Data Preprocessing

——Descriptive Data Summarization——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

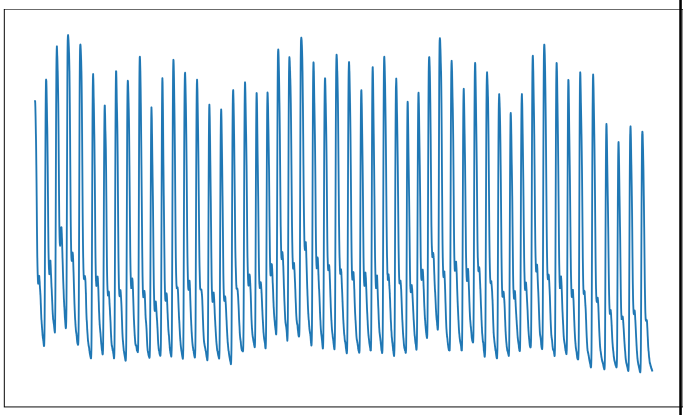
Data Preprocessing

- About data
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

2

Motivation

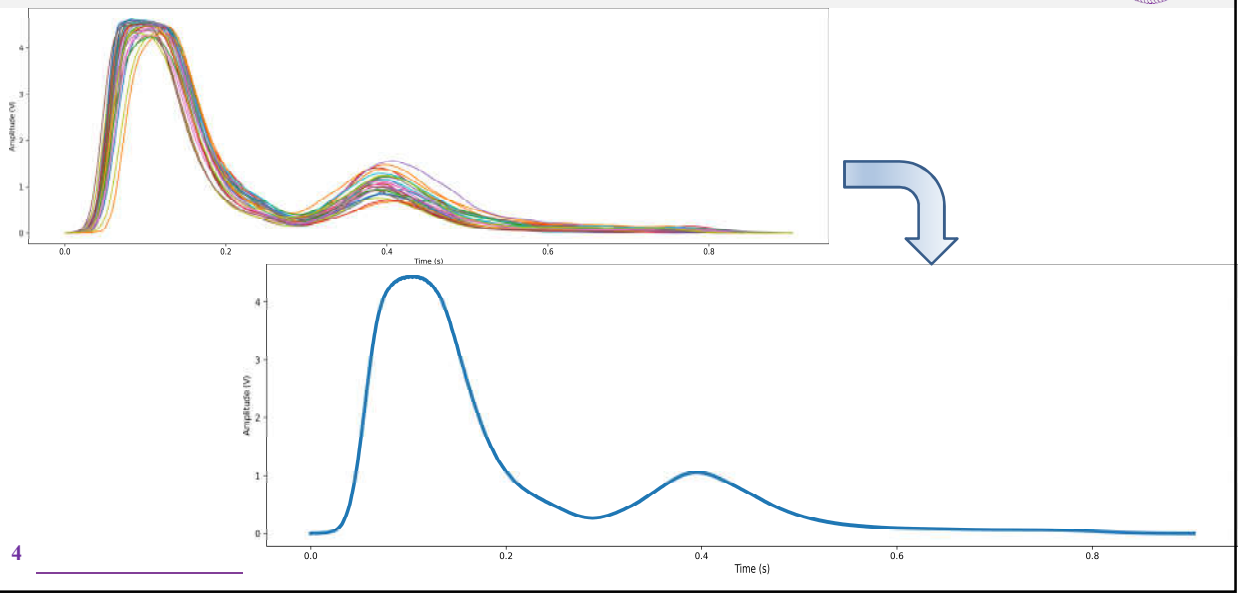
- **Motivation**
  - To better understand the data
  - To get overall picture of data
- **Descriptive data summarization**
  - Central tendency (集中趋势)
  - Dispersion (散布性)



Wrist Pulse Signal

3

Motivation



4

Three Categories of Measurement

- **Distributive（分布的）: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning.**
  - **count(), sum(), min(), max()**
- **Algebraic（代数的）: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function.**
  - **avg() 均值——聚集函数, min_N(), standard_deviation()——标准偏差**
- **Holistic（整体的）: if there is no constant bound on the storage size needed to describe a subaggregate.**
  - **median()——中位数, mode()——众数, rank()**

