1. Data, Graphical Descriptive Techniques

Introduction

Descriptive statistics involves the arrangement, summary, and presentation of data to enable meaningful interpretation and to support decision making.

Descriptive statistics methods make use of

- graphical techniques
- numerical descriptive measures

The methods presented apply to both

- the entire population
- the population sample

Types of data

A *variable* is a characteristic of population or sample that is of interest for us, for instance,

- Cereal choice
- Capital expenditure
- The waiting time for medical services

Data - the actual values of variables

- Quantitative data are numerical observations
- Qualitative data are categorical observations

Qualitative data

Person	Married/unmarried
1	yes
2	no
3	no
.	.
.	.

Professor	Rank
1	Lecturer
2	Full
3	Assistant
.	.

Quantitative data

Age	income
55	75000
42	68000
.	.
.	.

Weight gain
+10
+5
.
.

Cross-sectional and time-series data

Cross-sectional data is collected at a certain point in time, for example,

- Marketing survey (observe preferences by gender, age)
- Test score in a statistics course
- Starting salaries of an MBA program graduates

Time series data is collected over successive points in time, for instance,

- Weekly closing price of gold
- Amount of crude oil imported monthly

Type of analysis

Knowing the *type* of data is necessary to properly select the technique to be used.

Type of analysis allowed for each type of data

- Quantitative data - arithmetic calculations
- Qualitative data - counting the number of observation in each category

Qualitative data: frequency table

With qualitative data, all we can do is to calculate the count or proportion of data that falls into each category.

Qualitative data: frequency table

Example: faculty rank data

Lecturers	Assistant	Associate	Full	Total
15	25	5	15	60
25%	42%	8%	25%	100%

Pie charts, bar charts, line charts

- The graphical presentations shown here are used for qualitative data.
- These graphical tools are most appropriate when the raw data can be naturally categorized in a meaningful manner.

Pie charts

- Pie chart is a very popular tool used to represent the proportions of appearance for nominal data.
- The pie chart is a circle, subdivided into a number of slices that represent the various categories.
- The size of each slice is proportional to the percentage corresponding to the category it represents.

Bar charts

- Bar charts provide an alternative to pie charts.
- The frequency (or relative frequency) of each category is represented by a vertical bar.
- Use bar charts also when the *order* in which qualitative data are presented is meaningful.

Line charts

- Plot the frequency of a category above the point on the horizontal axis representing that category.
- Use line charts when the categories are points in time.

Graphical techniques for quantitative data. Histogram

1. Collect data
2. Prepare a frequency distribution
3. Draw a histogram

Histogram: more details

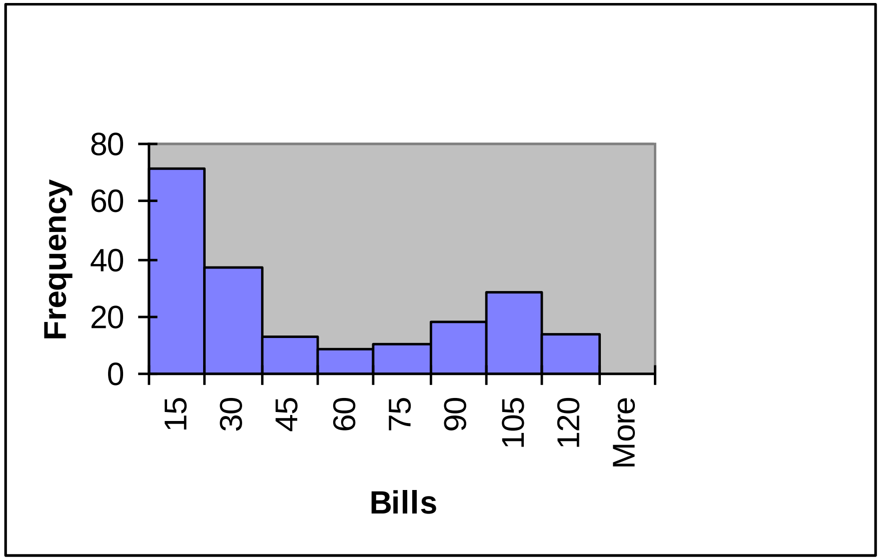
- **How many classes to use?**

# of observations	# of classes
Less than 50	5-10
50-200	7-12
200-500	9-15
More than 500	10-20

- ***Class width* = Range / # of classes**
- ***Range* = Largest Observation – Smallest Observation**
- ***Class frequency*= # of observations in the class**

Histogram

Example: Providing information concerning the monthly bills of new subscribers in the first month after signing on with a telephone company.



Histogram

What information can we extract from this histogram?

