Sampling and Descriptive Statistics

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 1 & Teaching Material

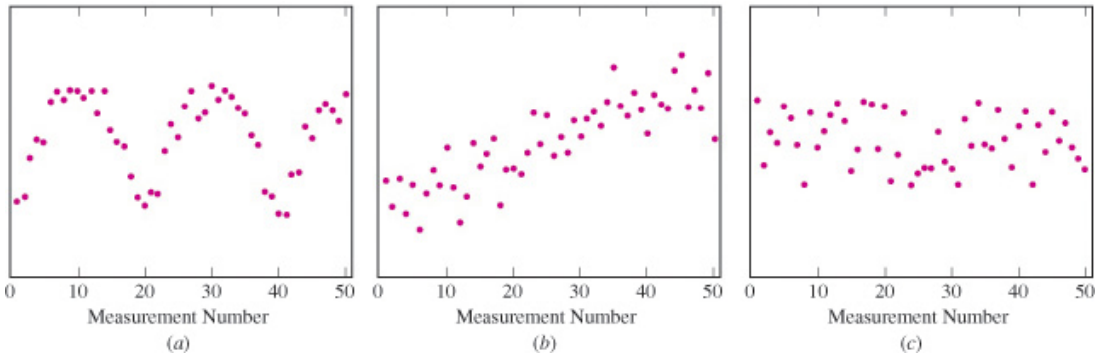
Sampling (1/2)

- Definition: A population is the entire collection of objects or outcomes about which information is sought
  - All NTNU students

- Definition: A sample is a subset of a population, containing the objects or outcomes that are actually observed
  - E.g., the study of the heights of NTNU students
    - Choose the 100 students from the rosters of football or basketball teams (appropriate?)
    - Choose the 100 students living a certain dorm or enrolled in the statistics course (appropriate?)

Sampling (2/2)

- Definition: A simple random sample (SRS) of size *n* is a sample chosen by a method in which each collection of *n* population items is equally likely to comprise the sample, just as in the lottery



- Definition: A sample of convenience is a sample that is not drawn by a well-defined random method
  - Things to consider with convenience samples:
    - Differ systematically in some way from the population
    - Only use when it is not feasible to draw a random sample

More on SRS (1/3)

- Definition: A conceptual population consists of all the values that might possibly have been observed
  - It is in contrast to "tangible (可觸之的) population"
  - E.g., a geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a slightly different reading
  - Here the population is conceptual. It consists of all the readings that the scale could in principle produce

More on SRS (2/3)

- A SRS is not guaranteed to reflect the population perfectly
- SRS's always differ in some ways from each other, occasionally a sample is substantially different from the population
- Two different samples from the same population will vary from each other as well
- This phenomenon is known as sampling variation

More on SRS (3/3)

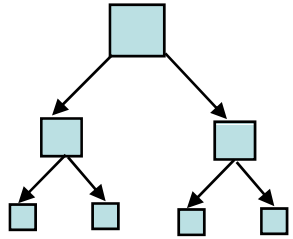
- The items in a sample are independent if knowing the values of some of the items does not help to predict the values of the others

- (A Rule of Thumb) Items in a simple random sample may be treated as independent in most cases encountered in practice
  - The exception occurs when the population is finite and the sample comprises a substantial fraction (more than 5%) of the population

- However, it is possible to make a population behave as though it were infinite large, by replacing each item after it is sampled
  - Sampling With Replacement

Other Sampling Methods

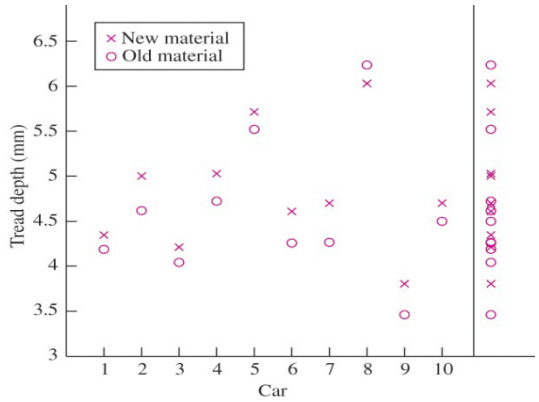
- Weighting Sampling
  - Some items are given a greater chance of being selected than others
  - E.g., a lottery in which some people have more tickets than others
- Stratified Sampling
  - The population is divided up into subpopulations, called strata
  - A simple random sample is drawn from each stratum
  - Supervised (?)
- Cluster Sampling
  - Items are drawn from the population in groups or clusters
  - E.g., the U.S. government agencies use cluster sampling to sample the U.S. population to measure sociological factors such as income and unemployment
  - Unsupervised (?)



Types of Experiments

- One-Sample Experiment
  - There is only one population of interest
  - A single sample is drawn from it

- Multi-Sample Experiment
  - There are two or more populations of interest
  - A simple is drawn from each population
  - The usual purpose of multi-sample experiments is to make comparisons among populations



Types of Data

- Numerical or quantitative if a numerical quantity is assigned to each item in the sample
  - Height
  - Weight
  - Age

- Categorical or qualitative if the sample items are placed into categories
  - Gender
  - Hair color
  - Blood type

Specimen	Torque (kN · m)	Failure Location
1	165	Weld
2	237	Beam
3	222	Beam
4	255	Beam
5	194	Weld

Summary Statistics

- The summary statistics are sometimes called descriptive statistics because they describe the data
  - Numerical Summaries
    - Sample mean, median, trimmed mean, mode
    - Sample standard deviation (variance), range
    - Percentiles, quartiles
    - Skewness, kurtosis
    - ….
  - Graphical Summaries
    - Stem and leaf plot
    - Dotplot
    - Histogram (more commonly used)
    - Boxplot (more commonly used)
    - Scatterplot
    - ….

