統計學

Fall 2025

政治大學統計系余清祥

2025年9月9日

第二章：敘述統計(圖表)

http://csyue.nccu.edu.tw

Chapter Contents

Introduction

This chapter introduces tabular and graphical displays for summarizing both categorical and quantitative data.

We begin with a discussion of tabular and graphical displays to summarize the data for a single variable.

This is followed by a discussion of the use of tabular and graphical displays to summarize the data for two variables to reveal the relationship between the two variables.

Data visualization is a term often used to describe graphical displays to summarize and present information about a data set.

The last section of this chapter introduces data visualization and provides guidelines for creating effective graphical displays.

Statistical software packages provide extensive capabilities for summarizing data and preparing visual presentations.

1. **Descriptive**. Traditional HR metrics are largely efficiency metrics (turnover rate, time to fill, cost of hire, number hired and trained, etc.). The primary focus here is on cost reduction and process improvement. Descriptive HR analytics reveal and describe *relationships* and *current and historical data patterns*. This is the foundation of your analytics effort. It includes, for example, dashboards and scorecards; workforce segmentation; data mining for basic patterns; and periodic reports.
2. **Predictive**. Predictive analysis covers a variety of techniques (statistics, modeling, data mining) that use current and historical facts to make predictions about the future. It's about probabilities and potential impact. It involves, for example, models used for increasing the probability of selecting the right people to hire, train, and promote.
3. **Prescriptive**. Prescriptive analytics goes beyond predictions and outlines decision options and workforce optimization. It is used to analyze complex data to predict outcomes, provide decision options, and show alternative business impacts. It involves, for example, models used for understanding how alternative learning investments impact the bottom line (rare in HR).

統計的分析觀點

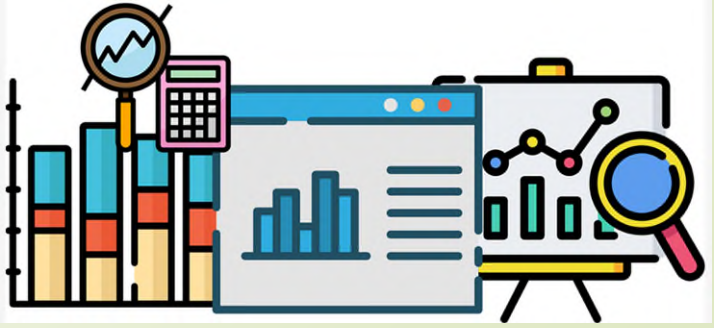
根據統計觀點，分析有以下兩類：

- 探索性資料分析(Exploratory Data Analysis)

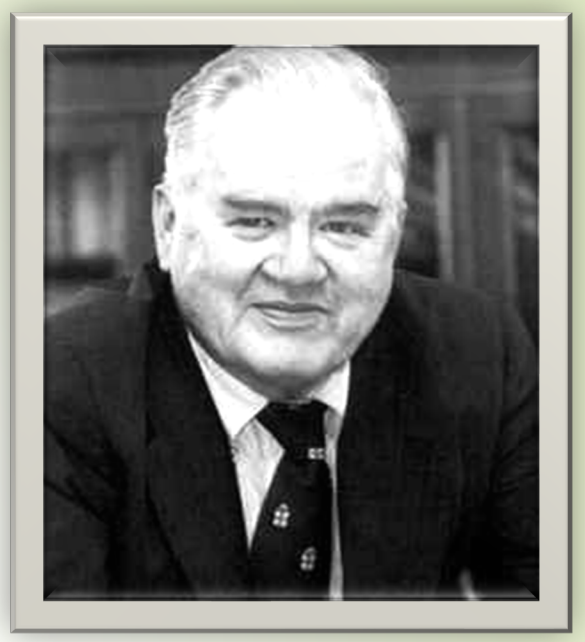
→The role of **EDA** is to figure out the essence of data and to develop research hypothesis,

- 驗證性資料分析(Confirmatory Data Analysis)

→While the role of **CDA** is to examine evidence and test hypothesis & build models.



https://towardsai.net/p/l/overview-of-exploratory-data-analysis-eda-with-haberman-dataset



EDA：讓資料說話

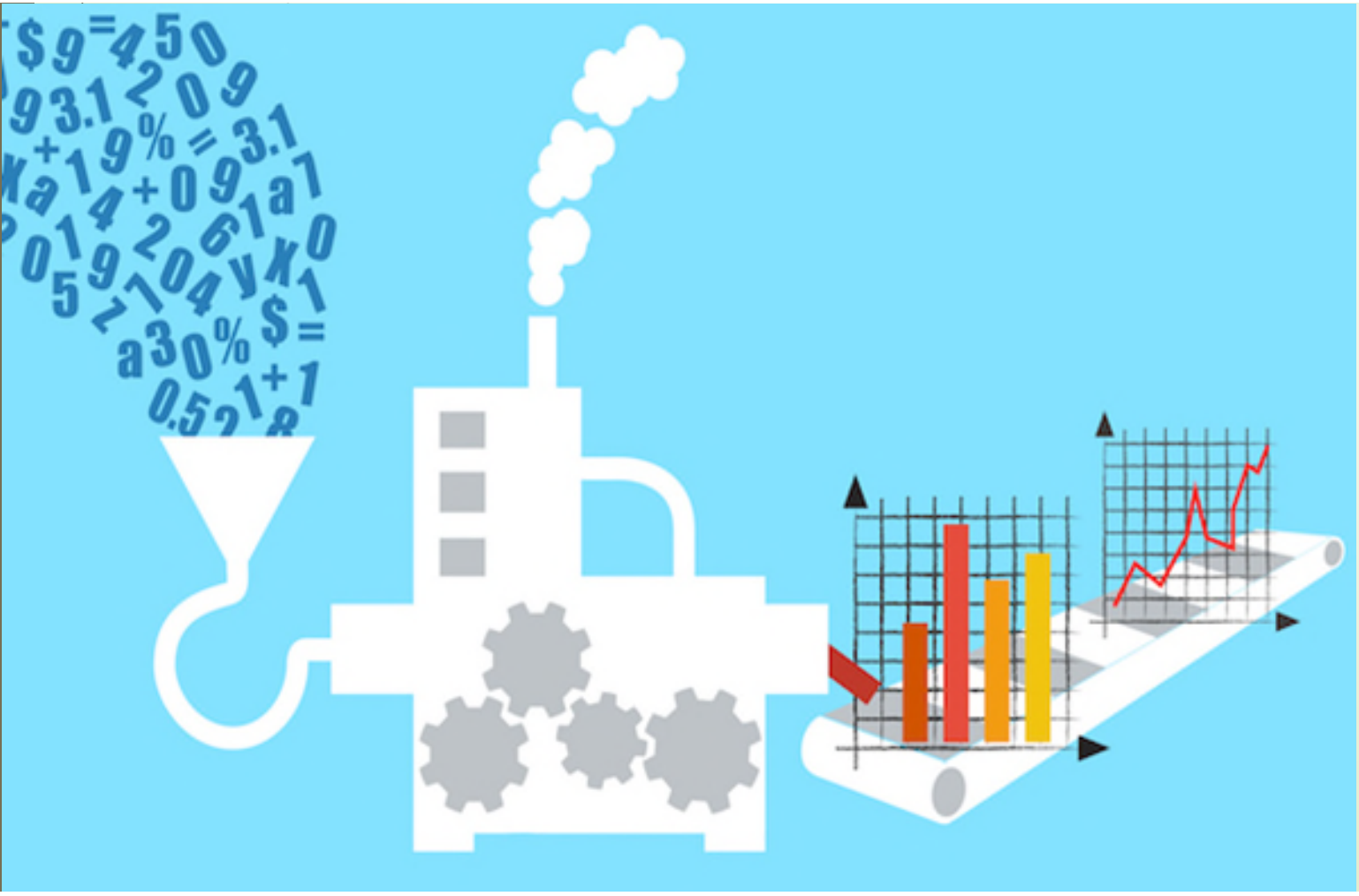
- 資料驅動(Data Driven)

