统计数据整理

授课教师： 周银香

浙江财经大学数据科学学院

统计数据整理的含义

根据统计研究的目的和要求，对统计调查所得到的大量的原始资料进行科学的加工、汇总、或对已经加工过的资料进行再加工，使之系统化、条理化、成为能够反映总体特征的综合资料的工作过程。

统计数据整理的步骤

- 整理方案的设计
- 数据预处理
- 统计分组和汇总
- 统计数据的显示保存与公布

资料的完整性、及时性、准确性

注意分组标志的选择

统计分组的含义与性质

含义

根据统计研究任务的要求和现象总体的内在特点，按照一定的标志将总体划分为若干性质不同而又有联系的几个部分的统计方法，称为 **统计分组** 。

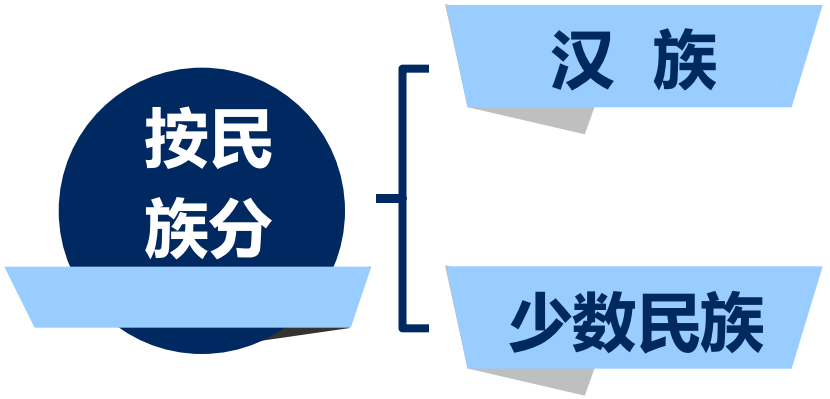
