

# Engineering Statistics



Copyright © 1993-2006 David  
M. Lane

## Descriptive Statistics

**Dr. Vvn Weian Chao (趙韋安)**

<https://ce.nctu.edu.tw/member/teachers/23>

Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan



# Descriptive statistics

# Describing your data

**Spread**

**Central  
tendency**

**Variability**

**Selecting right  
plot**

# Variables



Categorical

Numeric

# Important

N

Mean

Median

Mode

SD

IQR

Skewness

Kurtosis

## Some important terms:

- **Data (數據)**- collections of facts
- **Population (母體)** - a well-defined collection of objects
- **Census (普查)** – collecting desired information for all objects in the population
- **Sample (樣本)** - a subset of the population
- **Variable (變數)** - any characteristic whose value may change from one object to another in the population
- **Univariate data (單變量)** - observations on a single variable
- **Bivariate data (雙變量)** - observations on each of two variables  
籃球選手身高、體重
- **Multivariate data (多變量)** - observations on more than two variables  
舒張壓、收縮壓與血脂

## Example 1.1: Charity Business in the US (慈善機構) 籌款活動佔總費用支出的比例

- A sample of **5500** charitable organizations
- For some efficiently-operated charities, only a small percentage of total expenses are spent on fund-raising and administrative activities
- Others spend a high percentage of what they take in to perform the same activities
- Data on fund-raising expenses as a percentage of total expenditures for a random sample of **60** charities:

**Data:**

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

# Descriptive statistics

Display the data to find the answers:

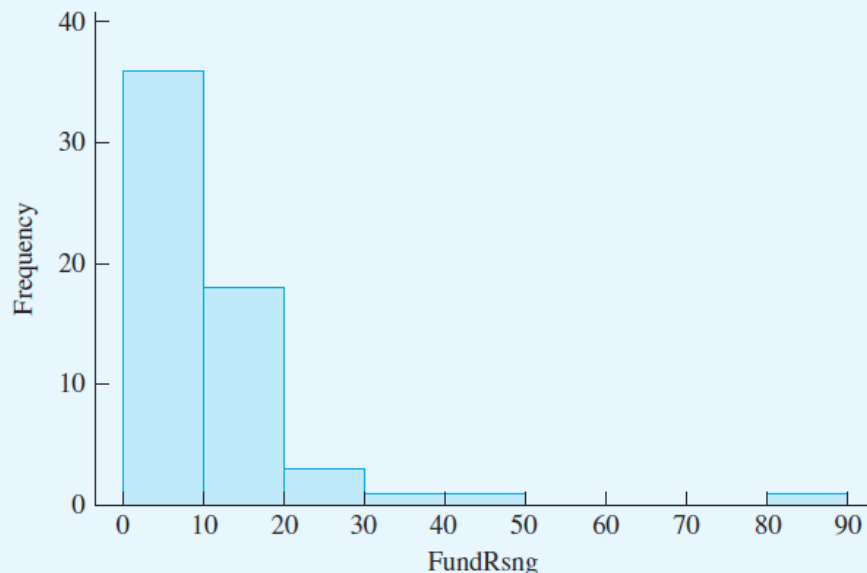
Stem-and-leaf of FundRsng N = 60

Leaf Unit = 1.0

```

0 | 0111112222333333344
0 | 555566666666778888
1 | 0001222244
1 | 55666789
2 | 01
2 | 6
3 | 4
3 |
4 |
4 | 8
5 |
5 |
6 |
6 |
7 |
7 |
8 | 3

```



- Observe how the percentages are distributed over the range of possible values from 0 to 100.
- A substantial majority of the charities in the sample obviously spend less than 20% on fund-raising.
- Only a few percentages might be viewed as beyond the bounds of sensible practice.



# Displays for univariate data



Stem-and-leaf  
Histograms

...

# Stem-and-Leaf Displays

- Separate each observation into two parts:
  - A **stem**: consists of one or more **leading** digits
  - A **leaf**: consists of the **remaining or trailing** digit(s)
- Stem values are listed in a single column
- Leaf of each observation are then placed on the row of the corresponding stem

**Example: Use of alcohol by college students**  
 (各所大學學生喝酒習慣比例)

0	4	
1	1345678889	
2	1223456666777889999	Stem: tens digit
3	011223334455566667777888899999	Leaf: ones digit
4	111222223344445566666677788888999	
5	00111222233455666667777888899	
6	01111244455666778	

- A stem-and-leaf display conveys
  - Identification of a **typical** or representative value
  - **Extent of spread** about the typical value
  - Presence of any **gaps** in the data
  - Extent of **symmetry** in the distribution of values
  - Number and location of **peaks**
  - Presence of any **outlying** values

# Stem-and-Leaf Displays

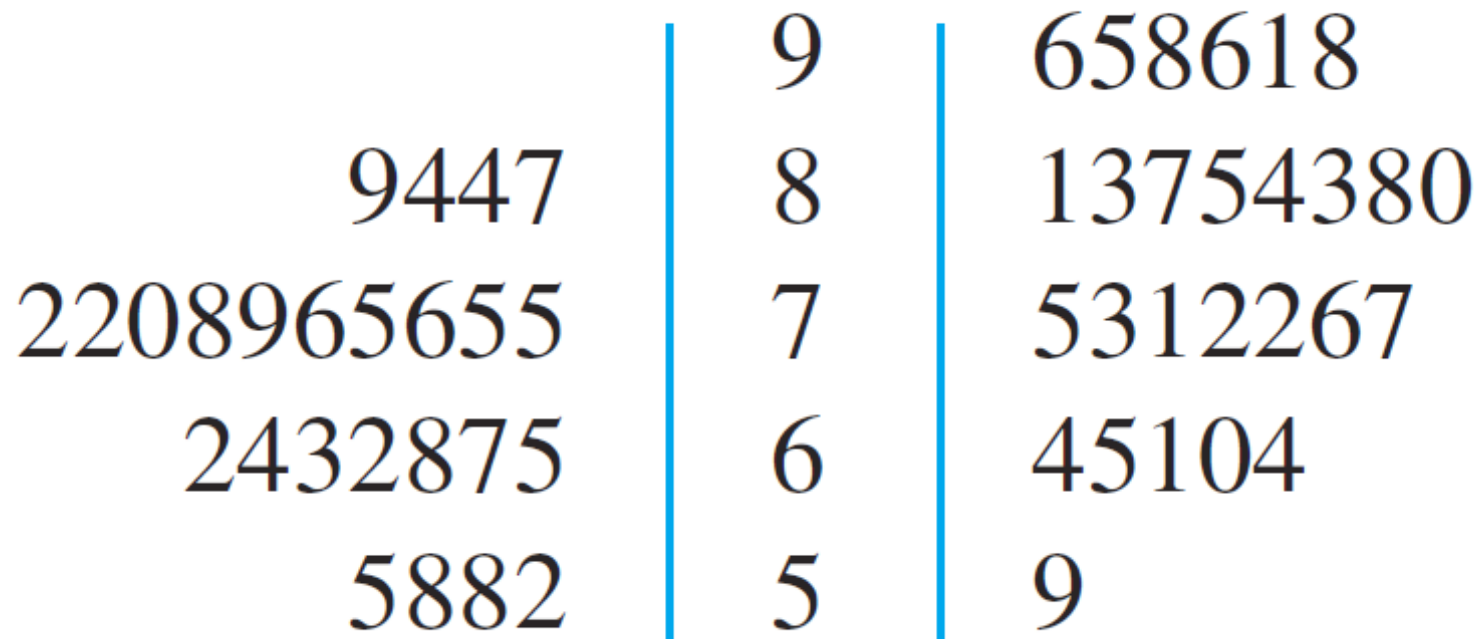
A display of the binge-drinking data with **repeated** stems

```
Stem-and-leaf of pct binge N = 140
Leaf Unit = 1.0
```

	1	0	4
	1	0	
	4	1	134
	11	1	5678889
	16	2	12234
	30	2	56666777889999
	40	3	0112233344
	61	3	555666677777888899999
median →	(14)	4	11122222334444
	65	4	5566666677788888999
	46	5	001112222334
	34	5	55666667777888899
	17	6	011112444
	8	6	55666778

