Describing Data:
Displaying and Exploring Data

Chapter 4

Dot Plots

- A **dot plot** groups the data as little as possible and the identity of an individual observation is not lost.
- To develop a dot plot, each observation is simply displayed as a dot along a horizontal number line indicating the possible values of the data.
- If there are identical observations or the observations are too close to be shown individually, the dots are "piled" on top of each other.

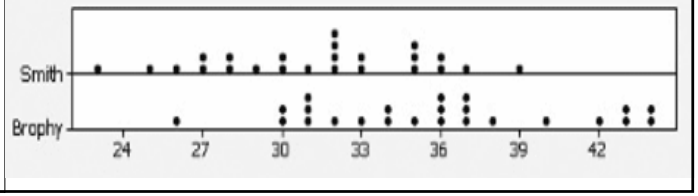
Dot Plot – Minitab Example

Smith Ford Mercury Jeep, Inc.

23	27	30	27	32	31	32	32	35	33
28	39	32	29	35	36	33	25	35	37
26	28	36	30						

Brophy Honda Volkswagen

31	44	30	36	37	34	43	38	37	35
36	34	31	32	40	36	31	44	26	30
37	43	42	33						



Dot Plot vs. Histogram

- The disadvantage of a histogram is that it loses the exact value of the observations.
- Dot Plots do not lose the identity of an individual.
- Dot plots are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets.

Stem-and-Leaf

- In Chapter 2, we showed how to organize data into a frequency distribution. The major advantage to organizing the data into a frequency distribution is that we get a quick visual picture of the shape of the distribution.
- A frequency distribution has two disadvantages :
  - Lose the exact identity of each value
  - How the values within each class are distributed is unknown.

5

Stem-and-Leaf

- One technique that is used to display quantitative information in a condensed form is the **stem-and-leaf display.**
- **Stem-and-leaf display** is a statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.
- Advantage of the stem-and-leaf display over a frequency distribution - the identity of each observation is not lost.

6

Stem-and-Leaf : Example

Suppose the seven observations in the 90 up to 100 class are: 96, 94, 93, 94, 95, 96, and 97.

The **stem** value is the leading digit or digits, in this case 9. The **leaves** are the trailing digits. The stem is placed to the left of a vertical line and the leaf values to the right. The values in the 90 up to 100 class would appear as:

9	6 4 3 4 5 6 7

Then, we sort the values within each stem from smallest to largest. Thus, the second row of the stem-and-leaf display would appear as follows:

9	3 4 4 5 6 6 7

7

Stem-and-leaf: Example

TABLE 4–1 Number of Advertising Spots Purchased by Members of the Greater Buffalo Automobile Dealers Association

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			

Stem	Leaf
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

8

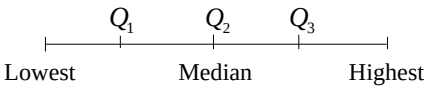
Other Measure of Dispersion

- The standard deviation is the most widely used measure of dispersion.
- Alternative ways of describing spread of data include determining the *location* of values that divide a set of observations into equal parts.
  - Quartile（四分位數）
  - Decile（十分位數）
  - Percentile（百分位數）

9

Quartiles（四分位數）

- Quartile: Divide a data set into 4 equal parts.

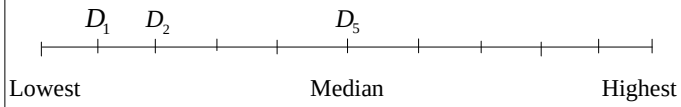


  - Q_1 denotes the 1st quartile below which 25% of the observations occurs.
  - Q_2 denotes the 2nd quartile below which 50% of the observations occurs. Q_2=Median.
  - Q_3 denotes the 3st quartile below which 75% of the observations occurs.

10

Deciles

- Decile: Divide a data set into 10 equal parts.

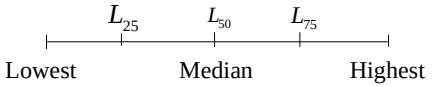


- D_1 denotes the 1st decile below which 10% of the observations occurs.

11

Percentiles

- Percentile: Divide a data set into 100 equal parts.



- L_n denotes the n-th percentile below which n% of the observations occurs.

