Lecture 2: Descriptive Statistics and Data Exploration

Objectives

- To learn how to explore a dataset through graphics, including box-plot, histogram, scatter plot, etc.
- To explore the distribution of a variable, and the inter-relations between two or more variables.

General

- **Graphics**:

  Univariate: box-plot, histogram, scatter plot, pi-chart, frequency table

  Bivariate: Scatter plot matrix, panel box-plots (with time-trend), etc.

- **Summary measures**

  (See Lecture 03)

Types of Numerical Data

- **Nominal data**
- **Ordinal data**
- **Ranked data**
- **Count data**

} Discrete (離散型)

- **Continuous data** (連續型)

Parametric (probability) models for different types of data: (Ref: Lecture 06)

- **Nominal** :

  Bernoulli and **binomial** distributions

  Example: alcohol drinking (yes or no)

- **Ordinal**:

  (Combined- or cumulative-) Bernoulli and/or multinomial,….

  Example: risk perception problem

Parametric (probability) models for different types of data: (cont.)

- **Count data and Poisson** distribution
  Example: mortality rate in vital stat.

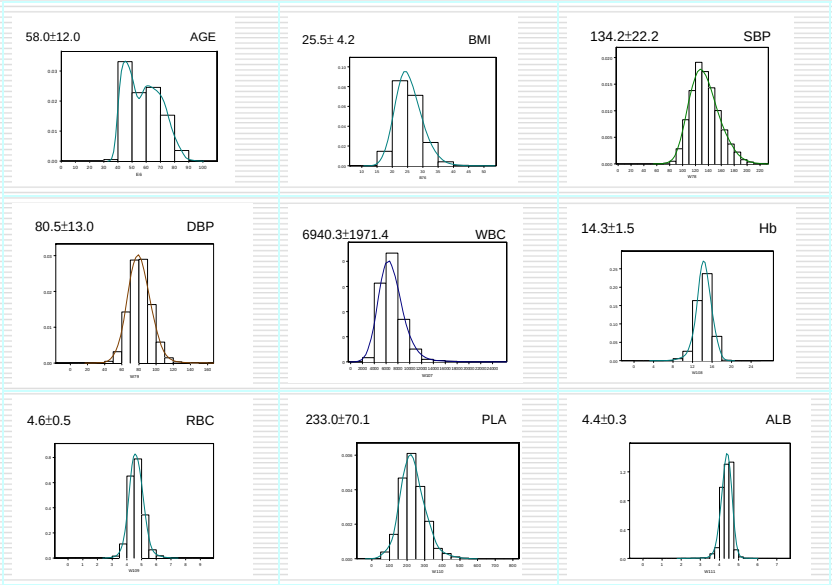
- **Continuous**: (Height, weight, blood pressure...)
  - **normal** (Gaussian) distribution,
  - **gamma** distribution,
  - **exponential** distribution, ...

- **Nonparametric statistics**:
  for ranked data (Ref: Lecture 12)

Example 1. Relationship between gender and several variables (1: chi-square test. 2.Source: "A model-based prediction on length of stay for rehabilitated stroke patients of mid-Taiwan," by Chien-Lin Lin et al., CMUH; preprint.)

	女性	男性	P值[1]
年齡(歲)			
<50	27 (10.4%)	87 (22.5%)	<.00001
50-64	77 (29.1%)	133 (34.5%)	
65-79	125 (48.46%)	150 (38.9%)	
>80	30 (11.6%)	10 (4.2%)	
共病症			
二者皆無	98 (37.8%)	149 (38.6%)	0.6532
只有糖尿病	15 (5.79%)	23 (5.96%)	
只有高血壓	107 (41.3%)	169 (43.8%)	
二者皆有	39 (15.1%)	45 (11.7%)	
物理治療			
無接受	1 (0.4%)	5 (1.3%)	0.2383
有接受	258 (99.6%)	381 (98.7%)	