→Tukey於1970年代提出EDA，他認為 *"more emphasis needed to be placed on using data to construct research hypotheses"*

→EDA is not a mere collection of techniques. EDA is a philosophy as to **how we dissect a data set**; **what we look for**; **how we look**; and **how we interpret**.

探索性資料分析(資料驅動)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics … EDA is for seeing what the data can tell us beyond the formal modeling. ---Wikipedia



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.aiche.org%2Facademy%2Fwebinars%2Fapplied-statistics-exploratory-data-analysis&psig=AOvVaw36ZuxAJqz27dLqU5IFzBMO&ust=1570108849384000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJC1qLXV_eQCFQAAAAAdAAAAABAJ

圖形有時描述地更傳神！

A Picture

Source: http://www.rapides.k12.la.us/nitro/visual_literacy.htm

is worth

=

Creamy, delicious, yummy, fudge ice cream, smooth, chocolate-chip mint ice cream, strawberry ice cream with real chunks of strawberry, colored sugar sprinkles, waffle sugar cone, sweet, wonderful, tastes great, cold, nice to eat, dessert, good yummy toppings, chocolate sprinkles, comforting, good fun, dripping, terrific,

A thousand words.

©2003 E. Aoyama

Source: http://rbonini.files.wordpress.com/2007/06/200706290853.jpg

2.1 Frequency Distribution

A **frequency distribution** is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.

Example DATAfile: *SoftDrink*

To develop a frequency distribution for the sample of 50 soft drink purchases, we count the number of times each soft drink appears.

The frequency distribution highlights Coca-Cola as the leader, followed by Pepsi, Diet Coke, Dr. Pepper, and Sprite.

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

2.1 Relative Frequency and Percent Frequency Distributions

The **relative frequency** of a class is the fraction or proportion of the total number of data items belonging to the class.

$$\text{Relative Frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

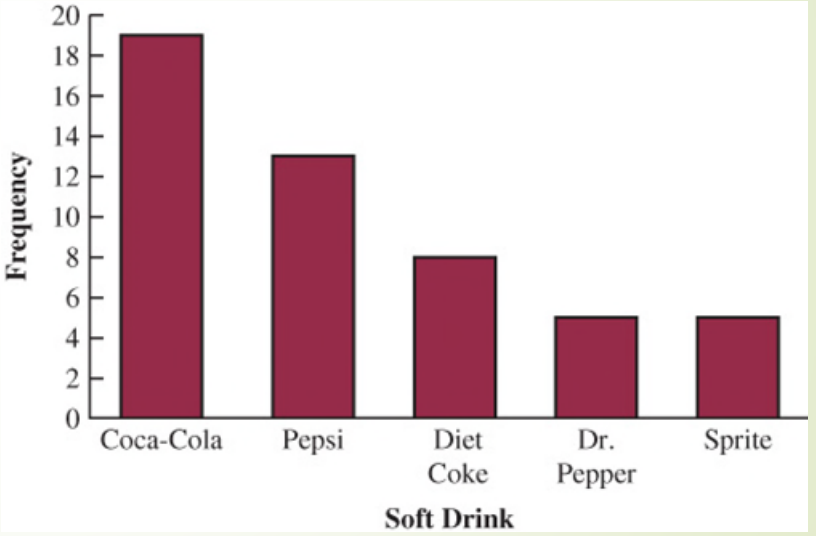
- The **percent frequency** of a class is the relative frequency multiplied by 100.
- The relative frequency distribution for the soft drink data shows a relative frequency of $19/50 = 0.38$ for Coca-Cola, $13/50 = 0.26$ for Pepsi, and so on.
- The percent frequency distribution, shows 38% Coca-Cola purchases, 16% Diet Coke purchases, and so on.

Soft Drink	Relative Frequency	Percent Frequency
Coca-Cola	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

2.1 Bar Chart

A **bar chart** is a graphical display for depicting categorical data summarized in a frequency, relative frequency, or percent frequency distribution.

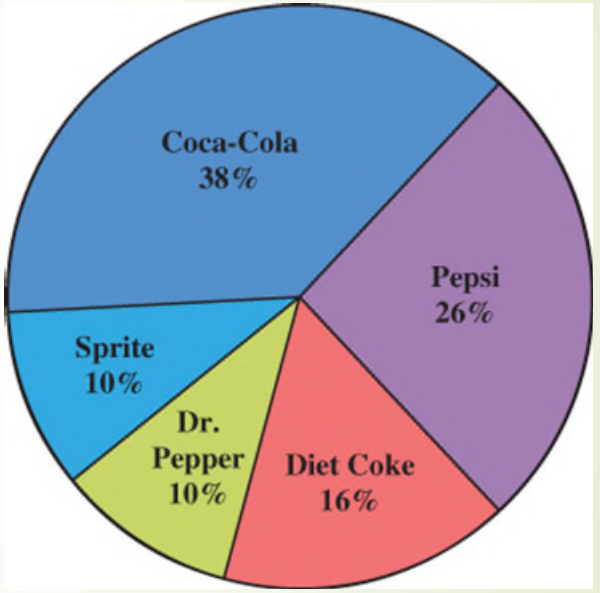
- On one axis (usually the horizontal axis), we specify the labels used for the classes (categories).
- A frequency, relative frequency, or percent frequency scale is used for the other axis (usually the vertical axis).
- Using a bar of fixed width, drawn above each class label, we extend the height appropriately.
- The bars are separated to emphasize the fact that each class is a separate category.
- In this example, the bars are also sorted.



