

卫生统计学实习

何平平

北京大学公共卫生学院
流行病学与卫生统计学系

Tel: 82801619

实习二

统计描述

第164 ~ 180页

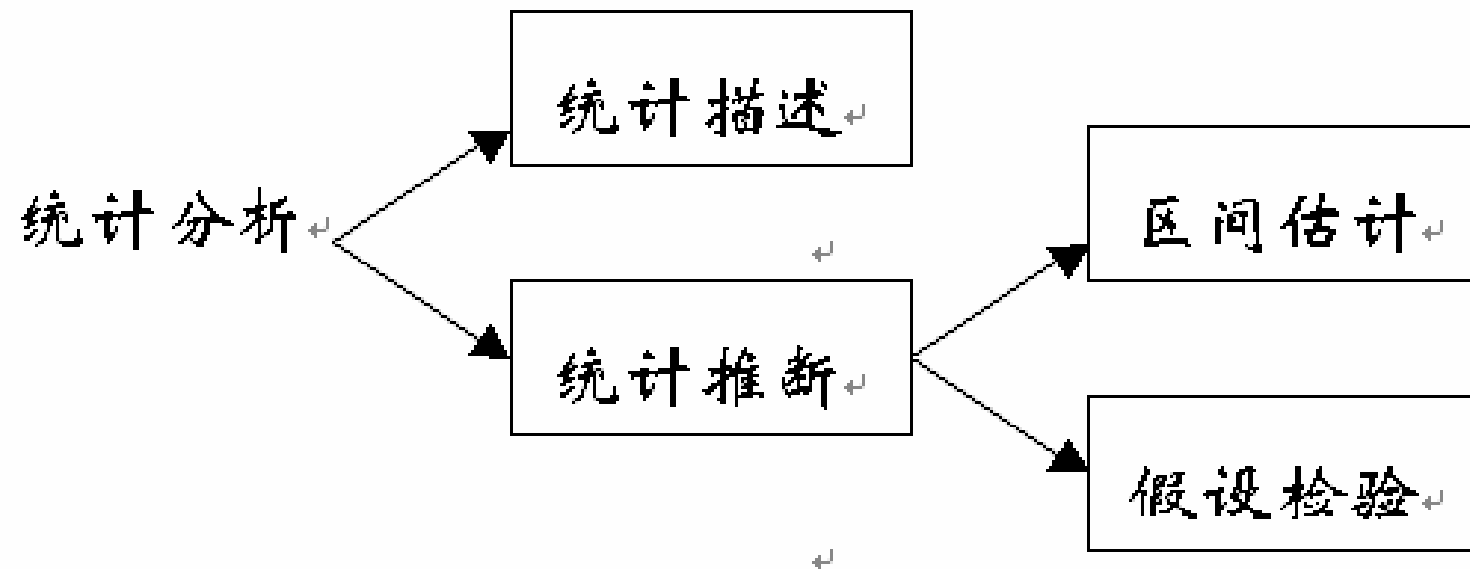
实习二 统计描述

医学统计资料类型

- **数值变量资料**：又称为计量资料。变量值是定量的，有单位的，表示为数值的大小。
- **无序分类资料**：又称为计数资料。变量值是定性的，没有单位，表示为相互独立的类别。
- **有序分类资料**：又称为等级资料。变量值是定性的，没有单位，各类别具有程度上的差异。

注：不同类型的资料，统计方法不同；各种类型的资料之间是可以相互转化的。

一、数值变量资料的统计描述



统计描述包括两个方面：集中趋势的描述和离散趋势的描述

一、数值变量资料的统计描述

(一) 数值变量资料的频数表

频数表 (frequency table)：当变量值或者观测值较多时，将变量值分为适当的组段，统计各组段中相应的频数（或者人数），以描述数值变量资料的分布特征和分布类型。