12

Relation

- Median = Q2 = D5 = L50
- Q1 = L25
- Q3 = L75

13

Computation

- Let L_p refer to the location of the n-th percentile.
  - L_{33} = the 33rd percentile
  - L_{50} = the 50th percentile (Median)
  - L_{10} = the 10th percentile (1st Decile)
- LOCATION OF A PERCENTILE $L_p = (n+1)\frac{P}{100}$ [4-1]
  - n = the number of observations
  - *p* is the desired percentile

14

Percentiles - Example

Listed below are the commissions earned last month by a sample of 15 brokers at Salomon Smith Barney's Oakland, California, office. Salomon Smith Barney is an investment company with offices located throughout the United States.

$2,038	$1,758	$1,721	$1,637
$2,097	$2,047	$2,205	$1,787
$2,287	$1,940	$2,311	$2,054
$2,406	$1,471	$1,460	

Locate the first quartile and the third quartile for the commissions earned.

15

Percentiles – Example (cont.)

Step 1: Organize the data from lowest to largest value

$1,460	$1,471	$1,637	$1,721
$1,758	$1,787	$1,940	$2,038
$2,047	$2,054	$2,097	$2,205
$2,287	$2,311	$2,406	

16

Percentiles – Example (cont.)

Step 2: Compute the first and third quartiles. Locate L_{25} and L_{75} using:

LOCATION OF A PERCENTILE $L_p = (n+1)\frac{P}{100}$ [4–1]

$$L_{25} = (15+1)\frac{25}{100} = 4 \qquad L_{75} = (15+1)\frac{75}{100} = 12$$

Therefore, the first and third quartiles are the 4th and 12th observation in the array, respectively

$L_{25} = \$1{,}721$

$L_{75} = \$2{,}205$

17

Percentiles – Another Example

- Data Set: {1,3,5,7,9,11,13,15,17,19 }, N=10.
- L25=(10+1) ×(25/100)=2.75
  Q1=3+(5-3) ×0.75 = 4.5
- L75 =(10+1) ×(75/100)=8.25
  Q3=15+(17-15) ×(25/100)=15.5
- The median, quartile, decile and percentile do not need to be one of the actual values in the data set.

18

Box Plot

- To construct box plot, we need five statistics:
  - the first quartile Q1
  - the third quartile Q3
  - median
  - Largest element
  - Lowest element

19

Boxplot - Example

Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

Q_1 = 15 minutes

Median = 18 minutes

Q_3 = 22 minutes

Maximum value = 30 minutes

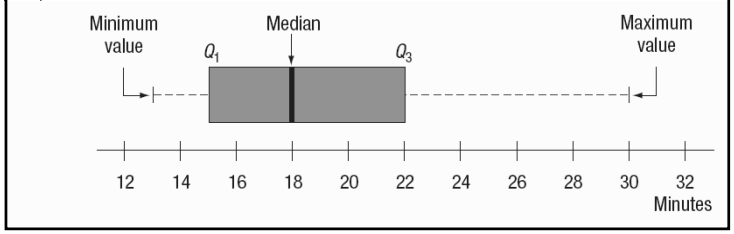
Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

20

Boxplot Example

The box plot reveals that the distribution of delivery times is positively skewed. The reasons are:

- The right dashed line is longer than the left dashed line.
- The median is not in the center of the box.



Moment (動差)

動差（Moment）:以原點爲中心

母體第 r 階動差: $\mathrm{M}_r = \sum_{i=1}^{N} (X_i)^r / N$

樣本第 r 階動差： $m_r = \sum_{i=1}^{n} (X_i)^r / n$

中央動差（Central Moment）:以平均數爲中心

母體第 r 階動差: $\mathrm{M}_r^c = \sum_{i=1}^{N} (X_i - \mu)^r / N$

樣本第 r 階動差： $m_r^c = \sum_{i=1}^{n} (X_i - \bar{X})^r / n$

22

Moment (動差)

- 各級動差用以描述資料不同的特性：
  - 一級動差可以用來描述資料平均數。
  - 二級動差可以用來描述資料變異程度。
  - 三級動差可用來描述資料的偏態程度。
  - 四級動差可以用來描述資料的峰度。
- $\mu = \mathrm{M}_1 = \sum_{i=1}^{N} (X_i)^1 / N$

  $\sigma^2 = \mathrm{M}_2^c = \sum_{i=1}^{N} (X_i - \mu)^2 / N = M_2 - (\mathrm{M}_1)^2$

23

描述資料的四種特徵

- 一般用來描述的資料特性，用四個數值來衡量：
  - 平均數 (Mean, μ)：描述資料的中點
  - 標準差 (Standard Deviation, σ)：描述資料偏離平均值的情形
  - 偏態係數 (Skewness, SK)：描述資料在平均數兩邊分佈的情形
  - 峰態係數 (Kurtosis, K)：描述資料偏離平均數的速度

