

Medical Statistics and Data Analysis

Lecture 2: Descriptive statistics

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

September 21, 2026

Learning Goals of This Lecture

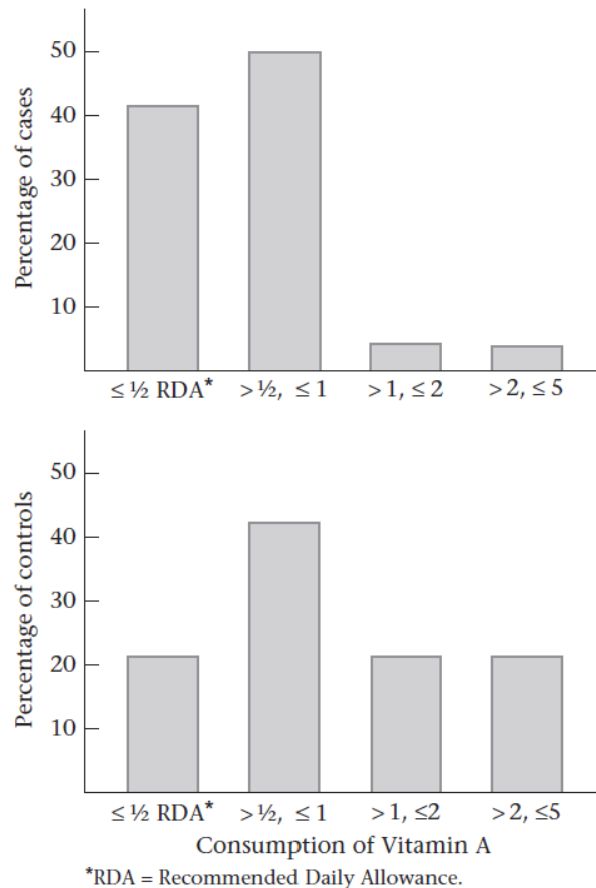
- Descriptive Statistics
 - Measure of location
 - Measure of spread
 - Graph methods
 - Statistical Description of Qualitative Data

Introduction

- The first step in data analysis is to describe the data in some concise manner.
- Descriptive statistics that involve numeric or graphic display are crucial in capturing and conveying the final results of studies in publications.
- Features of good numeric or graphic form of data summarization:
 - Self-contained
 - Understandable without reading the text
 - Clearly labeled of attributes with well-defined terms
 - Indicate principal trends in data

Example: Bar graphs

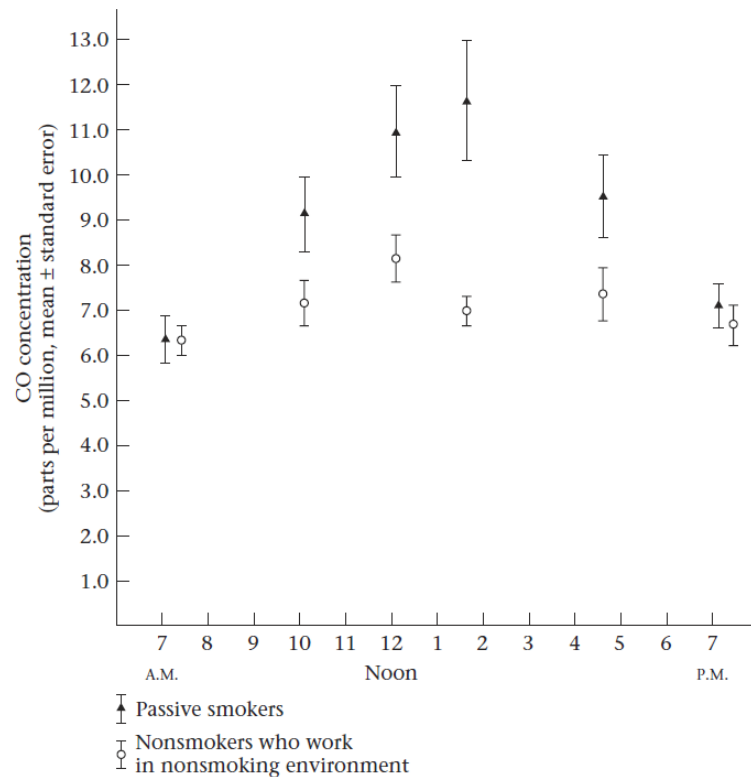
FIGURE 2.1 Daily vitamin-A consumption among cancer cases and controls



- Vitamin A consumption prevents cancer
- Total cancer cases: 200
- Total matched controls: 200
- The bar graphs show that the Vitamin A consumed by controls is more than that consumed by the patients with cancer. In some cases, the levels exceed the recommended daily allowance.