一、数值变量资料的统计描述

(一) 数值变量资料的频数表

频数表的用途

1. 描述数值变量资料的分布特征

集中趋势 (central tendency)：频数最多的组段代表了中心位置（平均水平），从两侧到中心，频数分布是逐渐增加的。

离散趋势 (tendency of dispersion)：从中心到两侧，频数分布是逐渐减少的。反映了数据的离散程度或者变异程度。

一、数值变量资料的统计描述

(一) 数值变量资料的频数表

频数表的用途

2. 描述数值变量资料的分布类型

正态分布：集中位置居中，左右两侧频数基本对称。常见近似正态分布。

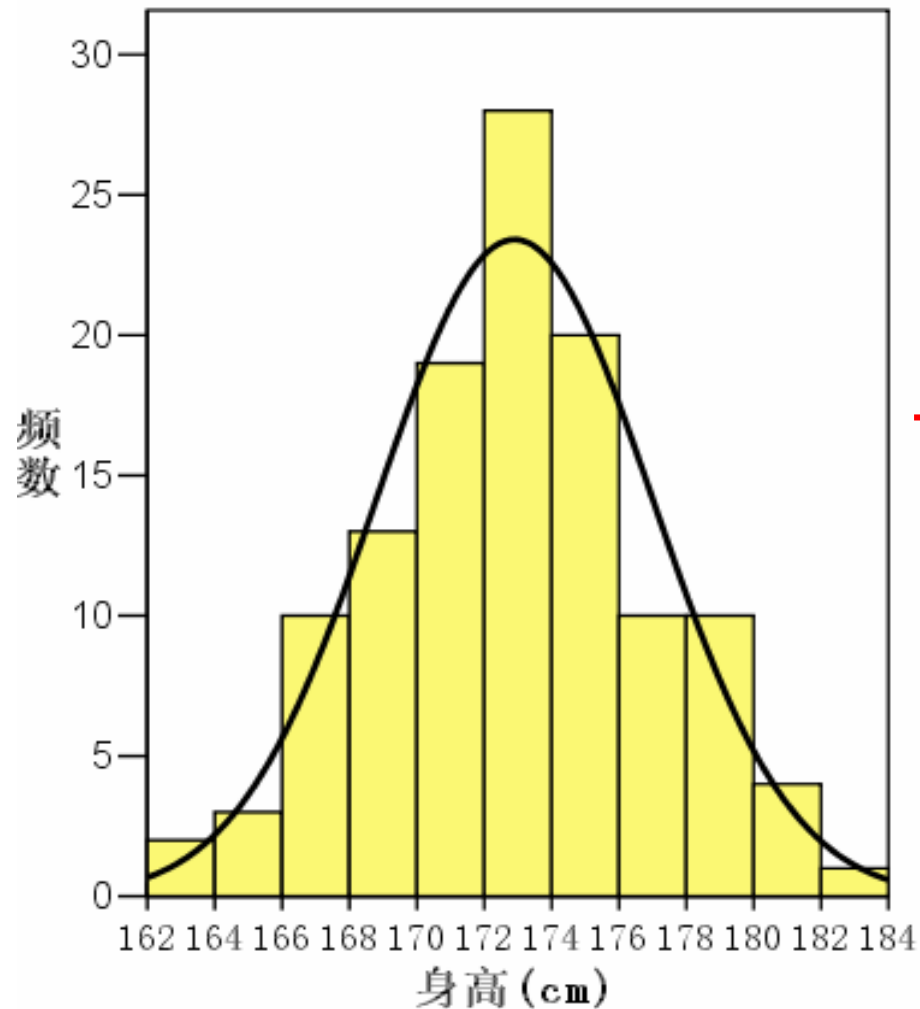
偏态分布：集中位置偏向一侧，频数分布不对称。

正偏态分布：集中位置偏向数值小的一侧或者左侧，有较长的右尾部。

负偏态分布：集中位置偏向数值大的一侧或者右侧，有较长的左尾部。

一、数值变量资料的统计描述

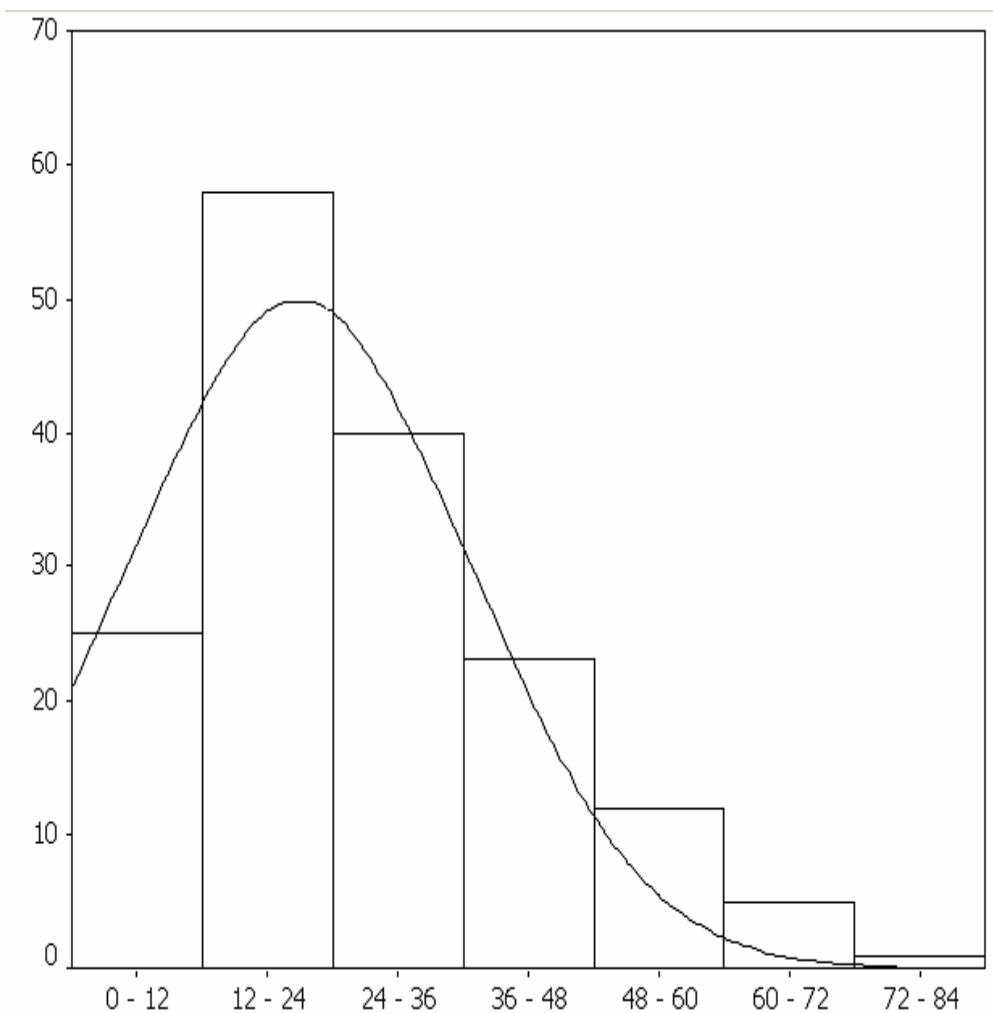
(二) 数值变量资料的频数分布图及正态曲线



→ 直方图及近似正态分布

一、数值变量资料的统计描述

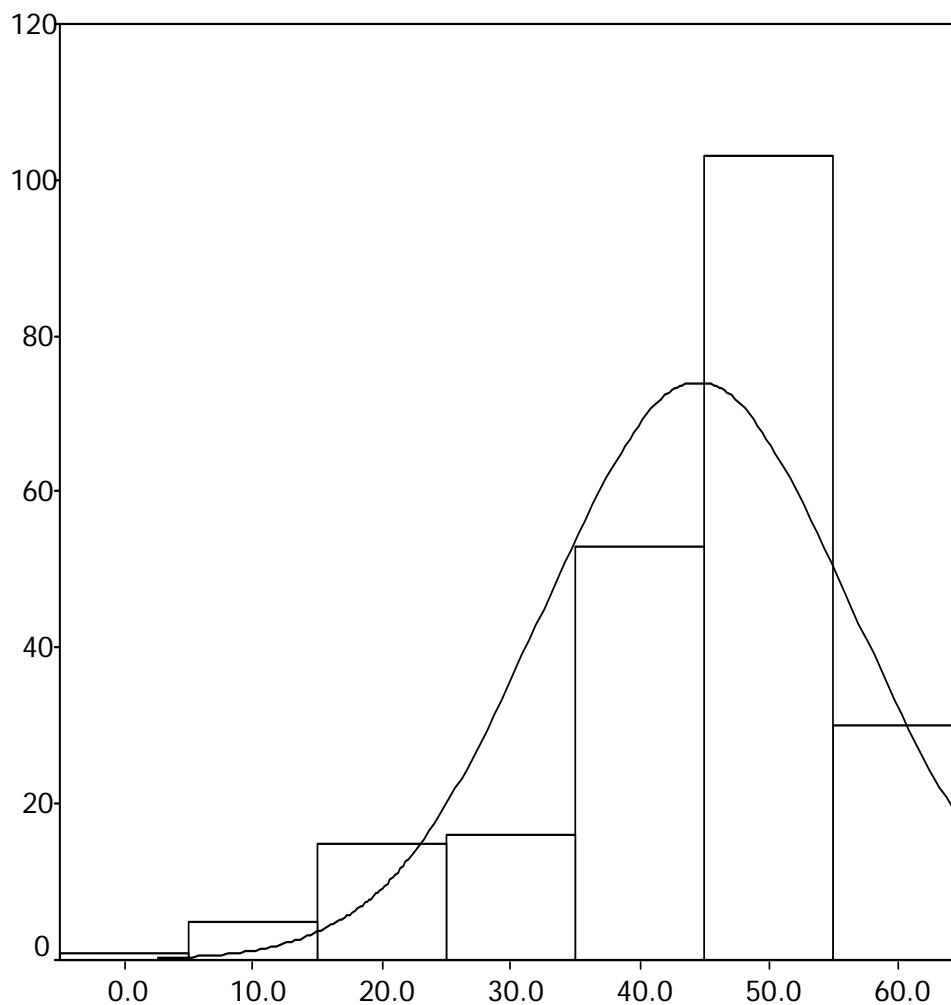
(二) 数值变量资料的频数分布图及正态曲线



→ 直方图及
正偏态分
布

一、数值变量资料的统计描述

(二) 数值变量资料的频数分布图及正态曲线



→ 直方图及
负偏态分
布

一、数值变量资料的统计描述

(三) 集中趋势指标描述

1. 算数均数 (均数 mean)

适用于正态分布或者近似正态分布

总体均数: μ ; 样本均数 \bar{X}

$$\bar{X} = \frac{x_1 + x_2 + x_3 \dots + x_n}{n} = \frac{\sum x}{n}$$

一、数值变量资料的统计描述

(三) 集中趋势指标描述

2. 几何均数 (geometric mean, G)

适用于一种特殊的偏态分布资料：等比资料（常见于抗体滴度）。此资料的原始数据为正偏态分布，取对数后，对数值为正态分布，所以又称为对数正态分布。

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

$$G = \lg^{-1} \left(\frac{\lg x_1 + \lg x_2 + \lg x_3 \cdots + \lg x_n}{n} \right) = \lg^{-1} \left(\frac{\sum \lg x}{n} \right)$$

一、数值变量资料的统计描述

(三) 集中趋势指标描述

3. 中位数 (median, M)

适用于偏态分布资料，或者分布类型未知，或者有不确切数据时。中位数是指将一组变量值从小到大排列，位次居中的变量值。

$$n \text{ 为奇数时, } M = x_{(\frac{n+1}{2})}$$

$$n \text{ 为偶数时, } M = (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) / 2$$

一、数值变量资料的统计描述

(三) 集中趋势指标描述

4. 注意事项

1. 对于偏态分布资料，中位数不受两端特大值和特小值的影响，**只和位置居中的观察值有关**。而均数受特大值和特小值的影响，会偏大或者偏小，所以对于偏态分布的资料，均数的代表性差，不适合描述偏态分布的集中趋势。