5

Three Categories of Measurement

- **Value Aggregation v.s. Data Granularity**
  - Coarse-Grained Data (粗粒度)
  - Fine-Grained Data (细粒度)

	A	B	C	D	E	F	G
1	日期	金额		求和项:金额	列标签		
2	20200102	1.21		行标签	收入	支出	总计
3	20200114	-102.64		1月	307584.73	-271410.27	36174.46
4	20200114	5.39		2月	77.86	-4088.04	-4010.18
5	20200115	-2.92		3月	20028.05	-50224.27	-30196.22
6	20200115	-248.97		4月	29699.81	-27272.86	2426.95
7	20200117	-41.47		5月	12071.66	-11480.57	591.09
8	20200117	200001.23		6月	11532.28	-11768.47	-236.19
9	20200117	-7389.3		7月	154699.44	-142374.08	12325.36
10	20200117	-4393.15		8月	36035.05	-16109.33	19925.72
11	20200117	-247.17		9月	248577.19	-214483.59	34093.6
12	20200117	-2487.67		10月	26249.09	-46681.34	-20432.25
13	20200117	-359.26		11月	3375.33	-18019.06	-14643.73
14	20200119	7391.77		12月	452971.72	-362871.36	90100.36
15	20200119	-7389.3		总计	1302902.21	-1176783.24	126118.97
16	20200119	-65.95					
17	20200119	-2416.82					
18	20200119	-716.14					

6

Measuring the Central Tendency(1)

- **Mean (均值 , Algebraic Measure)**
  - **Arithmetic mean:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - **Weighted arithmetic mean:**

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - **Trimmed mean(截断均值): chopping extreme values**

    e.g. Salary and grade

7

Measuring the Central Tendency(2)

- **Median (中位数, holistic measure)**
  - **Middle value if odd number of values, or average of the middle two values otherwise**
    - ***Data*** **57 55 85 24 33 49 94 2 8 51 71 30 91 6 47 50 65 43 41 7**
    - ***Ordered Data***
      - **2 6 7 8 24 30 33 41 43 *47 49* 50 51 55 57 65 71 85 91 94**
    - **Median 48**

8

Measuring the Central Tendency(4)

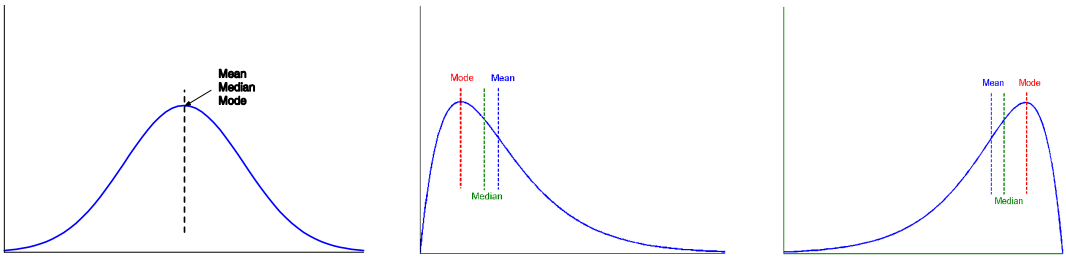
- **Mode**
  - **Value that occurs most frequently in the data**
  - **Unimodal (单峰) , bimodal (双峰) , trimodal (三峰)**
  - **Empirical formula:**
    - **For unimodal (单峰) frequency:**

$$mean - mode = 3 \times (mean - median)$$

9

Symmetric v.s. Skewed Data

- **Median, mean and mode of symmetric, positively and negatively skewed data**



10

Measuring the Dispersion of Data(1)

- Quartiles, outliers and boxplots
  - Quartiles(4分位数): Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range（中间四分位数极差）: IQR = Q3 – Q1
  - Five number summary（五数概括）: min, Q1, M, Q3, max
  - Boxplot（盒图）: ends of the box are the quartiles, median is marked, whiskers(外边界), and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR

11

Measuring the Dispersion of Data(2)