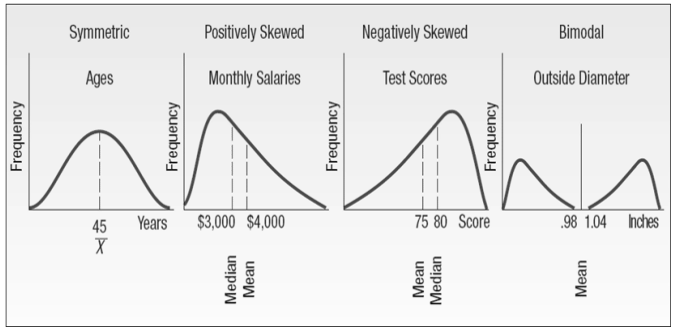
24

Skewness

- In Chapter 3, measures of central location for a set of observations (the mean, median, and mode) and measures of data dispersion (e.g. range and the standard deviation) were introduced.
- Another characteristic of a set of data is the shape.
- There are four shapes commonly observed:
  - symmetric,
  - positively skewed,
  - negatively skewed,
  - bimodal.

25

Commonly Observed Shapes



26

Skewness - Formulas for Computing

The coefficient of skewness can range from -3 up to 3.

- A value near -3, such as -2.57, indicates considerable negative skewness.
- A value such as 1.63 indicates moderate positive skewness.
- A value of 0, which will occur when the mean and median are equal, indicates the distribution is symmetrical and that there is no skewness present.

PEARSON'S COEFFICIENT OF SKEWNESS $$sk = \frac{3(\bar{X} - \text{Median})}{s}$$ [4–2]

SOFTWARE COEFFICIENT OF SKEWNESS $$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X - \bar{X}}{s}\right)^3\right]$$ [4–3]

27

偏態係數 SK

- 偏態係數可用來判斷資料在平均數兩邊分配的情形的
  - SK = 0 表示資料以平均數為中點對稱
  - SK > 0 表示資料右（正）偏，資料在平均數右方較多
  - SK < 0表示資料左（負）偏，資料在平均數左方較多
- |SK|越大，表示資料越偏向一方。
- 0≦|SK|≦0.5，表示資料幾乎對稱於平均數。
  0.5<|SK|<1，表示資料微偏向一方。
  |SK|>1，表示資料極偏向一方。

28

Skewness – An Example

- Following are the earnings per share for a sample of 15 software companies for the year 2005. The earnings per share are arranged from smallest to largest.

$0.09	$0.13	$0.41	$0.51	$ 1.12	$ 1.20	$ 1.49	$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

- Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson's estimate. What is your conclusion regarding the shape of the distribution?

29

Skewness – An Example

$$\overline{X} = \frac{\sum X}{n} = \frac{\$74.26}{15} = \$4.95$$

$$s = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + ... + (\$16.40 - \$4.95)^2)}{15-1}} = \$5.22$$

$$Pearson's\ sk = \frac{3(\overline{X} - Median)}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

$$Software's\ sk = \frac{n}{(n-1)(n-2)}\sum\left(\frac{X - \overline{X}}{s}\right)^3 = \frac{15}{(15-1)(15-2)}(11.8274) = 0.975$$

We conclude that the earnings per share values are somewhat positively skewed.

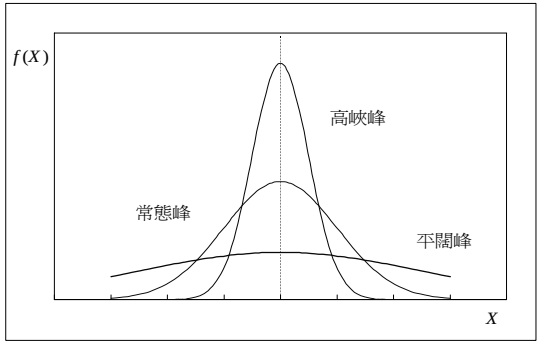
30

Kurtosis（峰態）

- 當分配有集中趨勢時，就會有『峰』出現。峰的型態，視資料集中於平均數或眾數的程度而分。
  - 常態峰 (Meso Kurtosis)：表示資料分佈呈現一般型態。
  - 高狹峰 (Lepto Kurtosis)：表示資料集中於眾數或平均數附近。
  - 平闊峰 (Platy Kurtosis): 表示資料較分散。

31

三種峰態的圖形



32

Describing Relationship between Two Variables

- Data
  - Univariate Data
  - Bivariate Data
  - Multivariate Data
- Introduce two techniques to portray the relationship between two variables.
  - Scatter Diagram
  - Contingency Table

