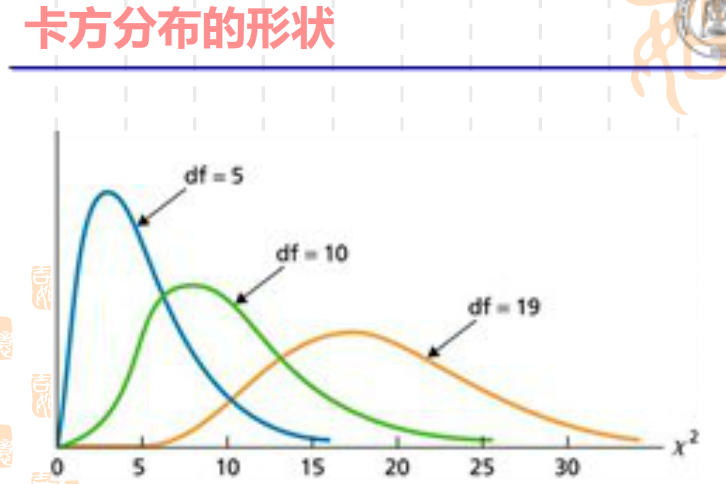
**22. 用PROC FREQ计算频数及卡方检验**

**（一）卡方检验**

**一、卡方分布**

k 个相互独立的标准正态分布变量的平方和服从自由度为k 的卡方分布。



**二、卡方检验概述**

卡方检验，由英国统计学家Karl Pearson得到，主要应用于计数数据（定性变量中的无序分类变量）的分析，对于总体的分布不作任何假设，因此它属于非参数检验法。

理论证明，实际观察频数（f0）与理论频数（fe, 又称期望频数）之差的平方再除以理论频数所得的统计量，近似服从卡方分布，可表示为：



这是卡方检验的原始公式，其中当fe越大，近似效果越好。显然fo与fe相差越大，卡方值就越大；fo与fe相差越小，卡方值就越小；因此它能够用来表示fo与fe相差的程度。

根据这个公式，卡方检验的一般问题是要检验名义型变量的实际观测频数和理论频数分布之间是否存在显著差异。

一般卡方检验要求：① 分类相互排斥，互不包容；② 观察值相互独立；③ 样本容量不宜太小，理论频数≥5，否则需要进行校正。如果个别单元格的理论频数小于5，处理方法有四种：

（1）单元格合并法；

（2）增加样本数；

（3）去除样本法；

（4）使用校正公式。

当期望次数小于5时，应该用校正公式计算卡方值：



**二、卡方检验的原理**

1. 卡方检验所检测的是样本观察频数与理论（或总体）频数的差异性；

2. 理论或总体的分布状况，可用统计的期望值（理论值）来体现；

3. 卡方的统计原理，是取观察频数与期望频数相比较。当观察频数与期望频数完全一致时，值为0；观察频数与期望频数越接近，两者之间的差异越小，值越小；观察频数与期望频数差别越大，两者之间的差异越大，值越大。一旦值大于某一个临界值，即可获得显著的统计结论。

4. 步骤：

原假设H0: = 0; 备择假设H1: ≠0;

根据数据计算卡方值、P值（右尾面积）；

若P值≤α，则拒绝H0; 若P值>α，则接受H0.

**三、卡方检验的应用**

1. 拟合优度检验

检验单个多项分类名义型变量的各分类间的实际观测次数（根据样本数据得到的实计数）与理论次数（根据理论或经验得到的期望次数）之间是否一致、或者服从理论上的某种分布？这一类检验称为拟合性检验。其自由度通常为分类数减去1。

2. 各变量间的独立性检验（定性变量列联表）

两个或两个以上因素多项分类的计数资料分析，也就是研究两类变量之间的关联性和依存性问题。如果两变量无关联即相互独立，说明对于其中一个变量而言，另一变量多项分类次数上的变化是在无差范围之内；如果两变量有关联即不独立，说明二者之间有交互作用存在。

独立性检验一般采用列联表的形式记录观察数据, 列联表是由两个以上的变量进行交叉分类的频数分布表，是用于提供基本调查结果的最常用形式，可以清楚地表示定类变量之间是否相互关联。

其自由度是：（行数-1）×（列数-1）

**（二）PROC FREQ过程步**

**一、基本语法：**

PROC FREQ data = 数据集;

TABLES *行变量 \* 列变量 / options*;

<WEIGHT 权重变量>;