Numerical Summaries (1/4)

- Definition: Sample Mean (the center of the data)
  - Let $X_1, \ldots, X_n$ be a sample. The sample mean is

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

    - It's customary to use a letter with a bar over it to denote a sample mean
- Definition: Sample Variance (how spread out the data are)
  - Let $X_1, \ldots, X_n$ be a sample. The sample variance is

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

Which is equivalent to $s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right)$

$\{20, 29, 30, 31, 32\}$

$\{10, 20, 30, 40, 50\}$

Numerical Summaries (2/4)

- Actually, we are interest in
  - Population mean
  - Population deviation: Measuring the spread of the population
    - The variations of population items around the population mean
- Practically, because population mean is unknown, we use sample mean to replace it
- Mathematically, the deviations around the sample mean tend to be a bit smaller than the deviations around the population mean
  - So when calculating sample variance, the quantity divided by $n-1$ rather than n provides the right correction
  - *To be proved later on !*

Numerical Summaries (3/4)

- Definition: Sample Standard Deviation
  - Let $X_1, \ldots, X_n$ be a sample. The sample deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Which is equivalent to

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)}$$

  - The sample deviation also measures the degree of spread in a sample (having the same units as the data)

Numerical Summaries (3/4)

- If $X_1,\ldots,X_n$ is a sample, and $Y_i = a + bX_i$,where a and b are constants, then $\overline{Y} = a + b\overline{X}$

- If $X_1,\ldots,X_n$ is a sample, and $Y_i = a + bX_i$,where a and b are constants, then $s_y^2 = b^2 s_x^2$ and $s_y = |b| s_x$

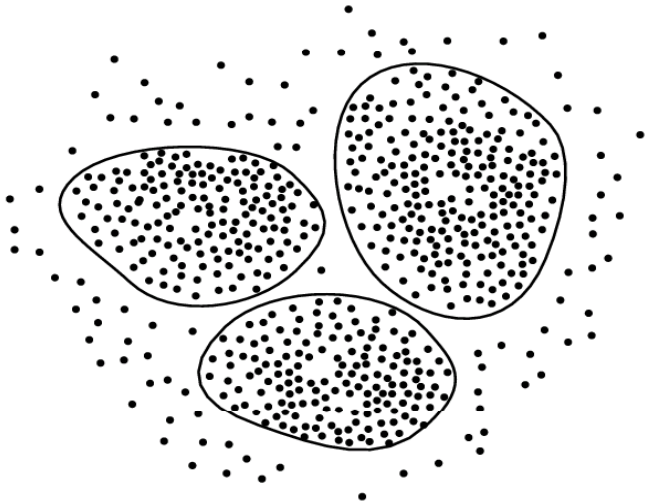
- Definition: Outliers
  - Sometimes a sample may contain a few points that are much larger or smaller than the rest (mainly resulting from data entry errors)
  - Such points are called outliers

More on Numerical Summaries (1/2)

- Definition: The median is another measure of center of a sample $X_1,\ldots,X_n$, like the mean
  - To compute the median items in the sample have to be ordered by their values
  - If n is odd, the sample median is the number in position $(n+1)/2$
  - If n is even, the sample median is the average of the numbers in positions $n/2$ and $(n/2)+1$
  - The median is an important (robust) measure of center for samples containing outliers

More on Numerical Summaries (2/2)

- Definition: The trimmed mean of one-dimensional data is computed by
  - First, arranging the sample values in (ascending or descending) order
  - Then, trimming an equal number of them from each end, say, *p*%
  - Finally, computing the sample mean of those remaining



More on "mean"

- Arithmetic mean $\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$
- Geometric mean $\overline{X} = \left(\prod_{i=1}^{n} X_i \right)^{\frac{1}{n}}$
- Harmonic mean $\overline{X} = n \cdot \left(\sum_{i=1}^{n} \frac{1}{X_i} \right)^{-1}$
- Power mean $\overline{X} = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^{\,m} \right)^{\frac{1}{m}}$

$m \to \infty$: maximum; $m \to -\infty$: minimum

$m = 1$: arithmetic mean; $m = -1$: harmonic mean

$m \to 0$: geometric mean

$m = 2$: quadratic mean

- Arithmetic mean ≥ Geometric mean ≥ Harmonic mean
- Weighted arithmetic mean $\overline{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$

http://en.wikipedia.org/wiki/Mean

Quartiles

- Definition: the quartiles of a sample $X_1, \ldots, X_n$ divides it as nearly as possible into quarters. The sample values have to be ordered from the smallest to the largest
  - To find the first quartile, compute the value 0.25(*n*+1)
  - The second quartile found by computing the value 0.5(*n*+1)
  - The third quartile found by computing the value 0.75(*n*+1)
- Example 1.14: Find the first and third quartiles of the data in Example 1.12

30 75 79 80 80 105 126 138 149 179 179 191

223 232 232 236 240 242 254 247 254 274 384 470

  - *n*=24
  - To find the first quartile, compute (*n*+1)25=6.25
    (105+126)/2=115.5
  - To find the third quartile, compute (n+1)75=18.75
    (242+245)/2=243.5

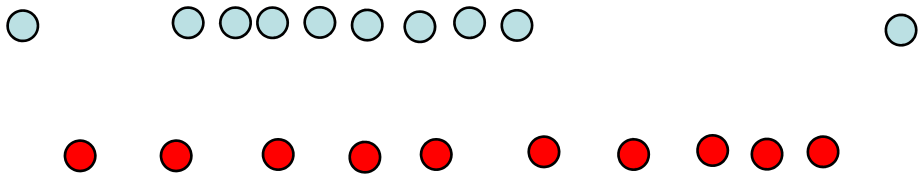
Percentiles

- Definition: The *p*th percentile of a sample $X_1, \ldots, X_n$, for a number between 0 and 100, divide the sample so that as nearly as possible *p*% of the sample values are less than the *p*th percentile. To find:
  - Order the sample values from smallest to largest
  - Then compute the quantity (*p*/100)(*n*+1), where *n* is the sample size
  - If this quantity is an integer, the sample value in this position is the pth percentile. Otherwise, average the two sample values on either side
- Note, the first quartile is the 25th percentile, the median is the 50th percentile, and the third quartile is the 75th percentile

Mode and Range

- Mode
  - The sample mode is the most frequently occurring values in a sample
  - Multiple modes: several values occur with equal counts

- Range
  - The difference between the largest and smallest values in a sample
  - A measure of spread that depends only on the two extreme values



Numerical Summaries for Categorical Data

- For categorical data, each sample item is assigned a category rather than a numerical value

- Two Numerical Summaries for Categorical Data
  - Definition: (Relative) Frequencies
    - The frequency of a given category is simply the number of sample items falling in that category
  - Definition: Sample Proportions (also called relative frequency)
    - The sample proportion is the frequency divided by the sample size

Sample Statistics and Population Parameters (1/2)