- Variance（方差） and standard deviation（标准差）
  - Variance s^2: (algebraic, scalable computation)

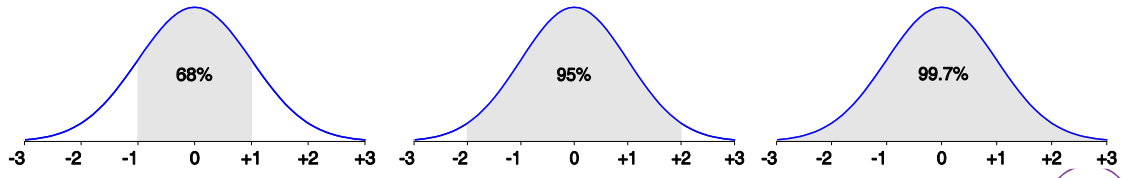
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

  - Standard deviation s is the square root of variance s^2

12

Properties of Normal Distribution Curve

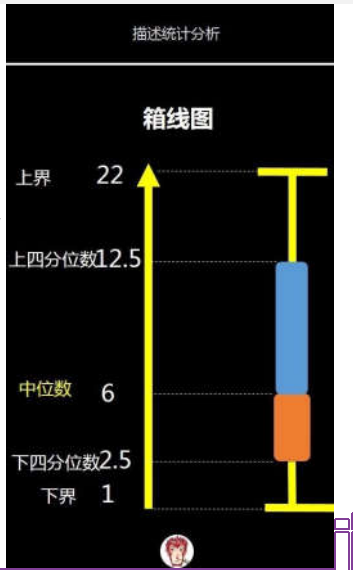
- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements
    (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



13

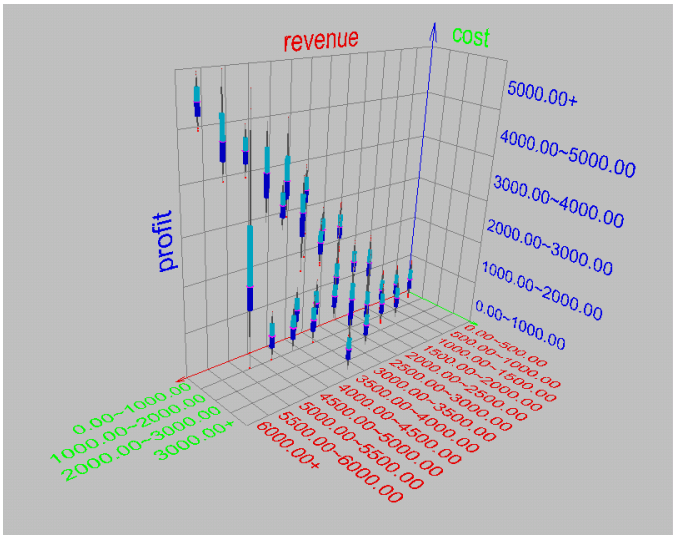
Boxplot Analysis

- **Five-number summary** of a distribution:
  Minimum, Q1, M, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum



14

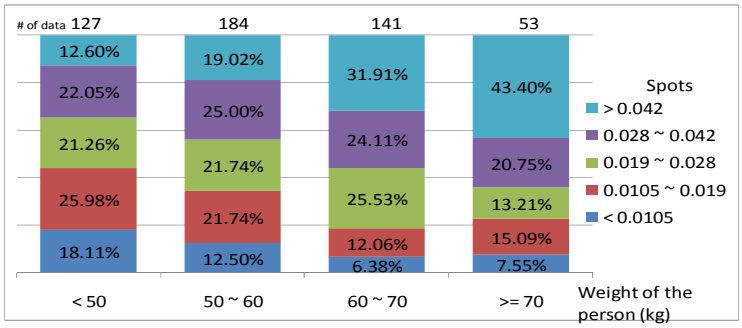
Visualization of Data Dispersion: Boxplot Analysis



15

Histogram Analysis

- **Graph displays the basic statistical class descriptions**
  - **Frequency histograms (频率直方图)**
    - **A univariate graphical method**
    - **Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data**