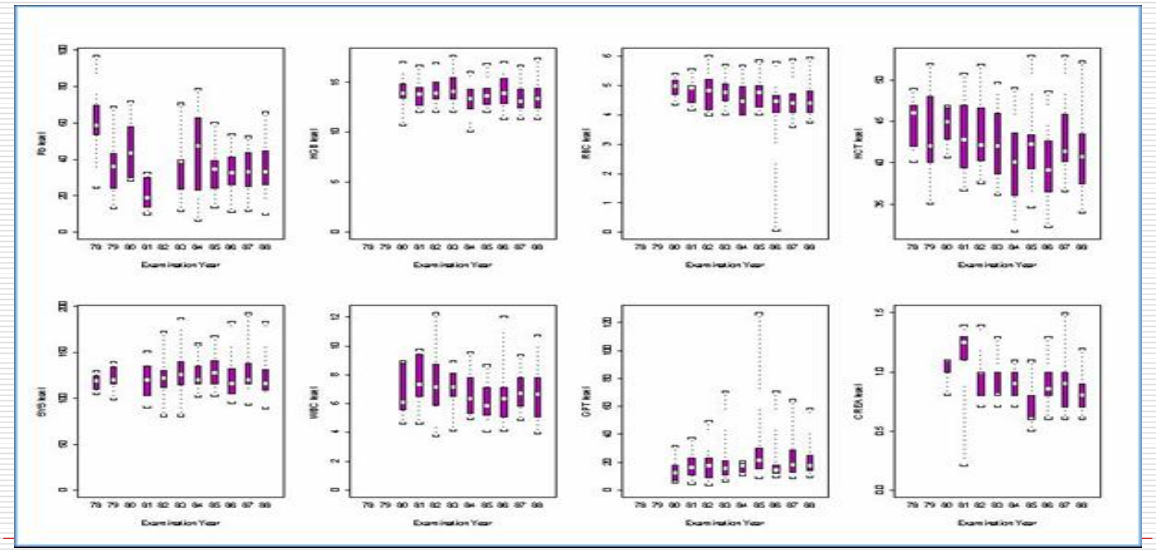
Example 2. The distributions of several **continuous variables of aged people of Hsin-Yi township** (1999~2000)



Source: Cuiwen Chang (2003). Master thesis,
Institute of Environmental Health, CMU, Taichung, Taiwan.

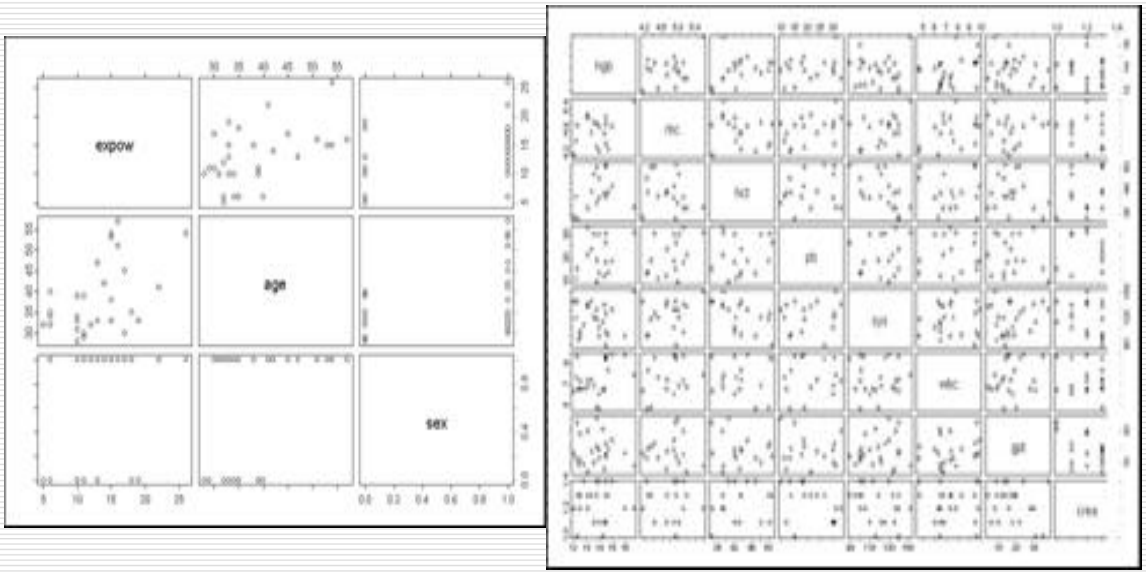
Example 3. A longitudinal study of the effects of long-term exposure to lead among lead battery factory workers in Taiwan (1989-1999).

- **Source**: C.-Y. Hsiao, H.-D. I. Wu, J.-S. Lai and H.-W. Kuo. (2001). *The Science of the Total Environment*, 279, 151-158.