- About half of all the bills are small
- A few bills are in the middle range
- Relatively large number of large bills

Relative Frequency

It is often preferable to show the relative frequency (proportion) of observations falling into each class, rather than the frequency itself.

Class relative frequency = Class frequency / Total # of observations

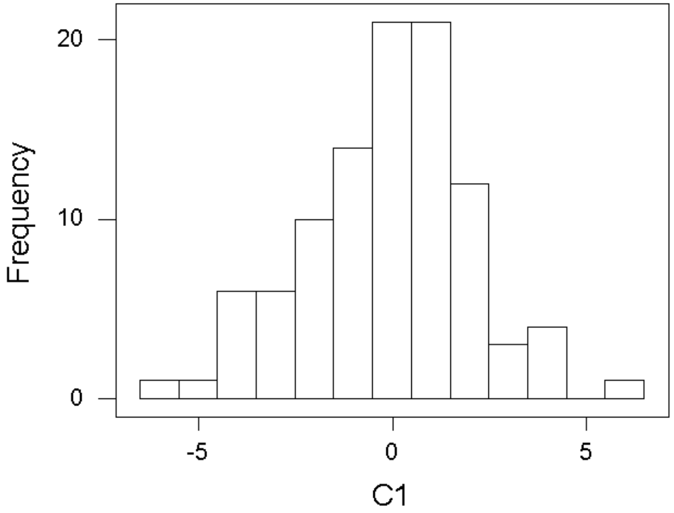
Relative frequencies should be used when

- the population relative frequencies are studied
- comparing two or more histograms
- the number of observations of the samples studied are different

