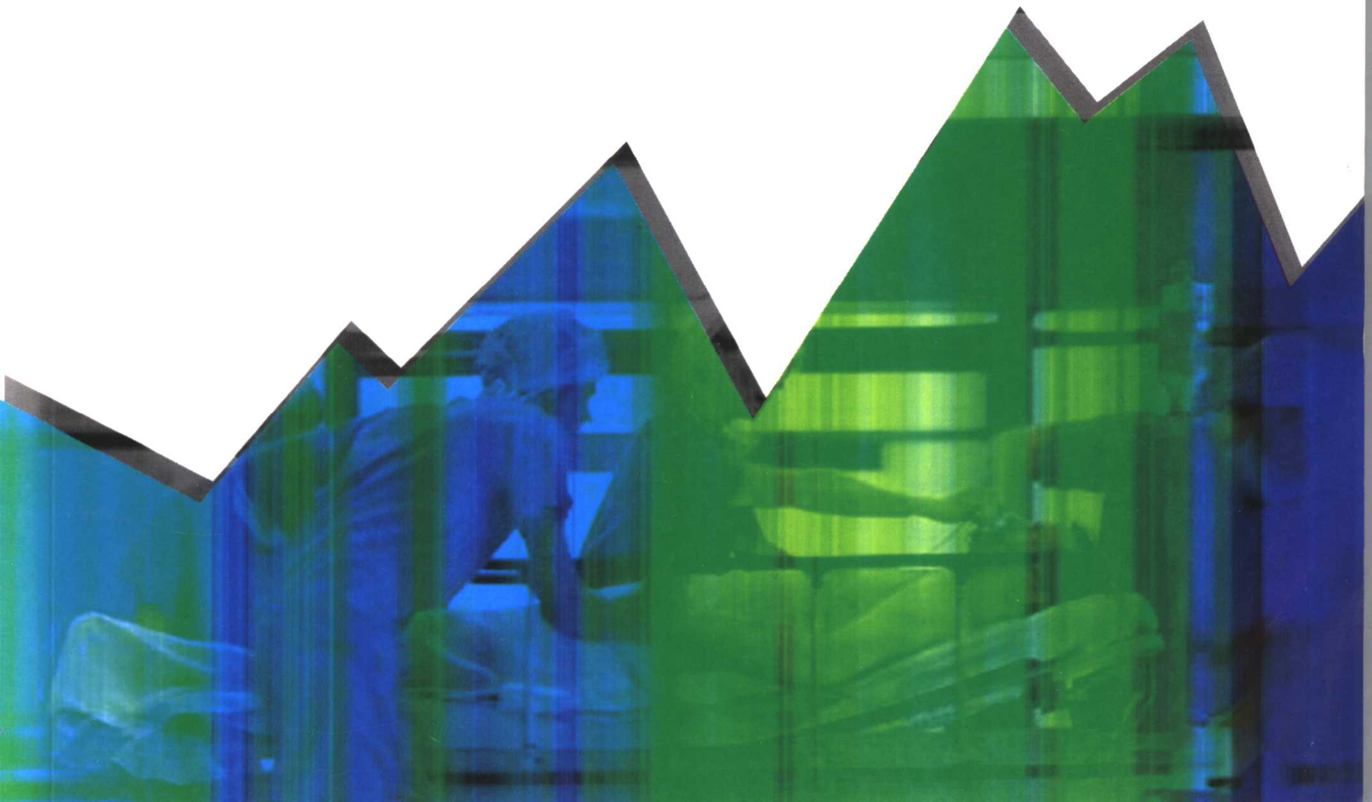
21世纪高等医学院校**双语教学**规划教材

医学统计与流行病学方法

METHODS OF MEDICAL STATISTICS & EPIDEMIOLOGY

主编 雷毅雄

广东科技出版社（全国优秀出版社）
GUANDONG SCIENCE & TECHNOLOGY PRESS



图书在版编目（CIP）数据

医学统计与流行病学方法／雷毅雄主编．—广州：广东科技出版社，2006.2
ISBN 7-5359-4054-4

Ⅰ．医…　Ⅱ．雷…　Ⅲ．①医学统计—双语教学—医学院校—教材—英、汉②流行病学—研究方法—双语教学—医学院校—教材—英、汉　Ⅳ．①R195.1②R181.2

中国版本图书馆 CIP 数据核字（2006）第 014823 号

出版发行：广东科技出版社
（广州市环市东路水荫路 11 号　邮码：510075）
E - mail：gdkjzbb@21cn.com
http：//www.gdstp.com.cn
经　　销：广东新华发行集团股份有限公司
排　　版：广东科电有限公司
印　　刷：肇庆市科建印刷有限公司
（广东省肇庆市星湖大道　邮码：526060）
规　　格：787mm×1 092mm　1/16　印张 13　字数 260 千
版　　次：2006 年 2 月第 1 版
2006 年 2 月第 1 次印刷
定　　价：30.00 元

如发现因印装质量问题影响阅读，请与承印厂联系调换。

前 言

为适应教育国际化的需要，切实加强本科教育工作，提高教学质量，教育部早在教高［2001］4号文件《关于加强高等学校本科教学工作提高教学质量的若干意见》中便指出要积极推动使用英语等双语教学（bilingual teaching），力争3年内双语教学达到所开课程的5%～10%。2002年教育部高教司在下发的《普通高等学校本科教学水平评估方案（试行)》中，第一次将双语教学的课程定义为使用外语教材并用外语授课的课时占该课程课时的50%以上的课程。因此，双语教学正逐渐步入各高校，已经成为中国高等教育发展的一个趋势。

高等医学院校实行双语教学，可为医学生这一特殊群体及时获取医学领域最新信息奠定了坚实的基础，是培养高级医学人才、使我国的生命科学尽快与国际接轨的需要。开展双语教学必须要有相应的教材，但目前国内尚没有统一的双语教材或英语班的外文教材，而外文原版教材与中文教材又不能相互取代，为使学生既能完成教学大纲所要求的内容，又能得到规范的医学专业外语训练，我们按照国家教学大纲的要求，参考了多本外文原版教材，并结合国内学科内容特点和发展方向，编写出这本“医学统计与流行病学方法”双语教学用书。

全书英汉对照，包括三编21章和附录，既涵盖教学大纲的要求，又尽量将目前的新观点、新技术讲授给学生，使学生在有限的时间内既掌握基本理论知识，又掌握最新、最实用的专业知识。本书的编写是作为高等医学院校本科生的医学统计与流行病学方法的双语教学之用书，也为广大医学院校教师和医疗、卫生工作者提供了一本切合实际的参考书。

本书在编写过程中，得到了多位专家教授的指导，特别是美国的统计学与流行病学博士Yuanhui Huang, M.P.H. & Ph.D.的大力支持和热情帮助。此外，广东科技出版社为本书的出版提供了良好的条件。在此，一并表示衷心的感谢！同时，也向本书有关参考文献的作者们致谢！由于时间仓促和水平所限，书中错误和疏漏之处在所难免，恳请同行专家和广大读者批评指正。

编 者

2006年1月于广州

Contents

Part Ⅰ Medical Statistics

Chapter 1
Introduction of Medical Statistics

1. Definition of Statistics

Auording to Oxford English Dictionary says that the word *statistics* came into use more than 200 years ago. At that time, statistics referred to a country's quantifiable political characteristics—characteristics such as population, taxes, and area. Statistics meant "state number." Tables and charts of those numbers turned out to be a most satisfactory method for understanding a country better, for comparing different countries, and for making projections about the future. Later, the word *statistics* was used to refer to tables and charts complied by people studying trade (economics) and natural phenomena (science).

Today the word *statistics* is as follows:

Statistics: The discipline of dealing with variation in data through collection, classification and analysis in such a way as to obtain reliable results

Medical statistics: Application of mathematical statistics in the field of medicine

2. Statistics and Medicine

Evidence-based practice is the new watchword in every profession concerned with the treatment and prevention of diseases and promotion of health and well-being. This requires both the gathering of evidence and its critical interpretation. The former brings more people into the practice of research, and the latter requirs the ability of all health professionals the ability to evaluate the research carried out. Much of this evidence is in the form of numerical data. The essential skill required for the collection, analysis, and evaluation of numerical data is statistics. Thus Statistics, the science of assembling and interpreting numerical data, is the core science of evidence-based practice.