2.1 Pie Chart

The **pie chart** provides another graphical display for presenting relative frequency and percent frequency distributions for categorical data

- First draw a circle.
- Then, use the percent frequencies to subdivide the circle into sectors that are proportional to the percent frequency for each class.
- Because there are 360 degrees in a circle, the sector of the pie chart labeled Coca-Cola consists of $0.38(360) = 136.8°$.
- Similar calculations for the other classes yield the pie chart shown.
- In general, pie charts are not the best way to present percentages for comparison.




Pie I have eaten
Pie I have not yet eaten

2.2 Frequency Distribution for a Quantitative Variable

To build a frequency distribution for quantitative data, we must be more careful in defining the nonoverlapping classes to be used in the frequency distribution.

The steps to define the classes for a frequency distribution with quantitative data are:

1. **Number of classes**: from 5, for small data sets, up to 20 for large data sets.
2. **Width of the classes**: $\text{Approx. class width} = \dfrac{\text{Largest Data Value} - \text{Smallest Data Value}}{\text{Number of classes}}$
3. **Class limits** are chosen so that classes do not overlap, and each data item belongs to only one class (*see notes.)

- **Example** DATAfile: *Audit*
- The data shows the time in days required to complete year-end audits for a sample of 20 clients of a small public accounting firm.
- Five classes were chosen, with class width = $(33 - 12)/5 = 4.2 \approx 5$. The resulting frequency distribution is shown to the right.

Audit Time (days)	Frequency
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
Total	20

2.2 Relative Frequency and Percent Frequency Distributions for a Quantitative Variable

We define the relative frequency and percent frequency distributions for quantitative data in the same manner as for categorical data

$$\text{Relative Frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

Remember that the percent frequency of a class is the relative frequency multiplied by 100.

- Based on the class frequencies for the 20 audit times, the table shows the relative frequency distribution and percent frequency distribution (*see notes.)
- Note that 40% of the audits required between 15 and 19 days, and only 5% of the audit required more 30 or more days.

Audit Time (days)	Relative Frequency	Percent Frequency
10-14	0.20	20
15-19	0.40	40
20-24	0.25	25
25-29	0.10	10
30-34	0.05	5
Total	1.00	100

2.2 Cumulative Distributions

A **cumulative frequency distribution** shows the number of items with values less than or equal to the upper limit of each class (*see notes.)

A **cumulative relative frequency distribution** shows the proportion of items with values less than or equal to the upper limit of each class.

A **cumulative percent frequency distribution** shows the percentage of items with values less than or equal to the upper limit of each class.

The distributions in the table show that 85% of the audits were completed in 24 days or less,

Audit Time (days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
Less than or equal to 14	4	0.20	20
Less than or equal to 19	12	0.60	60
Less than or equal to 24	17	0.85	85
Less than or equal to 29	19	0.95	95
Less than or equal to 34	20	1.00	100

2.2 Dot Plot

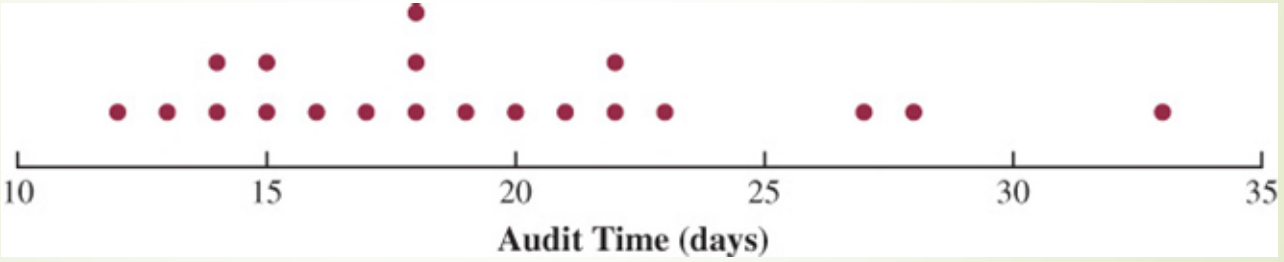
The **Dot Plot** is one of the simplest graphical summaries of quantitative data.

A horizontal axis shows the range for the data, and each data value is represented by a dot placed above the axis.

Dot plots show the details of the data and are useful for comparing the distribution of the data for two or more variables.

The dot plot for the audit time data is shown below.

- The three dots located above 18 on the horizontal axis indicate that an audit time of 18 days occurred three times.

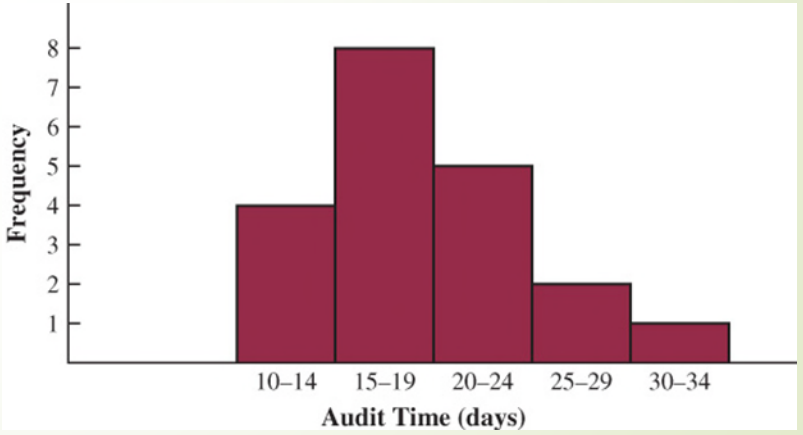


2.2 Histogram

The **histogram** is a common graphical display that can be prepared for quantitative data previously summarized in a frequency, relative frequency, or percent frequency distribution.