Example: Scatter plot

FIGURE 2.2 Mean carbon-monoxide concentration ($\pm$ standard error) by time of day as measured in the working environment of passive smokers and in nonsmokers who work in a nonsmoking environment



Source: Based on *The New England Journal of Medicine*, 302, 720–723, 1980.

- CO concentrations are about the same in the working environments of passive smokers and nonsmokers early in the day.
- CO concentrations change over the course of the day
- This supports the observation that passive smokers have lower pulmonary function than comparable nonsmokers

Measures of Location (位置度量)

- It is easy to lose track of the overall picture when there are too many sample points.
- Data summarization is important before any inferences can be made about the population from which the sample points have been obtained.
- Measures of location is a type of measure useful for data summarization that defines the center or middle of the sample.

Defining the Middle: The Arithmetic Mean

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- Arithmetic mean: the sum of all the observations divided by the number of observations.
- Statistically expressed as $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Limitation: Oversensitive to extreme values; in which case, it may not be representative of the location of the majority of sample points.

Arithmetic Mean Explained

- Sigma Σ is a summation sign.
- $\sum_{i=1}^n x_i$ implies $(x_1 + x_2 + \cdots + x_n)$
- If a and b are integers where $a \leq b$, then $\sum_{i=a}^b x_i$ meaning $(x_a + x_{a+1} + \cdots + x_b)$
- If $a = b$, then $\sum_{i=a}^b x_i = x_a$
- Factorization property (分解律): If c is some constant, then $\sum_{i=1}^n cx_i = c(\sum_{i=1}^n x_i)$

Median (中位数)

- Sample median is
- $\left(\frac{n+1}{2}\right)$ th the largest observation if n is odd
- Average of the $\left(\frac{n}{2}\right)$ th and the $\left(\frac{n}{2} + 1\right)$ th observation if n is even

TABLE 2.3 Sample of admission white-blood counts ($\times 1000$) for all patients entering a hospital in Allentown, Pennsylvania, on a given day

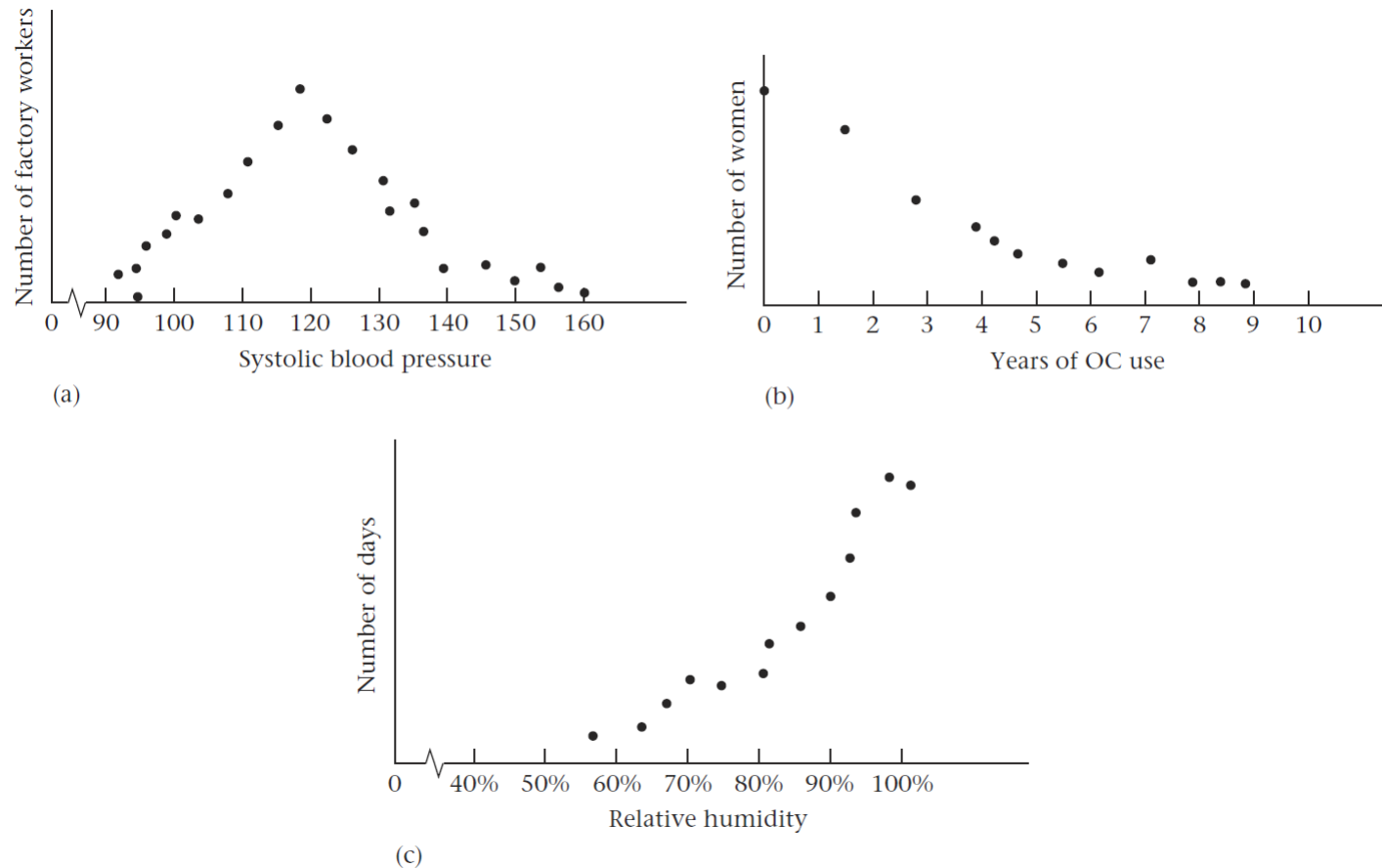
i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Example. Order the sample as follows:
3, 5, 7, 8, 8, 9, 10, 12, 35