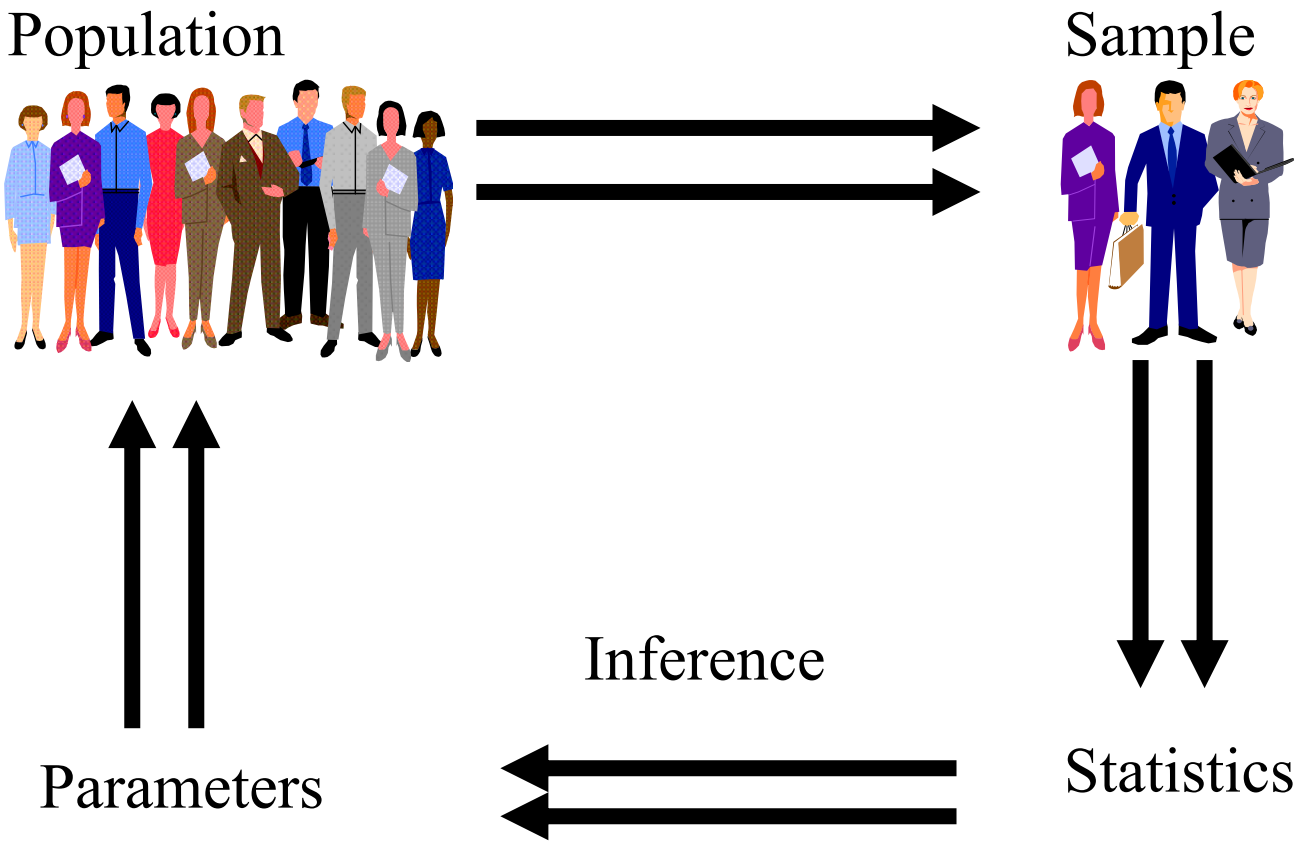
- A numerical summary of a sample is called a statistic
- A numerical summary of a population is called a parameter
  - If a population is finite, the methods used for calculating the numerical summaries of a sample can be applied for calculating the numerical summaries of the population (each value (or outcome) occurs with probability? See Chapter 2)
  - Exceptions are the variance and standard deviation (?)

$$\text{Normal} \ : f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- However, sample statistics are often used to estimate parameters (to be taken as estimators)
  - In practice, the entire population is never observed, so the population parameters cannot be calculated directly

Sample Statistics and Population Parameters (2/2)

- A Schematic Depiction



Graphical Summaries

- Recall that the mean, median and standard deviation, etc., are numerical summaries of a sample of of a population

- On the other hand, the graphical summaries are used as will to help visualize a list of numbers (or the sample items). Methods to be discussed include:
  - Stem and leaf plot
  - Dotplot
  - Histogram (more commonly used)
  - Boxplot (more commonly used)
  - Scatterplot

Stem-and-leaf Plot (1/3)

莖葉圖

- A simple way to summarize a data set
- Each item in the sample is divided into two parts
  - **stem**, consisting of the leftmost one or two digits
  - **leaf**, consisting of the next significant digit
- The stem-and-leaf plot is a compact way to represent the data
  - It also gives us some indication of the shape of our data

Stem-and-leaf Plot (2/3)

- Example: Duration of dormant (靜止) periods of the geyser (間歇泉) Old Faithful in Minutes

TABLE 1.3 Durations (in minutes) of dormant periods of the geyser Old Faithful

42	45	49	50	51	51	51	51	53	53
55	55	56	56	57	58	60	66	67	67
68	69	70	71	72	73	73	74	75	75
75	75	76	76	76	76	76	79	79	80
80	80	80	81	82	82	82	83	83	84
84	84	85	86	86	86	88	90	91	93

Stem	Leaf
4	259
5	0111133556678
6	067789
7	01233455556666699
8	000012223344456668
9	013

FIGURE 1.5 Stem-and-leaf plot for the geyser data in Table 1.3.

  - Let's look at the first line of the stem-and-leaf plot. This represents measurements of 42, 45, and 49 minutes
  - A good feature of these plots is that they display all the sample values. One can reconstruct the data in its entirety from a stem-and-leaf plot (however, the order information that items sampled is lost)

Stem-and-leaf Plot (3/3)

- Another Example: Particulate matter (PM) emissions for 62 vehicles driven at high altitude

TABLE 1.2 Particulate matter (PM) emissions (in g/gal) for 62 vehicles driven at high altitude

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02	23.38	9.24	3.22
2.06	4.04	17.11	12.26	19.91	8.50	7.81	7.18	6.95	18.64	7.10	6.04	5.66
8.86	4.40	3.57	4.35	3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14
8.67	9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01	5.91	5.60
5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32	1.83	7.56			

Contain the a count of number of items at or above this line

This stem contains the medium

Contain the a count of number of items at or below this line

```
Stem-and-leaf of HiAltitude    N = 62
Leaf Unit = 1.0

   4     0     1111
  19     0     222222223333333
(14)     0     44445555555555
  29     0     66666666777777
  15     0     8889999
   8     1     0
   7     1     233
   4     1
   4     1     7
   3     1     89
   1     2
   1     2     3
```

Leaf (ones digits)

Stem (tens digits)

Cumulative frequency column

FIGURE 1.6 Stem-and-leaf plot of the PM data in Table 1.2 in Section 1.2 as produced by MINITAB.