# Stem-and-Leaf Displays

- A **comparative** stem-and-leaf display



Displays for univariate  
data: most appropriate for  
smaller datasets

Package: **aplpack**

# Displays for univariate data: most appropriate for smaller datasets

```
stem.leaf(  
data, unit, m  
).
```

TRY  
it  
in  
R



# R: Descriptive statistics

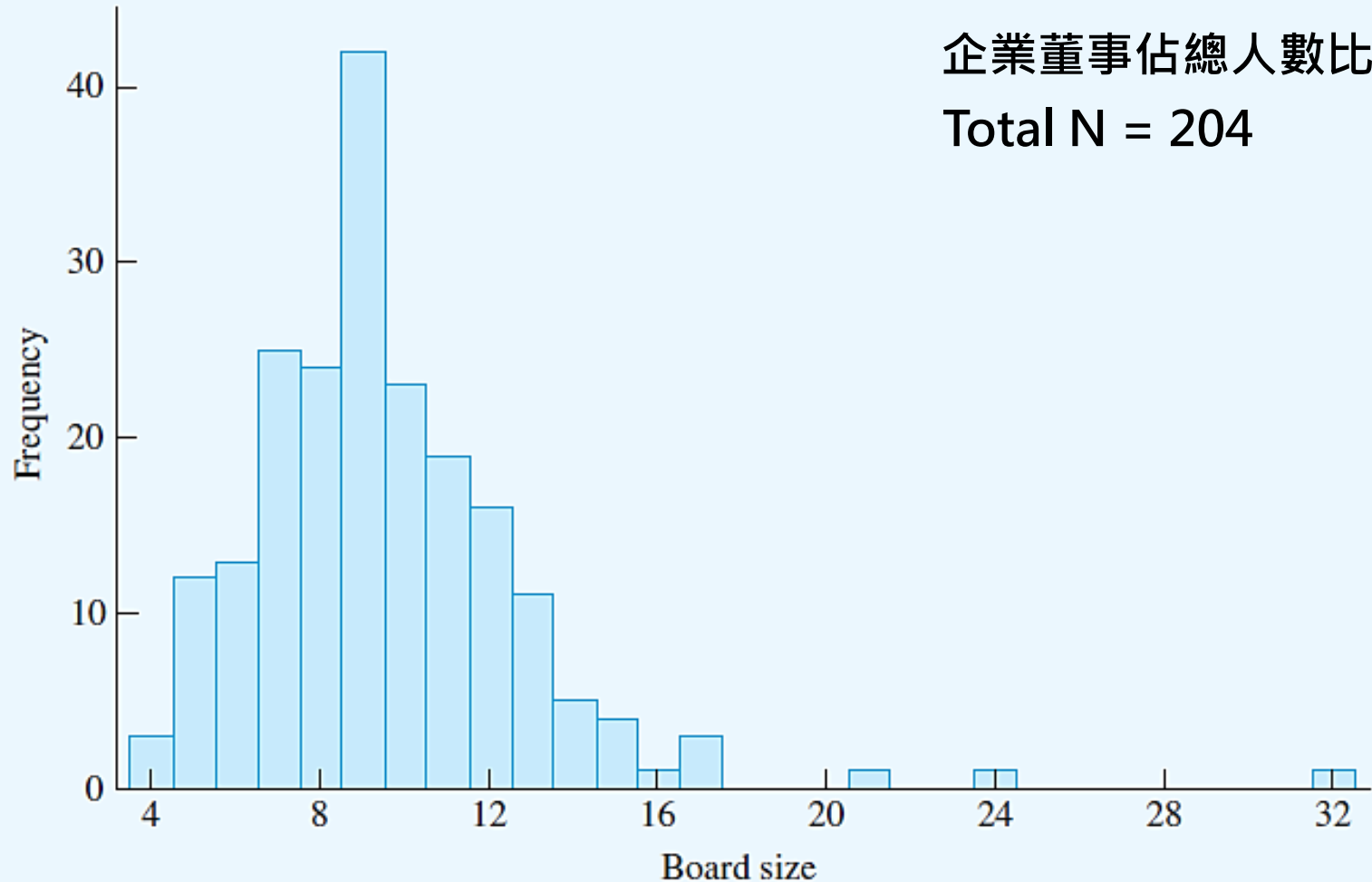


R\_descriptive\_a.R

## Definitions:

- **Discrete variable** - possible values either is finite or else can be listed in an infinite sequence  
(可以一一列舉)
- **Continuous variable** - possible values consist of an entire interval on the number line
- **Frequency** - the **number** of times a particular value occurs in the data set
- **Relative frequency** - the **fraction** or proportion of time the value occurs

## Positive Skew (正偏)



- 選定特定區間來計數，但若數值落在邊界上，只能重新調整區間。
- 區間數量約可以用總數據個數開根號來決定。

**number of classes  $\approx \sqrt{\text{number of observations}}$**

# Histograms- class interval

## 能源消耗數據

90 gas-heated homes

Power companies need information about customer usage to obtain accurate forecasts of demand. Investigators from Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. An adjusted consumption value was calculated as follows:

$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather, in degree days})(\text{house area})}$$

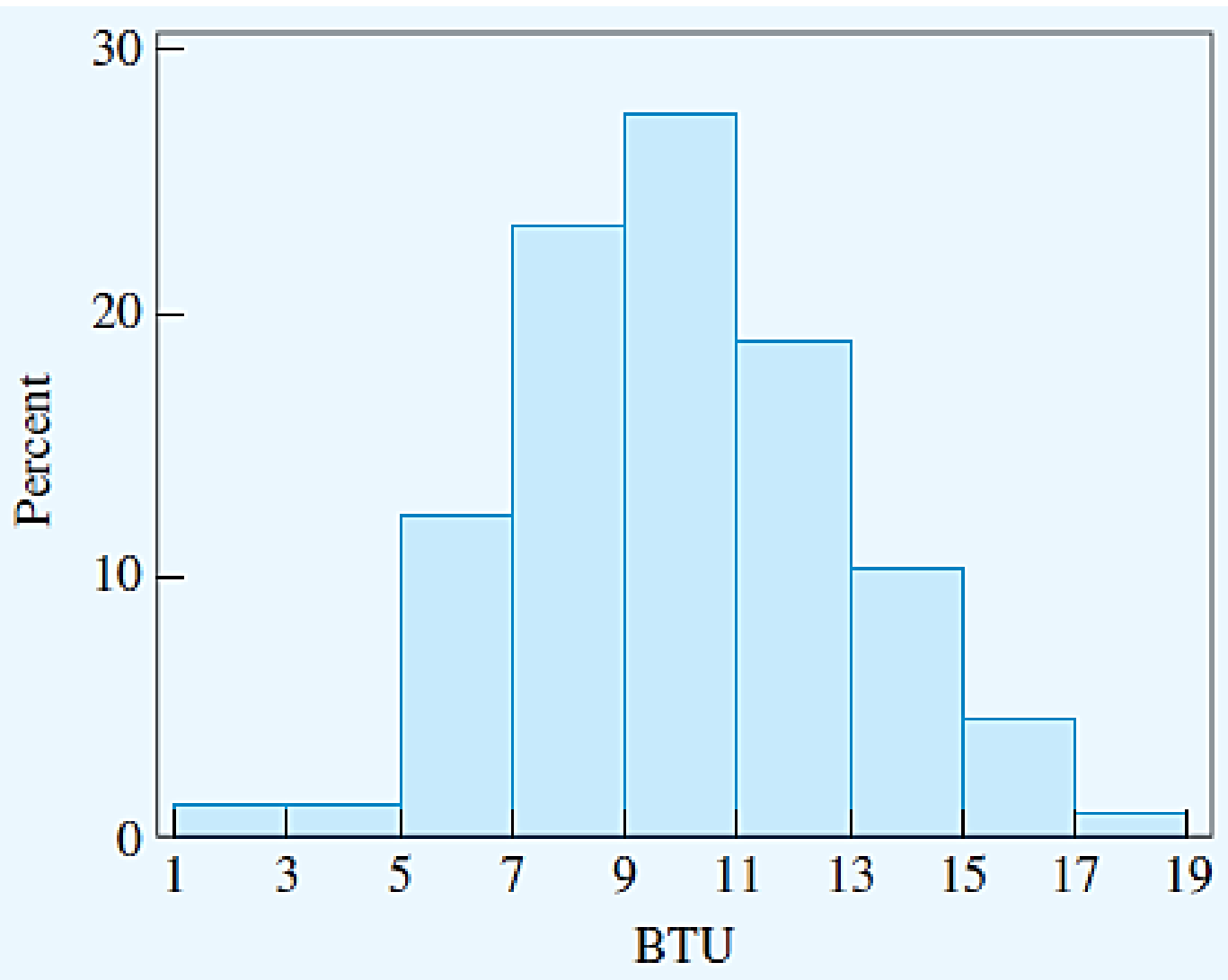
← normalization

This resulted in the accompanying data (part of the stored data set FURNACE.MTW available in Minitab, which we have ordered from smallest to largest):

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

↑  
Sorting data

# Histograms- class interval



90筆資料  
9 classes

# Histograms- class interval

Question: 若依照此分類統計結果，將如何計算BTU小於10的比例為何？

Class:	1 – <3	3 – <5	5 – <7	7 – <9	9 – <11	11 – <13	13 – <15	15 – <17	17 – <19
Frequency:	1	1	11	21	25	17	9	4	1
Relative frequency:	.011	.011	.122	.233	.278	.189	.100	.044	.011

From the histogram,

proportion of observations  
less than 9  $\approx .01 + .01 + .12 + .23 = .37$

(exact value =  $34/90 = .378$ )

The relative frequency for the 9 – <11 class is about .27, so roughly half of this, or .135, should be between 9 and 10. Thus

proportion of observations  
less than 10  $\approx .37 + .135 = .505$  (slightly more than 50%)

The exact value of this proportion is  $47/90 = .522$ .

