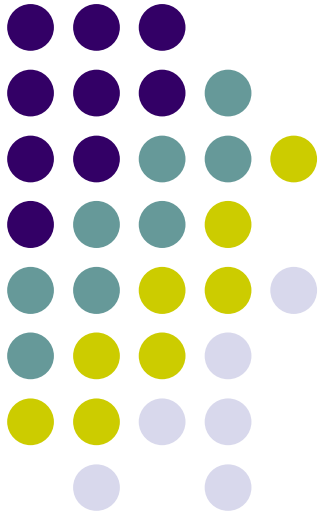
定性数据的建模分析

柯青

13660799897

493857517@qq.com



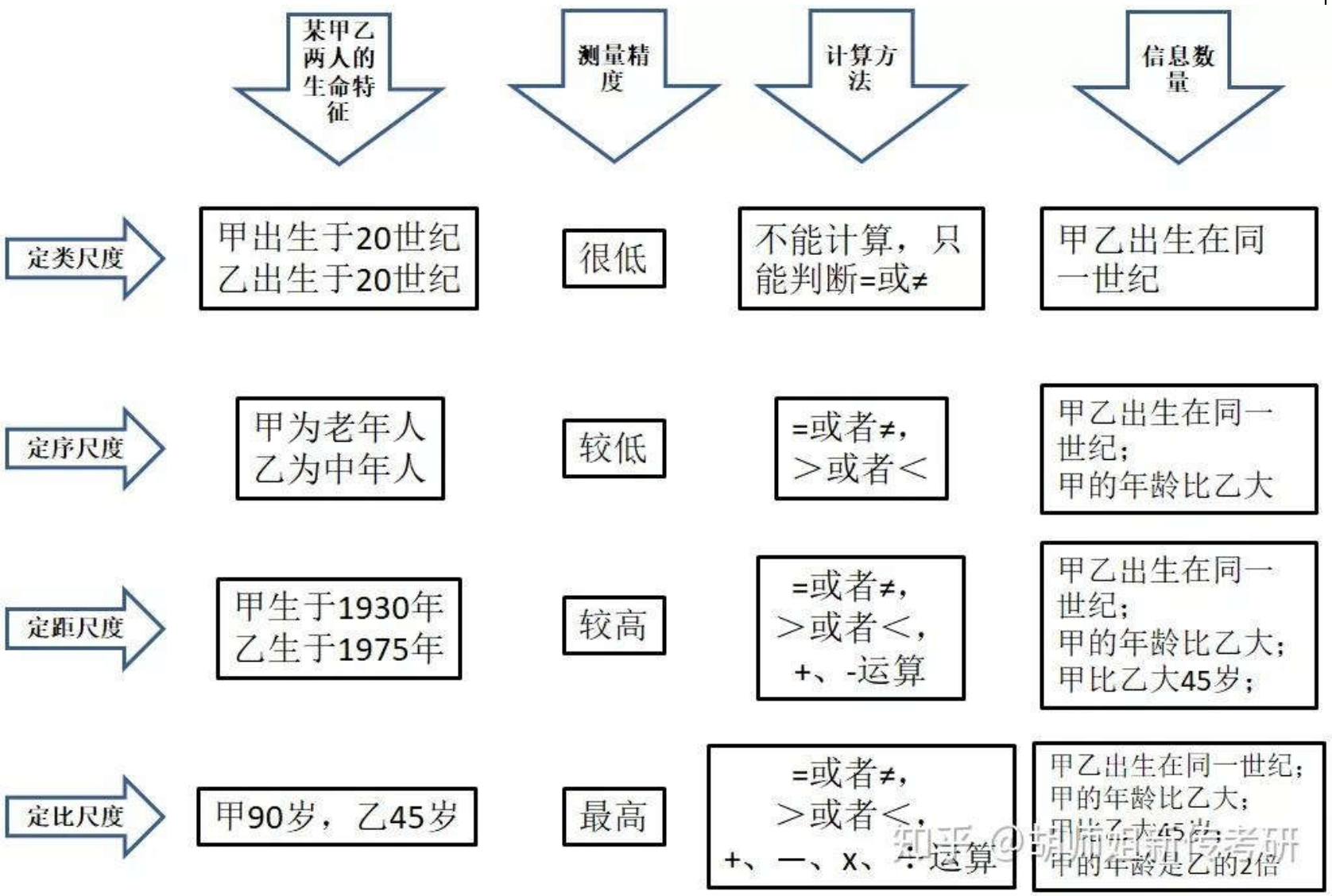
变量测量尺度

四种测量尺度的比较				
尺度	特征	数字性质	平均量度值	统计检验
分类数据 定类尺度	相互排斥且可辨别的类型	=，≠	众数	χ^2
分类数据 定序尺度	等级顺序大于或小于	>，<	中位数	符号秩检验
连续数据 定距尺度	尺度上的单位具有相等的意义	+，—	算术平均数	t 检验，F 检验
连续数据 定比尺度	有一个真正意义的零点	+，— ×，÷	几何平均数	t 检验，F 检验

变量测量尺度（续1）

	定类尺度	定序尺度	定距尺度	定比尺度
类别区分(=,≠)	有	有	有	有
次序区分(>,<)	-	有	有	有
距离区分(+,—)	-	-	有	有
比例区分(×,÷)	-	-	-	有

变量测量尺度（续2）



统计方法（看变量）

自变量X数据类型	因变量Y数据类型	研究方法
分类（且为两类）	连续	t检验