2. **中位数适合于任何分布类型的资料**，对于正态分布，理论上中位数等于均数。

一、数值变量资料的统计描述

(四) 离散趋势指标描述

1. 极差或者全距 (range, R) : 表示一组变量值中最大值和最小值之差。适合任何分布类型的资料。

$$R = \text{最大值} - \text{最小值}$$

计算简单, 但是不能反映所有变量值的变异程度, 易受最大值和最小值的影响, 不稳定

一、数值变量资料的统计描述

(四) 离散趋势指标描述

2. 方差 (variance) : 表示一组变量值的平均离散程度。方差越大, 离散或者变异程度越大。适合描述近似正态分布资料的离散趋势。

$$\text{总体方差: } \sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

$X - \mu$ 称为离均差; $\sum(X - \mu)^2$ 称为离均差平方和

$$\text{样本方差: } s^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

一、数值变量资料的统计描述

(四) 离散趋势指标描述

3. 标准差 (standard deviation) : 是方差的开方, 和均数的单位一致, 也表示一组变量值的平均离散程度。

适合描述近似正态分布资料的离散趋势

$$\text{总体标准差: } \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

$$\text{样本标准差: } s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

一、数值变量资料的统计描述

(四) 离散趋势指标描述

4. 四分位数间距 (quartile, Q) : P_{75} 、 P_{25} 分别表示第 75 百分位数和第 25 百分位数。

$$Q = P_{75} - P_{25}$$

注：适合描述任何分布类型资料的离散趋势，主要用于偏态分布资料。

一、数值变量资料的统计描述

(四) 离散趋势指标描述

5. 变异系数 (coefficient of variation, CV)

用于描述数据的相对离散程度。

$$CV = \frac{S}{\bar{X}} \times 100\%$$

CV : 单位不同, 均数相差悬殊

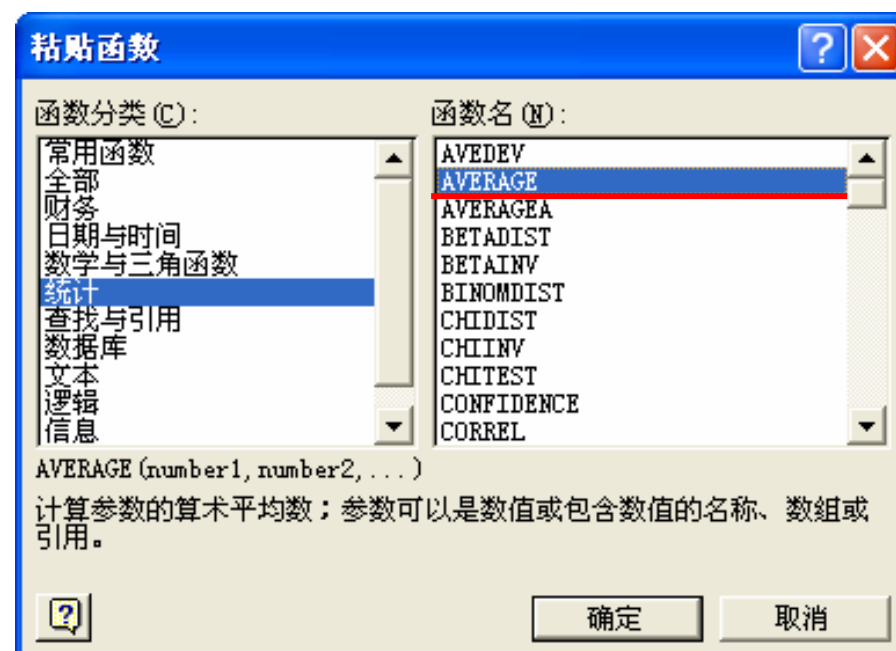
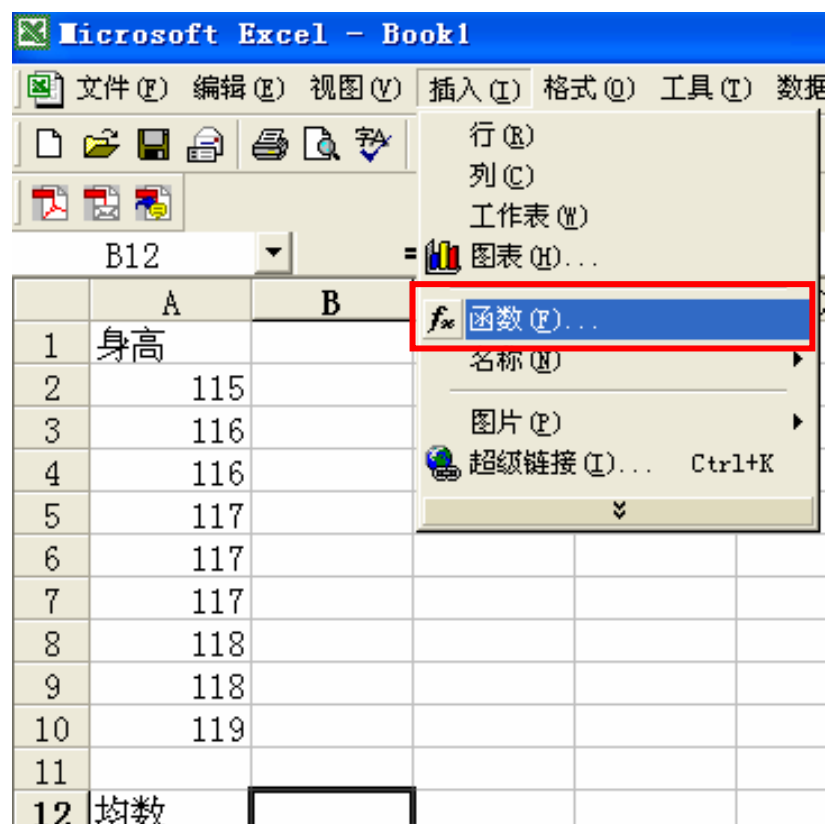
S : 单位相同, 均数相近