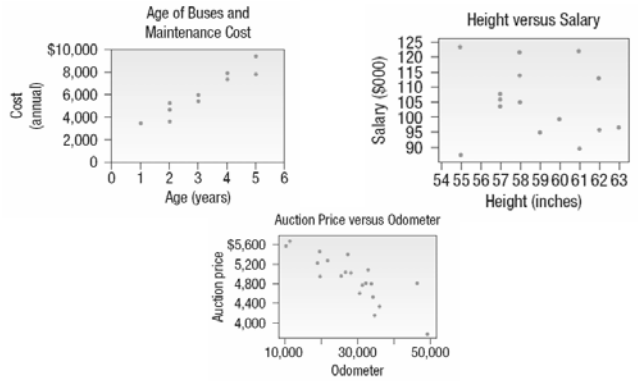
33

Describing Relationship between Two Variables

- One graphical technique we use to show the relationship between variables is called a **scatter diagram.**
- To draw a scatter diagram we need two variables. We scale one variable along the horizontal axis (*X*-axis) of a graph and the other variable along the vertical axis (*Y*-axis).

34

Describing Relationship between Two Variables – Scatter Diagram Examples



35

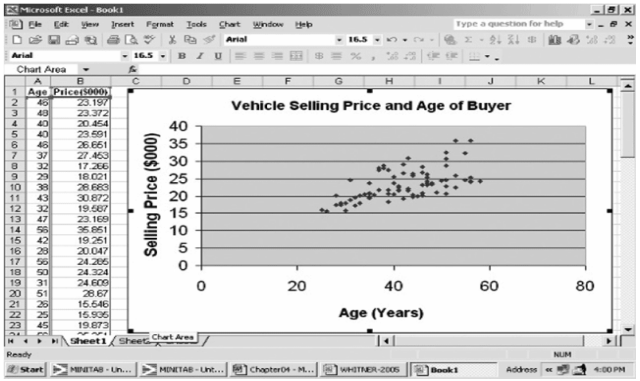
Describing Relationship between Two Variables – Scatter Diagram Excel Example

In the Introduction to Chapter 2 we presented data from AutoUSA. In this case the information concerned the prices of 80 vehicles sold last month at the Whitner Autoplex lot in Raytown, Missouri. The data shown include the selling price of the vehicle as well as the age of the purchaser.

Is there a relationship between the selling price of a vehicle and the age of the purchaser? Would it be reasonable to conclude that the more expensive vehicles are purchased by older buyers?

36

Describing Relationship between Two Variables – Scatter Diagram Excel Example



37

Contingency Tables

- A scatter diagram requires that both of the variables be at least interval scale.
- What if we wish to study the relationship between two variables when one or both are nominal or ordinal scale? In this case we tally the results in a **contingency table.**

CONTINGENCY TABLE A table used to classify observations according to two identifiable characteristics.

38

Contingency Tables – An Example

A manufacturer of preassembled windows produced 50 windows yesterday. This morning the quality assurance inspector reviewed each window for all quality aspects. Each was classified as acceptable or unacceptable and by the shift on which it was produced. Thus we reported two variables on a single item. The two variables are shift and quality. The results are reported in the following table.

	Shift			
	Day	Afternoon	Night	Total
Defective	3	2	1	6
Acceptable	17	13	14	44
Total	20	15	15	50

39

變異係數 (Coefficient of Variation)

- Range, Standard Deviation, Mean Deviation 只能用來衡量個別資料的絕對離散程度。
- 若欲比較兩組不同資料的離散程度時，無法直接以 Standard Deviation 比較，因為會受到平均數大小與單位不同而影響。
  - 在投資學，我們通常以標準差來描述股價的風險。在比較兩張不同股票的風險時，必須用變異係數來比較，因為高價股的變異數一般較低價股來的大，但不表示其風險較大。
  - 某班的身高體重資料，想比較身高資料與體重資料的離散程度大小。
- 比較兩組資料的離散程度，必須用變異係數來比較。

40

變異係數 (Coefficient of Variation)

- 變異係數 = 標準差 / 平均數
  - 母體資料： CV = σ / μ
  - 樣本資料： CV = S / $\bar{X}$
- 變異係數是將 標準差除以平均數，因此可將單位與平均數的因素去除，使兩筆資料可以比較離散程度。
- CV 越大，表示離散程度越高。

41

Exercises

- 3,7,11,13,15,19,23,25,27,29,33,41,43,

42

End of Chapter 4

43