分类	连续	方差分析
分类	分类	卡方分析
分类	连续（且不满足正态性）	非参数检验

差异关系研究

因变量数据类型	研究方法
连续	线性回归
分类（且为两类）	二元Logistic/Probit回归分析
分类（类别为2+且无序）	多元无序Logistic/Probit回归分析
分类（类别为2+且有序）	多元有序Logistic/Probit回归分析

影响关系研究

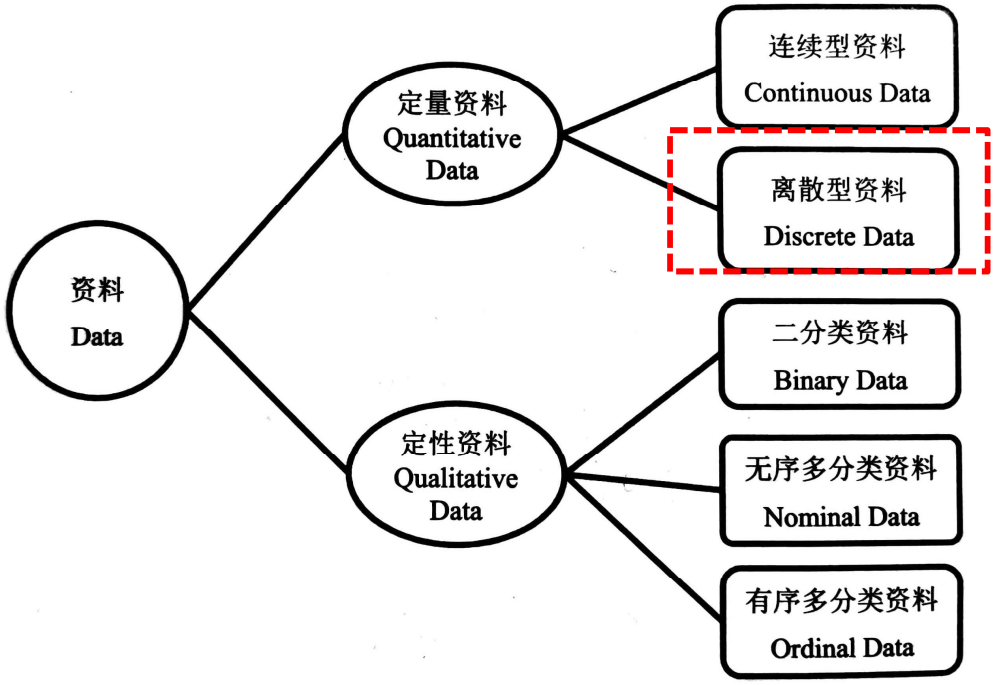
延伸思考1：

（1）**离散数据**（如计数资料）和分类数据有何区别？

（2）**离散数据**可否采用连续数据的方法进行分析？

“连续→离散→分类”
是信息损失方向

“分类→离散→连续”
是建模/度量升级方向



延伸思考2：

（1）**分类数据**使用**卡方分析**（列联表分析以及卡方检验）有何局限？

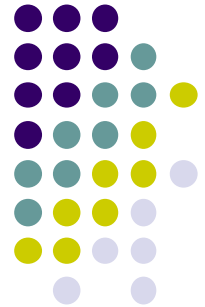
（2）**分类数据**建模能否借鉴**方差分析**模型的构造思想，对各变量的主效应及变量间的交互效应进行分析？