性质

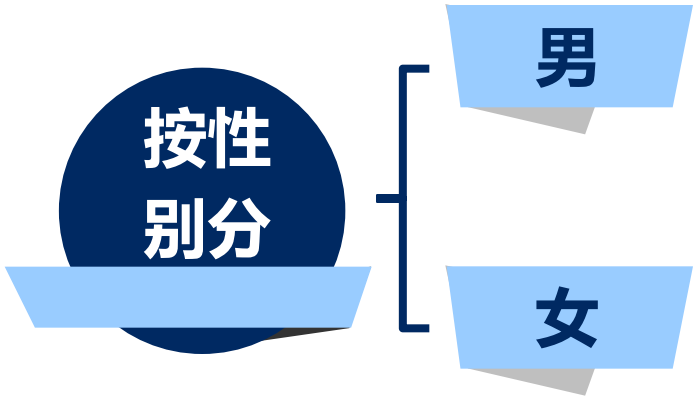
- **分与合**
- **穷尽与互斥**
- **组内同质与组间差异**

统计分组

统计分组的种类：按分组标志的性质分类

按品质标志分组

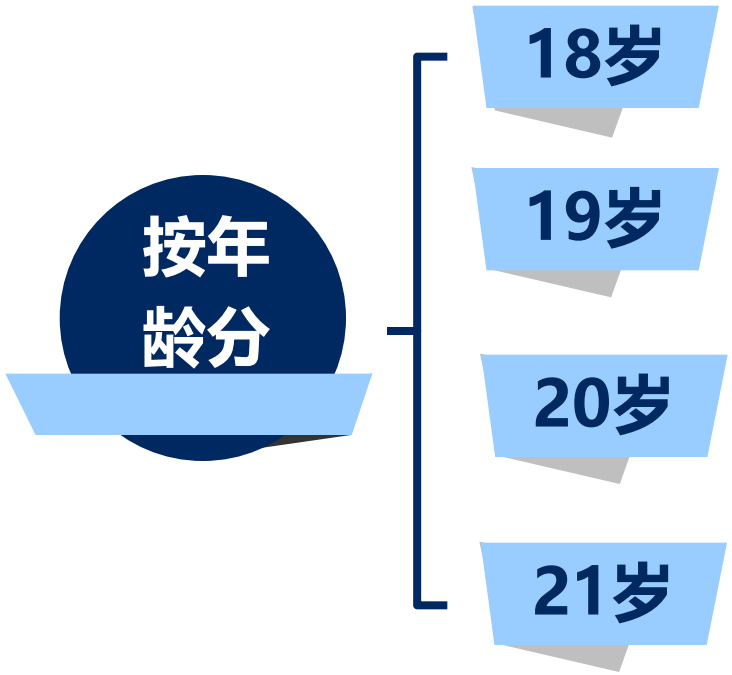


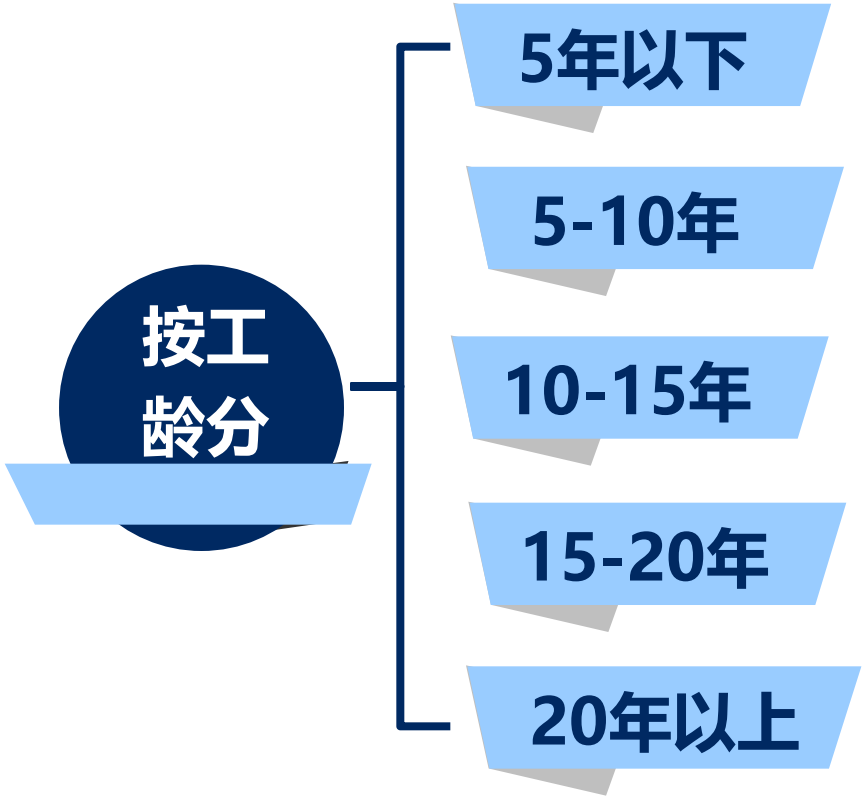


统计分组

统计分组的种类：按分组标志的性质分类

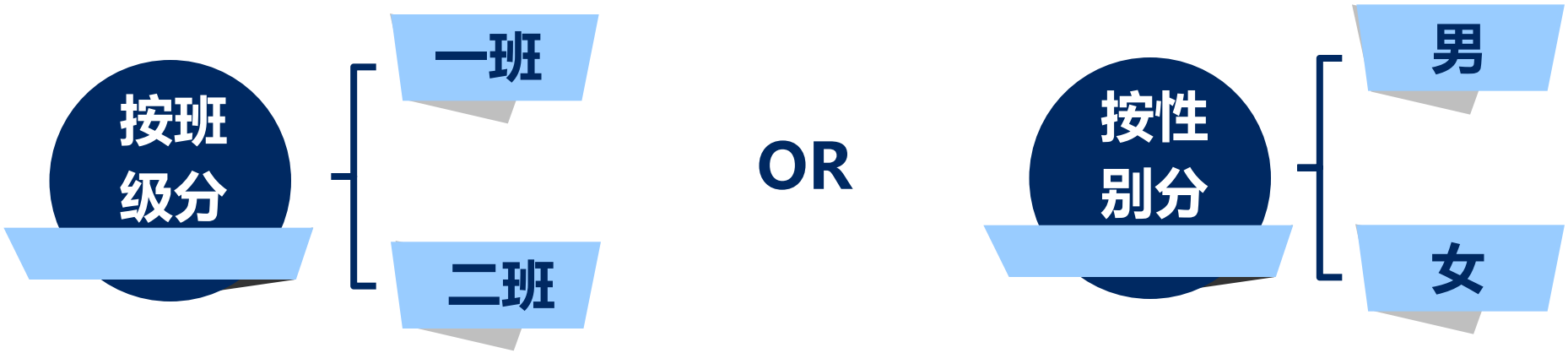
按数量标志分组





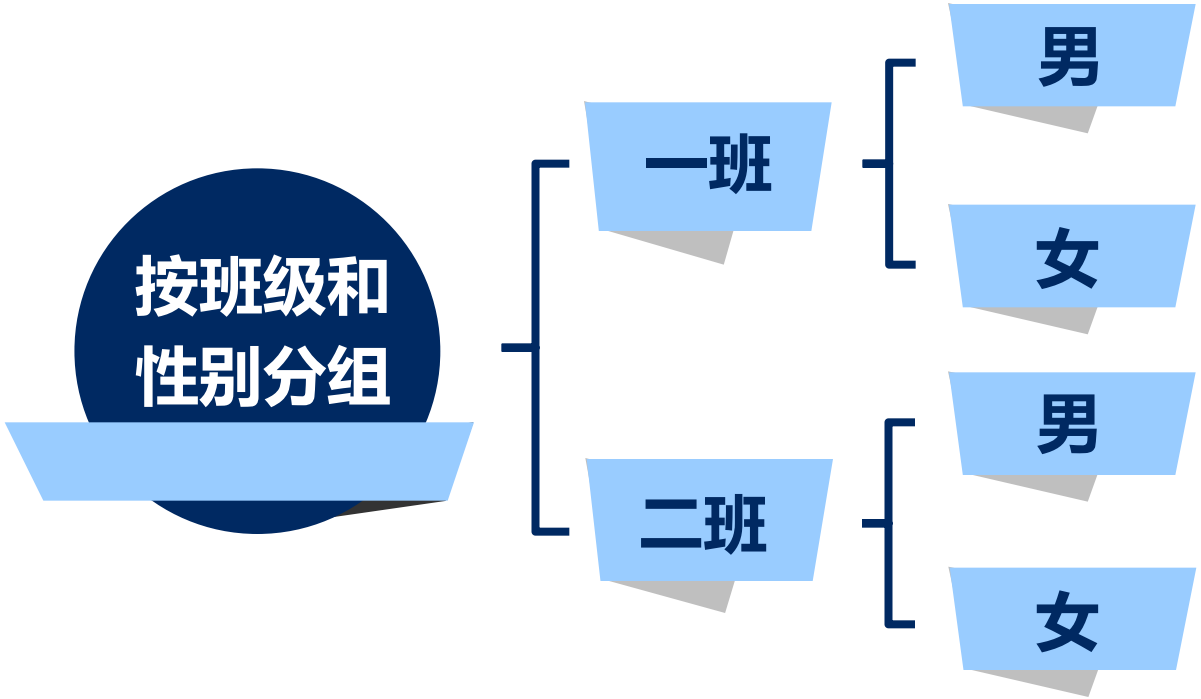