说明：结果将以表格形式（频数表）输出，

TABLES a—单向频数表；

TABLES a\*b—a为行，b为列的双向频数表；

TABLES a\*b\*c—a为分层，b为行，c为列的三维频数表；

TABLES a\*(b c)—等价于“TABLES a\*b a\*c”；

可选项：

（1）AGREE

做配对卡方检验；

（2）CHISQ

做独立性和关联度的卡方检验；

（3）CL

输出关联度的置信限；

（4）CMH

输出Cochran-Mantel-Haenszel统计量，特别对分层二维表；

（5）EXACT

做Fisher精确检验；

（6）MEASURES

输出Pearson and Spearman相关系数、gamma、

Kendall's tau-b、Stuart's tau-c、Somer's D、lambda、

odds ratios、risk ratios、置信区间的关联度；

（7）RELRISK

输出2×2表的相对风险度；

（8）TREND

对趋势做Cochran-Armitage检验；

（9）NOROW, NOCOL, NOPERCENT

不输出行百分比、列百分比、百分比；

**二、绘制PROC FREQ的图表**

默认也会输出PROC FREQ的图表，若要输出指定图表，需要在TABLES语句中，使用绘图可选项“PLOTS = (plot-list);”即可。可以绘制频数图、优势比图、Agreement图、偏差图、以及两类带Kappa统计量和置信限的图。

基本语法：

PROC FREQ data = 数据集;

TABLES *variable1 \* variable2 / options* PLOTS = *(plot-list)*;

可选绘图类型：

AGREEPLOT——双向（配对）表

CUMFREQPLOT——单向表

DEVIATIONPLOT——单向（卡方检验）表

FREQPLOT——（任意）

KAPPAPLOT——三维表

ODDSRATIOPLOT——h×2×2（MEASURES or RELRISK）

RELREISKPLOT——h×2×2（MEASURES or RELRISK）

RISKDIFFPLOT——h×2×2（RELRISK）

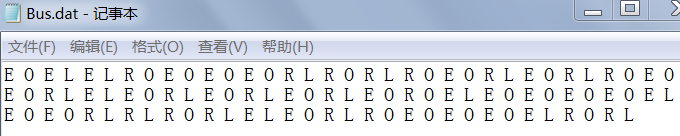
WTKAPPAPLOT——h×r×r (r>2) （配对表）

注：FREQPLOT可以加选项，例如分组条形图默认是竖直排列，若要改用水平排列，可以用：

TABLES variable1 \* variable2 / PLOTS = FREQPLOT(TWOWAY = GROUPHORIZONTAL);

若要堆叠分组条形，用“TWOWAY=STACKED”。

**例1** 一组常规公交车（R: Regular）和快速公交车（E: Express）的延误（L: Late）或准时（O: On Time）的数据（C:\MyRawData\Bus.dat）：



读入数据，用PROC FREQ过程步计算频数，并做卡方检验。

**代码：**

**data** bus;

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

input BusType $ OnTimeOrLate $ @@;

**run**;

**proc** **format**;

value $type 'R'='Regular'

'E'='Express';

value $late 'O'='On Time'

'L'='Late';

**run**;

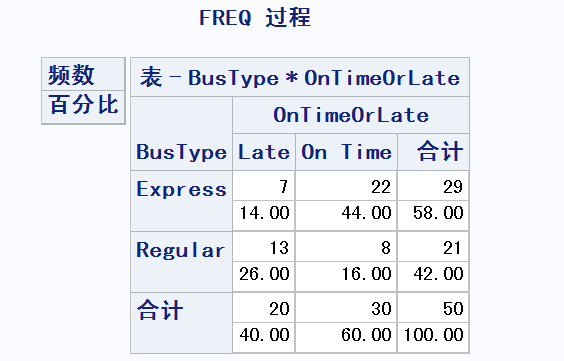
**proc** **freq** data = bus;

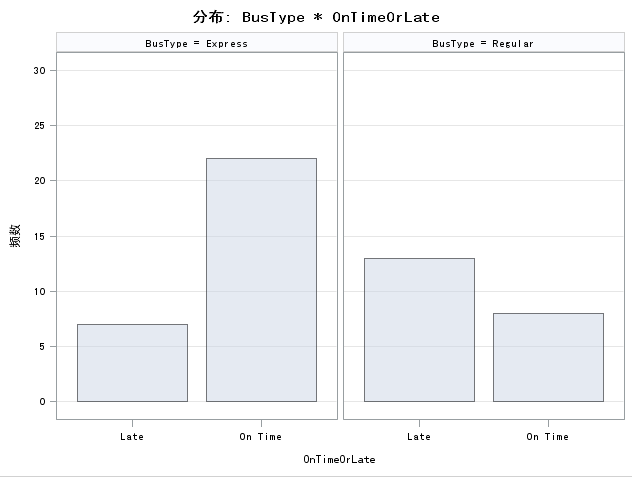
tables BusType \* OnTimeOrLate / NOROW NOCOL CHISQ PLOTS=FREQPLOT(TWOWAY=GROUPHORIZONTAL);

format BusType $Type. OnTimeOrLate $Late.;

**run**;

运行结果：









程序说明：

（1）常规公交车延迟率为61.9%, 快速公交车延迟率为24.14%;

（2）卡方检验的卡方值为7.2386，P值为0.0071<α=0.05; 说明两种公交车的延迟率有着明显差异，结果具有统计学意义；同时也说明“延误或准时与否”与选择哪种公交车是有关系的；另外，Fisher精确检验的结果也支持这一结论。