Because n is odd, the sample median is the fifth largest point, that is , 8

Comparing Mean and Median

FIGURE 2.3 Graphic displays of (a) symmetric, (b) positively skewed, and (c) negatively skewed distributions



- For symmetric distributions (对称分布), arithmetic mean is approximately the same as the median
- For positively (negatively) skewed distributions [正(负)偏态分布], arithmetic mean tends to be larger (smaller) than the median

Mode (众数)

- Mode: the most frequently occurring value among all the observations in a sample.
- Data distributions may have one or more modes.
- One mode = unimodal (单峰)
- Two modes = bimodal (双峰)
- Multiple modes = multimodal (多峰)



TABLE 2.4 Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

Example.

Mode is 28

Geometric Mean

- Many types of laboratory data can be expressed as multiples of 2 or a constant multiplied by a power of 2, that is
- $2^k c$, $k = 0, 1, \dots$ for some constant c

- **Example:**

TABLE 2.5

Distribution of minimum inhibitory concentration (MIC) of penicillin G for *N. gonorrhoeae*

淋病奈瑟菌 (*N. gonorrhoeae*) 对青霉素G的最低抑菌浓度 (MIC) 分布

Concentration ($\mu\text{g/mL}$)	Frequency	Concentration ($\mu\text{g/mL}$)	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Source: Based on *JAMA*, 220, 205–208, 1972.

- $2^k(0.03125)$, $k = 0, 1, \dots$
- For distributions that are skewed, the log concentrations may be considered.

- The geometric mean can be computed as

$$G(x) = \left(\prod_{i=1}^n x_i \right)^{1/n}$$

- Or the antilogarithm of the arithmetic mean of $\log x$

$$G(x) = \exp \left(\frac{1}{n} \sum_{i=1}^n \ln x_i \right)$$

- $\log_{10}$ may be used to replace $\ln$ for the calculations

Some Properties of Arithmetic Mean

- **Translation Property (平移律)**

- Original sample: $x_1, x_2, \dots, x_n$
- Translated sample: $x_1 + c, x_2 + c, \dots, x_n + c$ (where c is some constant)
- Let $y_i = x_i + c, i = 1, \dots, n$ then $\bar{y} = \bar{x} + c$

TABLE 2.6 Translated sample for the duration between successive menstrual periods in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
-4	5	1	96	6	7
-3	10	2	63	7	3
-2	28	3	24	8	2
-1	64	4	9	9	1
0	185	5	2	10	1