统计分组的种类：按分组标志的多少分类

简单分组：只按一个标志分组



统计分组的种类：按分组标志的多少分类

复合分组：按两个或两个以上的标志层叠分组



分布数列的概念

分布数列

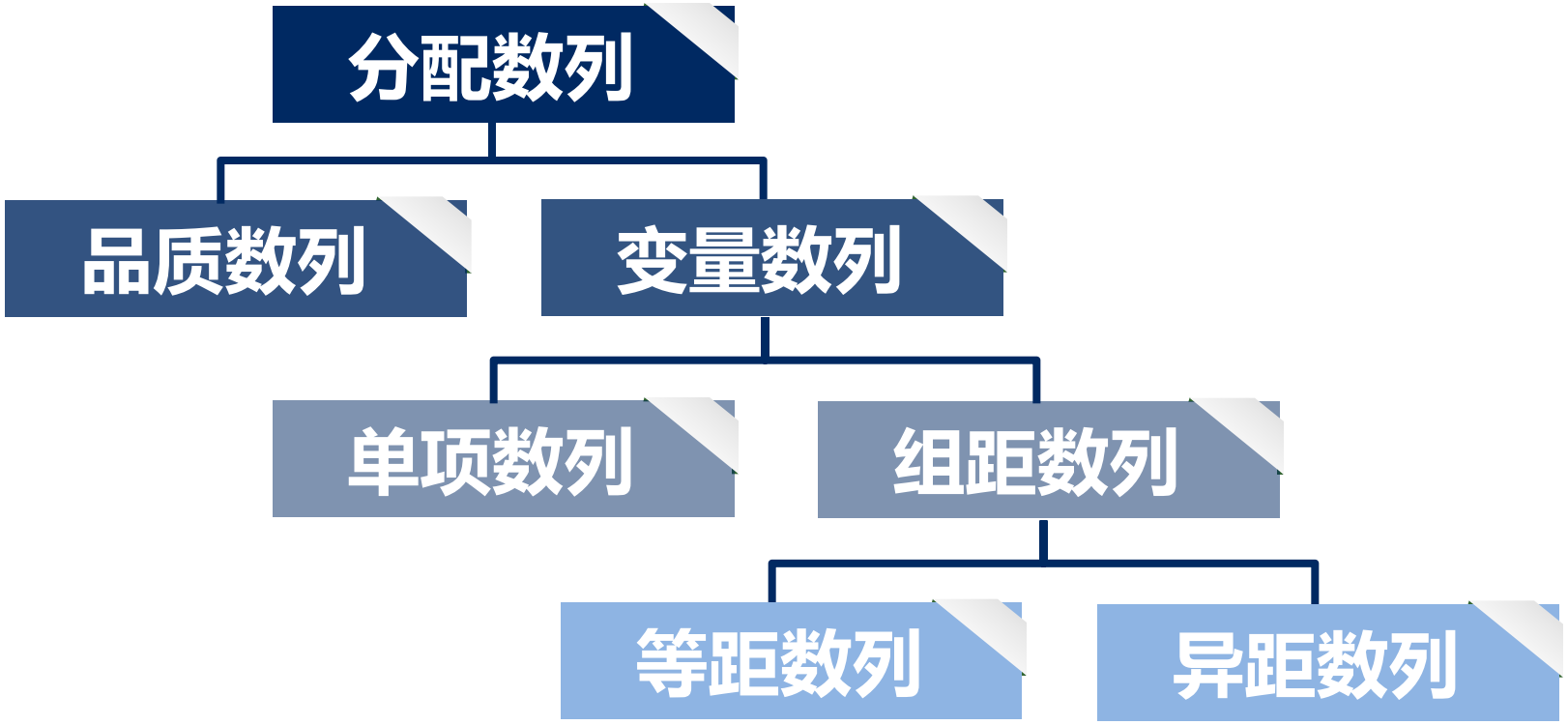
根据一定的分组标志将原始资料进行分组，并按一定顺序进行排列，说明总体各单位分布情况的数列。

基本形式

性别	人数	比重（%）
男	300	75
女	100	25
合计	400	100

各组名称　　次数、频数或权数(单位数)　　频率

分配数列的分类



分配数列的基本形式举例

品质数列

按性别分组	人数（人）(频数)	比率（%）(频率)
男	410	51.25
女	390	48.75
合　计	800	100.00

分配数列的基本形式举例

单项式数列

按家庭人口分组	职工户数（频数）	比率（%）（频率）
1	7	2.9
2	38	15.2
3	105	41.3
4	54	20.5
5	31	12.1
6	20	8.0
合计	255	100.0

分配数列的基本形式举例

组距数列：等距

按年龄分组	人数（人） (频数)	比率（%） (频率)
20 - 30	50	25
30 - 40	120	60
40 - 50	30	15
合　计	200	100

分配数列的基本形式举例

组距数列：异距

年收入（万元）	人数（频数）	比率（%） (频率)
5～10	100	25.0
10～20	250	62.5
20~50	50	12.5
合　计	400	100.0