- For Continuous Data (**Unequal** Class Widths):
  - After determining the frequencies and relative frequencies, calculate height of each rectangle:

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

- Resulting rectangle heights are called **densities**; the vertical scale is the **density scale**
- Will also work for equal class widths



# Density Histograms



Example: Corrosion of reinforcing steel in concrete structures ( $n = 48$ )

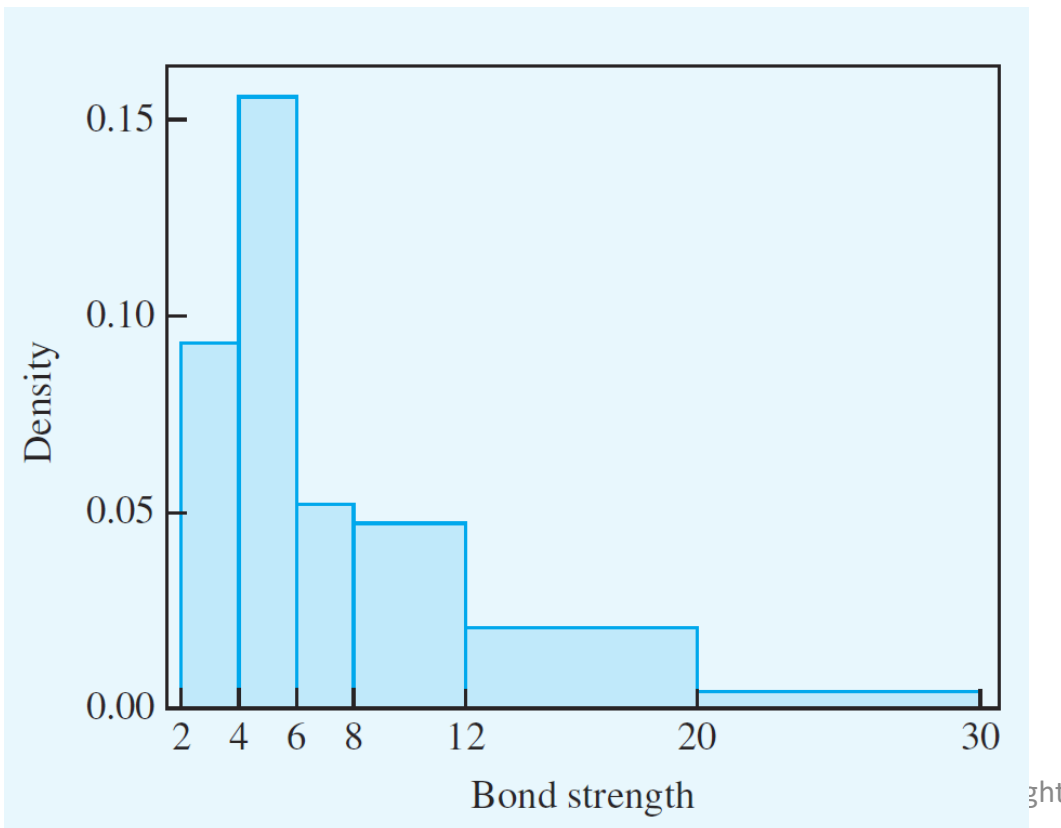
鋼筋混凝土建築的腐蝕問題，透過glass-fiber-reinforced plastic來包覆混凝土外圍的處理

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Class:	2— <4	4— <6	6— <8	8— <12	12— <20	20— <30
Frequency:	9	15	5	9	8	2
Relative frequency:	.1875	.3125	.1042	.1875	.1667	.0417
Density:	.094	.156	.052	.047	.021	.004

# Density Histograms

**relative frequency** = (class width) (density)  
= (rectangle width) (rectangle height)  
= **rectangle area**



底下面積總=1

- **Unimodal (單峰)** – rises to a single peak and then declines
  - **positively skewed:** if the right or upper tail is stretched out compared with the left or lower tail
  - **negatively skewed:** if the longer tail extends to the left
- **Bimodal (雙峰)** – has two different peaks
- **Multimodal (多峰)** – more than two peaks
- **Symmetric** – if the left half is a mirror image of the right half

