**19.统计学基础知识Ⅱ—统计表,统计图**

**（七）统计表**

统计表和统计图都是表达统计资料的重要工具。

**统计表，**表达统计资料具体、明确，可以用来表达定性资料、定量资料、混合型资料；

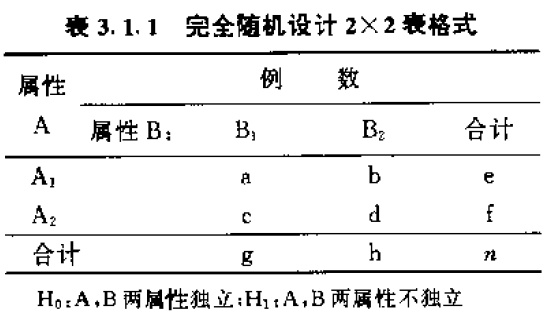
**统计图，**表达统计资料形象、生动，但只能要么表达定性资料、要么表达定量资料。

**一、表达定性资料的统计表**

**1. 二维2×2表列联表（2个定性指标）**

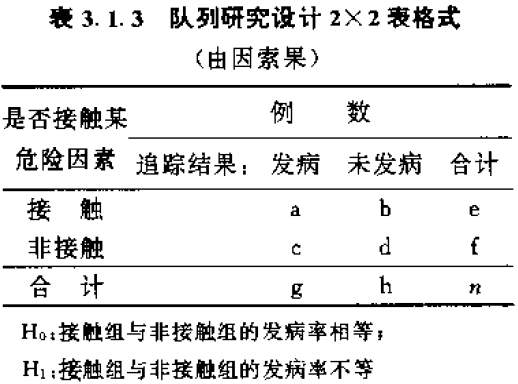
（1）完全随机设计

从全体对象中随机抽取n个个体，然后按属性A、B的两个分类进行两两组合分成四格，得到相应的2×2的频数表。



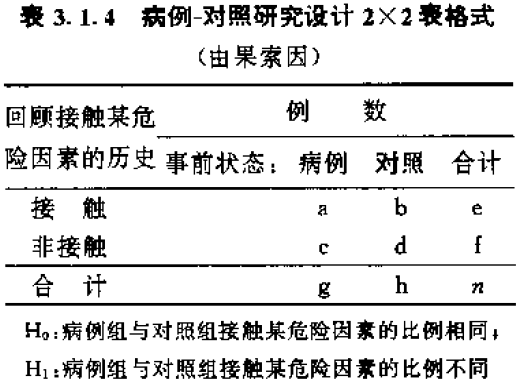
（2）队列研究设计

若把全体研究对象按因素A（是否接触某危险因素）分为两组（接触组、非接触组），再对每组的个体进行跟踪观察，并按因素B（是否患病）分成两组（患病组、未患病组），得到的2×2表。



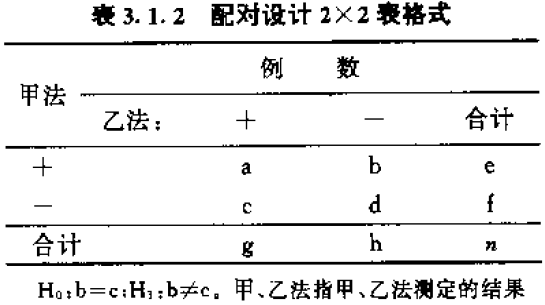
（3）病例-对照研究设计

若把全体研究对象按因素B（是否患某病）分成两组（病例组、对照组），再对每组中的个体进行回顾性调查，并按因素A（是否接触某危险因素）分成两组，得到的2×2表。



（4）配对研究设计

n个受试对象（或一组样品）分别用甲乙两种检测方法进行检测，并按测定结果（阳性+，阴性-）分类计数，即按“都为阳性，一阳一阴，一阴一阳，都为阴性”四种情况分别计数，得到的2×2表。