Dotplot

散點圖

- A dotplot is a graph that can be used to give a rough impression of the shape of a sample
  - Where the sample values are concentrated
  - Where the gaps are
- It is useful when the sample size is not too large and when the sample contains some repeated values
- Good method, along with the stem-and-leaf plot to informally examine a sample
- Not generally used in formal presentations

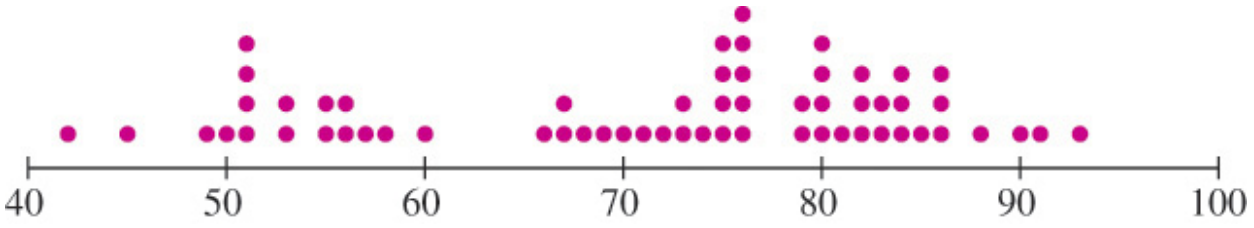


Figure 1.7 Dotplot of the geyser data in Table 1.3

Histogram (1/3)

直方圖

- A graph gives an idea of the shape of a sample
  - Indicate regions where samples are concentrated or sparse
- To have a histogram of a sample
  - The first step is to construct a frequency table
    - Choose boundary points for the class intervals
    - Compute the frequencies and relative frequencies for each class
      - Frequency: the number of items/points in the class
      - Relative frequencies: frequency/sample size
    - Compute the density for each class, according to the formula

      Density = relative frequency/class width

      - Density can be thought of as the relative frequency per unit

Histogram (2/3)

A frequency table

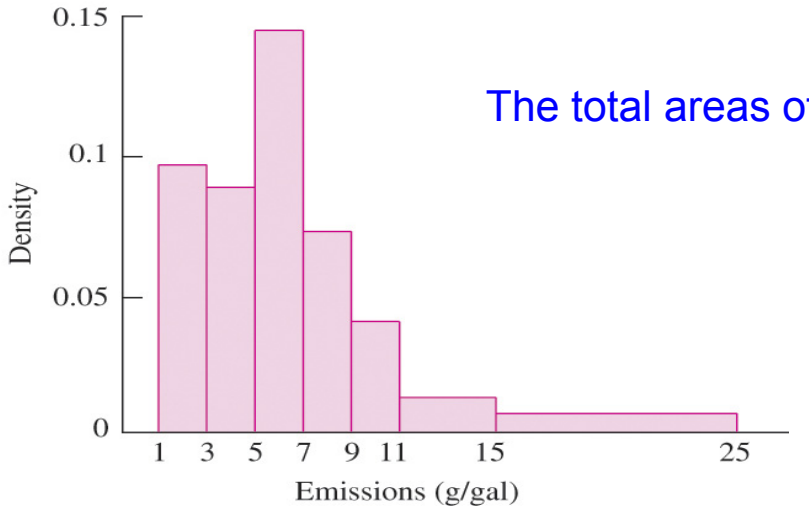
TABLE 1.4 Frequency table for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.194	0.0970
3–< 5	11	0.177	0.0885
5–< 7	18	0.290	0.1450
7–< 9	9	0.145	0.0725
9–< 11	5	0.081	0.0405
11–< 15	3	0.048	0.0120
15–< 25	4	0.065	0.0065

Table 1.4

– The second step is to draw a histogram for the table
  - Draw a rectangle for each class, whose height is equal to the density

A histogram



The total areas of rectangles is equal to 1

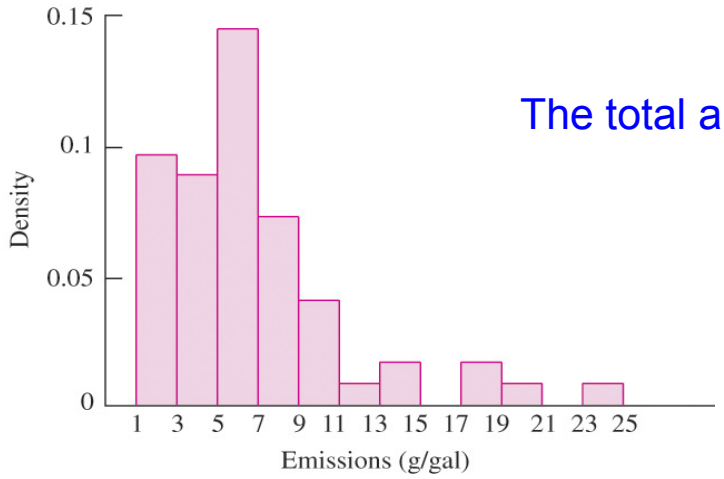
Figure 1.8

Histogram (3/3)

- A common rule of thumb for constructing the histogram of a sample
  - It is good to have more intervals rather than fewer
  - But it also to good to have large numbers of sample points in the intervals
  - Striking the proper balance between the above is a matter of judgment and of trial and error
    - It is reasonable to take the number of intervals roughly equal to the square root of the sample size

Histogram with Equal Class Widths

- Default setting of most software package
- Example: an histogram with equal class widths for Table 1.4



The total areas of rectangles is equal to 1

Figure 1.9

  - Devoted to too many (more than half) of the class intervals to few (7) data points
  - Compared to Figure 1.9, Figure 1.8 presents a smoother appearance and better enables the eye to appreciate the structure of the data set as a whole

Histogram, Sample Mean and Sample Variance (1/2)

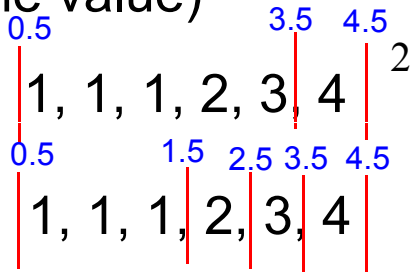
- Definition: The center of mass of the histogram is

$$\sum_i CenterOfClassInterval_i \times DensityOfClassInterval_i$$

  - An approximation to the sample mean
  - E.g., the center of mass of the histogram in Figure 1.8 is

$$2 \times 0.194 + 4 \times 0.177 + \cdots + 20 \times 0.065 = 6.730$$

    - While the sample mean is 6.596
  - The narrower the rectangles (intervals), the closer the approximation (the extreme case => each interval contains only items of the same value)



$$2 \times \frac{5}{6 \times 3} + 4 \times \frac{1}{6 \times 1} = \frac{22}{18} = 1.22$$

$$1 \times \frac{3}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} = \frac{12}{6} = 2$$