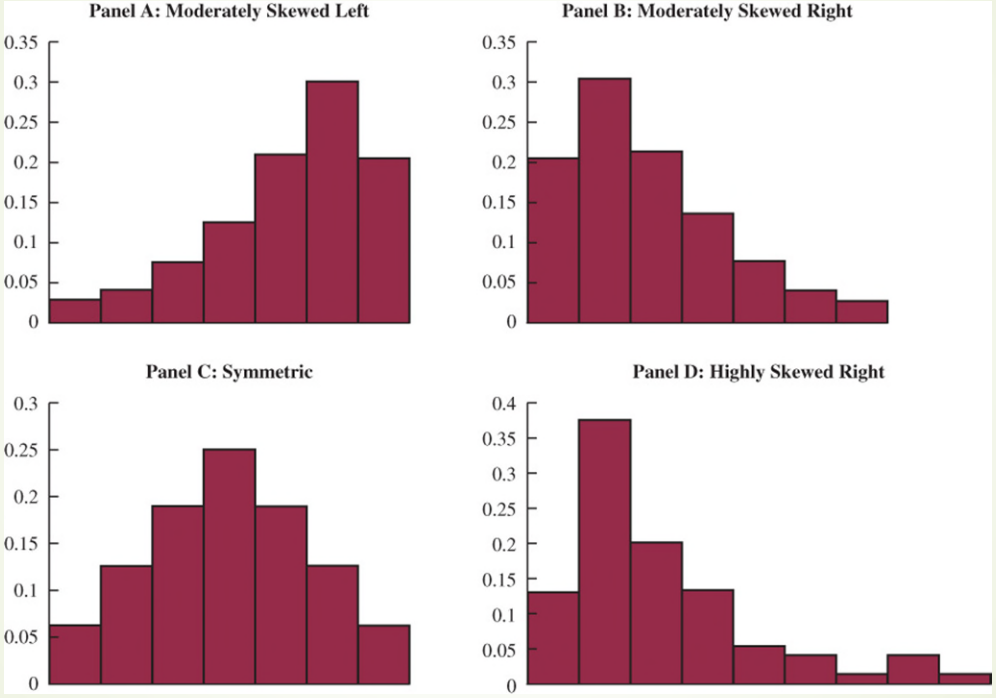
- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes (*see notes.)

The histogram for the audit time data is shown.



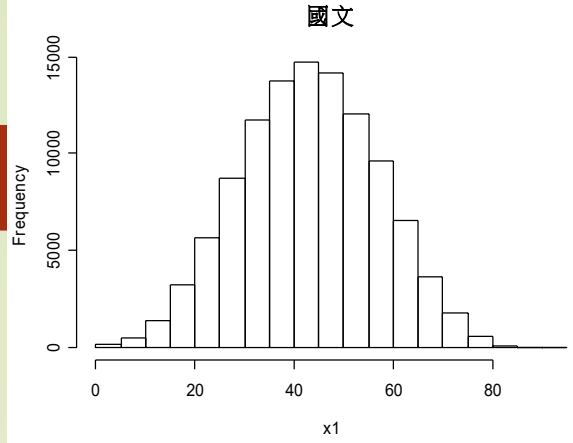
- Note that the class of 15–19 days shows the greatest frequency with 8 audits.
- The class of 30-34 days shows the lowest frequency with only 1 audit.

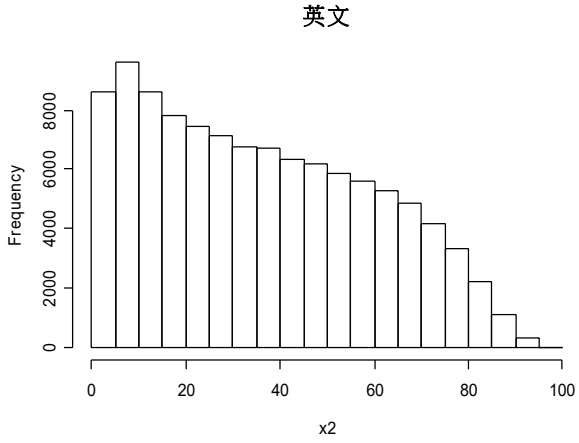
2.2 Histograms Showing Differing Levels of Skewness

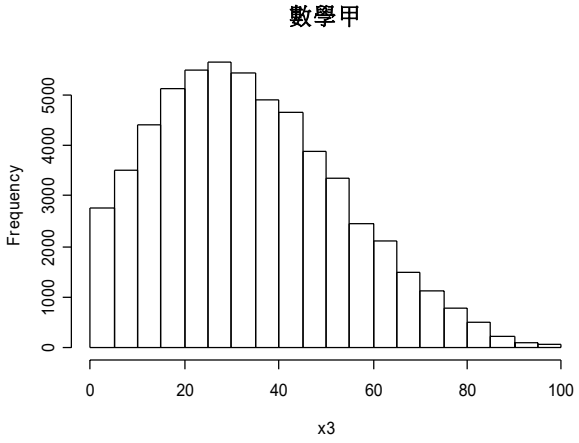


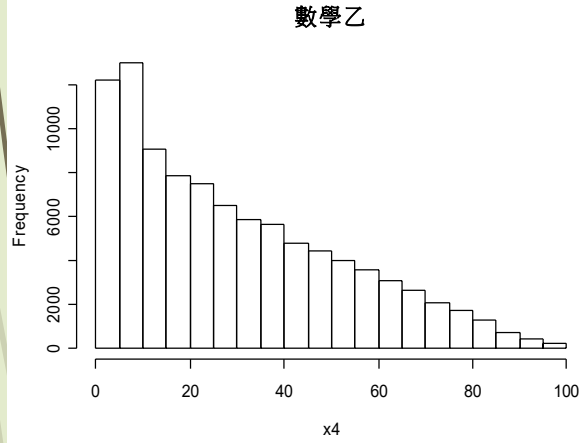
民國 94 年大學指定考試各科成績

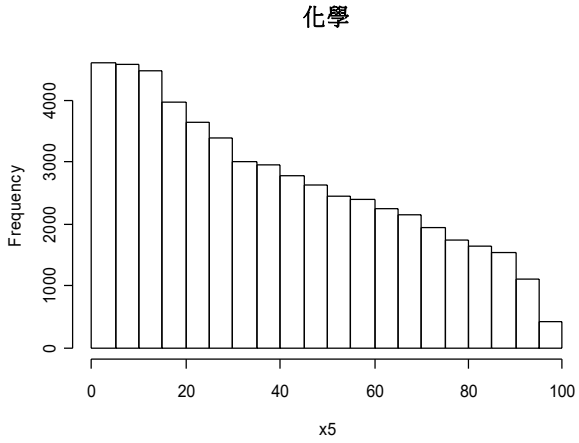
	國文	英文	數學甲	數學乙	化學	物理	生物	歷史	地理
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00
12%	27.00	8.00	11.00	4.00	8.00	6.00	22.00	13.0	18.00
1st Qu.	34.00	16.00	22.00	12.00	15.00	12.00	32.00	28.0	30.00
Median	44.00	34.00	34.00	29.00	34.00	23.00	45.00	39.0	39.00
Mean	43.56	36.68	36.36	34.36	38.88	28.75	46.16	38.7	39.51
3rd Qu.	53.00	56.00	49.00	56.00	60.00	41.00	60.00	50.0	49.00
88%	60.00	69.00	59.00	61.00	76.00	57.00	71.00	56.0	55.00
Max.	93.00	98.00	100.00	100.00	100.00	100.00	99.00	89.0	90.00
st.d.	13.88	23.88	18.72	25.97	27.00	21.50	19.39	16.20	14.46