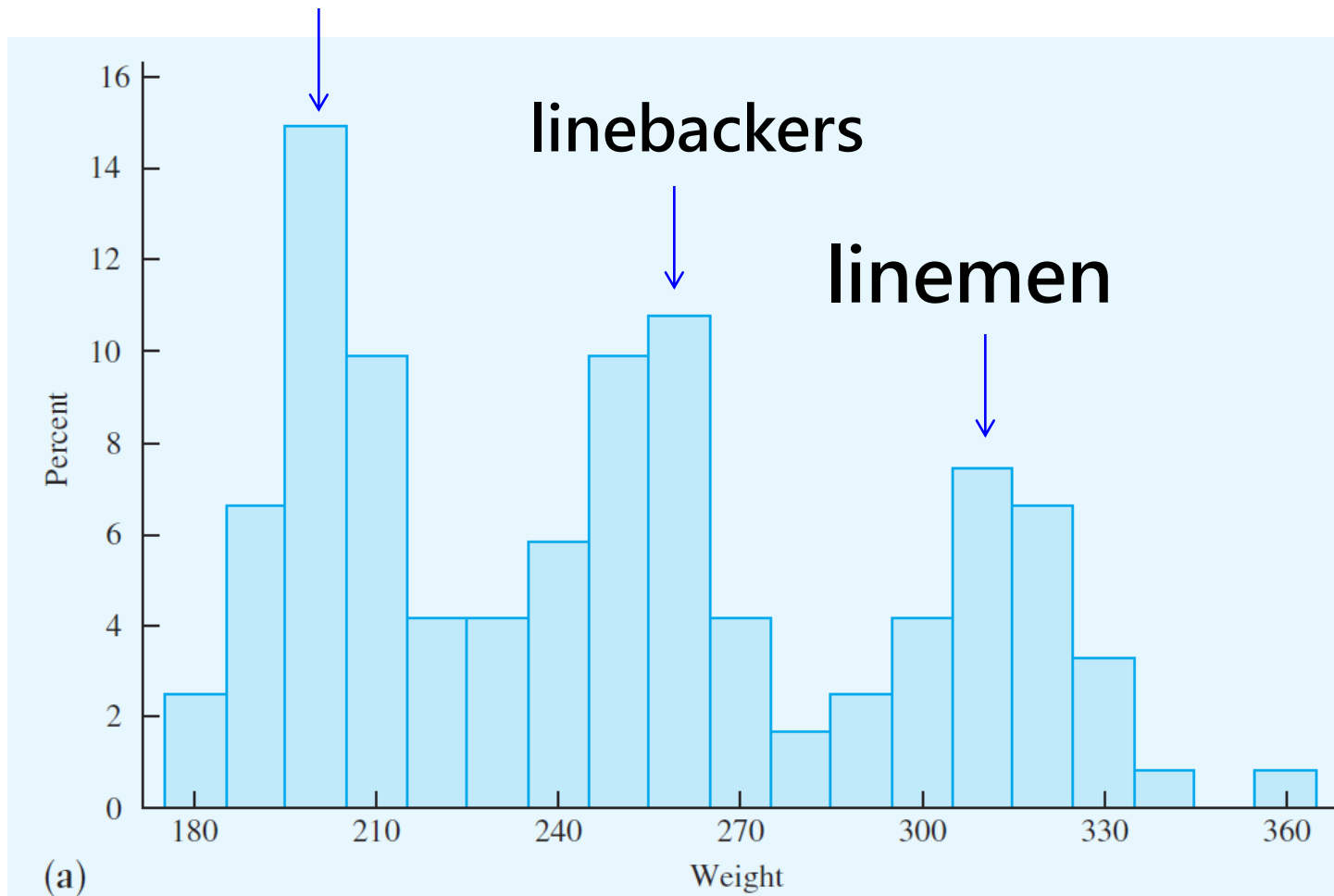
# Histograms Shapes

## Example: Histogram and smoothed histogram

- the weights (lbs) of 121 NFL players

others

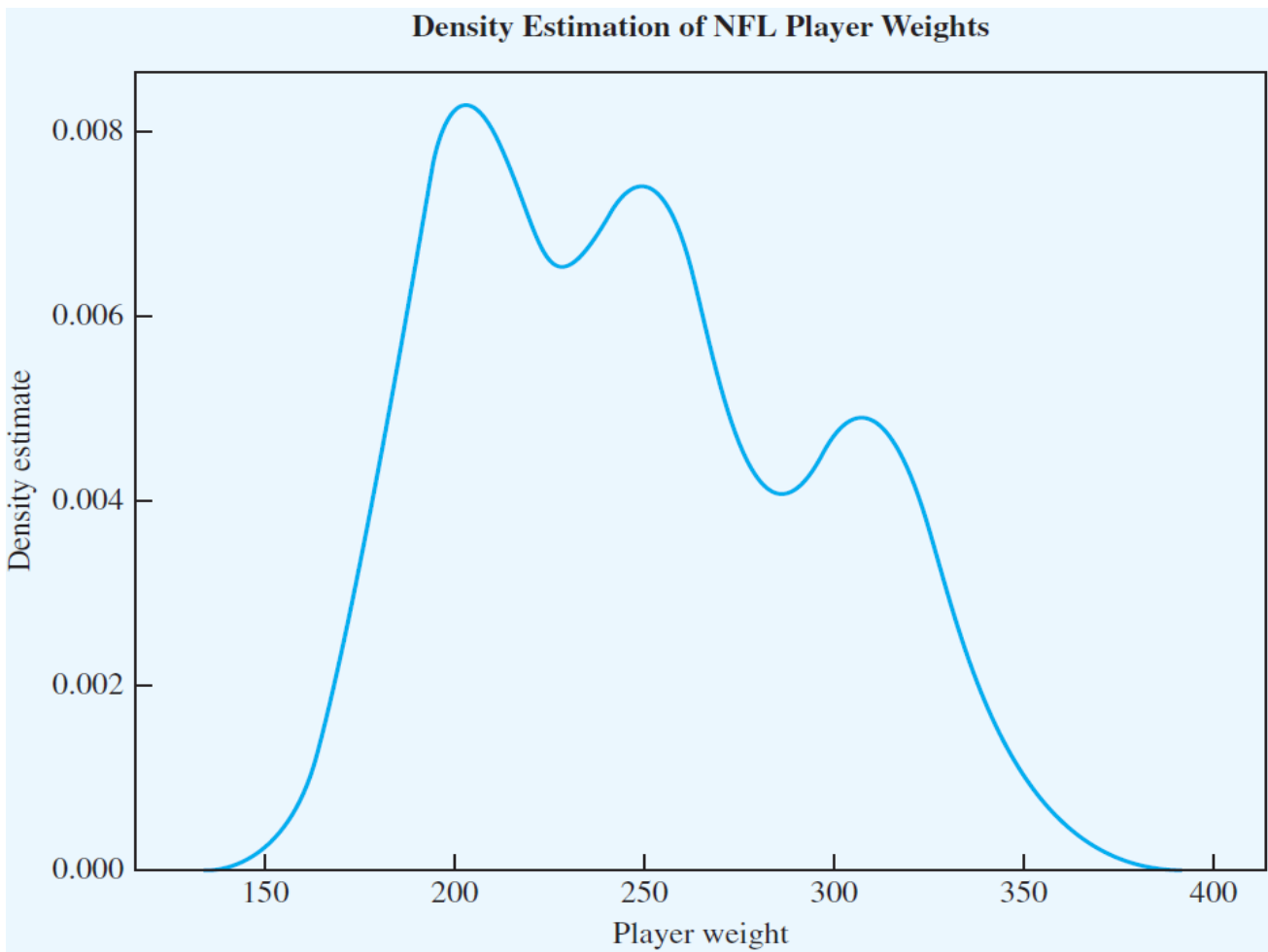
Multimodal



# Histograms Smoothed-density

Example: Histogram and smoothed histogram

- the weights (lbs) of 121 NFL players



**平滑化直方圖優勢:** 可以更清楚判釋直方圖的形貌特徵，易於定義說明資料特徵。

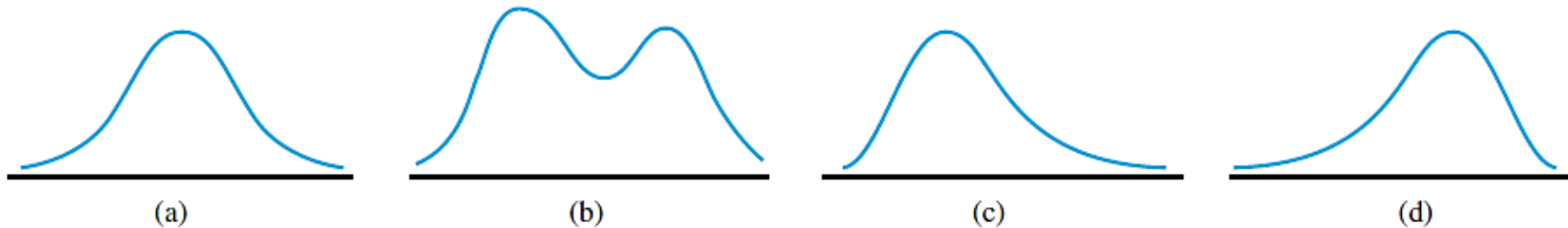


圖 1.11 平滑直方圖：(a) 對稱單峰；(b) 雙峰；(c) 正偏；(d) 負偏。

- **Density function**  $f(x)$ : used to describe population or process distribution of a continuous variable  $x$
- **Density curve**: graph of  $f(x)$
- The following properties must be satisfied:
  1.  $f(x) \geq 0$
  2.  $\int_{-\infty}^{\infty} f(x) dx = 1$
  3. For any two numbers  $a$  and  $b$  with  $a < b$ ,  
proportion of  $x$  values between  $a$  and  $b$  =

$$\int_a^b f(x) dx$$

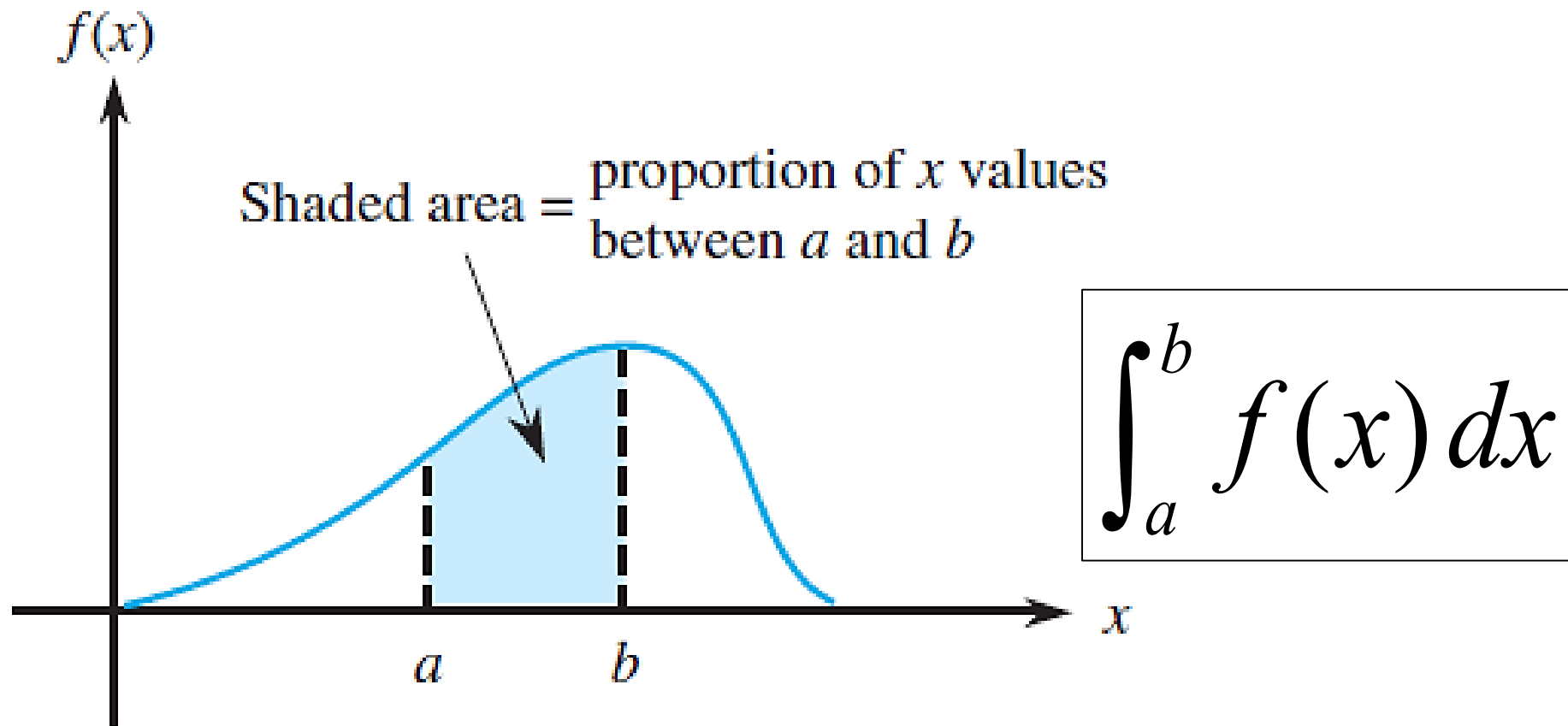


圖 1.14 密度曲線下方的面積等於區間內數值在整個數線上的比例。



Displays for univariate  
data: most appropriate for  
smaller datasets

`geom_histogram(  
bins, binwidth  
).`

TRY  
it  
in  
R

# R: Descriptive statistics



R\_descriptive\_b.R

length()	資料長度
mean()	平均數
median()	中位數
range()	全距
quantile()	四分位數
IQR()	四分位差
summary()	描述統計摘要
sd()	標準差
var()	變異數
skewness()	偏度
kurtosis()	峰度

加: +

減: -

乘: \*

除: /

整除: %/%

餘數: %%

冪次: ^

log()  
log10()  
exp()  
sin()  
cos()  
asin()  
acos()

- Measures of Center for Data
  - *The Sample Mean*
  - *The Sample Median*
  - *Trimmed Means*
- Measures of Center for Distributions
  - *Discrete Distributions*
  - *Continuous Distributions*
  - $\mu$  and  $\bar{x}$
  - *The Median of a Distribution*

- Suppose that the sample consists of observations on a numerical variable  $x$
- $n$  represents the sample size
- The individual observations:  $x_1, x_2, \dots, x_n$

## The Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$



# Measures of Center for Data: Example

✓ stem-and-leaf display of the data

0H	6
1L	0 2 4
1H	5 5 6 7 7 7 8 8 9
2L	
2H	
3L	0

Observation:

• Sample mean:  $\bar{x} = \frac{229.0}{14} = 16.36$

•  $x = 30.5$  is an **outlier**

Without this outlier:  $\bar{x} = 16.36 \square \bar{x} = 15.27$

✓ 樣本平均值容易受Outlier影響!!!

## The Sample Median:

- 相對於樣本平均值，樣本中位數較不受Outlier影響

$$\tilde{x} = \begin{cases} \frac{n+1}{2} \text{th value on ordered list} & n \text{ odd} \\ \text{average of } \frac{n}{2} \text{th and } \frac{n}{2} + 1 \text{th values} & n \text{ even} \end{cases}$$

## Music composer' s instructions

(12位演奏家呈現特定的曲目其演奏時間)

✓ Durations (min) are listed in increasing order:

62.3 62.8 63.6 65.2 65.7 66.4 67.4 68.4 68.8 70.8 75.7 79.0

- The sample median:  $\tilde{x} = (66.4 + 67.4) / 2 = 66.90$

✓ If the largest observation 79.0 had not been included:

$$\tilde{x} = 66.9 \quad \square \quad \tilde{x} = 66.4$$

## Trimmed Means

- A trimmed mean is a compromise between  $\bar{x}$  and  $\tilde{x}$ .
- **Less sensitive** to outliers than the mean but more sensitive than the median.
- The observations are **first ordered** from smallest to largest.
- A trimming percentage  $100r\%$  is chosen, where  $r$  is a number between **0 - 0.5**.
- The sample mean is a 0% trimmed mean.
- The median is a trimmed mean corresponding to the largest possible trimming percentage.