变量数列的编制

单项式数列的编制

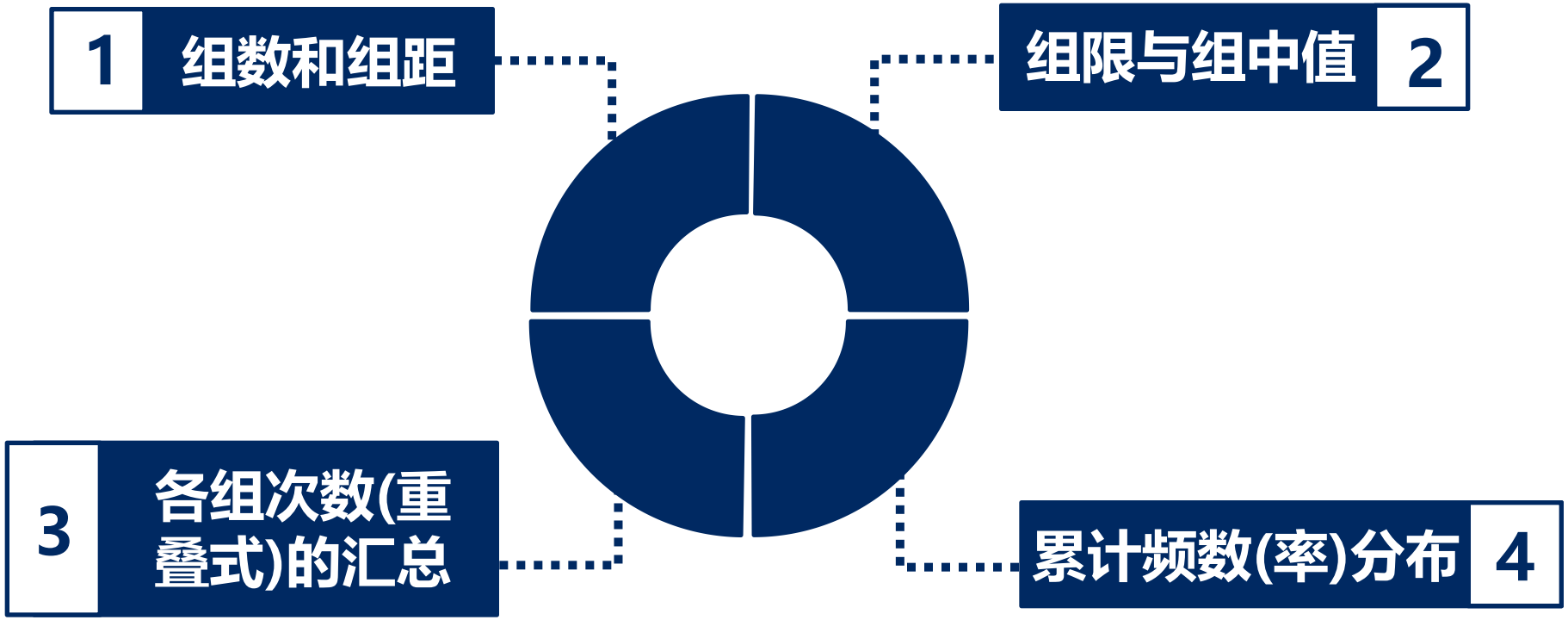
适用于离散型变量且变量值变化范围不大的情况

组距数列的编制

适用于连续型变量和变量值变动范围较大的离散变量

组距式数列的编制

编制组距式数列需处理好以下几个问题



组距式数列的编制

1.组数和组距

1 确定组数

组数的确定参考经验公式（如果总体大致呈正态分布）：

$n=1+3.33\lg N$ (N为单位数)

2 确定组距

组距 =全距/组数

=（最大值-最小值）/组数

组距式数列的编制

2.组限与组中值

- **下限** ▶ 一个组的最小值
- **上限** ▶ 一个组的最大值
- **组距** ▶ 上限与下限之差

➢ 组限的确定

最低组限 ≤ 数据的最小值
最大组限 ≥数据的最大值

对连续型数据，往往采用相邻两组组限重叠

对离散型数据，习惯上也采用组限重叠的分组方法

组距式数列的编制

组中值的计算

组中值 ▶ 下限与上限之间的中点值

闭口组

组中值
=（上限 + 下限）/2

开口组

缺下限开口组的组中值=上限-邻组组距/2

缺上限开口组的组中值=下限+邻组组距/2

◆ **如果中间非开口组的组距呈现某种规律，如各组组距呈等差或等比变化，则应按规律来确定开口组的组距。**

组距式数列的编制

3.各组次数（重叠式）的汇总

对于组限重叠的统计分组，一般按照“**上组限不在内**”的原则汇总单位数。

组距式数列的编制

4.累计频数（率）分布

向上累计 是将各组频数（率）由标志值低组向高组依次累计，说明至某组上限以下的累计频数（频率）的分布情况。

向下累计 是将各组频数（率）由标志值高组向低组依次累计，说明至某组下限以上的累计频数（频率）的分布情况。

累计频数（率）分布

被调查者月收入情况分布

按月收入分组（元）	次数		向上累计次数		向下累计次数	
	频数(人)	频率(%)	频数(人)	频率(%)	频数(人)	频率(%)
3000以下	50	10	50	10	500	100
3000-5000	90	18	140	28	450	90
5000-8000	120	24	260	52	360	72
8000-10000	160	32	420	84	240	48
10000以上	80	16	500	100	80	16
合计	500	100	-	-	-	-

累计频数（率）分布

被调查者月收入情况分布

按月收入分组（元）	人数(人)	比重(%)	向上累计		向下累计	
			频数(人)	频率(%)	频数(人)	频率(%)
3000以下	50	10	50	10	500	100
3000-5000	90	18	140	28	450	90
5000-8000	120	24	260	52	360	72
8000-10000	160	32	420	84	240	48
10000以上	80	16	500	100	80	16
合计	500	100	-	-	-	-

谢 谢！

日期：17/08/05