Statistics has not always been so popular with the medical profession. Statistical methods were first used in medical research in the 19th century by workers such as Pierre-Charles-Alexandre Louis, William Farr, Florence Nightingale and John Snow. Snow's studies of the modes of communication of cholera, for example, made use of epidemiological techniques upon which we have still made little improvement Despite the work of these pioneers, however, statistical methods did not become wildly used in clinical medicine until the middle of the 20th century. It was then the methods of randomized experimentation and statistical analysis based on sampling theory, which had been developed by Fisher and others were introduced into medical research, notably by Bradford Hill. It rapidly became apparent that research in medicine raised many new problems in both design and analysis, and much work has been done since towards solving these problems by clinicians, statisticians and epidemiologists.

Although considerable progress has been made in such fields as the design of clinical trials, there remains much to be done in developing research methodology in medicine. It seems likely that this will always be so, for every research project is something new, something which has never been done before. Under these circumstances we make mistakes. No research is perfect and there will always be something which, with hindsight, we would have changed. Furthermore, it is often from the flaws in a study that we learn most about research methods. For this reason, the work of several researchers is described in this book to illustrate the problems into which their designs or analyses led them. I do not wish to imply that these people were any more prone to error than the rest of the human race, or that their work was not a valuable and serious undertaking. Rather I want to learn from their experience of attempting something extremely difficult, trying to extend our knowledge, so that researchers and consumers of research may avoid these particular pitfalls in the future.

3. Basic Contents

- Statistical
  - Design of study
    - Professional
    - Statistical design
  - Data treatment
    - Descriptive statistics
    - Inferential statistics

Basic contents of statistics mainly include scientific design and data process (treatment), and data process mainly includes descriptive statistics and inferential statistics.

Descriptive statistics is the techniques we use to describe the main features of a sample. For example, we described the average number of times the children in the sample brushed their teeth.

Inference statistics is the process of using the value of a sample statistic to make an informed guess about the value of a population parameter. For example, we used the value of the sample statistical average number of times a day that teeth are brushed to make an informed guess of the value of the population parameter average number of times a day that teeth are brushed.

4. Basic Concept

(1) Population and sample:

1) Population is the whole collection of every member of a defined group of interest.

We might define a population as "all children aged between five and ten with caries living in Leeds". A particular characteristic (or variable) of the population that we wish to know about is called a *population parameter*. If we want to know how often they brush their teeth we could ask every child with caries in this age group how often they brush their teeth and calculate the average*; *average number of times a day that teeth are brushed* is thus the population parameter. This is clearly impractical, so we study a *sample* of them.

2) Sample is the section (a representative part) of a population that we actually study.

We might decide to select 50 children aged between five and ten in Leeds with caries and ask them how often they brush their teeth. (The value of a particular characteristic of a sample is called the *sample statistic*.) If the average number of times these children brushed their teeth was 1.7 then we might conclude that children in Leeds aged between five and ten brush their teeth on average *about* 1.7 times a day.

The way to make the sample representative is the following:

A: Randomization B: Enough members

(2) Parameter and statistic:

1) Parameter is the measure of the population or of the distribution of population.

Parameter is actually the index which is used to describe the population, and usually presented by Grecian letters such as μ, σ, π, etc.

2) Statistic is the measure of the sample or of the distribution of sample.

Statistic is actually the index that is used to describe the sample, and usually presented by Latin letters such as $\bar{x}$, s, p, etc.

(3) Variable and variable value:

Variable is the general characteristic being measured on a set of people, objects, or events, the members of which may take on different values. It means that variable is the index of observation unit. The different values of observation unit are usually called variable values.

(4) Error and sampling error:

Error is the difference between observation and true value. There are three kinds of error in statistics:

1) Systematic error, usually caused by artificial factors. It must be avoided in statistical work.
2) Random measurement error, caused by chance. It cannot be avoided but may be reduced by repeating measurement.
3) Sampling error, caused by sampling. It cannot be avoided but may be reduced with increasing members of samples, and may be estimated by statistics.