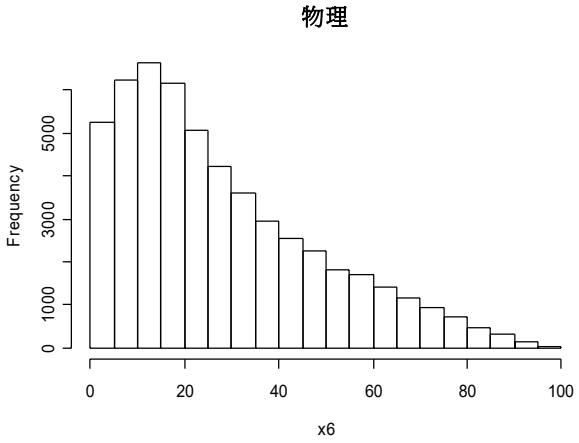

國文
Frequency
0
5000
10000
15000
0
20
40
60
80
x1

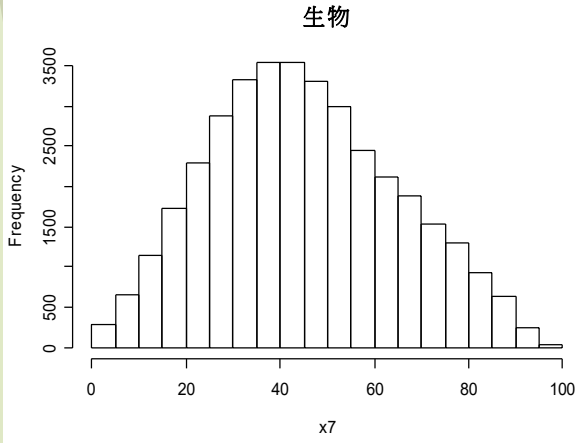

英文
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
100
x2

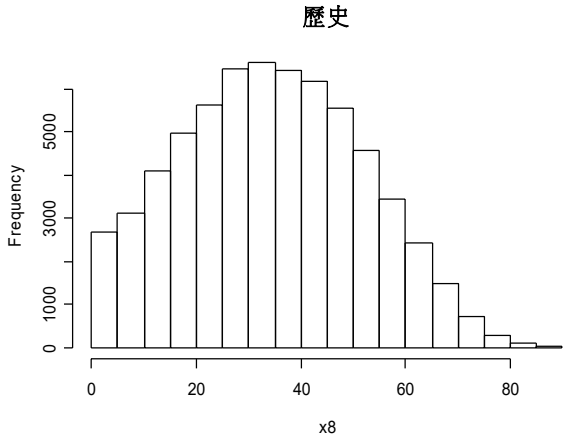

數學甲
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
x3

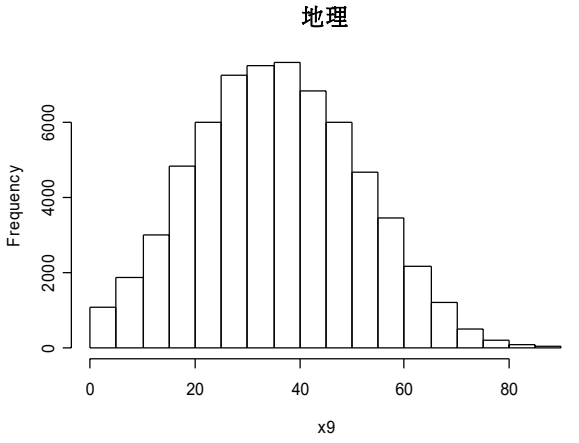

數學乙
Frequency
0
2000
6000
10000
0
20
40
60
80
100
x4


化學
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
x5


物理
Frequency
0
1000
3000
5000
0
20
40
60
80
100
x6


生物
Frequency
0
500
1500
2500
3500
0
20
40
60
80
100
x7


歷史
Frequency
0
1000
3000
5000
0
20
40
60
80
x8


地理
Frequency
0
2000
4000
6000
0
20
40
60
80
x9

2.2 Stem-and-Leaf Display

A **stem-and-leaf display** shows both the rank order and shape of a distribution of data.

It is similar to a histogram on its side, but it has the advantage of showing the actual values.

- The leading digits of each data item are arranged to the left of a vertical line.
- To the right of the vertical line, we record the last digit for each item in rank order.

  - Each line (row) in the display is referred to as a *stem*.
  - Each digit on a stem is a *leaf*.

- **Example** DATAfile: *AptitudeTest*
- The data indicate the number of questions answered correctly in a 150-question aptitude test given to 50 individuals recently interviewed for a position.
- Stem-and-leaf displays for data with more than three digits are possible.

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

2.3 Crosstabulation

Consider Zagat's review of 300 restaurants in the Los Angeles area. The data set includes measurements on quality rating and typical meal price (DATAfile: *Restaurant*.)

A **crosstabulation** of the data is shown below, where:

- the Quality Rating, a categorical variable with categories *good*, *very good*, and *excellent*, is shown in the left margin, and its frequency distribution in the right margin.
- the Meal Price, a quantitative variable ranging from $10 to $49 and grouped into four classes, is shown in the top margin, and its frequency distribution in the bottom margin.
- the cells in the body of the table provide the counts for all the combinations of Quality Rating and Meal Price.

	Meal Price				
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**	**Total**
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

2.3 Crosstabulation: Row Percentages

Converting the entries in a crosstabulation into row percentages or column percentages can provide more insight into the relationship between the two variables.

For row percentages, the results of dividing each frequency in the crosstabulation by its corresponding row total are shown in the table below.

Each row in the table is a percent frequency distribution of meal price for one of the quality rating categories.