2×2列联表，都可以作相关分析（两属性是否有相关关系）、独立性检验（或两总体率之间差异的显著性检验，用Fisher精确检验）；

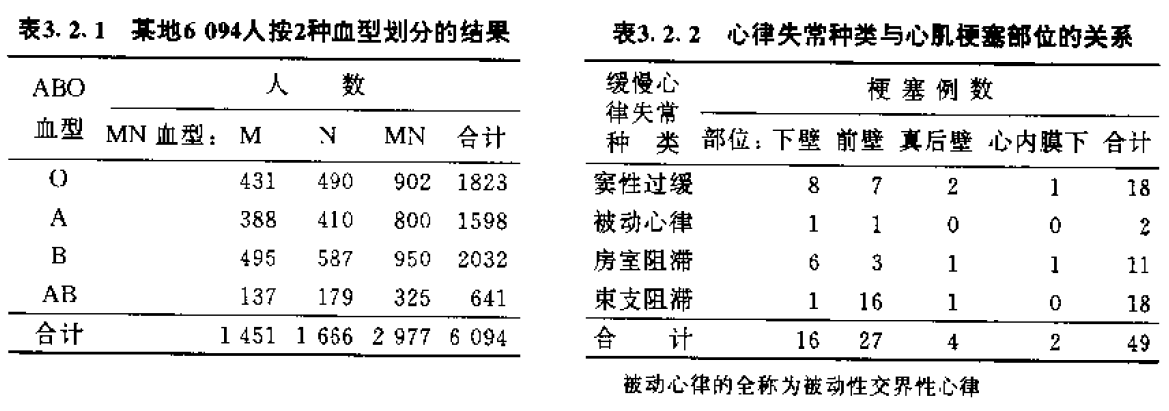
2×2队列研究设计，常需要先计算相对危险度RR（=af/ce），并做假设检验：H0：RR=1；H1：RR=0. 做该检验需要用Mantel-Haenszel统计量（U统计量）。