Sampling error

The statistic of sample is different from the parameter of population and this difference is the sampling error. The observations among individuals are various. The subjects sampled from the same population are various. So the statistics of different samples from the same population are various.

(5) Probability and random event:

1) Probability is a measurement for the possibility of occurrence of a random event.
2) Random event is that an event may occur in one experiment. Nobody is sure whether the event occurs or not before the experiment. However, there is a rule in a large number of experiments.

The probability of an event is denoted by p. Probabilities are usually expressed as decimal fractions, not as percentages, and must lie between zero (zero probability) and one (absolute certainty). The probability of an event cannot be negative.

If the event A always occurs, p (A) = 1; If the event A never occurs, p (A) = 0; If the event A occurs by chance, p (A) = 0 ~ 1.

In statistics, we usually use $p \leqslant 0.05$ or $p \leqslant 0.01$ mean that an event hardly occurs in one experiment.

Supposed n means the number of observations and m means the number of occurrence of random event A, m/n is the frequency. The m/n is close to p (A) only when n is large enough.

For example, if a fair coin was tossed an infinite number of times, heads would appear on 50% of the tosses; therefore, the probability of heads, or p (heads), is 0.50. If a random sample of 10 people was drawn an infinite number of timers from a population of 100 people, each person would be included in the sample 10% of the time; therefore, p (being included in any one sample) is 0.10.

5. Types of Data

The choice of an appropriate statistical technique depends on the types of data in question. So it is important to be able to distinguish different types of data from one another as we use different techniques to describe and analyses the different types. There are three kinds of data types:

(1) Numerical variable (measurement data): obtained by quantitatively measuring a characteristic of each individual.

(2) Categorical variable (enumeration data): obtained by classifying the individuals according to characteristic of individuals and then counting the number of individuals in each category.

(3) Ordinal categorical variable (ranked dada): obtained by the same way as enumeration data but the categories are of certain sequence category.

Types of data may be changed one another according to the research aim, demand, etc. For example: Protein in urine

Protein in urine = 0.01g/100ml (measurement data); 10% of the individuals are with protein in urine (enumeration data); Protein in urine is +++ (ranked dada)

6. Steps of Statistical Work

(1) Design of study: It is the key step in statistical work. It includes professional design (including as the research aim, subjects, measures, etc.) and statistical design (including sampling, size method, randomization, etc.)

(2) Data collection: Data source includes statistical report forms (regularity), medical records (regularity), and special study or investigation (temporary).

(3) Data sorting: It means data treatment and grouping. It is used for convenience of data analysis.

(4) Data analysis: It includes descriptive statistics and inferential statistics.

7. Methods of Study for Medical Students

(1) Making sense of basic knowledge of medical statistics.

(2) Knowing essential concepts and thinking of statistics.

(3) Training the ability of statistics thinking and the kill of treatment of medical data.

(4) Exercising the application of statistical methods and usage of some statistical software in the practice.

(5) Understanding the principle of investigation design and experiment design.

Chapter 2
Description of Measurement Data

1. Distribution of Frequency

One skill of statistics is to summarize and to present the data in a clear and easy understandable way. It is the work of statistical description.

To summarize the data or describe the distribution of frequency, frequency table or frequency graph is the common way

- Normal distribution (see Table 1-2-3 and Fig. 1-2-1)
- Skewed distribution
  - Positively Skewed distribution (see Table 1-2-1)
  - Negatively Skewed distribution (see Table 1-2-2)

Table 1-2-1 Frequency Distribution of Hair Hg Value (μg/g) among 238 Normal Adults

Value of hair Hg	Frequency	AF*	AF (%)
0.3~	20	20	8.4
0.7~	66	86	36.1
1.1~	60	146	61.3
1.5~	48	194	81.5
1.9~	18	212	89.1
2.3~	16	228	95.8
2.7~	6	234	98.3
3.1~	1	235	98.7
3.5~	0	235	98.7
3.9~	3	238	100.0

* AF: Accumulative Frequency

Table 1-2-2 Frequency Distribution of Patients who die of Malignant Tumors in some year and district.