- Of the restaurants with the lowest quality rating (good), 50% have low meal prices.
- Of the restaurants with an excellent quality rating, most are in the $30-$39 and $40-$49 higher price ranges.

	Meal Price				
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**	**Total**
Good	50.0%	47.6%	2.4%	0.0%	100.0%
Very Good	22.7%	42.7%	30.6%	4.0%	100.0%
Excellent	3.0%	21.2%	42.4%	33.4%	100.0%

2.3 Crosstabulation: Column Percentages

For column percentages, the results of dividing each frequency in the crosstabulation by its corresponding column total are shown in the table below.

Each column in the table is a percent frequency distribution of quality rating for one of the classes of meal prices.

- Only 2.6% of restaurants with low ($10-19) meal prices have an excellent quality rating.
- As meal price increases, percent frequency distribution shifts toward higher quality ratings.

	Meal Price			
Quality Rating	**$10-19**	**$20-29**	**$30-39**	**$40-49**
Good	53.8%	33.9%	2.6%	0.0%
Very Good	43.6%	54.2%	60.5%	21.4%
Excellent	2.6%	11.9%	36.8%	78.6%
Total	100.0%	100.0%	100.0%	100.0%

X 軸題目 : 2. 請問明年的總統大選，您認爲哪一位候選人最有可能當選？

Y 軸題目 : 8. 請問您的教育程度？

次數 / 百分比 / 列百分比 / 行百分比	連戰	陳水扁	宋楚瑜	許信良	李敖	鄭邦鎮	不知道/拒答	總計
國中及以下	103	43	75	0	0	0	181	402
	9.87	4.12	7.18	0.00	0.00	0.00	17.34	38.51
	25.62	10.70	18.66	0.00	0.00	0.00	45.02	
	22.64	43.88	33.19	0.00	0.00	0.00	68.30	
高中高職	192	34	70	0	0	0	57	353
	18.39	3.26	6.70	0.00	0.00	0.00	5.46	33.81
	54.39	9.63	19.83	0.00	0.00	0.00	16.15	
	42.20	34.69	30.97	0.00	0.00	0.00	21.51	
大專及以上	160	21	81	0	0	0	27	289
	15.33	2.01	7.76	0.00	0.00	0.00	2.59	27.68
	55.36	7.27	28.03	0.00	0.00	0.00	9.34	
	35.16	21.43	35.84	0.00	0.00	0.00	10.19	
拒答	0	0	0	0	0	0	0	0
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
總計	455	98	226	0	0	0	265	1044
	43.58	9.39	21.65	0.00	0.00	0.00	25.38	100

Chi Square : 101.760 P Value : 0.000

2.3 An Analysis of Two Judges' Verdicts

Judges Ron Luckett and Dennis Kendall presided over cases in Common Pleas Court and Municipal Court during the past three years. Some of the rendered verdicts were appealed.

A crosstabulation for a total of 275 appealed verdicts was developed based upon two variables: the Judge (Luckett or Kendall), and the verdict (upheld or reversed.)

The crosstabulation of the results, along with the column percentages in parentheses next to each value, shows that Judge Kendall has been doing the better job because of the greater percentage (88% vs. 86%) of upheld verdicts.

	Judge		
Verdict	**Luckett**	**Kendall**	**Total**
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
TotaI (%)	150 (100%)	125 (100%)	275

2.3 Simpson's Paradox

However, when the results are disaggregated by type of Court (Common Pleas vs. Municipal Court) into two a separate crosstabulation for each Judge, a different picture emerges:

- Judge Luckett leads Judge Kendall in Common Pleas Court upheld verdicts, 91% to 90%.
- Judge Luckett also leads Judge Kendall in Municipal Court upheld verdicts, 85% to 80%.

The reversal of conclusions based on aggregate and disaggregated data is called **Simpson's Paradox**.

Judge Luckett

Verdict	Common Pleas	Municipal Court	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
TotaI (%)	32 (100%)	118 (100%)	150

Judge Kendall

Verdict	Common Pleas	Municipal Court	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
TotaI (%)	100 (100%)	25 (100%)	125

Simpson's Paradox in University admission

UC Berkeley admitted 44% of males and 35% of females who applied in 1973. Data from the six largest departments.

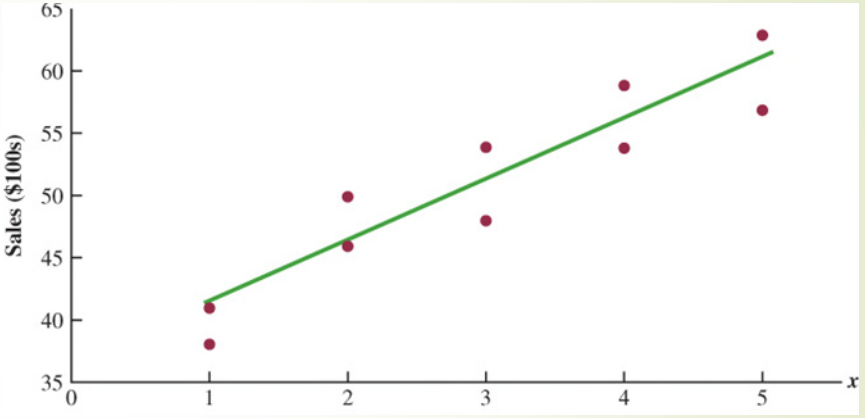
Department	Male acceptance rate	Female acceptance rate
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%

	Male		Female	
	Applicants	%	Applicants	%
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

2.4 Scatter Diagram and Trendline

A **scatter diagram** is a graphical presentation of the relationship between two quantitative variables.

A **trendline** is a line that provides an approximation of the relationship.