2×2队列研究设计和2×2病例-对照研究设计，可由经验Logistic变换公式计算Zi统计量，进而做U检验。

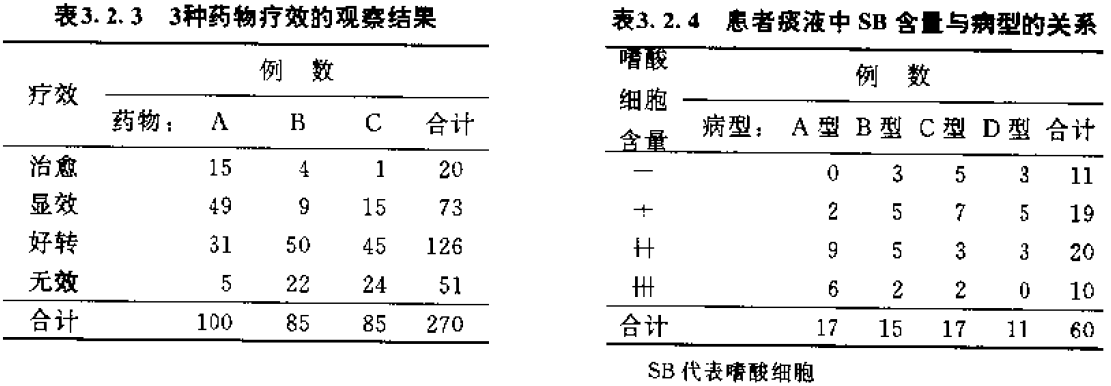
2×2配对研究设计，检验两种检测方法之间有无显著性差别。

**2. 二维R×C列联表（R，C≥2）**

（1）双向无序R×C表



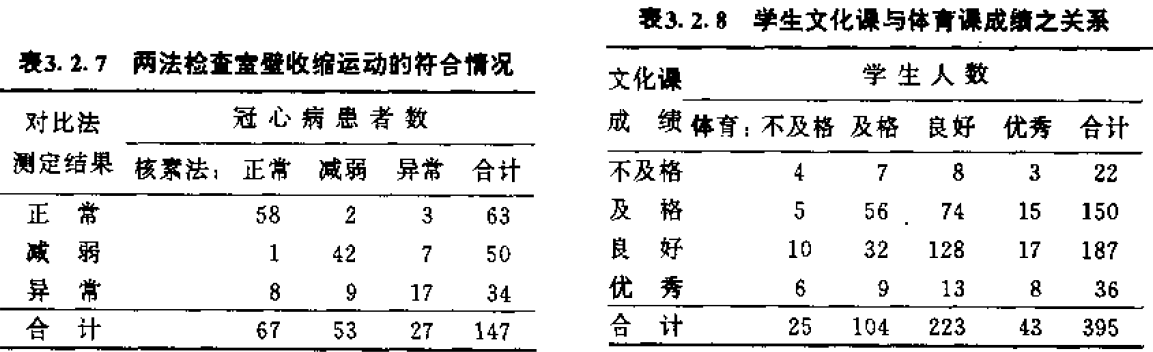
（2）单向有序R×C表



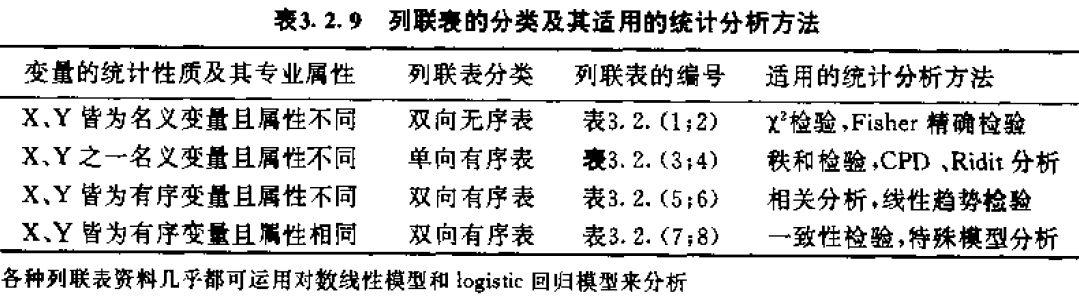
（3）双向有序属性不同R×C表



（4）双向有序属性相同R×C表



**适用的分析、检验方法：**



双向无序R×C表，所选用的统计方法应当与分组变量各水平的先后顺序或取值大小无关，仅与表中总频数、各行的合计、各列的合计有关。符合这些要求的方法有Pearson的拟合优度检验检验、基于似然函数（或熵的分解）导出的似然比检验、对数线性模型、SAS的FREQ过程步进行Fisher精确检验。