Age (yr.)	No. of Death	AF*	AF (%)
0~	5	5	0.42
10~	12	17	1.41
20~	15	32	2.66
30~	76	108	8.98
40~	189	297	24.69
50~	234	531	44.14
60~	386	917	76.23
70~	286	1 203	100.00

* AF: Accumulative Frequency

Normal distribution will be discussed in the following parts of this chapter. Here we will give you a definition of skewed distribution. A distribution that has the central location to the left and a tail off to the right is said to be "positively skewed" as in table 1-2-1. A distribution that has the central location to the right and a tail off to the left is said to be "negatively skewed" as in Table 1-2-2.

However, it is not enough. If we are faced with a large amount of data, we may want to describe its more important features more concisely. We need to summarize the data further. Usually we describe the data from two aspects. One is the Measures of Location, meaning the central tendency (central position), and another is the Measures of Spread, meaning the variation (tendency of dispersion).

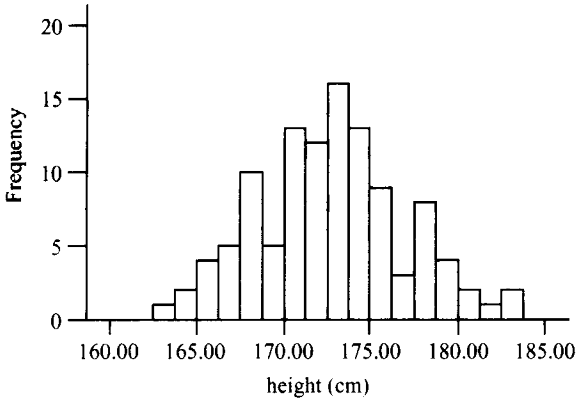


Fig. 1-2-1 Frequency Distribution of the heights (cm) of 110 health male students at the age of 20 at a small city in 2004

2. Average

What measure is used to describe the central tendency or central location (clustering)? It is the average including mean, geometric mean and median. *A measure of central location is the single value that best represents a characteristic such as age or height of a group of persons.*

$$\text{Average}\begin{cases}\text{Mean, symbolized by } \mu,\ \bar{x}\\ \text{Geometric mean, symbolized by G}\\ \text{Median, symbolized by M}\end{cases}$$

(1) Mean (μ, $\bar{x}$), also called arithmetic mean is the sum of all the values divided by the number of cases. It is suitable to the data distributed in normal distribution or at least symmetric distribution. Mean is symbolized by μ in a population, and by $\bar{x}$ (x-bar) in a sample. They can be calculated by direct method and weighing method.

The following formula is for original data (direct method):

$$\bar{X} = \frac{X_1 + X_2 + \cdots\cdots + X_n}{n} = \frac{\sum X}{n}$$

For example: There are 11 health male students. Their body heights are 174.9, 173.1, 171.8, 179.0, 173.9, 172.7, 166.2, 170.8, 171.8, 172.1, 168.5, respectively. Please calculate the average of student's heights.

$$\bar{X} = \frac{\sum X}{n} = \frac{174.9 + 173.1 + \cdots\cdots + 168.5}{11} = \frac{1894.8}{11} = 172.25(\text{cm})$$

The following formula is for frequency table (weighing method).

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \cdots\cdots + f_n X_n}{f_1 + f_1 + \cdots\cdots + f_n} = \frac{\sum fX}{\sum f}$$

For example: Someone randomly measure the heights (cm) of 110 health male students at the age of 20 at a city in 2004, the data is as follows, please calculate the mean of heights (cm) for these students.

