**14. 绘图Ⅱ—条形图,直方图,盒形图**

**（一）定性变量的图形**

定性变量（分类变量）也可以通过图形直观地描绘出它们各类的数量和所占比例，常用的有条形图、饼形图和马赛克图。

**一、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（GROUP可选项）；

基本语法：

PROC SGPLOT data = 数据集;

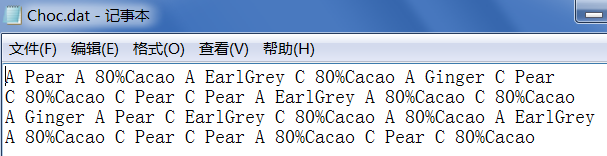
VBAR或HBAR *变量列表 </可选项>*;

注：“VBAR”绘制竖直方向的条形图，“HBAR”绘制水平方向的条形图。

**可选项：**

|  |  |
| --- | --- |
| GROUP = 变量名 | 指定分组变量 |
| GROUPDISPLAY =选项 | 指定分组条形的显示方式：  STACK（堆叠型，默认），CLUSTER（簇型）； |
| RESPONSE = 变量名 | 指定一个数值变量做统计分析； |
| STAT = 统计量 | 对RESPONSE变量做指定的统计分析；  有RESPONSE变量默认FREQ；否则默认SUM； |
| ALPHA = p | 设定置信限: [0,1]，默认为0.05（95%可信） |
| BARWIDTH=w | 设定“条形”宽度，介于0.1和1之间，默认为0.8； |
| DISCRETEOFFSET = x | 设定每个条形偏移中心的距离: [-0.5,0.5]，默认为0 |
| TRANSPARENCY = n | 指定条形的透明度: [0,1]，默认是 0（完全不透明） |
| DATALABEL = 变量名 | 指定用该变量的值作为每个“条形”的标签，  默认SAS自动计算适当的值 |
| MISSING | 绘制包含缺失值的条形； |
| LIMITSTAT = 选项 | 设定要显示的极限线的类型：CLM（均值左右95%的双侧置信区间）；STDDEV（样本标准差）；STDERR（标准误）； |

**例1** 巧克力生产商考虑是否增加 4 个新品种，公司找志愿者品尝它们的味道，得到数据（C:\MyRawData\Choc.dat），变量包括年龄组（A-成年；C-儿童）、喜欢吃的口味（注意每行有 6 个测试）：



读入数据，用 PROC SGPLOT 过程步生成关于年龄段分组的条形图。

代码：

**data** chocolate;

infile 'c:\MyRawData\Choc.dat';

input AgeGroup $ FavoriteFlavor $ @@;

**run**;

**proc** **format**;

value $AgeGp 'A' = 'Adult' 'C' = 'Child';

**run**;

\* Bar chart for favorite flavor;

**proc** **sgplot** data = chocolate;

vbar FavoriteFlavor / GROUP = AgeGroup GROUPDISPLAY = CLUSTER;

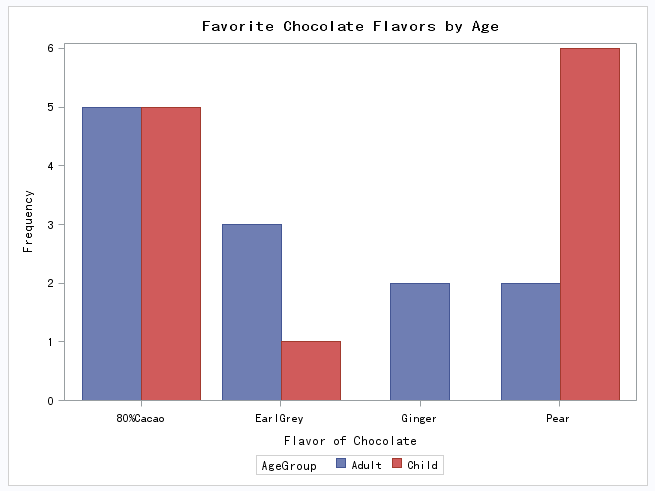
format AgeGroup $AgeGp.;

label FavoriteFlavor = 'Flavor of Chocolate';