单向有序R×C表，所选用的统计方法应当与有序的那个分组变量各水平的先后顺序或取值大小有关。显然，通常的检验已无能为力，应考虑把有序变量当作半定量指标在计算中加以考虑的方法：秩和检验、Ridit检验、CPD分析、有序变量的Logistic回归模型、有序变量的对数线性模型。

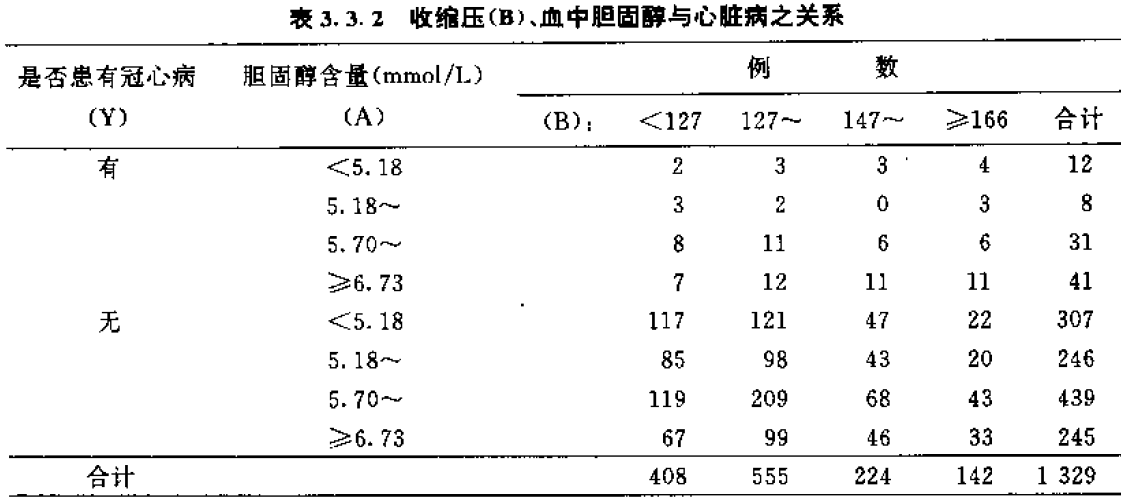
双向有序属性不同R×C表，当把两个属性看成地位平等的相互关系时，常需要考虑它们之间是否存在线性关系，即需要对资料进行相关分析（Spearman秩相关分析、典型相关分析）；若把一个属性当成自变量，另一个属性当成因变量，常需要考察它们之间是否存在直线变化趋势，即需要对资料进行线性趋势检验（U检验、检验：“利用回归思想产生的与线性回归有关的分量和偏离线性回归的分量”）。