Histogram, Sample Mean and Sample Variance (2/2)

- Definition: The moment of inertia (力矩慣量) for the entire histogram is

$$\sum_i \left(CenterOfClassInterval_i - CenterOfMassOfHistogram\right)^2 \times DensityOfClassInterval_i$$

  - An approximation to the sample variance
  - E.g., the moment of inertia for the entire histogram in Figure 1.8 is

$$(2-6.730)^2 \times 0.194 + (4-6.730)^2 \times 0.177 + \cdots + (20-6.730)^2 \times 0.065 = 20.25$$

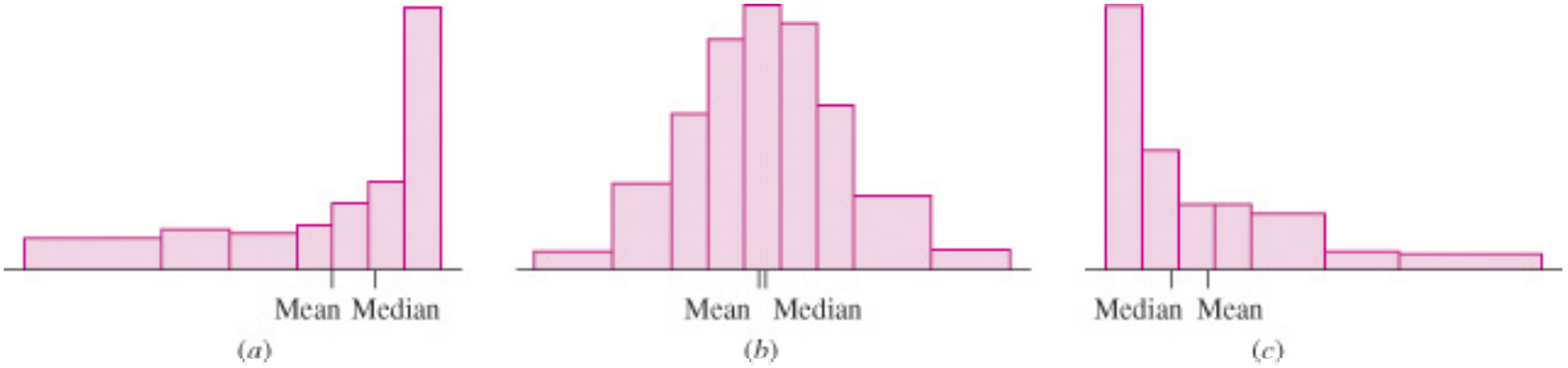
    - While the sample mean is 20.42
  - The narrower the rectangles (intervals) are, the closer the approximation is

Symmetry and Skewness (1/2)

- A histogram is perfectly symmetric if its right half is a mirror image of its left half
  - E.g., heights of random men

- Histograms that are not symmetric are referred to as skewed
- A histogram with a long right-hand tail is said to be skewed to the right, or positively skewed
  - E.g., incomes are right skewed (?)
- A histogram with a long left-hand tail is said to be skewed to the left, or negatively skewed
  - Grades on an easy test are left skewed (?)

Symmetry and Skewness (2/2)



skewed to the left nearly symmetric skewed to the right

- There is also another term called "kurtosis" that is also widely used for descriptive statistics
  - Kurtosis is the degree of peakedness (or contrarily, flatness) of the distribution of a population

More on Skewness and Kurtosis (1/3)

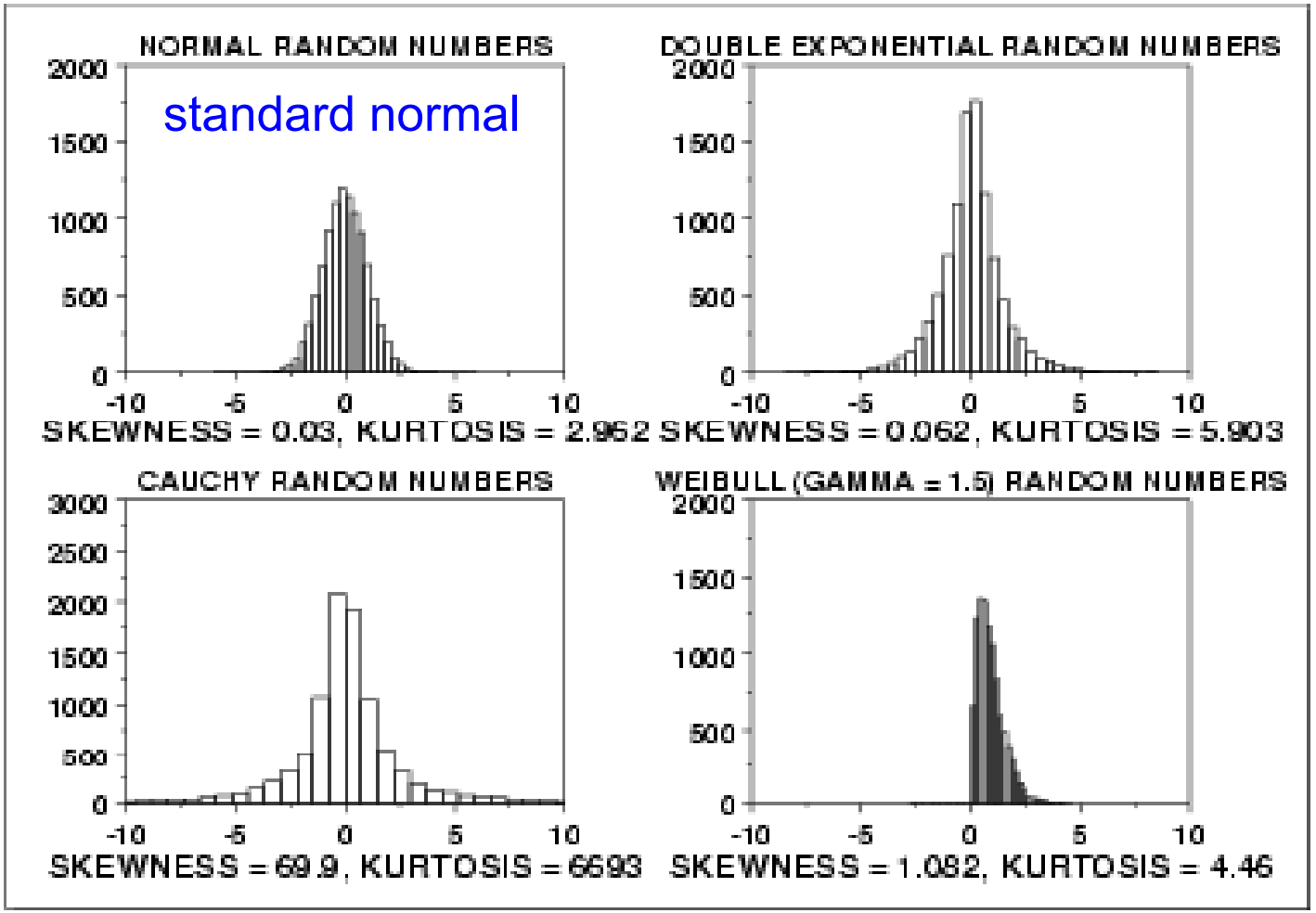
- Skewness can be used to characterize the symmetry of a data set (sample)