Note: $\bar{y} = [(-4)(5) + (-3)(10) + \dots + (10)(1)]/500 = 0.54$

$$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54 \text{ days}$$

- If the unit or scale changes, then using the rescaled sample

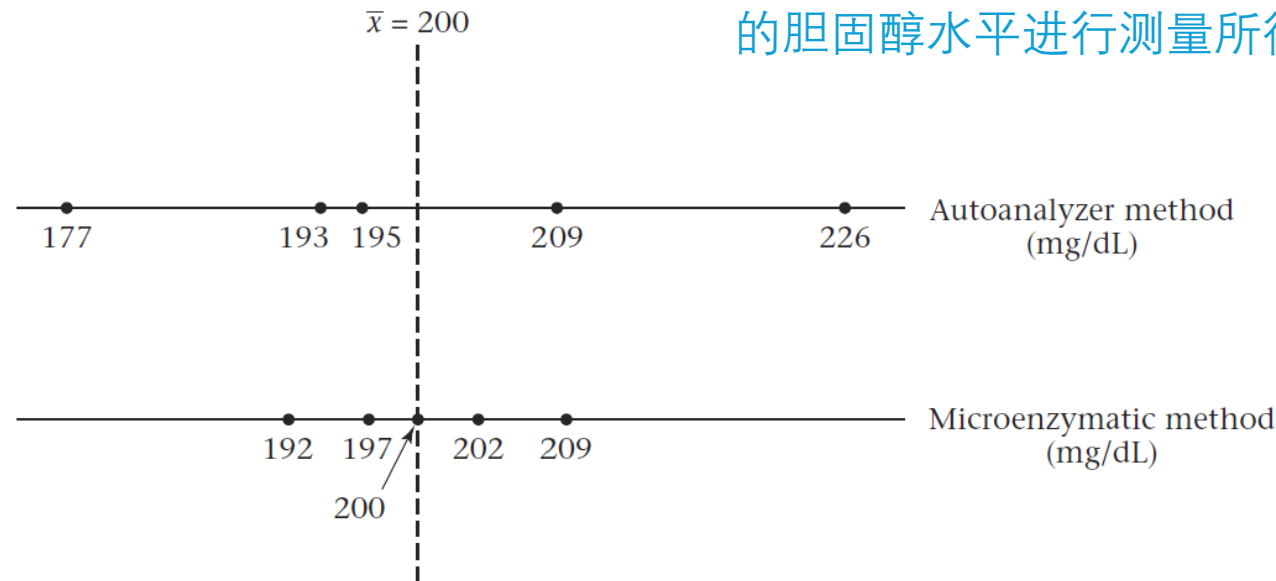
$$y_i = cx_i, i = 1, \dots, n$$

- Arithmetic mean is then $\bar{y} = c\bar{x}$
- Let $x_1, x_2, \dots, x_n$ be the original sample of data.
- Let $y_i = c_1x_i + c_2, i = 1, \dots, n$ represent a transformed sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2
- Then $\bar{y} = c_1\bar{x} + c_2$

Measures of Spread (离散程度的度量)

FIGURE 2.4 Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods

采用自动分析仪法和微量酶法对同一受试者的胆固醇水平进行测量所得的两组样本数据



- The mean obtained by the two methods is the same. However, the **variability** (变化程度) or **spread** (离散程度) of the Autoanalyzer method appears to be greater.

Range of variability

- Range (极差或全距) is the difference between the largest and smallest observations in a sample.
- Once the sample is ordered, it is very easy to compute the range.
- Range is very sensitive to extreme observations or outliers (离群值).
- Larger the sample size (n), the larger the range and the more difficult the comparison between ranges from datasets of varying sizes.
- A better approach to quantifying the spread in datasets is percentiles (百分位数) or quantiles (分位数).
- Percentiles are less sensitive to outliers and are not greatly affected by the sample size.

Quantiles or percentiles

- The p th percentile is the value V_p such that p percent of the sample points are less than or equal to V_p .
- Median is the 50th percentile, which is a special case of a quantile.
- The p th percentile is defined by
 - The $(k + 1)$ th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$)
 - The average of the $(np/100)$ and $(np/100 + 1)$ th largest observations if $np/100$ is an integer

Quantiles or percentiles (Cont.)

- Frequently used percentiles are
 - Quartiles (25th, 50th, and 75th percentiles)
 - Quintiles (20th, 40th, 60th, and 80th percentiles)
 - Deciles (10th, 20th, ..., 90th percentiles)
 - Ventiles (5th, 10th, ..., 95th percentiles)
- To complete percentiles, the sample points must be ordered.
- If n is large, a stem-and-leaf plot or a computer program may be used.

Variance and Standard Deviation

- If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean can be expressed as $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$
- That is,
$$d = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$
- The sum of the deviations (离差) of the individual observations of a sample about the sample mean is always zero
- Standard deviation d is a reasonable measure of spread if the distribution is bell-shaped.

Mean deviation

- The difference d does not help distinguish the difference in spreads between two methods.
- Mean deviation, express as $\sum_{i=1}^n |x_i - \bar{x}| / n$ may be used.
- Alternatively, **sample variance** or variance, which is the average of the squares of the deviations from the sample mean, may be used

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

- Another commonly used measure of spread is the **sample standard deviation**

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Properties of Variance and Standard Deviation

- Samples $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ where $y_i = x_i + c, i = 1, \dots, n$ if respective sample variances are s_x^2 and s_y^2 then $s_y^2 = s_x^2$
- Samples $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ where $y_i = cx_i, i = 1, \dots, n$ and $c > 0$ then $s_y^2 = c^2 s_x^2$ which is $s_y = cs_x$

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (cx_i - c\bar{x})^2}{n-1} = \frac{c^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = c^2 s_x^2$$

$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

Coefficient of Variation (CV, 变异系数)

- Defined as $100\% \times (s/\bar{x})$

TABLE 2.8 Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	<i>n</i>	Mean	<i>sd</i>	<i>CV (%)</i>
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

- Remains the same regardless of units used
- Useful in comparing variability of different samples with different arithmetic means
- Useful for comparing the reproducibility of different variables

Grouped Data

- When sample size is too large to display all the raw data, data can be frequently collected in grouped form

TABLE 2.9 Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

- The simplest way to display the data is to generate a frequency distribution using a statistical package.
- A frequency distribution is an ordered display of each value in a dataset together with its frequency, that is, the number of times that value occurs in the dataset.