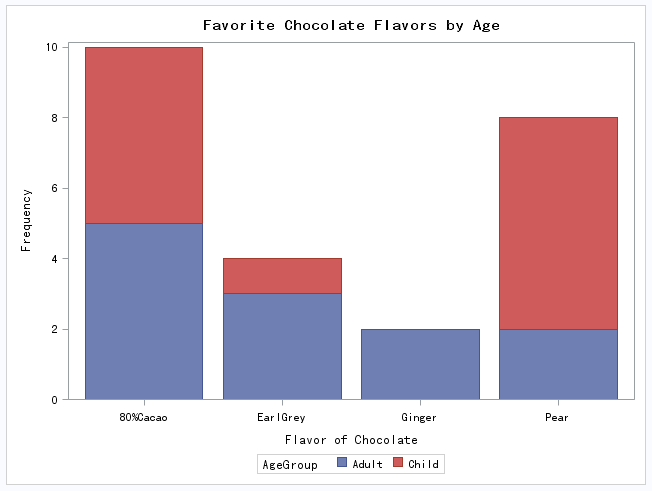
title 'Favorite Chocolate Flavors by Age';

**run**;

运行结果：



注：上图是“簇型”（CLUSTER），若改为“堆叠型”（STACK）输出：



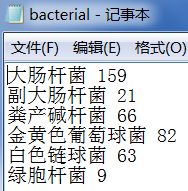
另外，可以用旧版的 GCHART 过程步绘制条形图；

**二、饼图（Pie Chart）**

**饼图，**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。

SGPLOT过程步没有提供绘制饼形图的命令，用GCHART过程步可以实现。

**例2** 细菌培养的数据（C:\MyRawData\bacterial.txt），变量包括细菌类别、例数：



读入数据，绘制饼形图来表达各类细菌的百分比关系。

**代码**：

**data** Bacterial;

infile 'C:\MyRawData\bacterial.txt';

input BacterialClass $ Number;

**run**;

\* Bar chart for Bacterial;

**proc** **gchart** data = Bacterial;

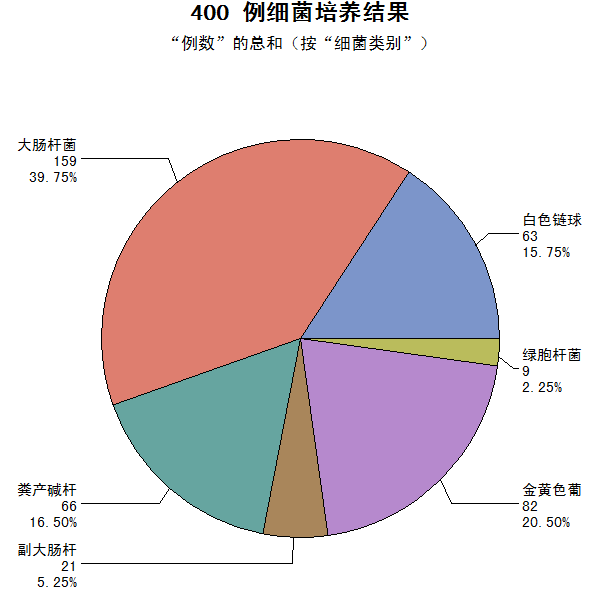
pie BacterialClass / sumvar = Number PERCENT = arraw;

label BacterialClass = '细菌类别' Number = '例数';

title '400 例细菌培养结果';

**run**;

**运行结果**：



程序说明：

（1）可选项“sumvar = Number”，指定变量Number为做统计计算的量（默认是FREQ）；

（2）可选项“PERCENT = arraw;”，指定按百分数显示，并连上线。

**（二）定量变量的图形**

**一、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。

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。

**用SGPLOT过程步绘制直方图以及密度曲线**

基本语法：

PROC SGPLOT data = 数据集;

HISTOGRAM *variable-name* / *options*;

DENSITY *variable-name / options*;

1. HISTOGRAM可选项

（1）BINSTART = n

指定第一个矩形中点的起始位置；

（2）BINWIDTH = n和NBINS = n

指定矩形的宽度，或矩形的个数（二者只给其中之一）

（3）SCALE = scalingtype

指定竖轴的刻度类型：PERCENT（默认，百分数）,

COUNT（频数）, PROPORTION（频率）；

（4）SHOWBINS

指定刻度标记标在每个矩形的中点位置，默认只标记将

取值范围分为四段的分点；

（5）TRANSPARENCY = n——设置透明度：[0,1].