- Given a sample : $X_1, X_2, \cdots, X_n$
  - Skewness is defined by $Skewness = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{(n-1)s^3}$
  - If X_i follows a normal distribution or other distributions with a symmetric distribution shape => $Skewness = 0$
  - $Skewness > 0$: Skewed to the right
  - $Skewness < 0$: Skewed to the left

More on Skewness and Kurtosis (2/3)

- kurtosis can be used to characterize the flatness of a data set (sample)
- Given a sample : $X_1, X_2, \cdots, X_n$
  - Kurtosis is defined by $Kurtosis = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{(n-1)s^4}$
  - A standard normal distribution has $Kurtosis = 3$
  - A larger kurtosis value indicates a “peaked” distribution
  - A smaller kurtosis value indicates a “flat” distribution

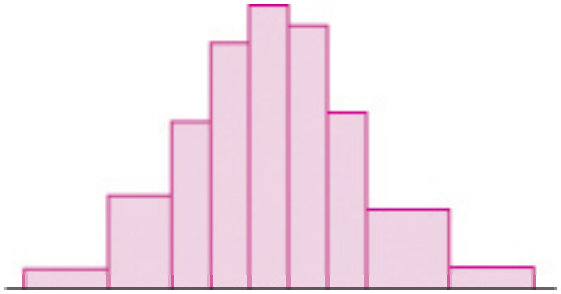
More on Skewness and Kurtosis (3/3)



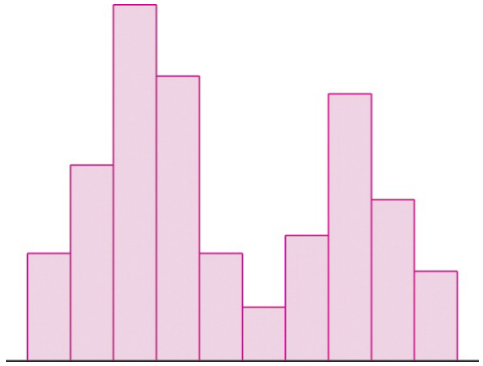
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

Unimodal and Bimodal Histograms (1/2)

- Definition: Mode
  - Can refer to the most frequently occurring value in a sample
  - Or refer to a peak or local maximum for a histogram or other curves



A unimodal histogram



A bimodal histogram

- A bimodal histogram, in some cases, indicates that the sample can be divides into two subsamples that differ from each other in some scientifically important way

Unimodal and Bimodal Histograms (2/2)

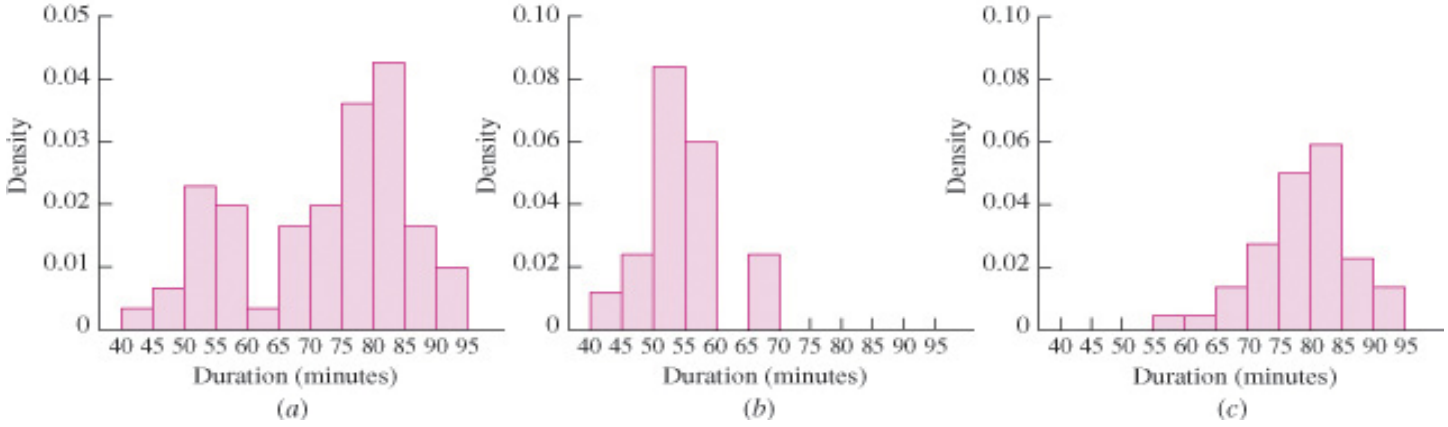
- Example: Durations of dormant periods (in minutes) and the previous eruptions of the geyser Old Faithful

TABLE 1.5 Durations of dormant periods (in minutes) and of the previous eruptions of the geyser Old Faithful

Dormant	Eruption	Dormant	Eruption	Dormant	Eruption	Dormant	Eruption
76	Long	90	Long	45	Short	84	Long
80	Long	42	Short	88	Long	70	Long
84	Long	91	Long	51	Short	79	Long
50	Short	51	Short	80	Long	60	Long
93	Long	79	Long	49	Short	86	Long
55	Short	53	Short	82	Long	71	Long
76	Long	82	Long	75	Long	67	Short
58	Short	51	Short	73	Long	81	Long
74	Long	76	Long	67	Long	76	Long
75	Long	82	Long	68	Long	83	Long
80	Long	84	Long	86	Long	76	Long
56	Short	53	Short	72	Long	55	Short
80	Long	86	Long	75	Long	73	Long
69	Long	51	Short	75	Long	56	Short
57	Long	85	Long	66	Short	83	Long

long: more than 3 minutes

short: less than 3 minutes



Histogram of all 60 durations

Histogram of the durations following short eruption

Histogram of the durations following long eruption

Histogram with Height Equal to Frequency

- Till now, we refer the term "histogram" to a graph in which the heights of rectangles represent densities
- However, some people draw histograms with the heights of rectangles equal to the frequencies
- Example: The histogram of the sample in Table 1.4 with the heights equal to the frequencies