- If the data is too large, then the data is categorized into broader groups.

TABLE 2.12 **Grouped frequency distribution of the birthweight (oz) from 100 consecutive deliveries**

The FREQ Procedure				
Group_interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$29.5 \leq x < 69.5$	5	5.00	5	5.00
$69.5 \leq x < 89.5$	10	10.00	15	15.00
$89.5 \leq x < 99.5$	11	11.00	26	26.00
$99.5 \leq x < 109.5$	19	19.00	45	45.00
$109.5 \leq x < 119.5$	17	17.00	62	62.00
$119.5 \leq x < 129.5$	20	20.00	82	82.00
$129.5 \leq x < 139.5$	12	12.00	94	94.00
$139.5 \leq x < 169.5$	6	6.00	100	100.00

Graphic Methods

- Graphic methods of displaying data give a quick overall impression of data. The following are some graphic methods.
- Bar graphs (柱状图或条形图):
 - Used to display grouped data
 - Difficult to construct
 - Identity of the sample points within the respective groups is lost
- Stem-and-Leaf plots (茎叶图):
 - Easy to compute the median and other quantiles
 - Each data point is converted into stem and leaf, e.g., 438 (stem:43; leaf:8)
- Box plots (箱线图或箱形图):
 - Uses the relationships among the median, upper quartile, and lower quartile to describe the skewness or symmetry of a distribution

Stem-and-leaf plots

FIGURE 2.6 Stem-and-leaf plot for the birthweight data (oz) in Table 2.9 using R

```
> stem.leaf(bwt$birthweight, unit=1, trim.outliers=FALSE)
```

1 | 2: represents 12

leaf unit: 1

```

              n: 100
1      3 | 2
      4
2      5 | 8
5      6 | 478
      7
15     8 | 3556788999
26     9 | 12344568889
45    10 | 0123444445567888899
(17)  11 | 00122235555556889
38    12 | 011122223444445567788
18    13 | 222334557888
6     14 | 0146
2     15 | 5
1     16 | 1