一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

1. 计算均数

AVERAGE

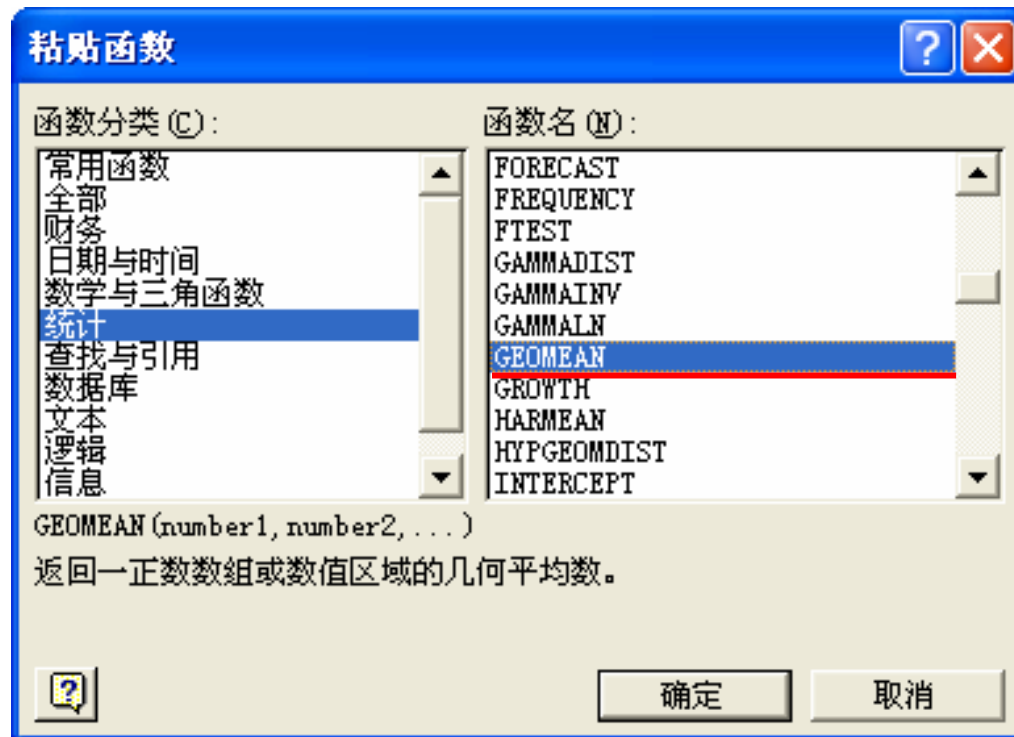


一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

2. 计算几何均数

GEOMEAN

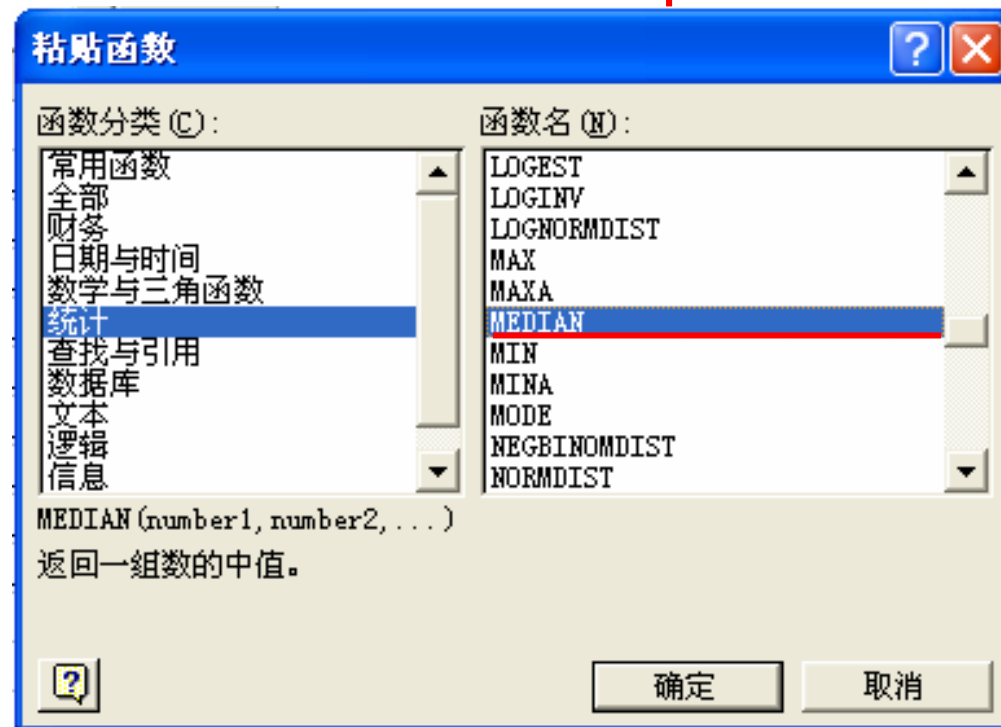


一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

3. 计算中位数

MEDIAN

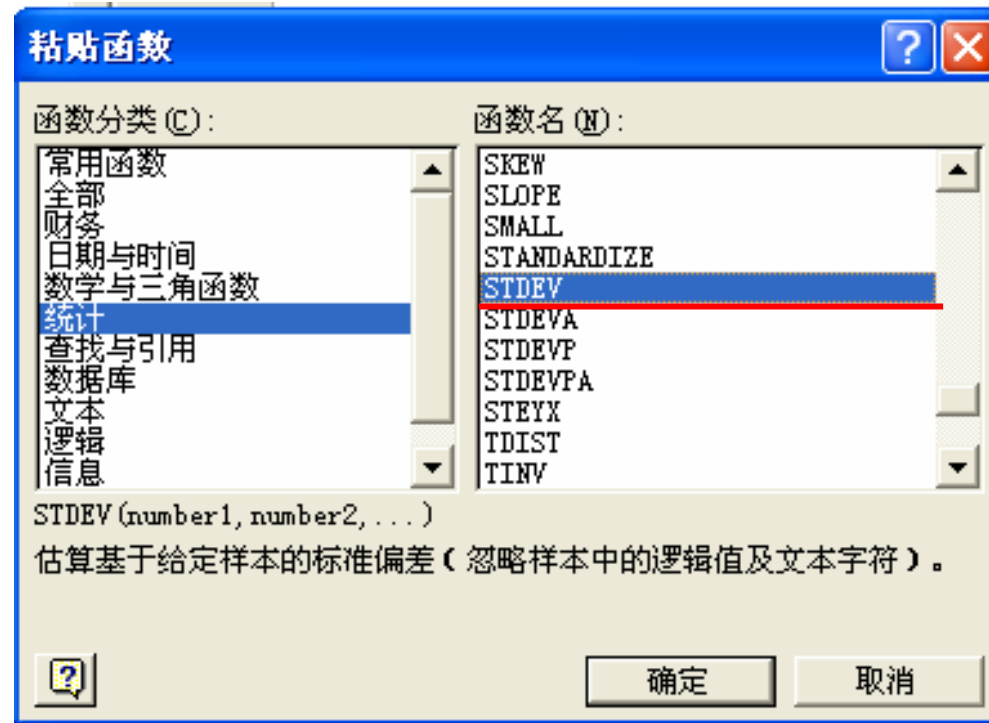


一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

4. 计算样本标准差

STDEV