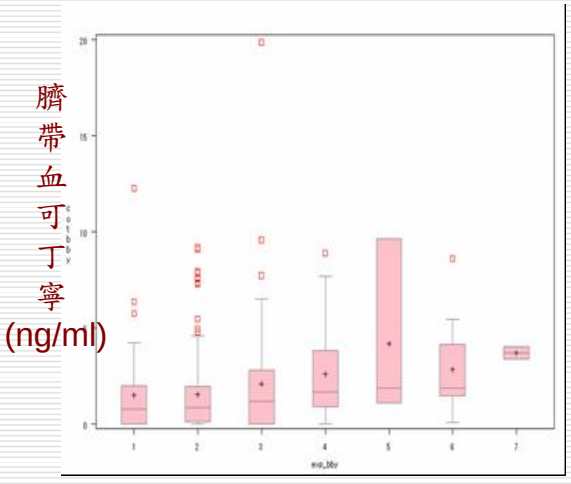


Example 3. (cont.)

The scatter-plot matrix



Example 4. Different exposure levels versus cotinine

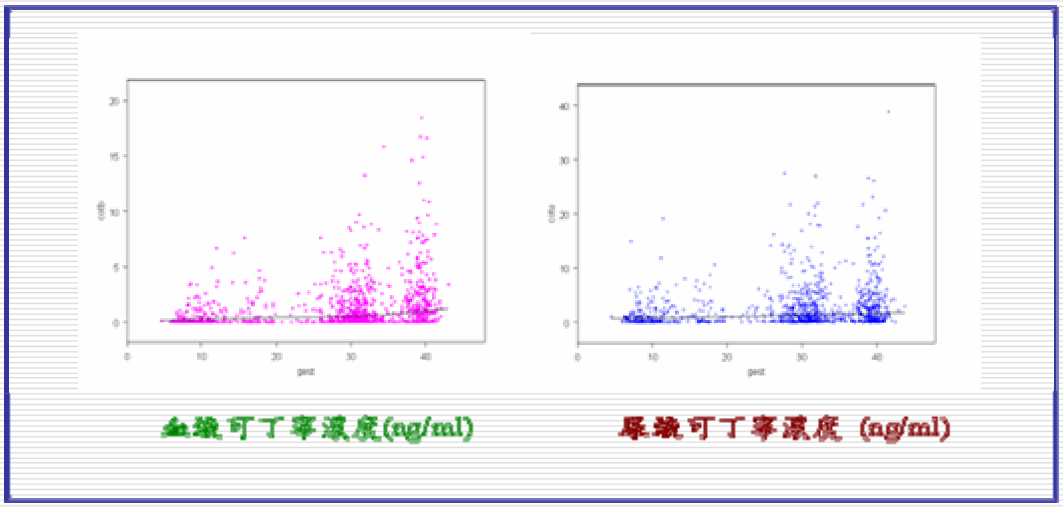


依照菸害暴露分組後，臍帶血可丁寧濃度之群集盒狀圖（以各組可丁寧濃度中位數的大小排序，各組的分類方式參照**表3.**）

Table. 依照菸害暴露分組後，臍帶血可丁寧濃度平均值與中位數

菸害暴露分組		臍帶血可丁寧 (ng/ml)			
吸菸狀態	二手菸	人數	平均值	中位數	等級
從未吸菸	無	116	1.544	0.876	2
從未吸菸	只有家中	70	2.108	1.187	3
從未吸菸	只有職場	50	1.501	0.781	1
從未吸菸	家中與職場	26	2.618	1.696	4
已戒煙	無	2	3.699	3.699	7
已戒煙	家中或職場	12	2.855	1.903	6
目前吸菸	無	--	--	--	--
目前吸菸	家中或職場	3	4.204	1.875	5

Cotinine level versus gestational age of pregnant women



Chebyshev's (or Tchebychev's) Inequality

- Markov inequality: $\Pr(|X|>a) \leq E|X|/a$

  (Prove it! Exercise !)

- $\Pr(|X-\mu| \leq k\sigma) \geq 1-1/k^2$

  or $\Pr(|X-\mu| \geq k\sigma) \leq 1/k^2$

  Proof: Using the Markov inequality with the random variable X being replaced by $(X-\mu)^2/\sigma^2$

Concluding remarks

- **Box-plot, histogram, and one-way or two-way scatter plot, etc., are very useful tools for data exploration. They usually reveal the most important information contained in the data.**

- **Displaying graphics is usually, and should better be, contrasted with a simple analysis (e.g., those using SAS procedures: PROC univariate, PROC corr,…etc.) in order to get a better understanding on the data structure, and also help formulating our statistical and scientific questions.**

Homework and exercises

1. Prove the Markov inequality
2. What is the implication of Chebyshev's inequality on the scatter (or dispersion) of the data?
3. Textbook exercises: (pp. 41~54) problems 24, 25.