总结：

- 卡方分析更多地应用于二维列联表，无法同时研究多个分类变量间的关系，更无法研究它们的交互作用。
- 方差分析的因变量是连续变量，且对其分布有特定要求。
- 卡方分析及方差分析无法对连续自变量的影响进行深入分析。

- 因变量为分类资料服从二项分布或多项分布，可用**Logistic回归模型**。
- 因变量为计数资料服从多项分布或泊松分布，可用**对数线性模型**。

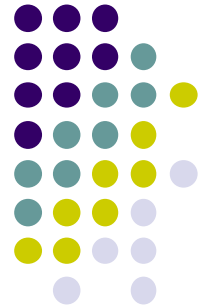
Logistic回归模型和对数线性模型比较快速入门（1）



- 把“调查数据”想成一盒彩色玻璃球，每颗球有颜色、大小、花纹三个标签，盒子边放着一台“天平”和一台“概率机”。

- **对数线性模型——用“天平”看格子频数**
  - 先把球按“颜色×大小×花纹”分成很多小格。
  - 天平称的是**每格里球的“个数”**。
  - 问：哪几格的球明显多/少？→ 就知道“颜色、大小、花纹谁和谁关联”。

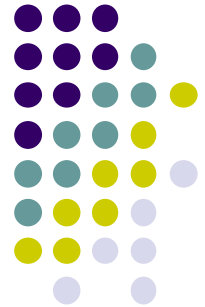
Logistic回归模型和对数线性模型比较快速入门（2）



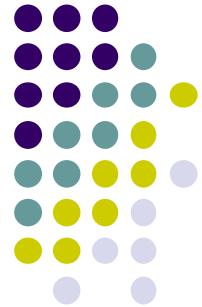
- 把“调查数据”想成一盒彩色玻璃球，每颗球有颜色、大小、花纹三个标签，盒子边放着一台“天平”和一台“概率机”。

- **Logistic回归——用“概率机”看结果发生概率**
  - 先指定一个标签当“结果”（比如“花纹”）。
  - 概率机只关心：**在同样颜色、同样大小下，球出现某花纹的几率是多少？**
  - 问：颜色或大小一变，几率涨多少？→ 就知道“谁影响了花纹”。

Logistic回归模型和对数线性模型比较快速入门（3）



- 同：
  - 都用“**log**”把乘法变加法，方便计算
- 异：
  - 对数线性——看**格子频数**，研究“谁跟谁有关联”
  - Logistic——**看结果概率**，研究“谁影响谁”
- 关联：
  - Logistic是**对数线性**在你有明确目标时的一种特殊应用和视角转换。