173.9	173.9	166.9	179.5	171.2	167.8	177.1	174.7	173.8	182.5
173.6	165.8	168.7	173.6	173.7	177.8	180.3	173.1	173.0	172.6
173.6	175.3	178.4	181.5	170.5	176.4	170.8	171.8	180.7	170.7
173.8	164.4	170.0	175.0	177.7	171.4	162.9	179.0	174.9	178.3
174.5	174.3	170.4	173.2	174.5	173.7	173.4	173.9	172.9	177.9
168.3	175.0	172.1	166.9	172.7	172.2	168.0	172.7	172.3	175.2
171.9	168.6	167.6	169.1	166.8	172.0	168.4	166.2	172.8	166.1
173.5	168.6	172.4	175.7	178.8	169.1	175.5	170.3	171.7	164.6
171.2	169.1	170.7	173.6	167.2	170.7	174.7	171.8	167.3	174.8
168.5	178.7	177.3	165.9	174.0	170.2	169.5	172.1	178.2	170.9
171.3	176.1	169.7	177.9	171.1	179.3	183.5	168.5	175.5	175.9

(a) Making frequency table

a) Seeking range (R): R = Maximum value-Minimum value = 183.5 - 162.9 = 20.6

b) Deciding interval and group: i = R/k = 20.6/10 = 2.6≈2.0

i means interval, usually 1/10 of range; k means numbers of groups, usually 8 ~ 15 groups. The starting point (lower limit) of the first group must include the minimum value; the ending point (upper limit) of the last group must include the maximum value.

c) Marking sign and frequency: As in Table 1-2-3, columns (1), (2) and (3)

(b) Calculating mean using weighing method:

As in Table 1-2-3, columns (4) and (5).

Midpoint: meaning middle value within group. It is a half of upper limit plus lower limit

$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1900}{110} = 172.73$ (cm) The mean of heights (cm) among 110 health male students at the age of 20 at a city in 2004 is 172.73 (cm)

Table 1-2-3 Frequency Distribution of the heights (cm) of 110 health male students at the age of 20 at a small city in 2004

Height (1)	Mark (2)	Frequency (f) (3)	Midpoint (X) (4)	f x (5)
162 ~	|	1	163	163
164 ~	| | | |	4	165	660
166 ~	正 | | | |	9	167	1 503
168 ~	正正 | | |	13	169	2 197
170 ~	正正正 | | | |	19	171	3 249
172 ~	正正正正正 | |	27	173	4 671
174 ~	正正正 |	16	175	2 800
176 ~	正 | | |	8	177	1 416
178 ~	正 | | |	8	179	1 432
180 ~	| | |	3	181	543
182 ~ 184	| |	2	183	366
Total	—	110	—	19 000 ($\sum$f x)

(2) Geometric mean (G), is calculated by the same formula as for mean and the only difference is transform the value into logarithm when the calculation. It is suitable to the data distributed in positive skewed distribution or logarithm normal distribution. Geometric mean can be calculated by direct method and weighing method.

The following formula is for original data (direct method):

$$G = \sqrt[n]{X_1 \cdot X_2 \cdot \cdots\cdots \cdot X_n} \quad \text{or} \quad G = \lg^{-1}\left(\frac{\lg X_1 + \lg X_2 + \cdots\cdots + \lg X_n}{n}\right) = \lg^{-1}\left(\frac{\sum \lg X}{n}\right)$$

For example: There are 6 items of serum antibodies, their concentrations respectively are 1:10, 1:20, 1:40, 1:80, 1:80, 1:160, please calculate their average concentration.

$$G = \lg^{-1}\left(\frac{\lg 10 + \lg 20 + \lg 40 + \lg 80 + \lg 160}{6}\right) = \lg^{-1}(1.6522) = 45$$

Geometric mean (G) of serum antibodies' concentrations is 1:45

The following formula is for frequency table (weighing method).

$$G = \lg^{-1}\left(\frac{\sum f \lg X}{\sum f}\right)$$

For example: One month later when 30 susceptible children immunized with measles vaccine, their antibodies' concentrations of blood coagulation inhibition are as follows (see Table 1-2-4), please calculate their average antibodies' concentration.

$$G = \lg^{-1}\left(\frac{\sum f\lg X}{\sum f}\right) = \lg^{-1}\left(\frac{50.5728}{30}\right) = 48.5$$

Geometriç mean (G) of the antibodies' concentrations of blood coagulation inhibition after one month when 30 susceptible children immunized with measles vaccine is 1:48.5