```

Stem | Leaf

- The collection of leaves indicates the shape of the data distribution

FIGURE 2.7 Stem-and-leaf plot for the birthweight data (g) in Table 2.1

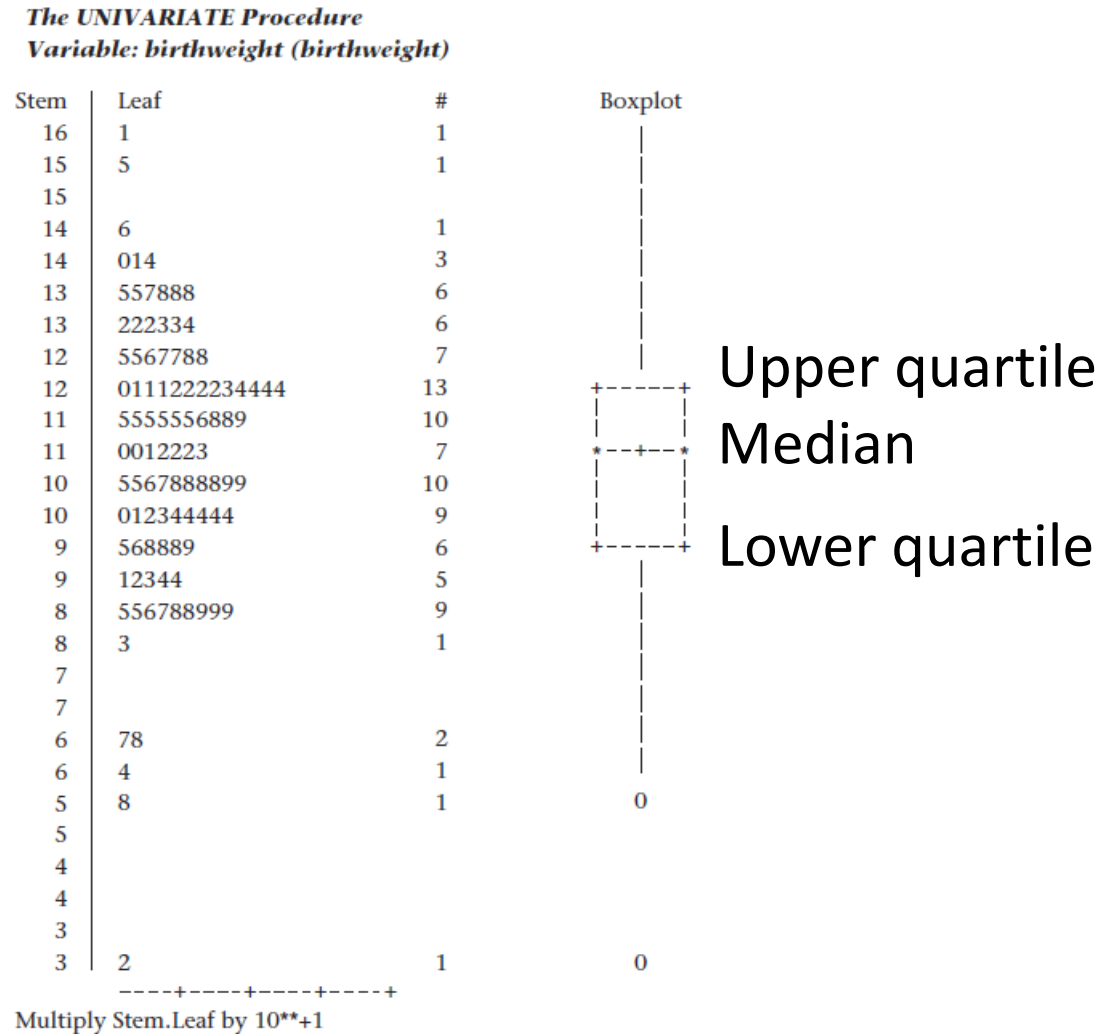
```

20 | 69
21 |
22 |
23 |
24 |
25 | 81
26 |
27 | 59
28 | 41 38 34
29 |
30 | 31
31 | 01
32 | 65 60 45 00 48
33 | 23 14
34 | 84
35 | 41
36 | 49 09
37 |
38 |
39 |
40 |
41 | 46

```

Stem-and-leaf vs. Box Plot

FIGURE 2.8 Stem-and-leaf and box plot for the birthweight data in Table 2.9 as generated by the SAS Univariate Procedure



Box plot

- If the distribution is symmetric, then upper and lower quartiles should be approximately equally spaced from the median
- If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed
- If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed

- An outlying value is a value x such that either
$$x > \text{upper quartile} + 1.5 \times (\text{upper quartile} - \text{lower quartile})$$
$$x < \text{lower quartile} - 1.5 \times (\text{upper quartile} - \text{lower quartile})$$
- An extreme outlying value is a value x such that either
$$x > \text{upper quartile} + 3.0 \times (\text{upper quartile} - \text{lower quartile})$$
$$x < \text{lower quartile} - 3.0 \times (\text{upper quartile} - \text{lower quartile})$$
- A vertical bar connects the upper quartile to the largest nonoutlying value in the sample
- A vertical bar connects the lower quartile to the smallest nonoutlying value in the sample

Case study 1: Effects of lead exposure on neurological and psychological function in children

FIGURE 2.9 Number of finger-wrist taps in the dominant hand for exposed and control groups, El Paso Lead Study