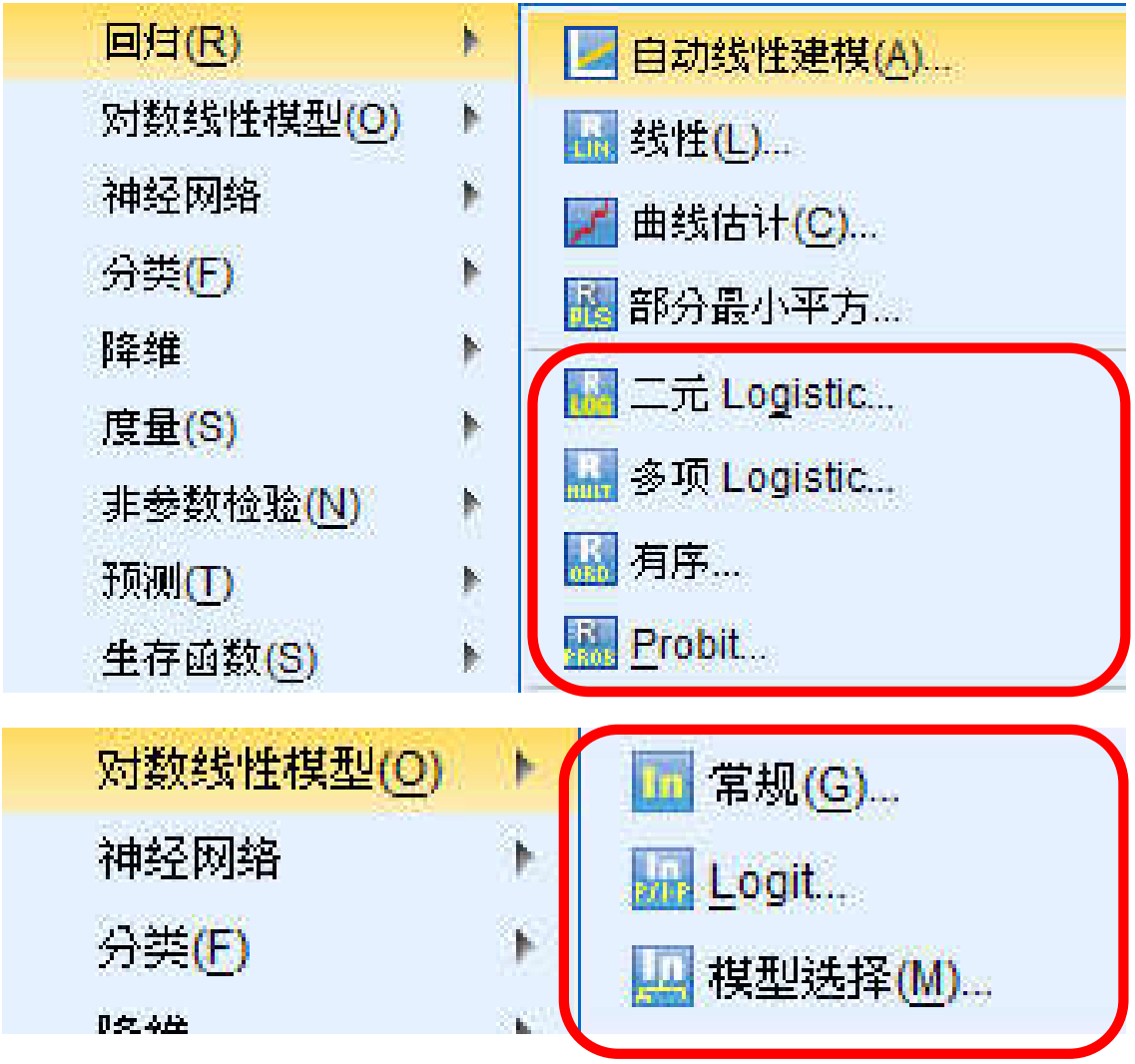
- **例：**分析不同年级/不同性别学生的考试成绩，选择卡方分析、方差分析、Logistic回归模型，还是对数线性模型？
  - 不同年级还是不同性别？
  - 考试成绩是百分制还是等级制？
  - 百分制成绩是否服从正态分布？
  - 等级制成绩是否合格？合格及以上具体等级？
  - 等级制成绩是否优秀？不同年级/性别的优秀等级人次？

  。。。。。。

目录（1）

目录（2）

分析工具——SPSS



Logistic回归模型

对数线性模型

分析工具——R语言（1）

Generalized Linear Model

- glm()函数

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

二项Logistic回归模型

mymodel1 <- glm (Y ~ X1+X2+X3 , family = binomial (link=logit), data = mydata)

二项Probit回归模型

mymodel2 <- glm (Y ~ X1+X2+X3 , family = binomial (link=probit), data = mydata)

Poisson回归模型

mymodel3 <- glm (Y ~ X1+X2+X3 , family = poisson ( ), data = mydata)

分析工具——R语言（2）

Proportional Odds Logistic Regression

- polr()函数（MASS包）

```
polr(formula, data, weights, start, ..., subset, na.action,
     contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "loglog", "cloglog", "cauchit"))
```

