1-10.0	10	20	25.1-30	2	4
10.1-15	4	8	30.1-35	0	0
15.1-20	2	4	35.1-40	1	2



8. Histogram through an example

👉 R command: `hist(data, breaks, freq)` and `hist(data, nclass, freq)`.

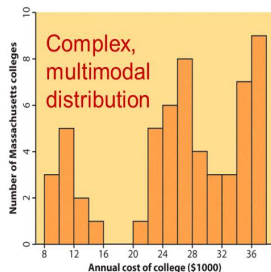
- Specify `breaks` as a vector for unequal widths or
- specify `nclass` for equal widths.
- Use counts or percentages through specifying `freq=T` or `F`.



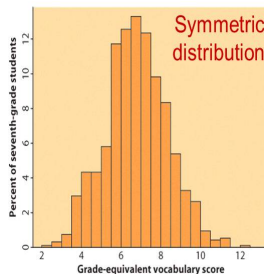
9. Examining the distribution of a variable

Determine the **pattern** through describing **shape, center and spread**.

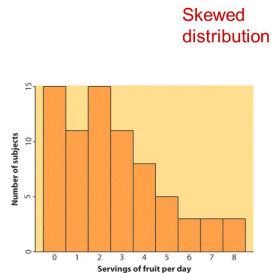
✌ Shape: number of modes (peaks), symmetric or skewed in one direction (right/left tail longer).



(d) Multi-peaks



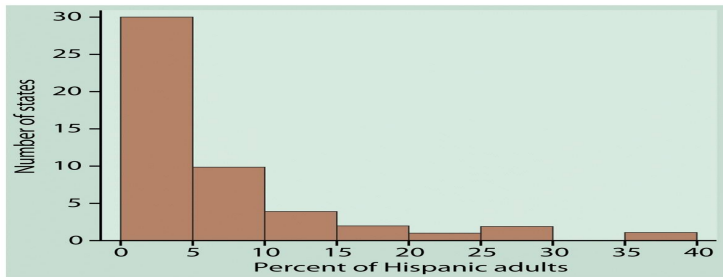
(e) Symmetric



(f) Right tail skewed



10. Example: Distribution of Hispanic Adults



✌ Shape: Right skewed, unimodal.

✌ Center: about 5%.

✌ Spread : 0 – 40% with only one state (NM) more than 30%.

✌ Is the extreme observation on the right an **outlier**?

Histograms are only meaningful for quantitative data.



11. Distribution of quantitative variables

✌ Deviations from 24,800 nanoseconds.

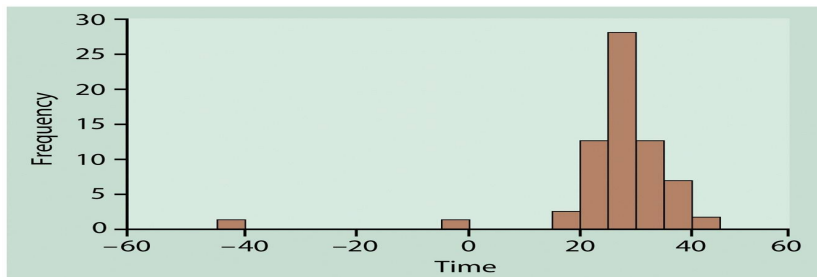
TABLE 1.1 Newcomb's measurements of the passage time of light					
28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

- 66 observations taken in July-Sept, 1882.
- Variable: passage time, scaled and centered.
- Observations are different due to variation of the environment of every measurement.

✌ We can further examine the nature of the variation by using graphs.



11. Distribution of quantitative variables - histogram

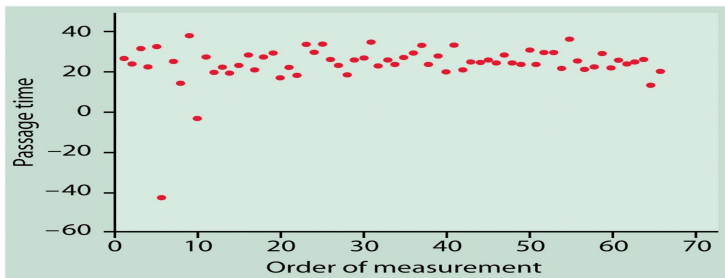


✌ For the outlier,

- Check for recording errors.
- Violation of the experimental condition.
- Discard it only for a valid practical or statistical reason.



11. Distribution of quantitative variables - time series



✌ It is observed that the measurement

- is more variant at the beginning, and then
- gets stabilized or less variant as time elapsed.

HOMEWORK 4: Find R commands to produce scatter plot.

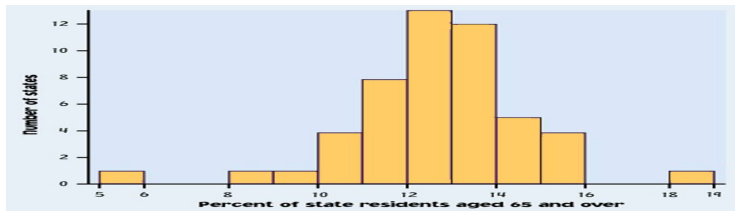


12. Concluding remarks – histogram

✌️ Outliers are those observations that lie outside the overall pattern of a distribution.

Always look for outliers and try to explain them.

✌️ A large gap in the distribution is a typical sign of an outlier.



Fairly symmetric overall except for 2 states clearly not belonging to the main trend. Alaska and Florida have unusual representation of the elderly in their population.

✌️ Data are the way they are. Never try to force them into a particular shape!



12. Concluding remarks – time series

- ✌ Plot observations over time (time on the x axis).
- ✌ **Trend**: persistent, long-term rise or fall.
- ✌ **Seasonal variation**: a pattern that repeats itself at a regular intervals of time.
- ✌ Example: For gas price, we observe the increasing trend, small seasonal variations, increase in spring and summer, slump in fall.

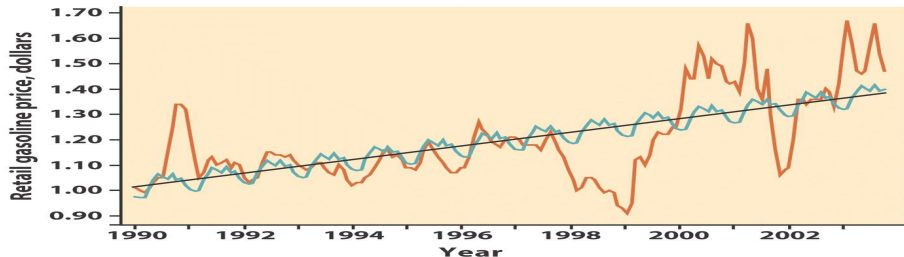


Figure 1-10
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company