双向有序属性相同R×C表，两个分组变量都是“测定结果”，且档次划分也相同（有序）。对这种资料，研究者关心的不是两变量之间是否存在相关性或线性趋势，而是这两种测定结果的一致性如何。适合采用一致性检验：Kappa检验。特别地，如表3.2.8，研究者可能更关心：文化课成绩与体育锻炼达标的等级之间是否呈现某种特殊的变化趋势（直线趋势模型，即主对角线上的频数占大多数，之外的频数之和趋于0），可以用前面线性趋势检验、或一些特殊模型：对称模型、条件对称模型、对角线模型。

注：有时候若想要更多的挖掘R×C表的统计信息，还可以对R×C表进行分割，得到若干2×2表做进一步研究。

**3. 高维列联表（变量个数≥3）**

例如，下面是三维列联表：



处理高维列联表资料的方法，借助统计软件有Logistic回归模型（只能分析自变量对因变量的影响）、对数线性模型（可分析全部变量及其交互作用对列联表中格网格上理论频数之对数的影响）；也可以将高维列联表压缩为二维列联表再进行处理。

**4. 具有重复测量设计的定性资料（做方差分析）**

（1）单因素重复测量资料：



方差分析的总思想：

将总变异分解为：个体间（between subjects）变异与个体内的变异。其中个体内变异是与重复因素有关的变量。

（2）双因素重复测量资料

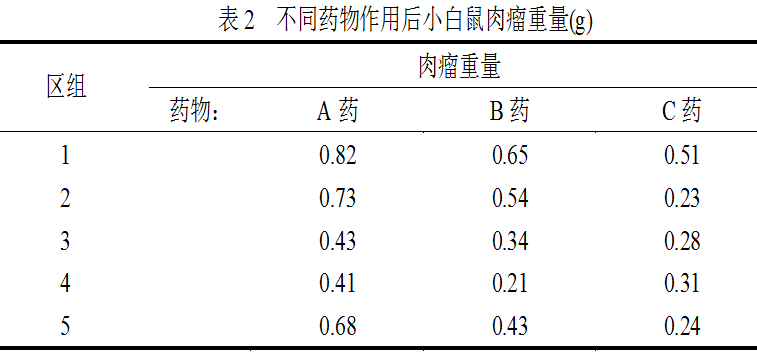


方差分析总思想：

将总变异分解为：对象间（between subjects）变异与对象内（within subject）变异，其中对象内变异是与重复因素有关的变量。

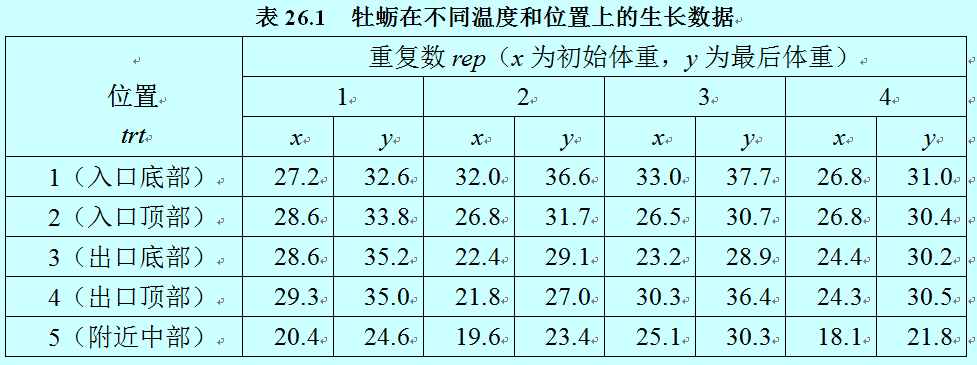
**二、定量资料的统计表**

1. 随机区组设计一元定量资料



适合做方差分析，事先需要检验正态性和方差齐性（若不满足则要数据变换和非参数检验）。

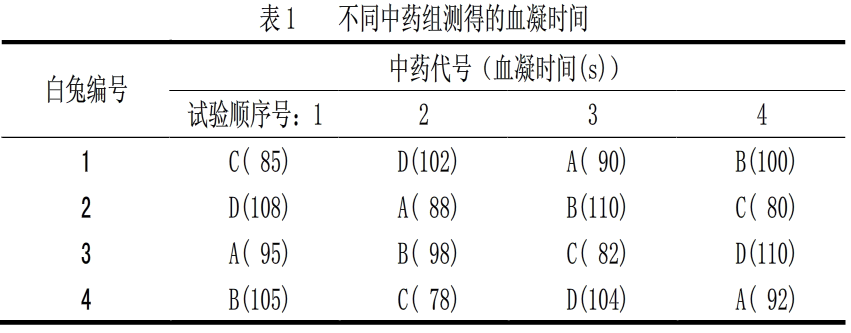
2. 含一个协变量的随机区组设计一元定量资料



初始体重x是协变量，适合做一元协方差分析（若含有多个协变量，则要做多元协方差分析）。

3. 拉丁方设计一元定量资料

4×4拉丁方设计：



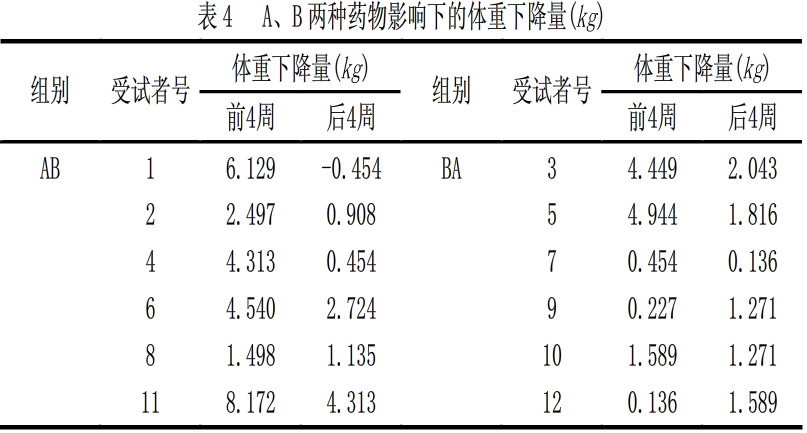
注：拉丁方要求分组数和处理数相同，在每行每列各种处理都出现且仅出现一次。

既要控制系统误差且试验动物的数量又较少，则常采用拉丁方设计（处理数不能太多，一般以4～10个为宜）。

基本要求：（1）必须是三个因素（分组、处理、结果）的实验，且三个因素的水平数相等（若三因素的水平数略有不同，应以主要处理因素的水平数为主，其它两因素的水平数可进行适当调整）；（2） 三因素间是相互独立的，均无交互作用；（3）各行、列、字母所得实验数据的方差齐。