注：计算 Hessian 矩阵，后续 summary()
才能给出标准误与 p 值

多项有序Logistic回归模型

mymodel4 <- polr (Y ~ X1+X2+X3 , method =“logistic”, Hess=T, data = mydata)

多项有序Probit回归模型

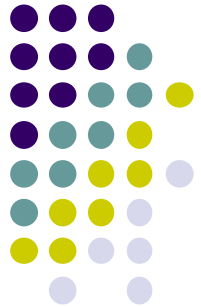
mymodel5 <- polr (Y ~ X1+X2+X3 , method =“probit”, Hess=T, data = mydata)

参考书籍

- **贾俊平**等. 统计学(第8版). 北京: 中国人民大学出版社, 2021

- **薛薇**. 基于**SPSS**的数据分析(第6版). 北京:中国人民大学出版社, 2025
- **张文彤**等. **SPSS**统计分析高级教程(第3版). 北京:高等教育出版社, 2018

- **王斌会**. 多元统计分析及**R**语言建模(第5版). 北京:高等教育出版社, 2020
- **王汉生**等.商务数据分析与应用—基于**R**(第3版). 北京: 中国人民大学出版社, 2022



1. 铺垫知识

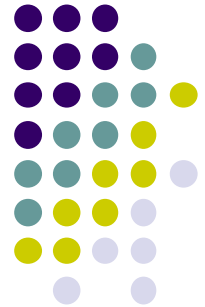
1.1 多项分布与Poisson分布

1.2 卡方分析

1.3 方差分析

1.4 常用统计模型串讲

1.1 多项分布与Poisson分布



1.1.1 离散型随机变量

1.1.2 二项分布

1.1.3 多项分布

1.1.4 Poisson分布

1.1.1 离散型随机变量

- 在同一组条件下，如果每次试验可能出现这样或那样的结果，并且所有结果都能列举出来，即X的所有可能值$x_1, x_2, \ldots, x_n$都能列举出来，而且X的可能值$x_1, x_2, \ldots, x_n$具有确定概率 $P(x_1)$，$P(x_2)$，$\ldots$，$P(x_n)$，其中$P(x_i) = P(X=x_i)$，称为概率函数，则 X 称为 $P(X)$ 的**随机变量**，$P(X)$ 称为随机变量 X 的**概率函数**。

- ★**离散型随机变量**——随机变量X的所有取值都可以**逐个列举**出来（如一批产品中取到次品的个数，单位时间内某公交站台的候车人数）

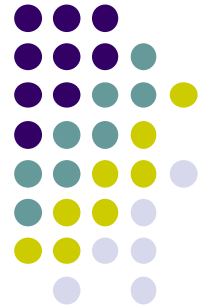
- **连续型随机变量**——随机变量X的所有取值无法逐个列举出来，而是取**数轴上某一区间**内的任一点（如一批电子元件的寿命，实际工作中常遇到的测量误差）

（1）离散型随机变量的概率分布

$X=x_i$	x_1，x_2，…，x_n
$P(X=x_i)=p_i$	p_1，p_2，…，p_n

$P(X=x_i)=p_i$是X的概率函数，且 $\sum_{i=1}^{n} p_i = 1$

（2）离散型随机变量的期望值（*E(X)*或*μ*）

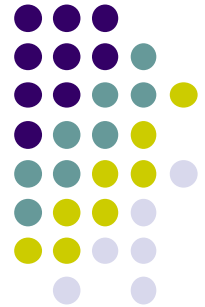


- 在离散型随机变量X的一切可能值的完备组中，各可能取值x_i与其对应概率p_i的乘积之和。

$$E(X)=\sum_{i=1}^{n}x_i p_i\ (X\text{取有限个值})$$

$$E(X)=\sum_{i=1}^{\infty}x_i p_i\ (X\text{取无穷个值})$$

（3）离散型随机变量的方差（$D(X)$,$V(X)$或σ_X^2）



- 随机变量X的每一个取值与期望值的离差平方的数学期望。

$$D(X)=E[X-E(X)]^2=\sum_{i=1}^{\infty}[x_i-E(X)]^2 p_i,$$

式中，$p_i=P\{X=x_i\}(i=1,2,\cdots)$

或者，$D(X)=E(X^2)-[E(X)]^2$

1.1.2 二项分布

- 二项分布与**伯努利试验**有关：
  - 包含了n个相同的试验
  - 每次试验只有两个可能的结果：“成功”和“失败”
  - 出现“成功”的概率p对每次试验是相同的，“失败”的概率q也相同，且$p+q=1$
  - 试验是相互独立的
  - 试验“成功”或“失败”可以计数