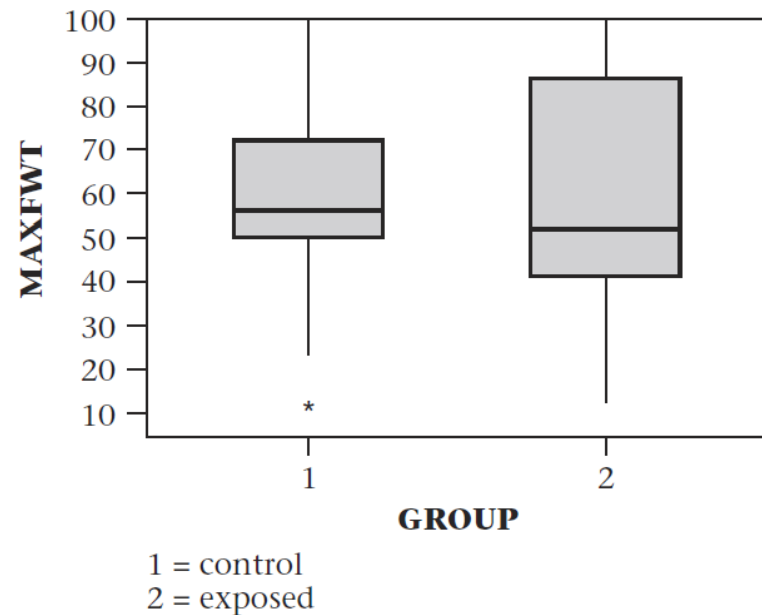
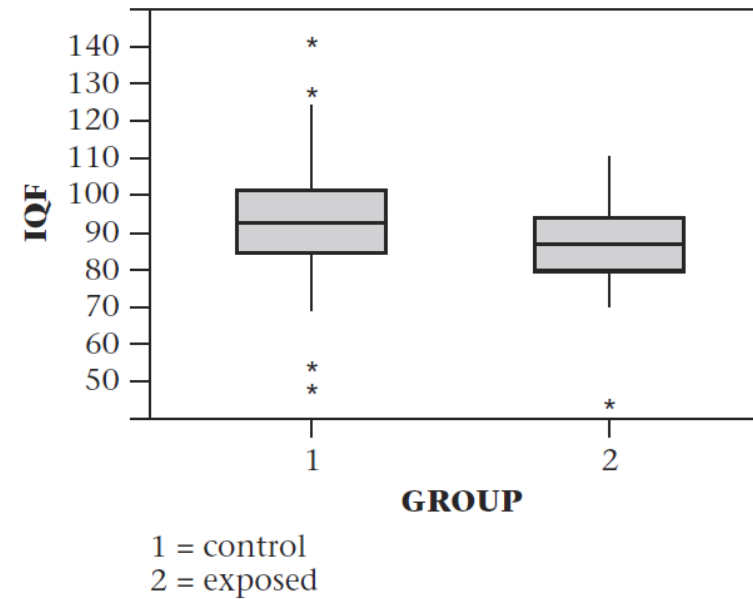
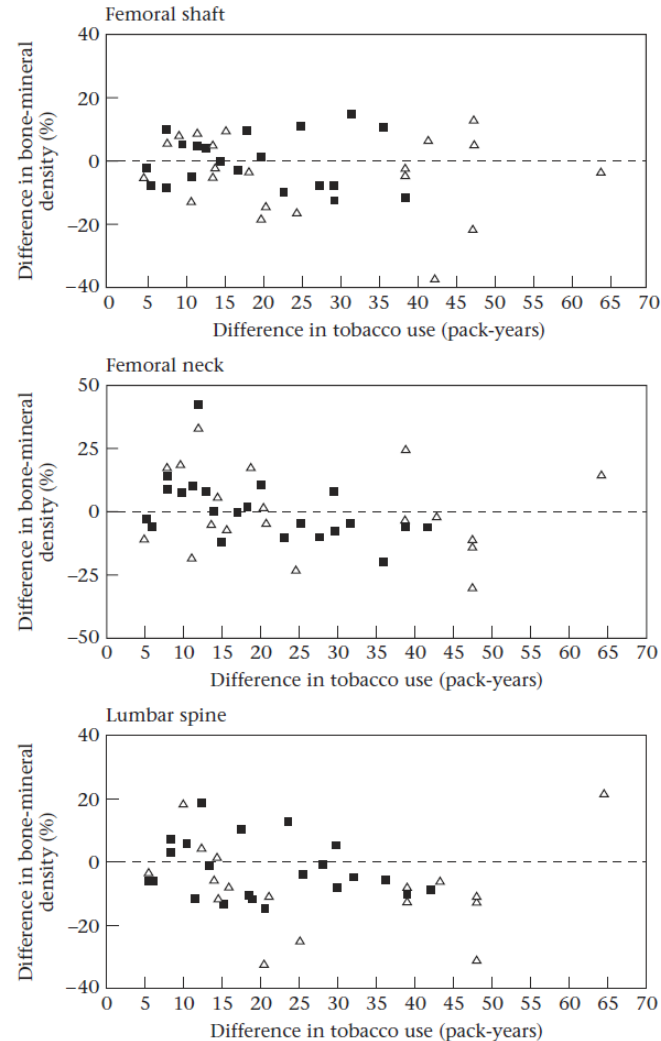


FIGURE 2.10 Wechsler full-scale IQ scores for exposed and control groups, El Paso Lead Study



The finger-wrist tapping scores (MAXFWT) and full-scale IQ scores (IQF) seem slightly lower in the exposed group than in the control group.

FIGURE 2.11 Within-pair differences in bone density at the femoral shaft, femoral neck, and lumbar spine as a function of within-pair differences in pack-years of tobacco use in 41 pairs of female twins. Monozygotic (identical) twins are represented by solid circles and dizygotic (fraternal) twins by open circles. The difference in bone density between members of a pair is expressed as the percentage of the mean bone density for the pair.



Source: Based on "The bone density of female twins discordant for tobacco use," by J. H. Hopper and E. Seeman, 1994, *The New England Journal of Medicine*, 330, 387–392.