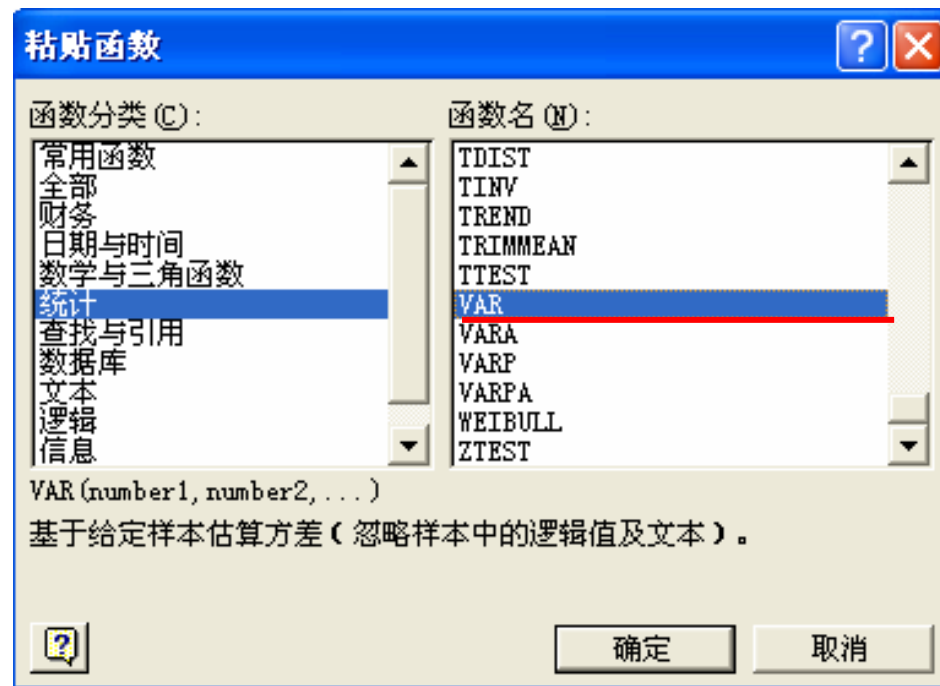


一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

5. 计算样本方差

VAR

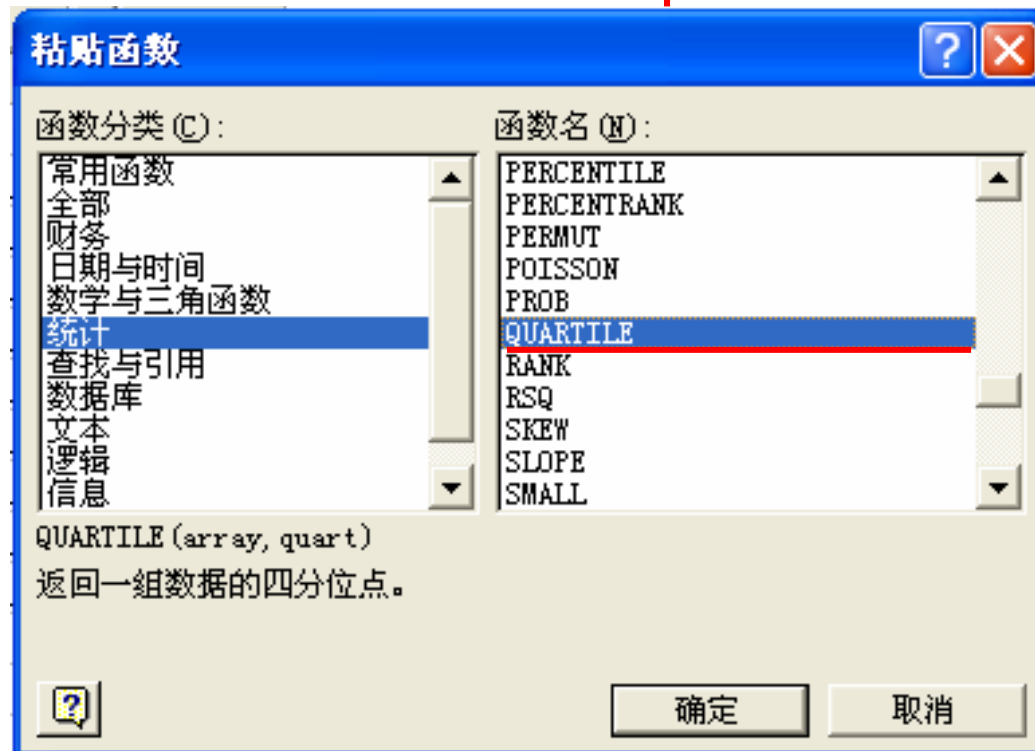


一、数值变量资料的统计描述

(五) 用EXCEL软件实现统计描述

6. 计算四分位数间距

QUARTILE



选项中,

0: 最小值

1: P_{25}

2: P_{50}

3: P_{75}

4: 最大值

一、数值变量资料的统计描述

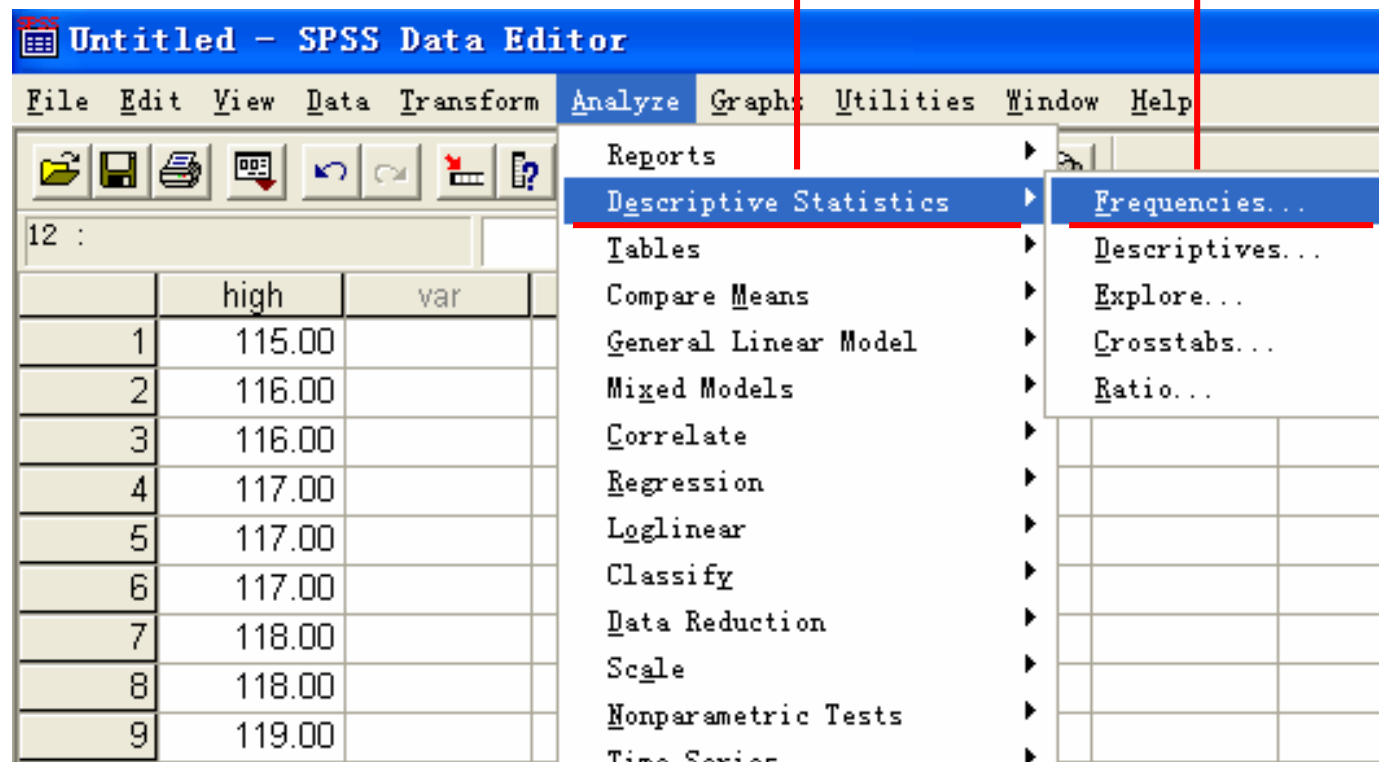
(六) 用SPSS软件实现统计描述

操作步骤:

1. 选择“Frequencies”

描述性统计

频数



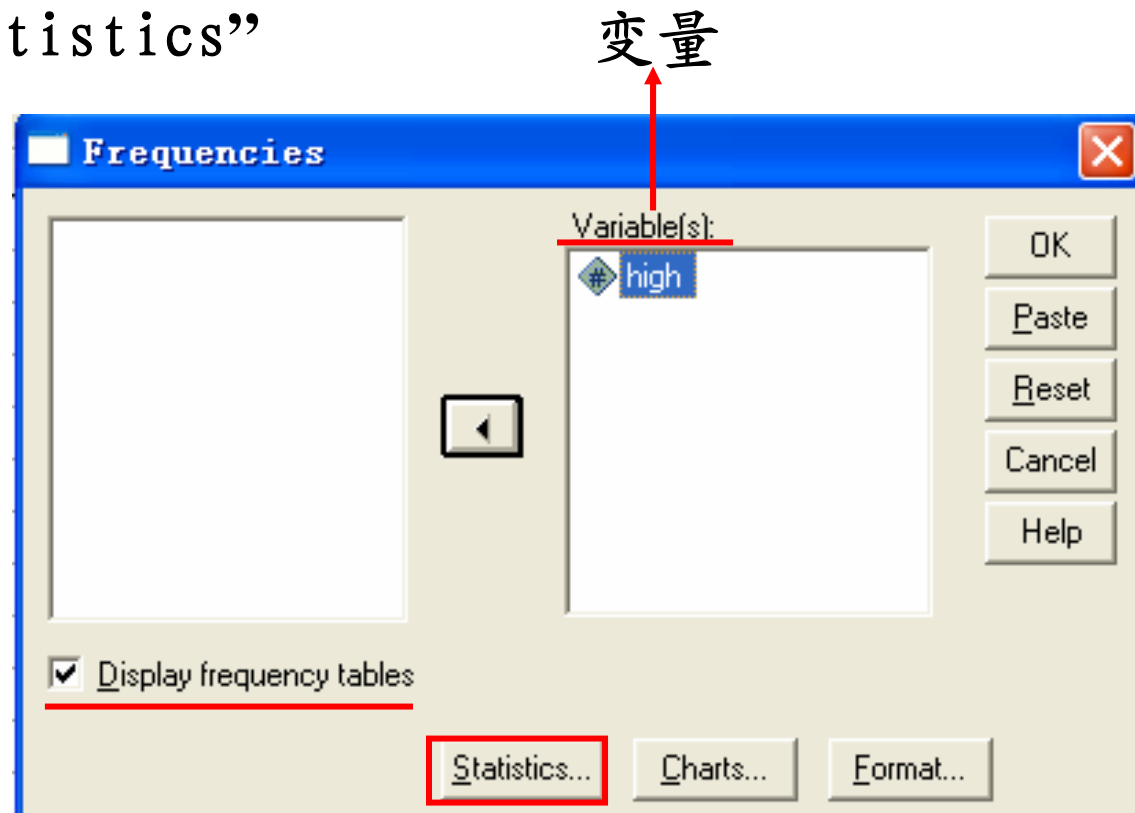
一、数值变量资料的统计描述

(六) 用SPSS软件实现统计描述

操作步骤:

2. 将变量选入变量框,
点击“Statistics”

列出
频数表

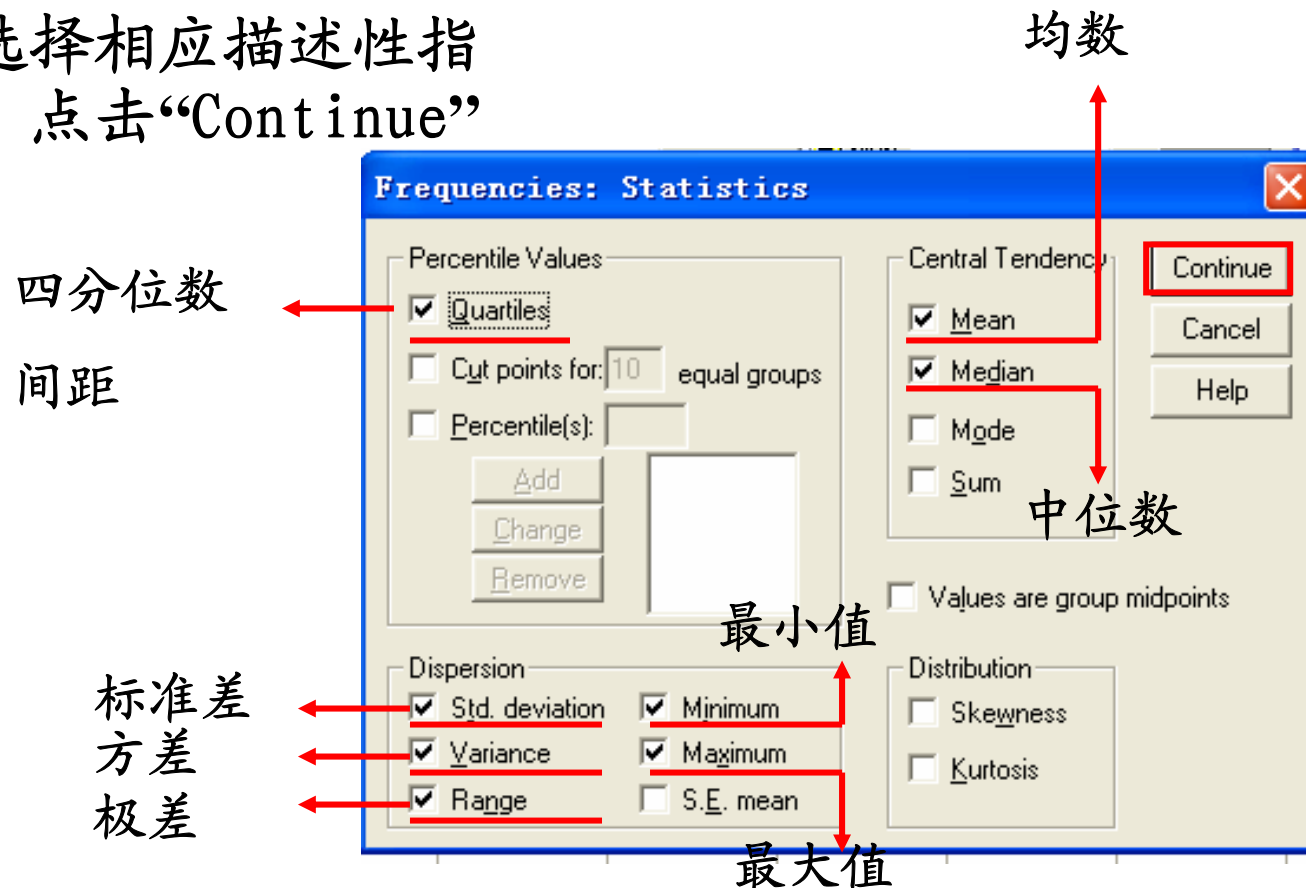


一、数值变量资料的统计描述

(六) 用SPSS软件实现统计描述

操作步骤:

2. 选择相应描述性指标, 点击“Continue”



统计结果

统计描述指标

Statistics		
high		
N	Valid	9
	Missing	0
Mean		117.0000
Median		117.0000
Std. Deviation		1.22474
Variance		1.500
Range		4.00
Minimum		115.00
Maximum		119.00
Percentiles	25	116.0000
	50	117.0000
	75	118.0000

严格来说，本例的例数太少，不适合计算四分位数间距。在此仅为举例

频数

百分比

有效百分比

累计百分比

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	115.00	1	11.1	11.1	11.1
	116.00	2	22.2	22.2	33.3
	117.00	3	33.3	33.3	66.7
	118.00	2	22.2	22.2	88.9
	119.00	1	11.1	11.1	100.0
	Total	9	100.0	100.0	

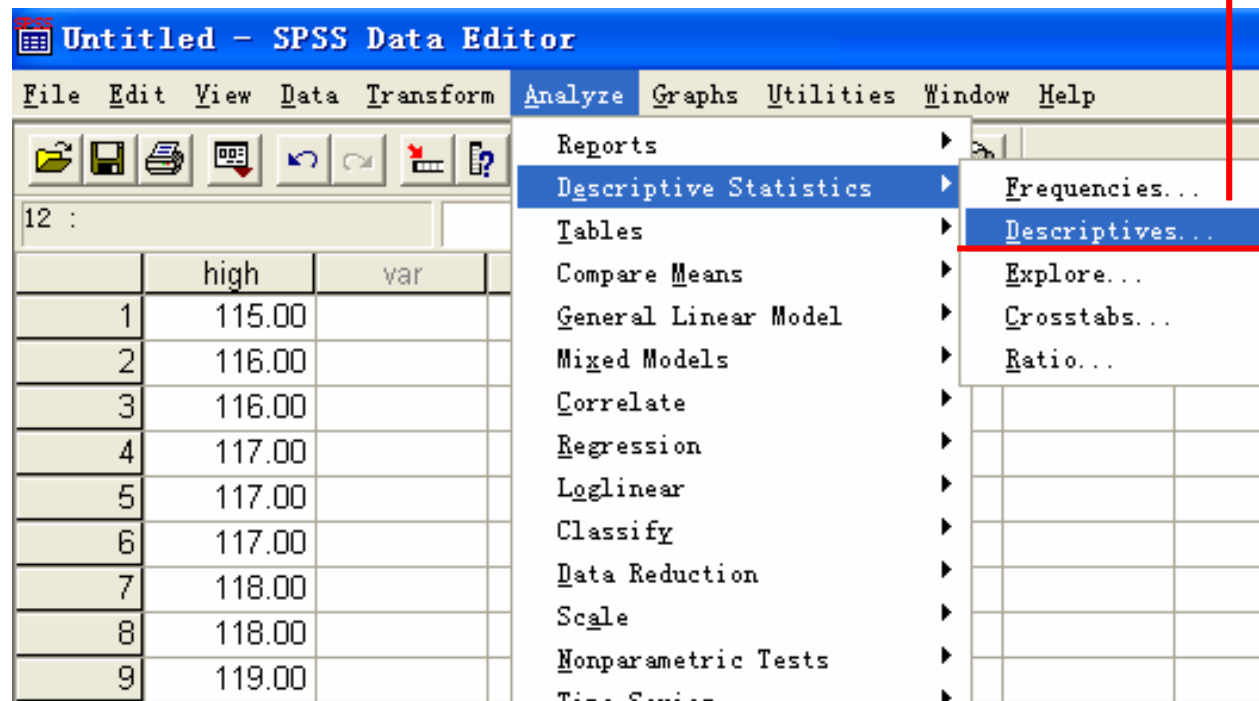
注：对于数值变量资料的原始数据，很少做频数表。在此仅为举例

一、数值变量资料的统计描述

(六) 用SPSS软件实现统计描述

注：除了用“Frequencies”外，还可以使用“Descriptives”进行统计描述

描述



一、数值变量资料的统计描述

(七) 正态分布和医学参考值范围的估计

1. 正态分布的性质

任何正态分布经过 u 变换, 都可以变换为标准正态分布 (u 分布)