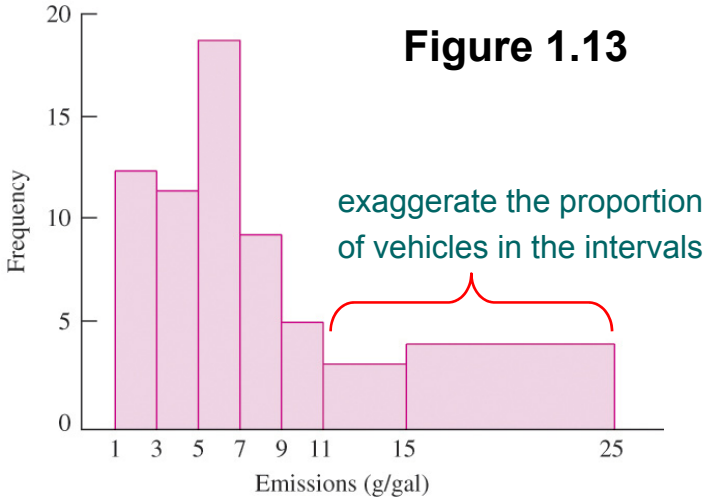


Figure 1.13

cf.

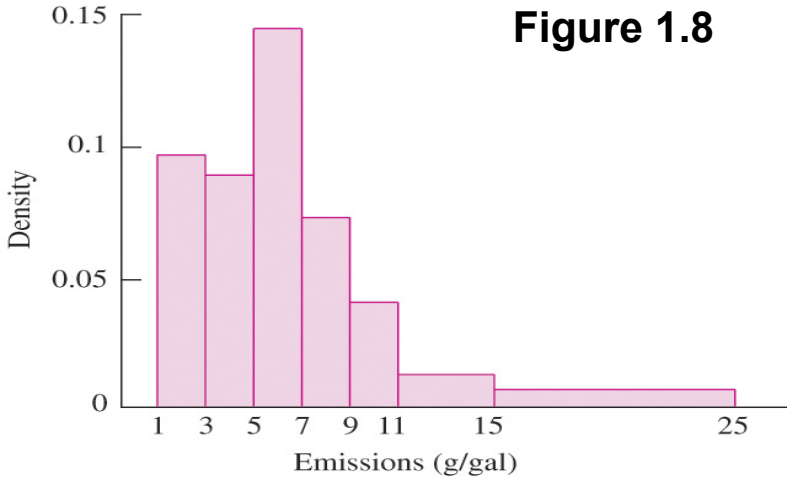


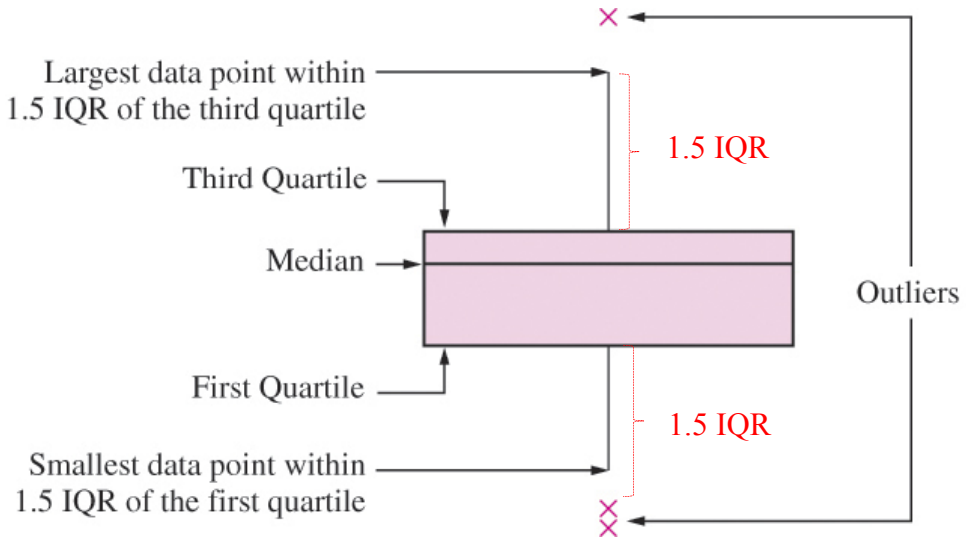
Figure 1.8

Boxplot (1/4)

箱型圖
(盒鬚圖)

1.5 IQR

- A boxplot is a graph that presents the median, the first and third quartiles, and any outliers present in the sample



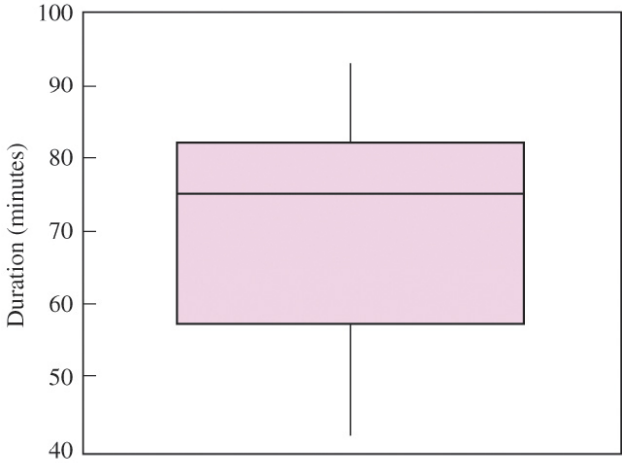
  - The interquartile range (IQR) is the difference between the third and first quartile. This is the distance needed to span the middle half of the data

Boxplot (2/4)

- Steps in the Construction of a Boxplot
  - Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box
  - Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is not more than 1.5 IQR below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points
  - Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually

Boxplot (3/4)