4. 配对交叉设计一元定量资料

**示例：两阶段交叉设计**



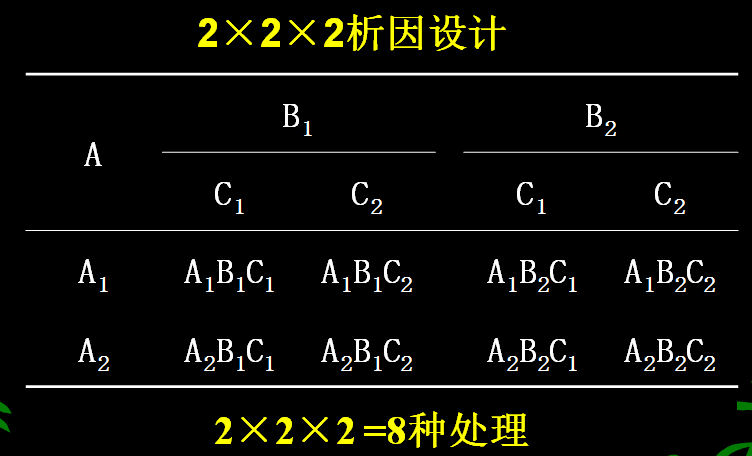
当试验中涉及到一个具有2水平的试验因素，这两个水平要先后作用于同一个受试对象，并且这两个水平要么在条件相近的同一对受试者中交叉实施，就称为配对二阶段交叉设计；要么在两组受试者中交叉实施，就称为成组二阶段交叉设计。

可以考察一个具有两水平的试验因素和两个受试对象组因素（即个体差异、测定顺序）对观测结果的影响；试验因素的两个水平在两组受试对象中施加的顺序呈交叉状；对于每一个受试者而言，均有一个“洗脱期”，从而消除“携带效应”的影响。

另外，还有三阶段交叉设计、3×3交叉设计（将三种处理或药物分三个时期先后给予同一个受试者，观察受试者接受每种处理后的反应；处理A、B、C施加的顺序共有6种排列方式，即ABC、ACB、BAC、BCA、CAB、CBA，故至少要将受试者分为6个组，每组中至少要有一位受试者）。

5. 析因设计

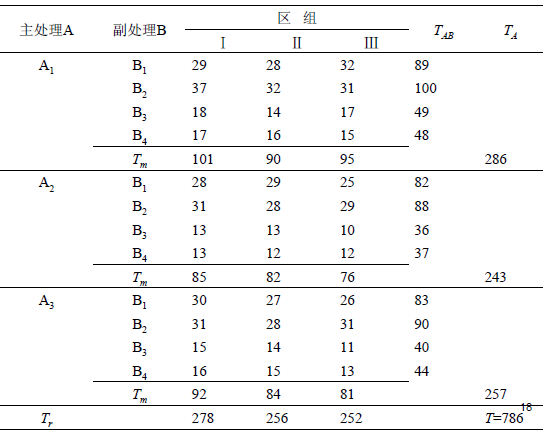
析因设计（factorial design）是一种多因素的交叉分组设计。它不仅可检验每个因素各水平间的差异，而且可检验各因素间的交互作用。两个或多个因素如存在交互作用，表示各因素不是各自独立的，而是一个因素的水平有改变时，另一个或几个因素的效应也相应有所改变；反之，如不存在交互作用，表示各因素具有独立性，一个因素的水平有所改变时不影响其他因素的效应。





6. 裂区试验设计

先将每一区组按第一因素的处理数划分小区（称为主区），在主区里随机安排主处理，然后在每个主处理里按第二因素的处理数再划分小区（称为副区），在副区里随机排列副处理。



注：有时候在表中使用“”或“M(Q1~Q3)”简化形式表示数据。

**（八）统计图**

**一、定量变量的统计图形**

**1. 直方图（histogram, 也称质量分布图）**

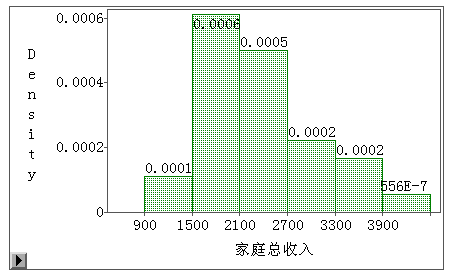
对于定量变量（数值型变量），常用直方图来展示某个变量取值的分布，利用直方图可以估计总体的概率密度，观察其大致的分布曲线。