Lifetime (hr) of a certain type of incandescent lamp  
(燈泡壽命)

✓ Consider 20 observations, ordered from smallest to largest:

612	623	666	744	883	898	964	970	983	1003
1016	1022	1029	1058	1085	1088	1122	1135	1197	1201

$$\bar{x} = 19,299 / 20 = 965.0 \text{ and } \tilde{x} = (1003 + 1201) / 2 = 1102$$

• The 10% trimmed mean ( $20 \times 0.1 = 2$ ):

$$\bar{x}_{\text{tr}(10)} = \frac{19,299 - 612 - 1201}{16} = 979.1$$

The primary measure of center for a discrete distribution is the mean value; for continuous distributions, both the mean value and the median are frequently used.

## Definitions:

- The mean value of a discrete variable  $x$

$$\mu_x = \sum x \cdot p(x)$$

描述隨機變數水準的統計量稱為  
期望值(expected value)。

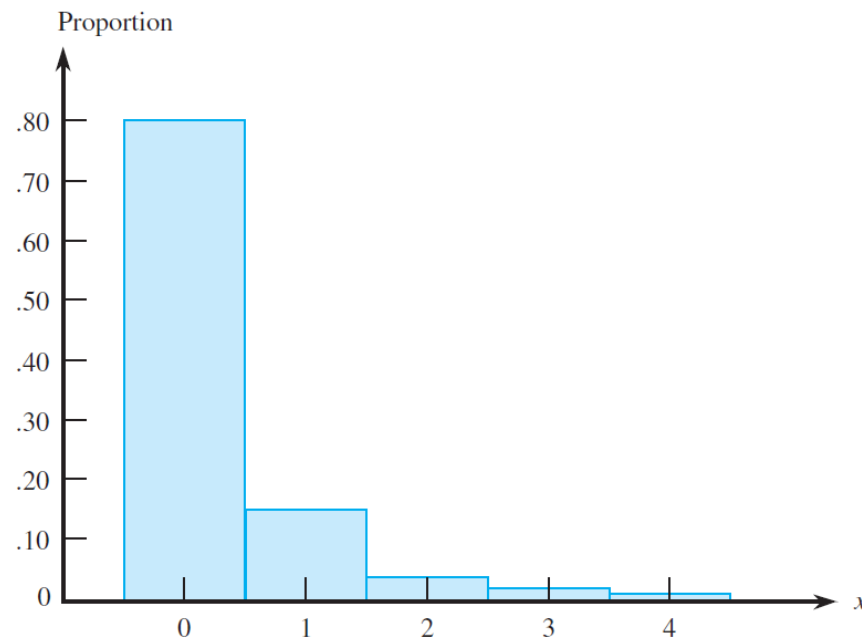
## Discrete Distributions

### Example: Manufacturing plastic parts

- Let  $x$  represent the number of defects on a single part

$x$ :	0	1	2	3	4
$p(x)$ :	.80	.14	.03	.02	.01

- Where is this distribution centered?
- What is the mean or long-run average value of  $x$ ?





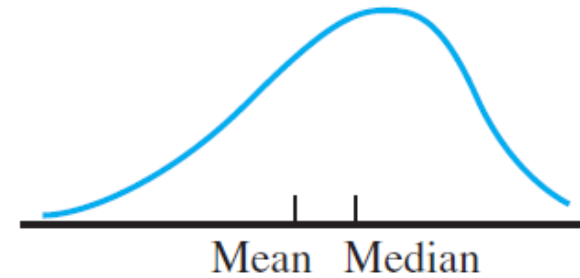
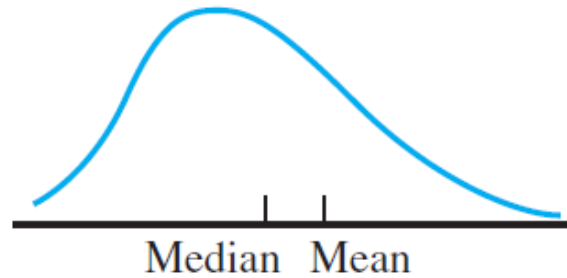
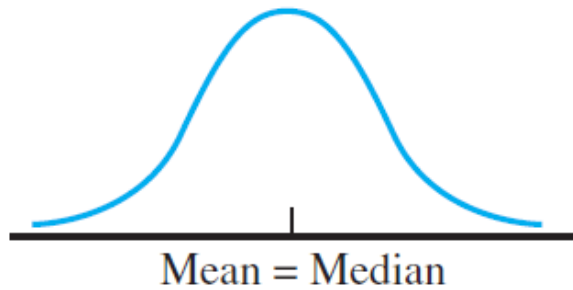
## Example 2.4

We return now to the plastic part scenario introduced at the outset of this subsection. The mean value of  $x$ , the number of defects on a part, is

$$\begin{aligned}\mu_x &= \sum_{x=0}^4 x \cdot p(x) \\ &= 0(p(0)) + 1(p(1)) + 2(p(2)) + 3(p(3)) + 4(p(4)) \\ &= (0)(.80) + (1)(.14) + (2)(.03) + (3)(.02) + (4)(.01) \\ &= .30\end{aligned}$$

When we consider the population of all such parts, the population mean value of  $x$  is .30. Alternatively, .30 is the long-run average value of  $x$  when part after part is monitored. It can also be shown that the histogram of the distribution of Figure 2.4

**產品平均缺陷數量 = 0.3 is not a possible value of  $x$**



偏度(skewness)是指資料分配的不對稱性。測度資料分配不對稱性的統計量稱為偏度係數(coefficient of skewness, SK)。

$$k_3 = \frac{(x - \bar{x})^3}{n}$$

$$k_2 = \frac{(x - \bar{x})^2}{n}$$

$$SK = \frac{k_3}{(k_2)^{3/2}}$$

- ✓ 負值表示左偏分配
- ✓ 正值表示右偏分配
- ✓ SK等於0: 對稱分配
- ✓ 若SK大於1或小於-1，視為嚴重偏斜分配

- Measures of Variability for Sample Data
- The Variance and Standard Deviation of a Discrete Distribution
- The Variance and Standard Deviation of a Continuous Distribution
  - *The Case of a Normal Distribution*
  - *Other Continuous Distributions*
  - $\sigma^2$  and  $s^2$

# Measures of Variability for Sample Data



- **Range:** the difference between the largest and smallest sample values
- **Deviations from the mean:**  $x_1 - \bar{x}, x_2 - \bar{x}, \dots$
- **Sample variance: (樣本變異數)**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1} \quad S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2$$

- **Sample standard deviation: (樣本標準差)**

$$s = \sqrt{s^2}$$

- The **variance of a discrete distribution** for a variable  $x$ , mass function  $p(x)$ , is

$$\sigma^2 = \sum (x - \mu)^2 \cdot p(x)$$

- The **standard deviation** is  $\sigma$ , the positive square root of the variance

- The **variance** of a continuous distribution specified by density function  $f(x)$  is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

- The **standard deviation** is  $\sigma$ , the positive square root of the variance

峰度(kurtosis)是指資料分佈峰值高低。通常是與標準常態分佈比較而言。由於標準常態分配的峰度係數為3，當K大於3時為尖峰分配，資料的分佈相對集中；當K小於3時為扁平分佈，資料的分佈相對分散。