$$N(\mu, \sigma) \Rightarrow N(0, 1)$$

$$u = \frac{X - \mu}{\sigma}$$

一、数值变量资料的统计描述

(七) 正态分布和医学参考值范围的估计

2. 医学参考值范围的估计

(1) 定义：同质总体中某研究指标大多数变量值的波动范围。常取95%的医学参考值范围。

(2) 计算：正态分布法

双侧： $\bar{X} \pm 1.96S_x$

单侧： $\bar{X} + 1.64s$ 或 $\bar{X} - 1.64s$

二、分类资料的统计描述

(一) 相对数

1.构成比：某事物中各部分所占的比重。

$$\text{某组成部分的构成比} = \frac{\text{某一组成部分的观察单位数}}{\text{同一事物各组成部分的观察单位总数}} \times 100\%$$

构成比的性质：各部分之和为**100%**；某一部分的比重增加，则相应其它部分的比重减少。

二、分类资料的统计描述

(一) 相对数

2.相对比：又称为比，是两个有关的指标之比。

$$\text{相对比} = \frac{A}{B}$$

3.率：某现象发生的频率或强度。

$$\text{率} = \frac{\text{发生某现象的观察单位数}}{\text{可能发生该现象的观察单位数}} \times 100\%$$

二、分类资料的统计描述

(一) 相对数

4.应用相对数的注意事项

➤ 率和构成比的区别

率：某现象发生的频率或强度。

构成比：某事物中各部分所占的比重。

构成比不能反映事物发生的频率或强度，因为它未考虑人口基数的影响。

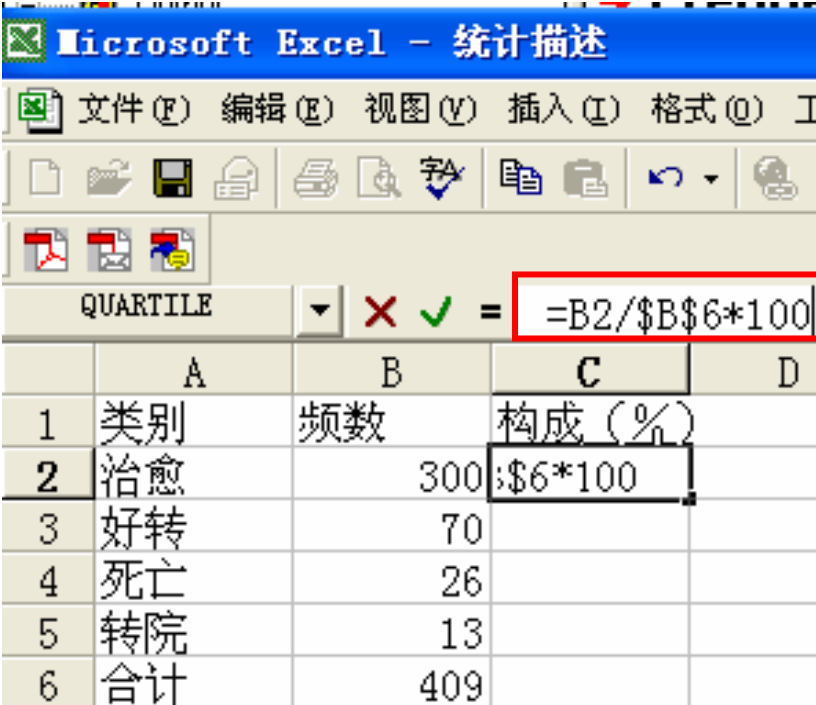
➤ 计算相对数的分母不宜过小。

➤ 率不能直接相加求平均。

二、分类资料的统计描述

(一) 相对数

5.用EXCEL软件实现分类资料的统计描述



	A	B	C	D
1	类别	频数	构成 (%)	
2	治愈	300	$=B2/\$B\$6*100$	
3	好转	70		
4	死亡	26		
5	转院	13		
6	合计	409		

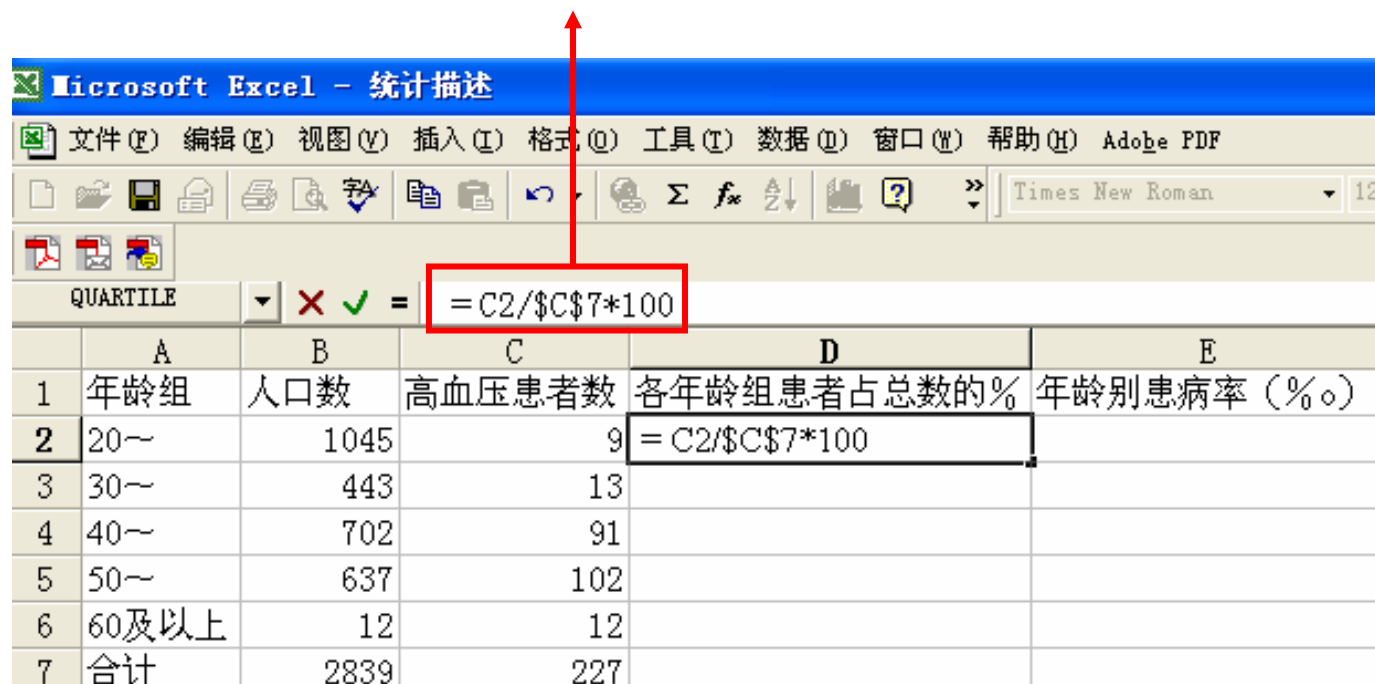
计算构成比，
其中用到“固
定地址\$B\$6”