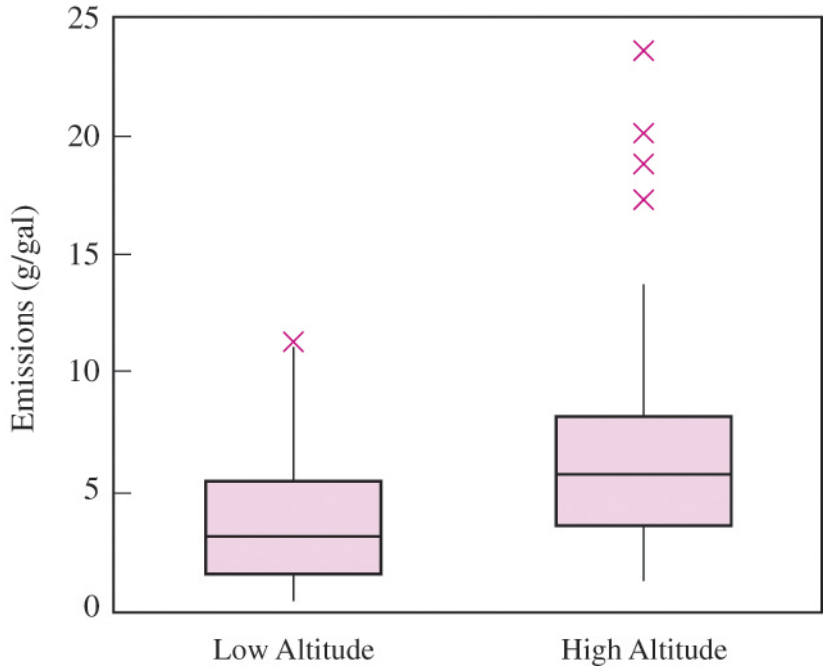
- Example: A boxplot for the geyser data presented in Table 1.5
  - Notice there are no outliers in these data
  - The sample values are comparatively densely packed between the median and the third quartile
  - The lower whisker is a bit longer than the upper one, indicating that the data has a slightly longer lower tail than an upper tail



  - The distance between the first quartile and the median is greater than the distance between the median and the third quartile
  - This boxplot suggests that the data are skewed to the left (?)

Boxplot (4/4)

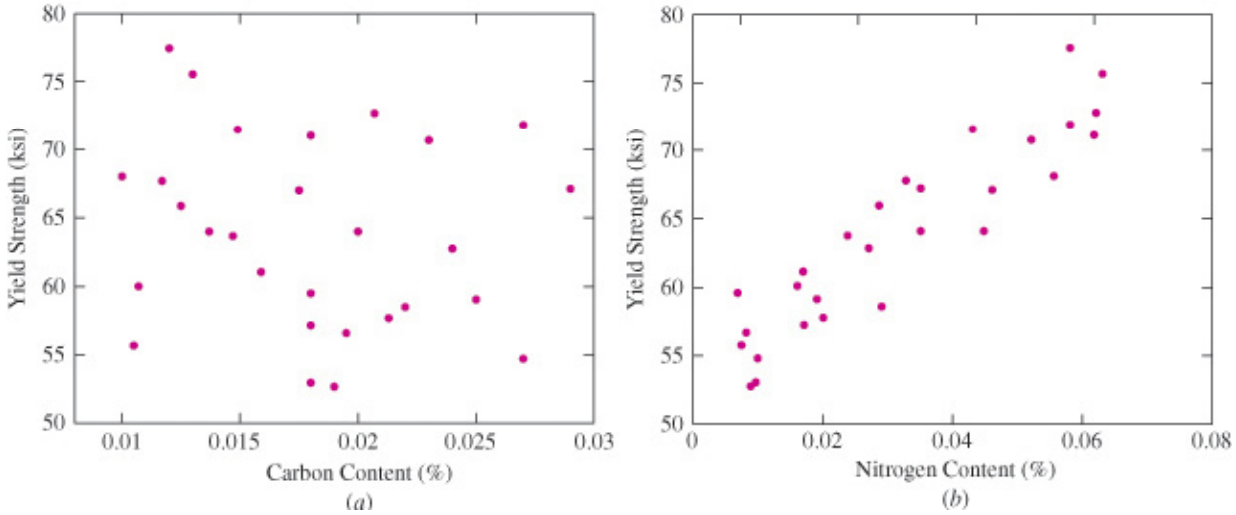
- Another Example: Comparative boxplots for PM emissions data for vehicle driving at high versus low altitudes



Scatterplot (1/2)

散佈圖

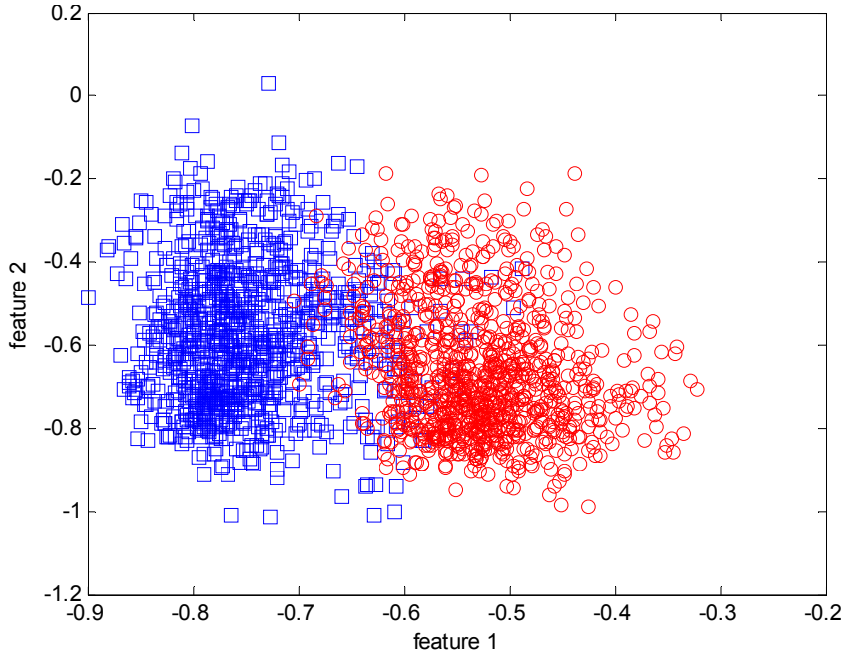
- Data for which item consists of a pair of values is called bivariate
- The graphical summary for bivariate data is a scatterplot
- Display of a scatterplot (strength of Titanium (鈦) welds vs. its chemical contents)



碳　　氮

Scatterplot (2/2)

- Example: Speech feature sample (Dimensions 1 & 2) of male (blue) and female (red) speakers after LDA transformation



Summary

- We discussed types of data
- We looked at sampling, mostly SRS
- We studied summary (descriptive) statistics
  - We learned about numeric summaries
  - We examined graphical summaries (displays of data)