$$m_4 = \frac{(x - \bar{x})^4}{n}$$

$$m_2 = \frac{(x - \bar{x})^2}{n}$$

$$KT = \frac{m_4}{(m_2)^2}$$

# R: Descriptive data:

## Data shape

# Package: moments



# R: Descriptive data:

Data shape

**skewness().**

**kurtosis().**

TRY  
it  
in  
R

# R: Descriptive data:

## Data shape

R\_descriptive\_c.R

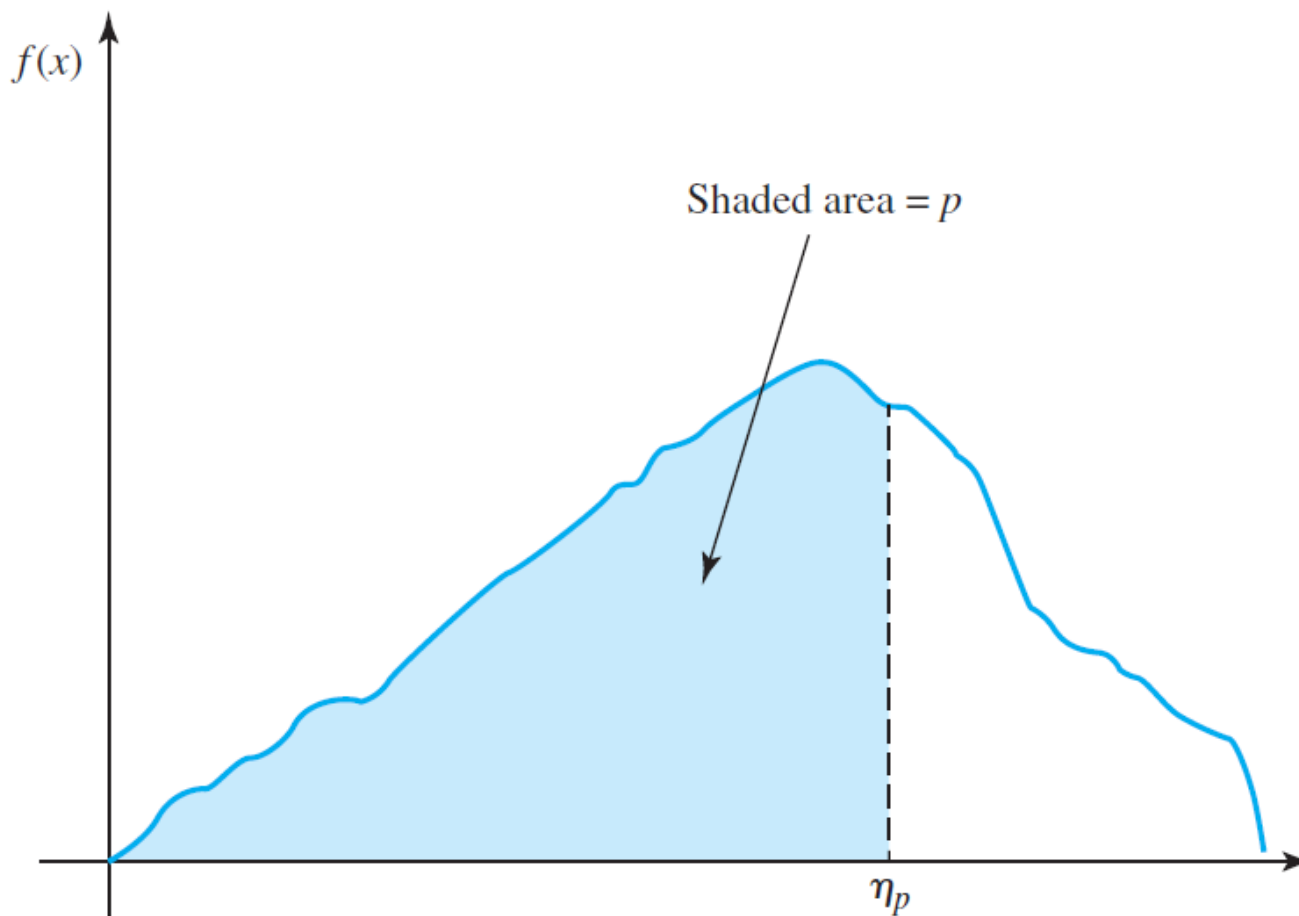
描述樣本資料離散程度的統計量，除了變異數和標準差，主要還有四分位距  
(quartile deviation or inter-quartile range)。

$$IQR = Q_{75\%} - Q_{25\%}$$

$$\int_{-\infty}^{\eta_p} f(x) dx = p$$

**Percentiles (百分位數)**

$p$  is the area under the density curve to the left of  $\eta_p$ .

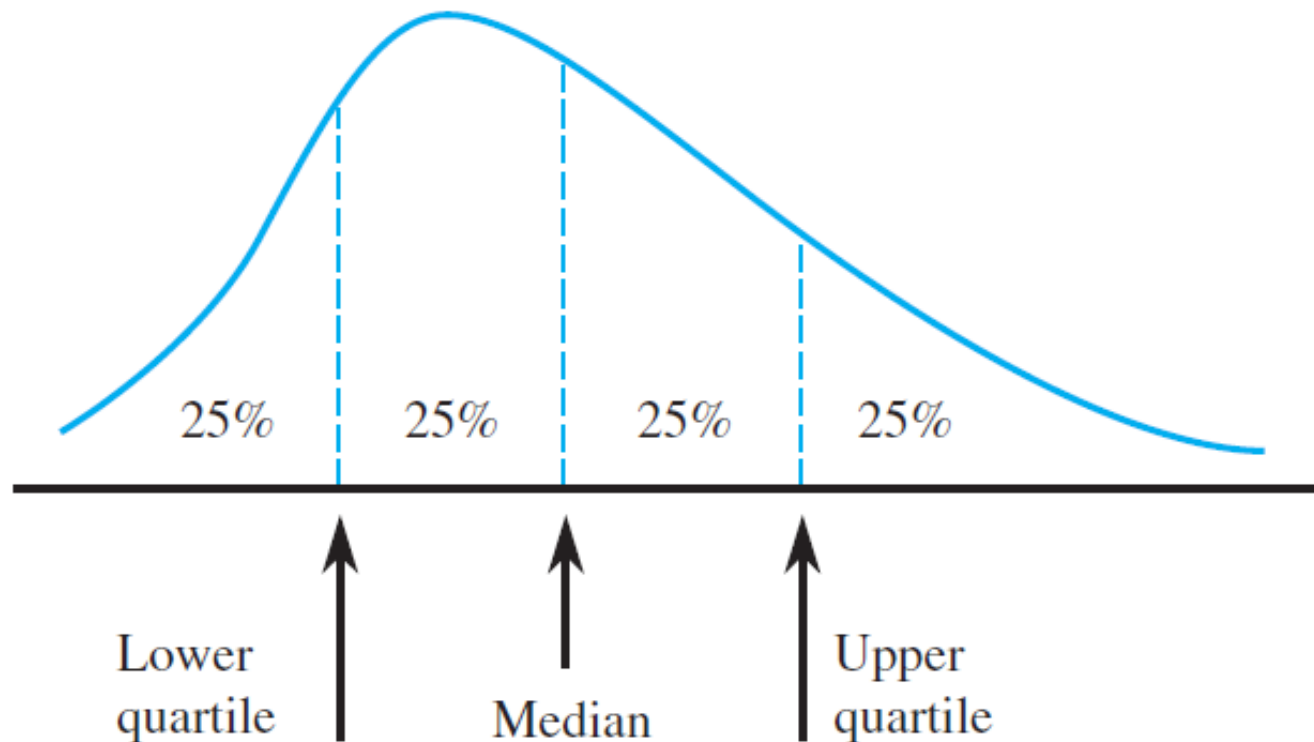


The median and IQR can be used together to give a concise yet informative visual summary of sample data called a

✓ **boxplot (盒鬚圖)**

# More Detailed Summary Quantities

- Quartiles and the Interquartile Range (IQR)
- Boxplots
- Boxplots That Show Outliers
- Percentiles



# Quartiles & the IQR: Example

- A stem-and-leaf display of the 27 observations follows:

[illegible]

- Because  $n = 27$  is odd, the median  $\tilde{x} = 7.7$  is included in each half of the data:

Lower half: 5.9 6.3 6.3 6.5 6.8 6.8 7.0 7.0 7.2 7.3 7.4 7.6 7.7 7.7

Upper half: 7.7 7.8 7.8 7.9 8.1 8.2 8.7 9.0 9.7 9.7 10.7 11.3 11.6 11.8

$$\text{lower quartile} = \frac{7.0 + 7.0}{2} = 7.0 \quad \text{upper quartile} = \frac{8.7 + 9.0}{2} = 8.85$$

$$\text{IQR} = 8.85 - 7.0 = 1.85$$

✓ 數列最大及最小值若改變將不影響IQR，  
但卻會影響變異數及標準差



## Definition:

- A continuous variable  $x$  is said to have a **normal distribution with parameters  $\mu$  and  $\sigma$** , where  $-\infty < \mu < \infty$  and  $\sigma > 0$ , if the density function of  $x$  is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty$$

- Many population and process variables have distributions that can be closely fit by a normal curve

# Quartiles & the IQR: normal distribution

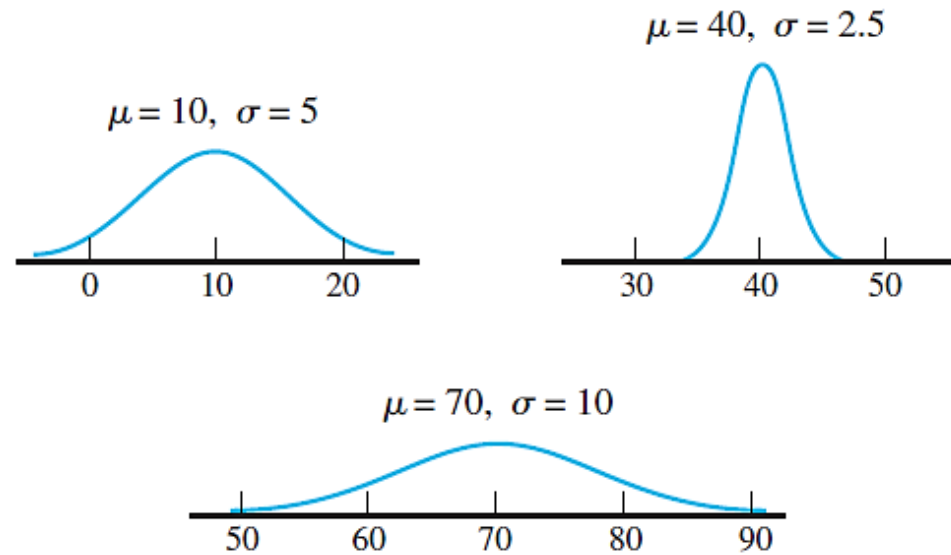


圖 1.19 數種常態密度曲線。

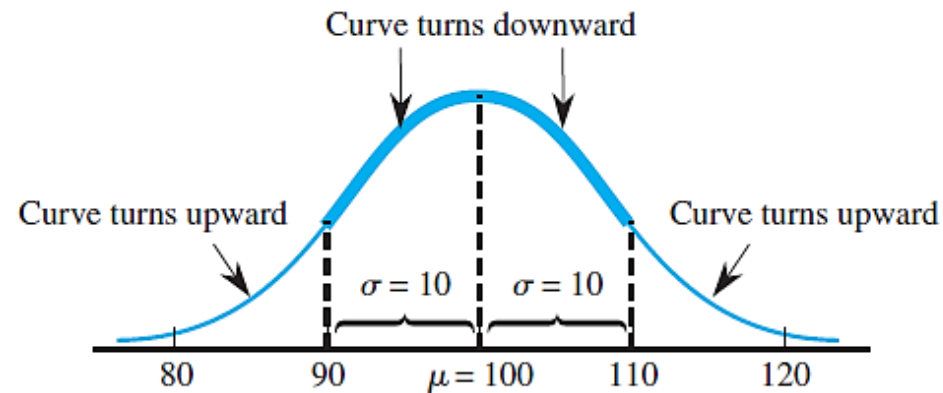


圖 1.20  $\mu$  跟  $\sigma$  的視覺化識別。

# Quartiles & the IQR: Example

常態分佈尋找下四分位數(25%)及上四分位數(75%)