Shapes of histograms

There are four typical shape characteristics

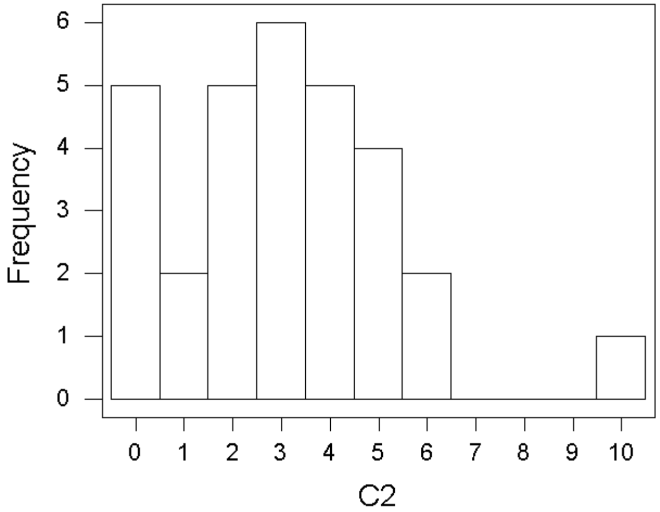
Symmetry



Shapes of histograms

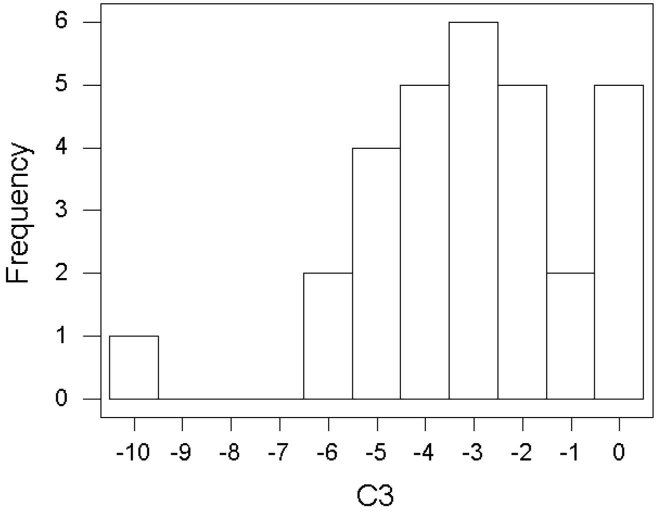
Skewness

Positively skewed: longer, heavier tail on the positive side



Shapes of histograms

Negatively skewed: longer, heavier tail on the negative side

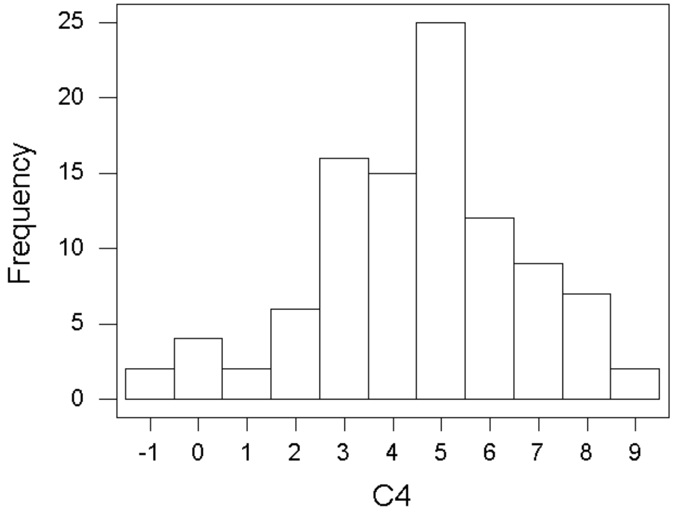


Shapes of histograms

Number of modal classes

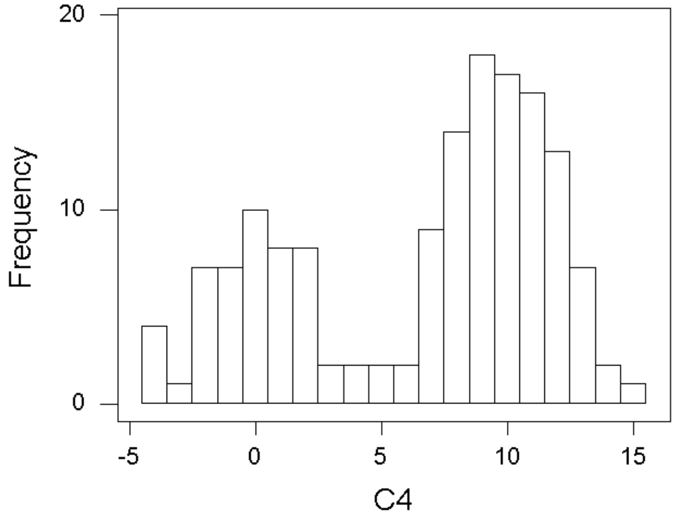
A modal class is the one with the largest number of observations.

Unimodal histogram



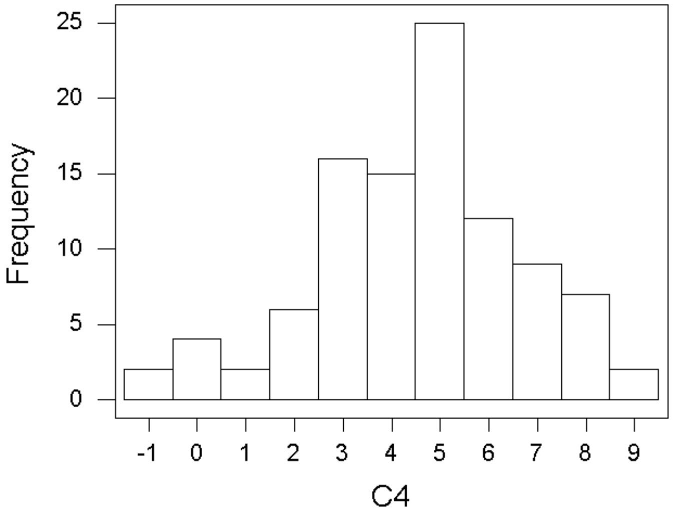
Shapes of histograms

Bimodal histogram



Shapes of histograms

Bell shaped histogram



Shapes of histograms

- Many statistical techniques require that the population be bell shaped.
- Drawing the histogram helps verify the shape of the population in question.

Stem and Leaf Display

- This is an interval-scaled display, most useful in preliminary analysis.
- Stem and leaf diagram shows the value of the original observations (whereas the histogram “loses” them).
- A *stem-and leaf display* is a way to summarize data. Each number in the data set is broken into two pieces: a *stem* and *a leaf.* The *stem* is the first part of the number and consist of the beginning digits. The *leaf* is the last part of the number and consists of the final digits.

Creating a stem and leaf display

Observe the data in the table below

19.1	19.8	18.0	19.2	19.5	17.3	20.0	20.3
19.6	18.5	18.1	19.7	18.4	17.6	21.2	20.6
22.2	19.1	21.1	19.3	20.8	21.2	21.0	18.7
19.9	18.7	22.1	17.2	18.4	21.4		

Determine what constitutes a stem and a leaf (there is more than one way). For example:

- the digits to the left of the decimal point is the stem
- the digits to the right of the decimal point is the leaf

Stem and Leaf Display

List the stems in a column from smallest to largest. Place each leaf at the same row as its stem.

The complete display is:

Stem	Leaf
17	236
18	0144577
19	1123567789
20	038
21	01224
22	12

Note: 17 | 2=17.2

Conclusions from the stem and leaf display

- The observations range from 17.2 to 22.2.
- Most of the observations fall between 18.0 and 20.0.
- The shape of the distribution is not symmetrical.
- Half the observations are below 19.5 and half above it.