二、分类资料的统计描述

(一) 相对数

5.用EXCEL软件实现分类资料的统计描述

计算构成比，其中用到“固定地址\$C\$7”



Microsoft Excel - 统计描述

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W) 帮助(H) Adobe PDF

QUARTILE X ✓ = **=C2/\$C\$7*100**

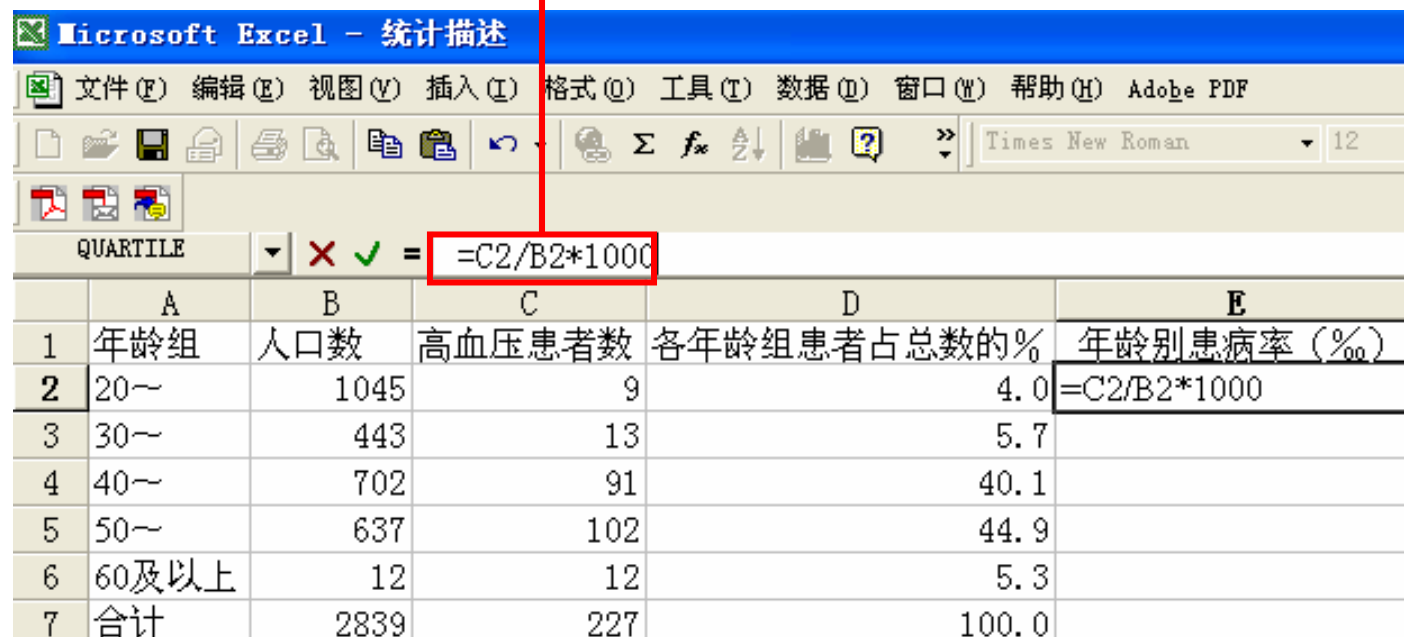
	A	B	C	D	E
1	年龄组	人口数	高血压患者数	各年龄组患者占总数的%	年龄别患病率(‰)
2	20~	1045	9	=C2/\$C\$7*100	
3	30~	443	13		
4	40~	702	91		
5	50~	637	102		
6	60及以上	12	12		
7	合计	2839	227		

二、分类资料的统计描述

(一) 相对数

5.用EXCEL软件实现分类资料的统计描述

计算率，其中用到相对地址B2



Microsoft Excel - 统计描述

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W) 帮助(H) Adobe PDF

QUARTILE X ✓ = **=C2/B2*1000**

	A	B	C	D	E
1	年龄组	人口数	高血压患者数	各年龄组患者占总数的%	年龄别患病率(‰)
2	20~	1045	9	4.0	=C2/B2*1000
3	30~	443	13	5.7	
4	40~	702	91	40.1	
5	50~	637	102	44.9	
6	60及以上	12	12	5.3	
7	合计	2839	227	100.0	

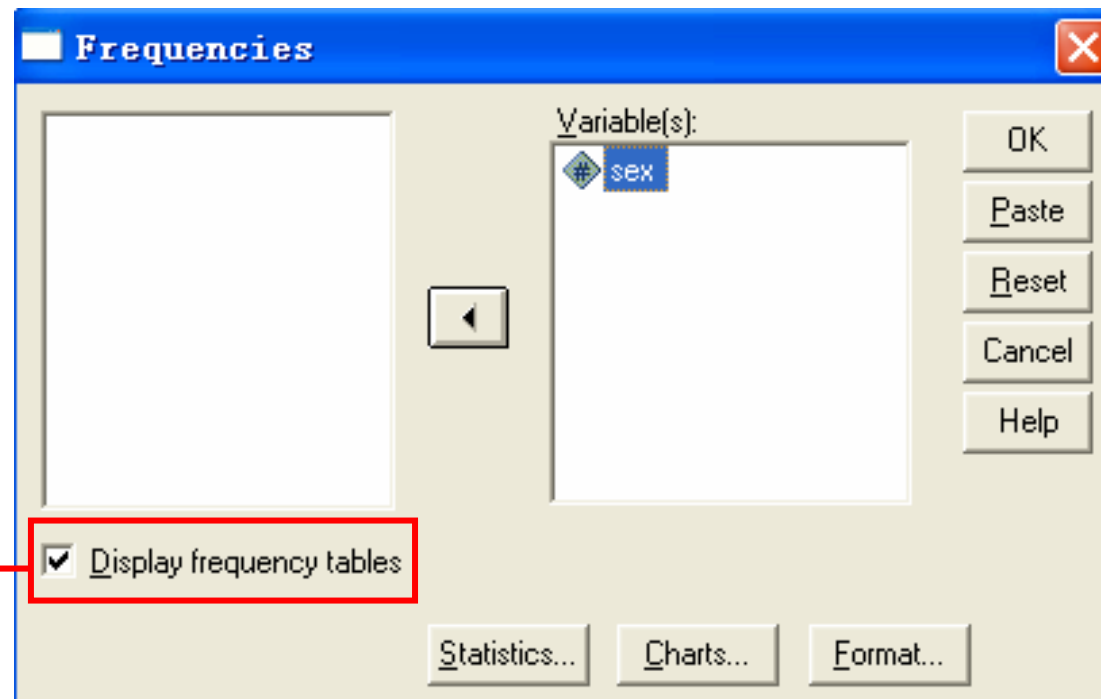
二、分类资料的统计描述

(一) 相对数

5.用SPSS软件实现分类资料的统计描述

使用“Frequencies”命令

列出的频数
表中的百分
比即为构成
比或者率



二、分类资料的统计描述

(二) 标准化法

1.意义：要正确对各组进行比较，必须先按照统一的标准对各组的人口构成进行校正，然后计算出校正后的标准化率再进行比较。

2.基本思想：采用统一的标准人口构成，以消除人口构成不同对总率的影响。

二、分类资料的统计描述

(二) 标准化法

3. 标准化率的计算:

$$p' = \frac{N_1 p_1 + N_2 p_2}{N} = \frac{N_1}{N} p_1 + \frac{N_2}{N} p_2$$

$$p' = \frac{\sum N_i p_i}{N} = \sum \left(\frac{N_i}{N} \right) p_i$$

$N_i p_i$ 为小组的预期（治愈、发病或者死亡）人数； $\sum N_i p_i$ 为总预期（治愈、发病或者死亡）人数。

二、分类资料的统计描述

(二) 标准化法

4.应用标准化法的注意事项:

- 选取的标准不同，标准化率的大小也不同，但选取同一标准的各组的标准化率的相对水平不变。
- 标准化率不能反映实际的发病或死亡水平，只是为了比较各组的标准化率的相对水平。