## Example 2.12

The quartiles of a normal distribution are easily expressed in terms of  $\mu$  and  $\sigma$ . First, consider a variable  $z$  having the standard normal distribution. Symmetry of the standard normal curve about 0 implies that  $\tilde{\mu} = 0$ . Looking for .2500 inside Appendix Table I, we obtain the following information:

area to the left of  $-.67$ : .2514

area to the left of  $-.68$ : .2483

Since .25 is roughly halfway between these two tabled areas, we take  $-.675$  as the lower quartile. By symmetry,  $.675$  is the upper quartile.

It is then easily verified that if  $x$  has a normal distribution with mean value  $\mu$  and standard deviation  $\sigma$ ,

$$\text{upper quartile} = \mu + .675\sigma \quad \text{lower quartile} = \mu - .675\sigma$$

我們可以發現常態分佈的IQR會等於**1.35倍的標準差**。  
換句話說，可檢驗樣本資料的IQR值是否等於1.35倍樣本標準差來看資料是否符合常態分佈。

# Quartiles & the IQR: Example

常態分佈尋找下四分位數(25%)及上四分位數(75%)

## Example 2.12

The quartiles of a normal distribution are easily expressed in terms of  $\mu$  and  $\sigma$ . First, consider a variable  $z$  having the standard normal distribution. Symmetry of the standard normal curve about 0 implies that  $\tilde{\mu} = 0$ . Looking for .2500 inside Appendix Table I, we obtain the following information:

area to the left of  $-.67$ : .2514

area to the left of  $-.68$ : .2483

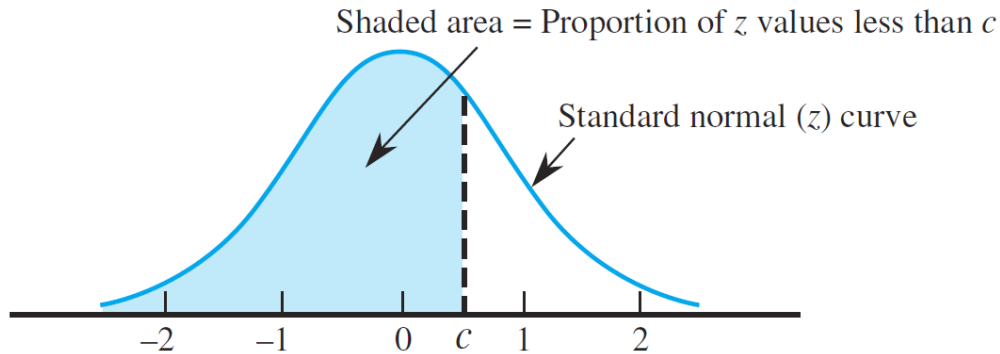
Since .25 is roughly halfway between these two tabled areas, we take  $-.675$  as the lower quartile. By symmetry,  $.675$  is the upper quartile.

It is then easily verified that if  $x$  has a normal distribution with mean value  $\mu$  and standard deviation  $\sigma$ ,

$$\text{upper quartile} = \mu + .675\sigma \quad \text{lower quartile} = \mu - .675\sigma$$

我們可以發現常態分佈的IQR會等於**1.35倍的標準差**。  
換句話說，可檢驗樣本資料的IQR值是否等於1.35倍樣本標準差來看資料是否符合常態分佈。

# Quartiles & the IQR: Example

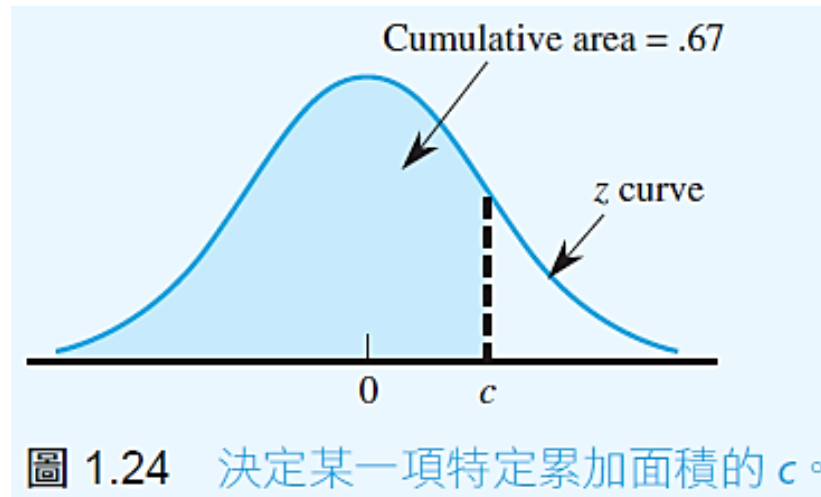


**`pnorm`**( $c$ , mean =, sd =)

Here:

mean = 算數平均數

sd = 標準差



**`qnorm`**( $p$ , mean =, sd =)

Here:

$p = 0.67$

mean = 算數平均數

sd = 標準差

圖 1.24 決定某一項特定累加面積的  $c$ 。

# R: Descriptive data:

## IQR in normal distribution

**pnorm().**

**qnorm().**

TRY  
it  
in  
R

# R: Descriptive data: IQR in normal distribution

R\_descriptive\_d.R



- A **boxplot** is a visual display of data based on the following **five-number** summary:
  - smallest  $x_i$
  - lower quartile
  - median
  - upper quartile
  - largest  $x_i$

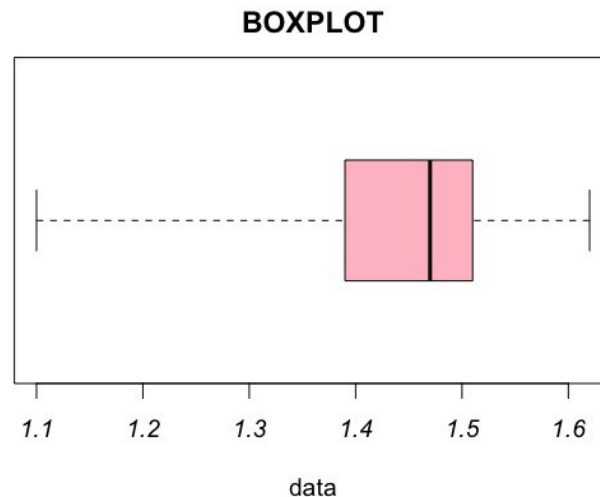
# Boxplot: Example

## 材料樣品之測量比重值

1.10	1.29	1.38	1.39	1.40	1.45	1.46
1.48	1.49	1.50	1.51	1.51	1.56	1.62

## The five-number summary

Smallest  $x_i = 1.10$    lower quartile = 1.39    $\tilde{x} = 1.47$    upper quartile = 1.51  
 Largest  $x_i = 1.62$



- A boxplot can be embellished to indicate explicitly the presence of outliers.
- Any observation farther than **1.5 IQR** from the closest quartile is an **outlier**.
- An outlier is **extreme** if it is more than **3 IQR** from the nearest quartile.

Thus, any observation smaller than  $48 - 34.5 = 13.5$  or larger than  $71 + 34.5 = 105.5$  is an outlier. There is one outlier at the lower end of the sample and two at the upper end. Because  $71 + 69 = 140$ , the largest observation of 144 is an extreme outlier; the other outlier is mild. The whiskers extend out to 32 and 76, the most extreme observations that are not outliers. The resulting boxplot is in Figure 2.11.