将变量取值的范围分成若干区间。直方图是用面积而不是用高度来表示数，总面积是100%. 每个区间矩形的面积恰是落在该区间内的百分数（频率），所以

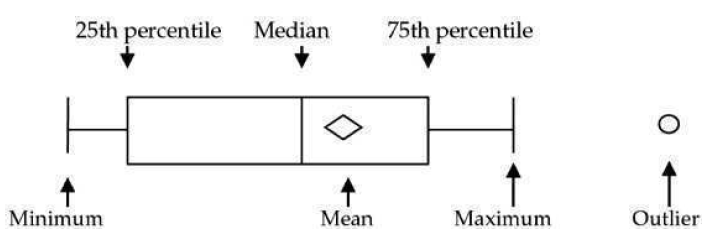
**矩形的高 = 频率 / 区间长度 = 密度**

特别地，若区间是等长的，则“**矩形的高”**就是频率。

注意：直方图矩形之间是没有间隔的。



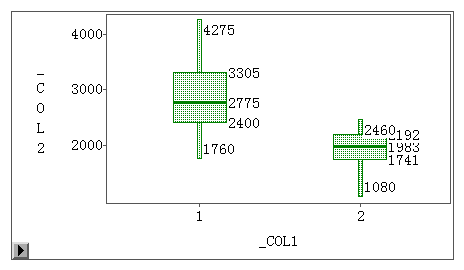
**2. 盒形图（boxplot），也称箱线图**



在一条数轴上，以数据的上下四分位数（Q1-Q3）为界画一个矩形盒子（中间50%的数据落在盒内）；在数据的中位数位置画一条线段为中位线；用◇标记数据的均值；默认延长线不超过盒长的1.5倍，之外的点认为是异常值（用○标记）。

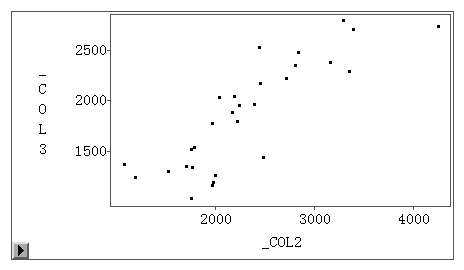
盒形图的主要应用就是，剔除数据的异常值、判断数据的偏态和尾重。

例如，下图是分地区情况对家庭收入所绘的盒形图：



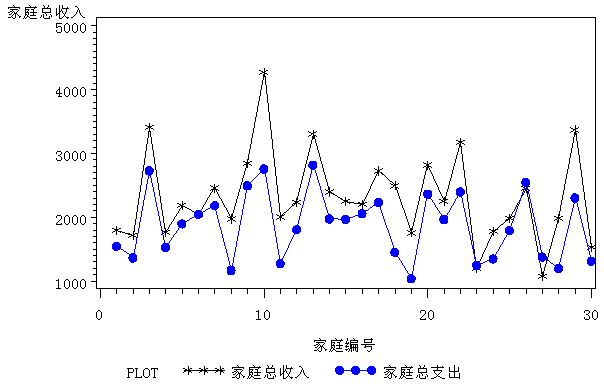
**3. 散点图**（Scatter Plot）

通常得到的数据可能有两个变量，比如家庭收入和家庭支出。希望通过图形了解家庭收入和家庭支出的关系，这时可以用一个变量为横坐标（如家庭收入），另一个为纵坐标（如家庭支出）对数据进行描点来作图。这种图称为散点图。例如，