Table 1-2-4 The antibodies' concentrations of blood coagulation inhibition after one month when 30 susceptible children immunized with measles vaccine

Antibodies' concentrations (1)	Children (f) (2)	Reciprocal of concentrations X (3)	lgX (4)	flgX = (2) × (4) (5)
1:8	2	8	0.9031	1.8062
1:16	6	16	1.2041	7.2246
1:32	5	32	1.5051	7.5255
1:64	10	64	1.8062	18.0620
1:128	4	128	2.1072	8.4288
1:256	2	256	2.4082	4.8164
1:512	1	512	2.7093	2.7093
Total	30 (Σf)			50.5728 (ΣflgX)

(3) Median (M), is the value of observation located in the middle of value sequence of observations. Median is suitable to all kinds of data but it is poor attribution for further analysis comparing to mean. Median can be calculated by direct method and frequency table method or percentile method.

The following formula is for original data (direct method):

$M = X_{(\frac{n+1}{2})}$ n = odd number

$M = \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}]$ n = even number

For example: There are 9 cases, the latent period of them is 2, 3, 3, 3, 4, 5, 6, 9, 16 days, please calculate their average latent period.

In this case, n = 9 is an odd number, so we can calculate the Median:

$$M = X_{(\frac{n+1}{2})} = X_5 = 4\text{days}$$

For example, in the example above, adding one case whose latent period is 20 days, so that the average latent period is 4.5 days.

$$M = \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] = \frac{1}{2}(X_5 + X_6) = \frac{1}{2}(4 + 5) = 4.5\text{days}$$

Note: we must sort the date before calculate the Median.

For the data from a frequency table, we do not know the exactly value of median, so the following formula of median or percentile is for frequency table (frequency table method or percentile method)

$$P_X = L_X + \frac{i_X}{f_X}(n \times X\% - \sum fL)$$

In this formula: X means percentile; L_X means the low limit of group where percentile located in; i_x means the interval of the x group; fx means frequency in the x group; n means the total cases; ΣfL means accumulative frequency that less than L.

If Px = 50% = M, we can calculate the median by following formula:

$$M = L_M + \frac{i_M}{f_M}(\frac{n}{2} - \sum fL)$$

For example, as in Table 1-2-2, Frequency Distribution of Hair Hg Value (μg/g) among 238 Normal Adults, we can calculate the median and percentile of Hair Hg Value (μg/g).

Median calculation, from Table 1-2-2, accumulative frequency 50% is within the group "1.1 ~", $L_M = 1.1$, $i_M = 0.4$, $f_M = 60$, ΣfL = 86, n = 238

$$M = L_M + \frac{i_M}{f_M}(\frac{n}{2} - \sum fL) = 1.1 + \frac{0.4}{60}(\frac{238}{2} - 86) = 1.32(\mu g/g)$$

Percentile calculation, when P_{95}, x = 95, accumulative frequency 95% is within the group "2.3 ~", L = 2.3, i = 0.4, fx = 16, ΣfL = 212, n = 238

$$P_{95} = 2.3 + \frac{0.4}{16}(238 \times 95\% - 212) = 2.65(\mu g/g)$$

3. Measures of Spread or Dispersion

There are some features to describe the distribution of different data. Two common features we might be interested in are:

What is the typical (average) value of a variable (what is its location)?

How much variability is there in the data (how much does it spread out)?

Except the central tendency, we need to describe the dispersion of data by the measures of variability. Measure of dispersion quantifies how much persons in the group vary from each other and from our measure of central location. The common variations are as follows.

Variations
- Range, symbolized by R
- Interval of quartile, symbolized by Q
- Variance, symbolized by σ^2, S^2
- Standard deviation, symbolized by σ, SD or S
- Coefficient of variation, symbolized by CV

(1) Range (R), is the simplest measure of variability, from the smallest to the largest value. Range is suitable to all kinds of data but it is a poor measure of variability because it is based on only two extreme observations, it ignores the distribution of observation within the two extremes, and the greater the number of observation, the greater the range.

For example, there are three groups of boys at the same age, their average weights are the same 30 (kg), please analyses their variations using Range (R).