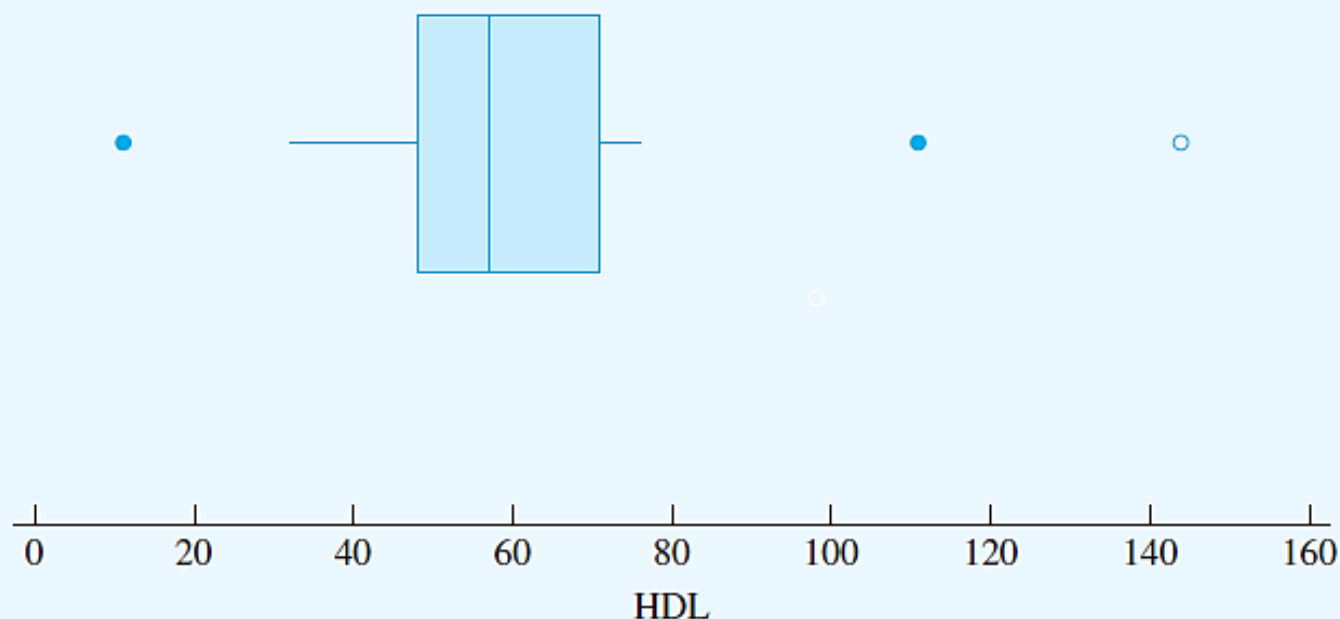


圖 2.11 HDL 膽固醇數據的盒型圖顯示有中度與極度離群值。

R: Descriptive data:  
display data distribution  
**geom\_boxplot(  
coef =).**

TRY  
it  
in  
R

# R: Descriptive data: display data distribution

R\_descriptive\_e.R

- Bar chart (長條圖): often used to describe histogram for categorical data
- Pareto (柏拉圖) diagram: a bar chart where categories appear in order of decreasing frequency; if a **miscellaneous** category is required, it is placed last

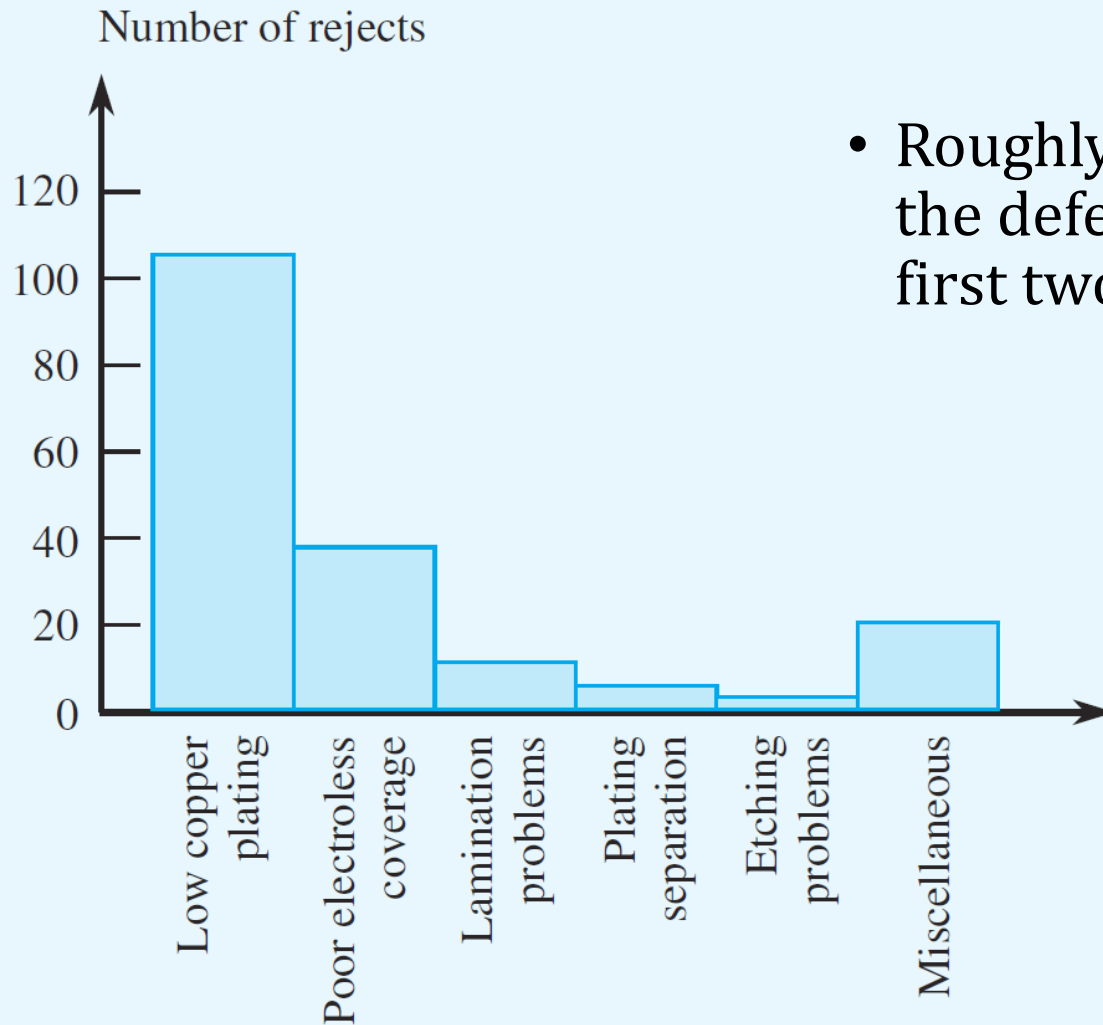


# Categorical Data: An Example

- Example: Manufacture of printed circuit boards
- Type of defect for each board rejected:  
(產品的製成缺陷分類項目，包覆塗料的缺陷、電鍍不完整、層狀問題、塗料剝落、侵蝕剝落)

Type of defect	Frequency	Relative frequency
Low copper plating	112	.615
Poor electroless coverage	35	.192
Lamination problems	10	.055
Plating separation	8	.044
Etching problems	5	.027
Miscellaneous	12	.066

# Categorical Data: An Example



- Roughly 80% ( $.615 \times 1 + .192$ ) of the defects were of one of the first two types

R: Descriptive data:  
quick to get more details

Package:  
psych

R: Descriptive data:  
quick to get more details  
**describe()**.

# 課堂練習: 學號-姓名-ch5-Descriptive.R

25個圓柱狀試體的彈性模數大小數據:

37.0 37.5 38.1 40.0 40.2 40.8 41.0 42.0 43.1 43.9  
44.1 44.6 45.0 46.1 47.0 50.2 55.0 56.0 57.0 58.0  
62.0 64.3 68.8 70.1 74.5

單位: MPa

試著回答以下問題:

(1) Construct a histogram with the density curve.

(hist (breaks = 5))

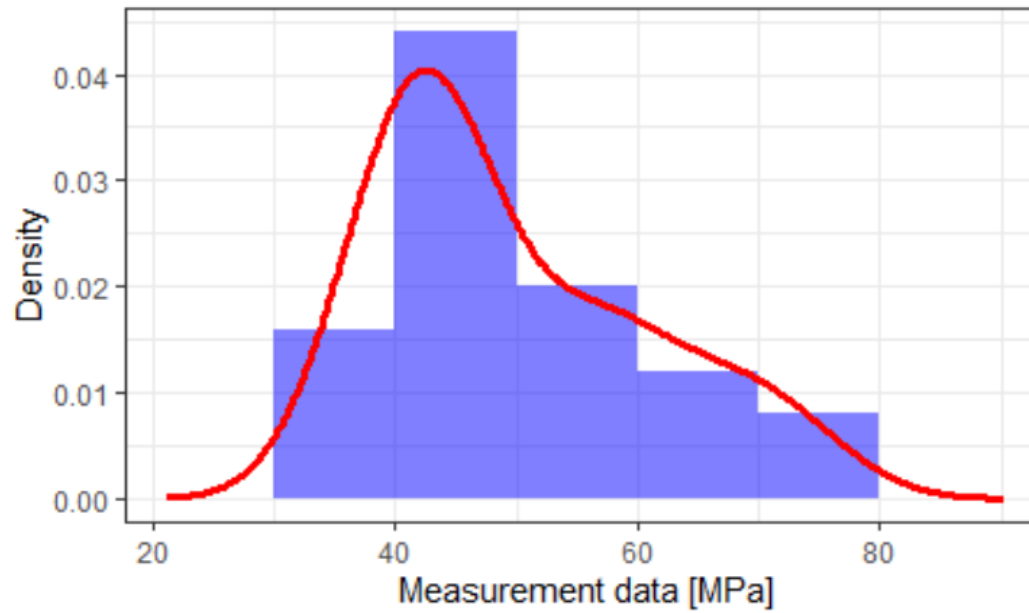
(2) What are the skewness and kurtosis values? Please make a short conclusion.

(3) Construct a boxplot display of the data. Does there appear to be any outlying values? (coef = 1.0)

(去除y軸座標指令: scale\_y\_discrete())



Histogram and Density curve of data



Boxplot of data