16

Quantile Plot (分位数图)

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i

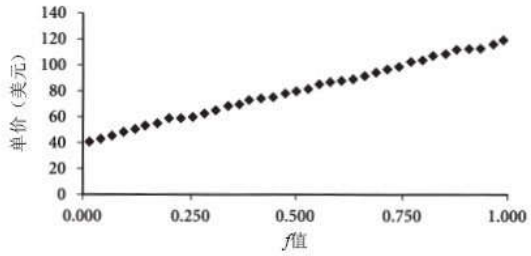


图2-5 表2-1单价数据的分位数图cnrepair.com

17

Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles (分位数) of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another

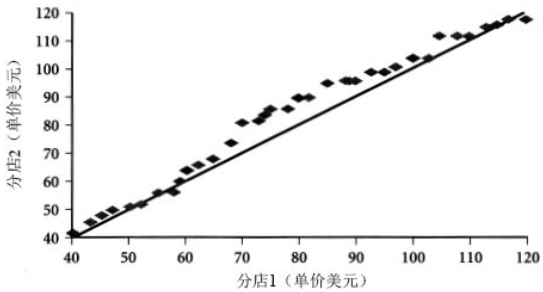
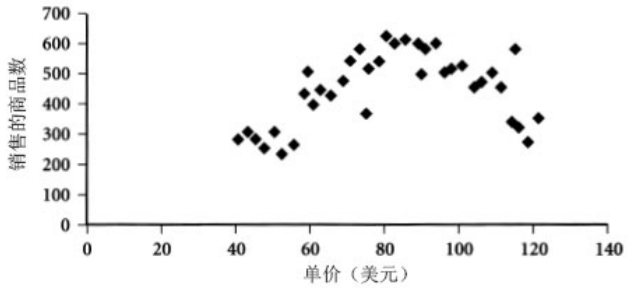


图2-6 两个不同分店的单价数据的分位数 - 分位数图

18

Scatter plot(散布图)

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



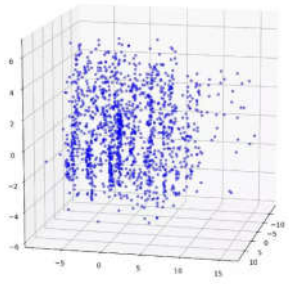
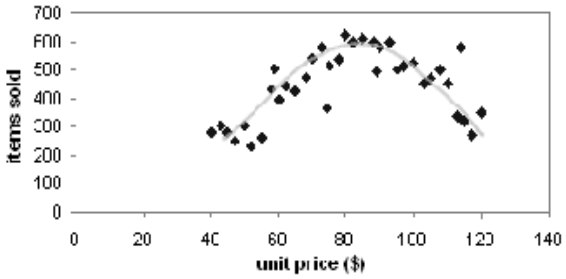


图2-7 表2-1中数据的散布图

19

Loess Curve（局部回归曲线）

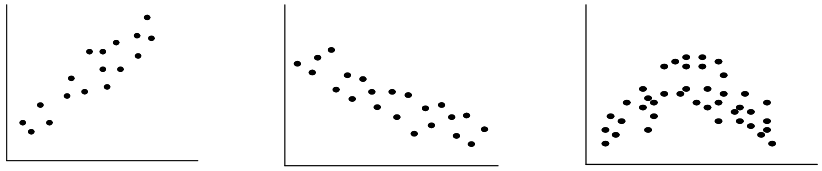
- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



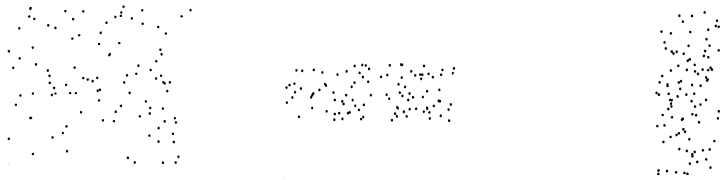
20

Correlated Data

- Positively and Negatively Correlated Data



- Not Correlated Data



21

Graphic Displays of Basic Statistical Descriptions

- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence.

22

Thanks !