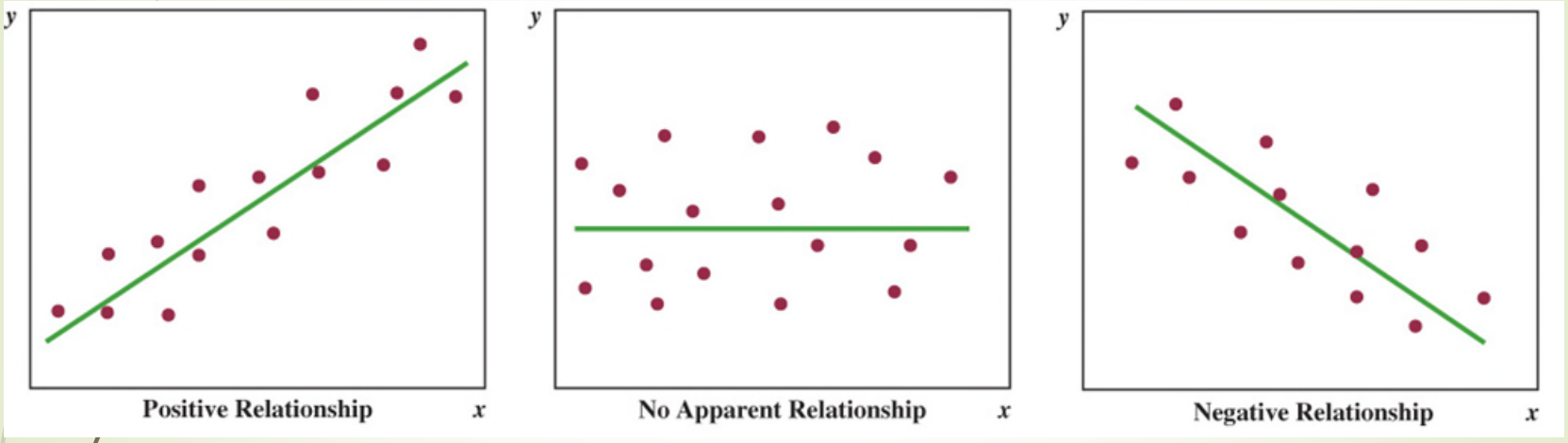


Example DATAfile: *Electronics*

The horizontal axis (x) shows the number of commercials; the vertical axis (y), sales for an electronic store in San Francisco.

The relationship between the number of commercials and sales does not form a perfectly straight line. However, the general pattern of the points and the trendline suggest that the overall relationship is positive.

2.4 Types of Relationships Depicted by Scatter Diagrams



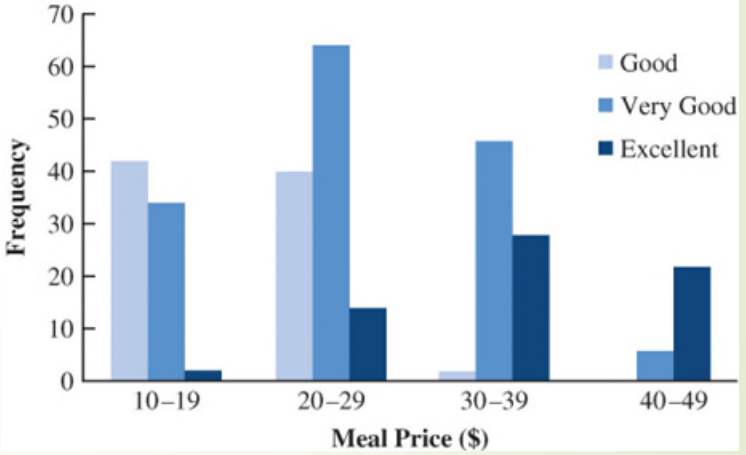
2.4 Side-by-Side Bar Chart

A **side-by-side bar chart** is a graphical display for depicting multiple bar charts on the same display.

We can use a side-by-side bar chart to represent the column percentages of the Zagat's restaurant reviews.

- Each cluster of bars represents one of the four classes of meal prices.
- Each bar within a cluster represents one of the three quality rating categories.

The chart makes even more apparent the shift toward higher quality ratings as the meal prices increase.



2.4 Stacked Bar Chart

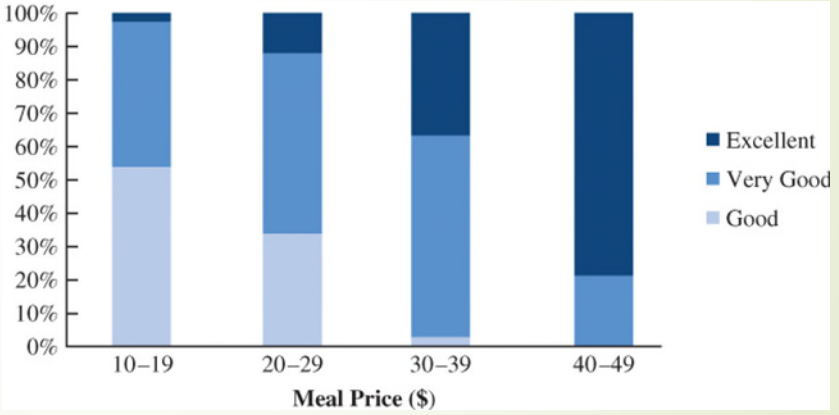
We can also use a **stacked bar chart** to compare two variables on the same display.

In a stacked bar chart, each bar is broken into rectangular segments of a different color showing the relative frequency of each class in a manner similar to a pie chart.

We can also use a stacked bar chart to represent the column percentages of the Zagat's restaurant reviews.

Note that, because percentage frequencies are displayed, all bars are of the same height, extending to the 100% mark (*see notes.)

The stacked bar chart shows even more clearly the shift toward higher quality ratings as meal prices increase.



2.5 Creating Effective Graphical Displays

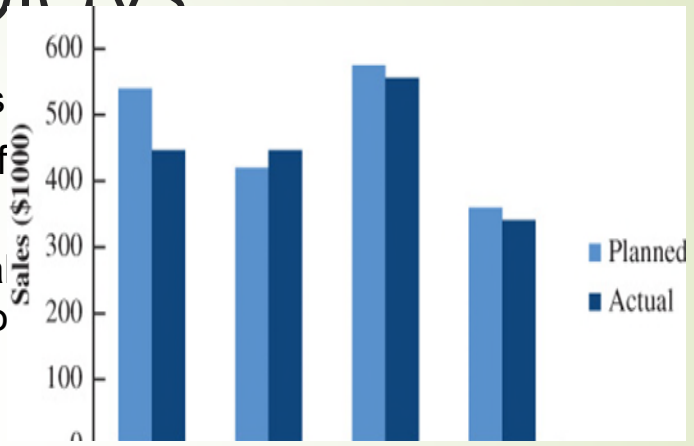
Data visualization is the use of graphical displays to s

The goal of data visualization is to communicate, as eff key information about the data.

The side-by-side bar chart of the planned versus actua United States for Gustin Chemical is a good example o

The chart is an effective graphical display because it follows these guidelines:

- It displays a clear and concise title.
- It keeps the display simple, yet informative.
- Each axis is clearly labeled, and units are provided.
- The two colors used are distinct, and a legend is provided to explain their use.