2. DENSITY可选项

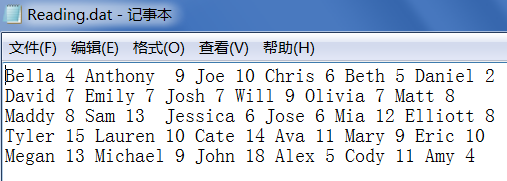
（1）TYPE =distribution-type

指定密度曲线的分布类型：NORMAL（默认，正态分布）、

KERNEL（核密度曲线）；

（2）TRANSPARENCY = n——设置透明度。

**例3** 四年级学生在某月阅读书籍的数据（C:\MyRawData\Reading.dat），变量包括学生姓名、书籍数，注意每行有多组数据：



读入数据，绘制关于书籍数的直方图以及密度曲线，并与正态分布的密度曲线做对比，竖轴显示频数，直方图每个矩形下面都作刻度标记。

代码：

**data** contest;

infile 'c:\MyRawData\Reading.dat';

input Name $ NumberBooks @@;

**run**;

\* Create histogram and density curves;

**proc** **sgplot** data = contest;

histogram NumberBooks / BINWIDTH = **2** SHOWBINS SCALE = COUNT;

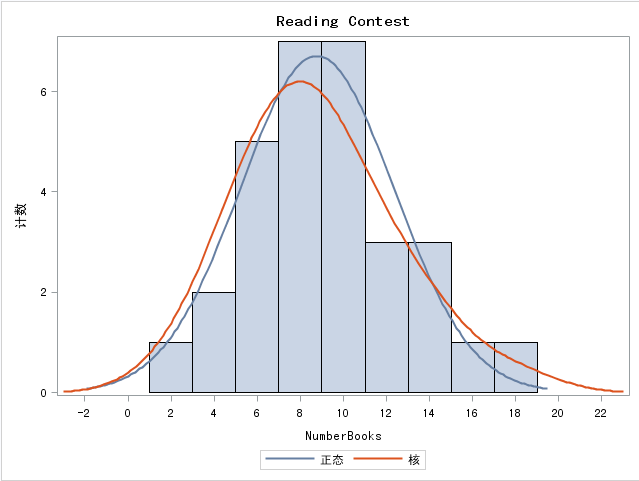
density NumberBooks;

density NumberBooks / type = KERNEL;

title 'Reading Contest';

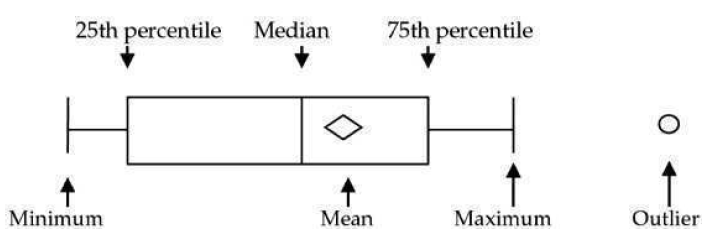
**run**;

运行结果：



**二、盒形图（Box Plots），也叫箱线图**

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。

**用SGPLOT过程步绘制盒形图**

基本语法：

PROC SGPLOT;

VBOX或HBOX *variable-name* / *options*;

注：VBOX——竖直盒形图；HBOX——水平盒形图。

可选项：

（1）CATEGORY = variable-name

指定分类变量，给它的每个值分别绘制一个盒形图；

（2）EXTREME

指定延长线到真正的最大、最小值（相当于没有异常值）；

（3）GROUP = variablename

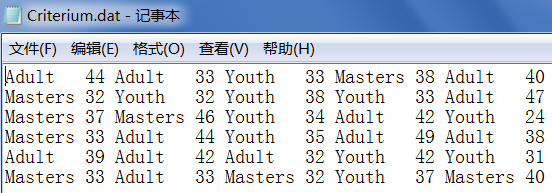
指定CATEGORY变量的下一级分类变量；

（4）MISSING

包含CATEGORY变量或GROUP变量的缺失值；

（5）TRANSPARENCY = n——设定透明度。

**例4** 自行车环形公路赛的数据（C:\MyRawData\Criterium.dat），变量包括年龄组（青年组、成人组、选手组）、圈数，注意每行有多组数据：



读入数据，用SGPLOT过程步为每个年龄组分别绘制盒形图。

代码：

**data** bikerace;

infile 'c:\MyRawData\Criterium.dat';

input Division $ NumberLaps @@;

**run**;

\* Create box plot;

**proc** **sgplot** data = bikerace;

vbox NumberLaps / CATEGORY = Division;

title 'Bicycle Criterium Results by Division';

**run**;

运行结果：