Scatter plot

- Matched-pair study
- Exposed and control are matched on other characteristics related to the outcome BMD.
- In this case, matching is based on having similar genes.
- An inverse relationship is seemed between difference in BMD and difference in tobacco use.

Statistical Description of Qualitative (Categorical) Data

- Concepts and Definitions of relative numbers
 - Ratio (比)
 - Proportion (比例)
 - Rate (率)
- Frequently used measures in healthcare

Ratio (比)

- Ratio is the ratio of two professionally related metrics.
- Example:
- Sex Ratio: Male population / Female population $\times 100\%$
- BMI (Body Mass Index): Weight (kg) / Height² (m²)
- RR (Relative Risk): P (an event occurring in the exposed group) / P (the same event occurring in the unexposed group)
- OR (Odds Ratio): Odds of an event occurring in the exposed group / the odds of the same event occurring in the unexposed group

2x2 Contingency Table (列联表)

	Cancer	No Cancer	Total
Smoking	a	b	$a + b$
No Smoking	c	d	$c + d$

$$RR = \frac{a/(a + b)}{c/(c + d)}$$

$$OR = \frac{a/b}{c/d}$$

Proportion (比例或构成比)

- A proportion is the percentage of a part to the whole, indicating the relative contribution of each component to the total.
 - The denominator (分母) must cover the numerator (分子).
 - The denominator and the numerator must be the same type of metric.

• Example:

Age (years old)	# of patients	Proportion of patients
20-39	140	33.3%
40-49	180	42.9%
50-60	100	23.8%
Total	420	100.0%

Prevalence rate (PR, 患病率)

- Prevalence is the **proportion** of individuals in a population who have a particular disease or condition at a specified point in time or during a specified period.
- Example:

Age (years old)	# of examined people	# of patients	Proportion of patients
20-39	2000	140	7.0%
40-49	1500	180	12.0%
50-60	500	100	20.0%
Total	4000	420	10.5%

Rate (率)

- Rate is the frequency of an event occurring per unit of time.
- The numerator is the number of occurrences of the event during the observation period.
- The denominator is the total person-time of observation.

Incidence Density (ID, 发病率)

- Incidence Density (Incidence Rate, IR) is the number of new cases of a disease during the observation period divided by the total person-time at risk during that period.
- Example: Incidence Density of Breast Cancer. A study investigated the incidence of breast cancer among women aged 45–49 who used oral contraceptives. During the study, 9 new cases of breast cancer were identified, with a total observation time of 2,935 person-years. The incidence density of breast cancer in this study population was:
- Incidence Density = $\frac{9}{2935} = 0.00307$ per person-year

Other commonly used statistical metrics in healthcare

- **Detection Rate (检出率)**

- The ratio of the number of individuals found to have a particular disease during an examination to the total number of individuals examined at that time.

- **Infection Rate (感染率)**

- The ratio of the number of individuals found to be infected with a particular pathogen to the total number of individuals examined at that time.

Metrics indicating disease severity

- **Cause-Specific Death Rate (CSDR, 死因别死亡率)**

- The mortality rate attributable to a particular disease or cause of death.

$$\text{CSDR} = \frac{\text{Number of deaths due to a specific cause}}{\text{Total population during the same period}}$$

- **Fatality Rate (FR, 病死率)**

- The proportion of individuals with a particular disease who die from that disease during a specified period.

Metrics indicating disease prevention and treatment effectiveness

- Cure Rate (治愈率): the proportion of patients with a particular disease who are cured following treatment.

$$\text{Cure Rate} = \frac{\text{Number of patients cured}}{\text{Total number of patients treated}}$$

- Treatment Effectiveness Rate (治疗有效率) : the proportion of patients with a particular disease who achieve a predefined favorable treatment outcome.

$$\text{Treatment Effectiveness Rate} = \frac{\text{Number of patients with effective treatment outcomes}}{\text{Total number of patients treated}}$$

- Survival Rate (生存率) : the proportion of patients with a particular disease or receiving a specific treatment who remain alive after a specified follow-up period of n years.

$$\text{Survival Rate} = \frac{\text{Number of patients alive after } n \text{ years}}{\text{Total number of patients followed up for } n \text{ years}}$$

Summary

- Numeric or graphic methods for displaying data help in
 - Quickly summarizing a dataset
 - And/or presenting results to others
- A dataset can be described numerically in terms of measures of location and a measure of spread
 - Measure of location (Arithmetic mean, Median, Mode, Geometric mean)
 - Measure of spread (Standard deviation, Quantiles, Range)
- Graphic methods include bar graphs and more exploratory methods such as stem-and-leaf plots and box plots
- Ratio, Proportion, and Rate