2.5 Summary of Graphical Displays Used to Show the Distribution of Data

Bar Chart: used to show the frequency distribution and relative frequency distribution for categorical data

Pie Chart: used to show the relative frequency and percent frequency for categorical data. A pie chart is generally not preferred to the use of a bar chart

Dot Plot: used to show the distribution of quantitative data over the entire range of the data

Histogram: used to show the frequency distribution for quantitative data over a set of class intervals

Stem-and-Leaf Display: used to show both the rank order and shape of the distribution for quantitative data

2.5 Summary of Graphical Displays Used to Make Comparisons and Show Relationships

To Make Comparisons

Side-by-Side Bar Chart: used to compare two variables

Stacked Bar Charts: used to compare the relative frequency or percent frequency of two categorical variables

To Show Relationships

- *Scatter diagram*: used to show the relationship between two quantitative variables
- *Trendline*: used to approximate the relationship of data in a scatter diagram

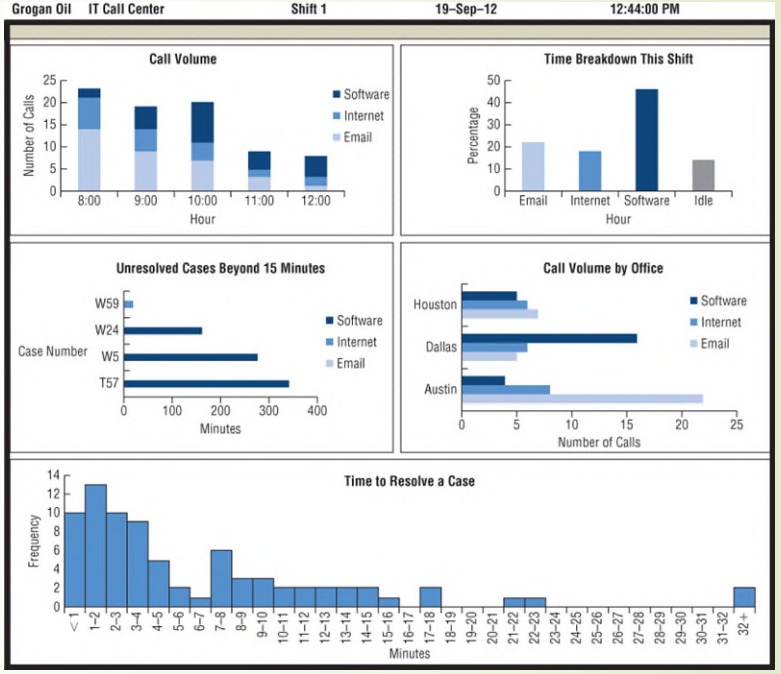
2.5 Data Dashboards

A **data dashboard** is a widely used data visualization tool that organizes and presents key performance indicators (KPIs) used to monitor an organization or process.

A data dashboard provides timely summary information that is easy to read, understand, and interpret.

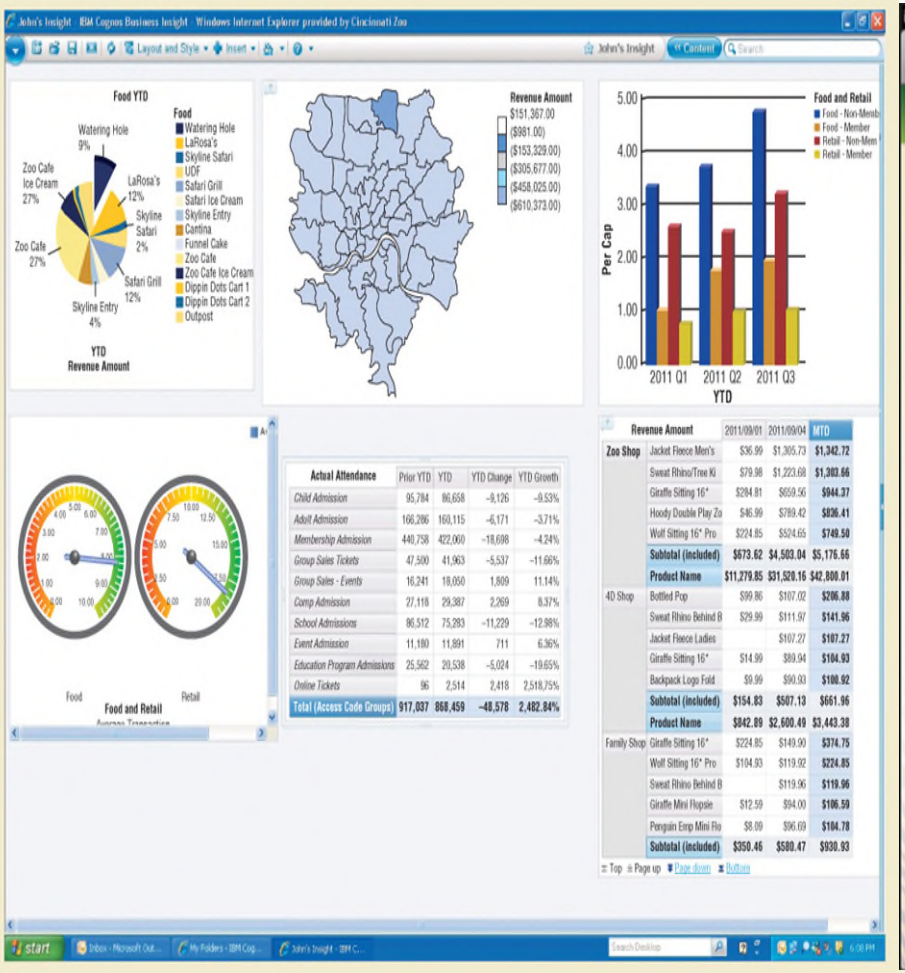
Additional guidelines for data dashboards include:

- Minimize the need for screen scrolling.
- Avoid unnecessary use of color or 3D displays.
- Use borders between charts to improve readability.



2.5 Data Visualization in Practice

Cincinnati Zoo Data Dashboard



Botanical Garden Data Dashboard

Summary

- A set of data, even if modest in size, is often difficult to interpret directly in the form in which it is gathered.
- Tabular and graphical displays can be used to summarize and present data so that patterns are revealed, and the data are more easily interpreted.
- We described tabular and graphical displays to summarize the data for:
  - a single categorical variable
  - a single quantitative variable
  - two variables, categorical or quantitative
  - two quantitative variables
- We discussed guidelines for creating effective graphical displays and how to choose the most appropriate type of display.
- We introduced how data dashboards can be used to monitor a company's performance in a manner that is easy to read, understand, and interpret.

A Summary of Tabular and Graphical Displays of Data