**4. 线图（Line Plot）**

将散点进一步用线段连接起来，就是线图。可以表示变量间的取值变化情况，有单式和复式两种。在复式线图中可用不同颜色的实线来标志区别，例如，



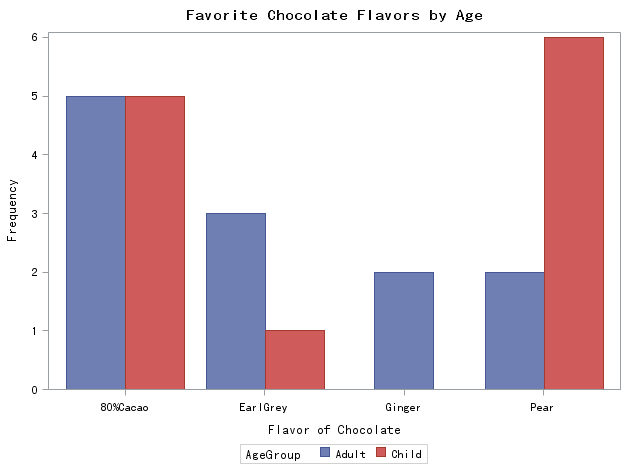
**二、分类变量的统计图形**

**1. 条形图**（Bar Chart）

用若干个细长的矩形条的高度（不是宽度，也不是面积）来表示定性变量各水平组的频数。分为

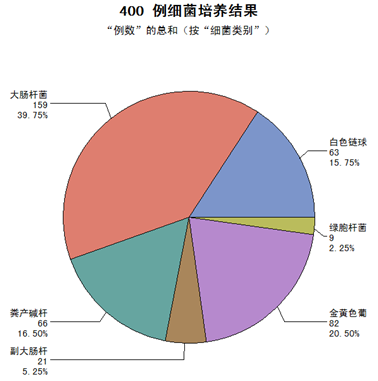
单式条形图——横轴上只有一个定性变量；

复式条形图——横轴上有两个或多个定性变量例如，



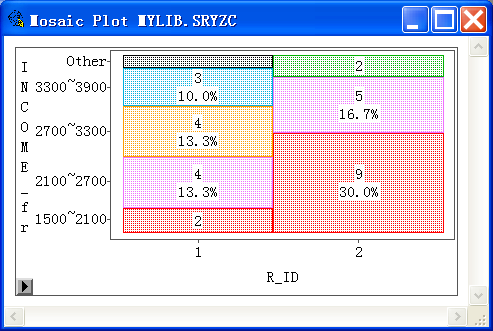
**2. 饼图（Pie Chart）**

用圆的各扇形面积的大小来表示定性变量各水平组的频数。例如，



**3. 马赛克图（Mosaic Plot）**

马赛克图一般不对单个变量作，而是对两个分类变量来做，好处是直观显示了两个变量每种取值组合的观测个数和比例。例如，



**三、其它图**

1. **经验分布图**

经验分布图，是根据样本观测值做出的经验分布函数而绘制的，经验分布函数是样本对总体累积分布函数的一个估计。

**2. 概率图**

概率图，在直方图基础上我们可以猜想变量是否服从某类型的分布？通过绘制概率图可以来证实我们的想法。概率图将数值排序，给出每个数值对应的分位数，然后打点作图。如果这些点呈现线性特征，说明他们与理论分布相符，同时又在图像上加上一条给定分布的曲线，并给出分布的系数。

**3. PP图**

PP 图，是根据变量的累积概率对应于所指定的理论分布累积概率绘制的散点图，用于直观地检测样本数据是否符合正态分布分布。如果被检验的数据符合所指定的分布，则代表样本数据的点应当基本在代表理论分布的对角线上。

**4. QQ图**

绘制QQ图，鉴别样本分布是否近服从正态分布的一种直观简便的图形，它是以某种分布的分位数为横坐标，以样本值为纵坐标绘制的散点图。Q-Q图的结果与P-P图非常相似，只是P-P图是用分布的累计比，而Q-Q图用的是分布的分位数来做检验。和P-P图一样，如果数据为正态分布，则在Q-Q正态分布图中，数据点应基本在图中对角线上。

注：Q-Q图与P-P图判断正态分布不具准确性，一般不太用。