★0-1分布(伯努利分布，当n=1)：某次试验是否成功

$$P\{X=k\}=p^k(1-p)^{1-k},\quad k=0,1$$

- 进行 n 次重复独立试验，出现“成功”的次数 X 的概率分布称为**二项分布**（binomial distribution），记作 $X \sim B(n, p)$

$$P\{X = x\} = C_n^x p^x q^{n-x}, x = 0, 1, 2, \cdots, n$$

式中：$C_n^x = \dfrac{n!}{x!(n-x)!}$

且：$\sum_{x=0}^{n} C_n^x p^x q^{n-x} = (p+q)^n = 1$

- 二项分布的**期望值**：

$$E(X) = np$$

- 二项分布的**方差**：

$$D(X) = npq$$

1.1.3 多项分布

- 多项分布（Multinomial distribution）是二项分布的推广（**如扔骰子，足球比赛等**）：
  - 每次试验有多种可能的结果，但是每种结果只会出现一个；
  - 每种结果都有各自发生的概率，所有结果的发生概率之和为1；
  - 各次试验相互独立，每次试验结果都不受其他各次试验结果的影响。

- 假设某个多项分布试验可能发生结果的数量为k，依据历史数据，每种结果发生的统计概率分别为$p_1, p_2, \ldots, p_k$。现在进行n次多项分布试验，假设观测到结果a_1的次数为x_1次，结果a_2的次数为x_2次，…，结果a_k的次数为x_k次，那么多项分布的**联合概率函数**为：

$$P(X_1 = x_1, X_2 = x_2, \cdots X_k = x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

- 多项分布每一个结果的**期望值**：

$$E(X_i) = np_i$$

- 多项分布每一个结果的**方差**：

$$D(X_i) = np_i(1-p_i)$$

1.1.4 Poisson分布

- 泊松分布（Poisson Distribution）用于描述在一指定时间范围内或在指定的面积或体积之内某一事件出现次数的分布。
  - 某企业每月发生事故的次数
  - 单位时间内到达某一服务柜台需要服务的顾客人数
  - 人寿保险公司每天收到的死亡声明的个数
  - 某种仪器每月出现故障的次数
  - 。 。 。 。 。 。

- Poisson分布的**概率函数**（λ为给定的时间范围内事件的平均数）：

$$P(X)=\frac{\lambda^{x}e^{-\lambda}}{x!},x=0,1,2,\cdots$$

- Poisson分布的**期望值**和**方差**：

$$E(X)=D(X)=\lambda$$

二项分布、多项分布与Poisson分布

- 若$n = 1$，则二项分布化为**0-1分布**
- 若$k = 2$，则多项分布即为**二项分布**。
- 若n很大，p很小，且 $np \to \lambda$ ，则二项分布近似**Poisson 分布**（实际应用中p≤0.25，n>20，np≤5近似效果良好）。

$$C_n^x p^x q^{n-x} \approx \frac{\lambda^x e^{-\lambda}}{x!}$$

1.2 卡方分析

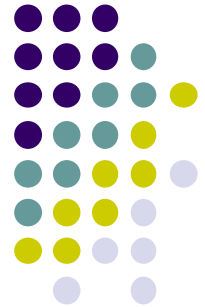
1.2.1 卡方分析的目的和基本任务

1.2.2 列联表的主要内容（任务之一）

1.2.3 列联表行列变量间关系的分析（任务之二）

1.2.4 卡方分析应用举例

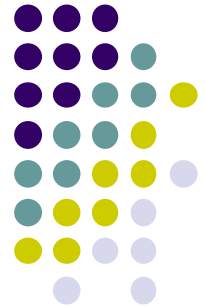
1.2.1 卡方分析的目的和基本任务



- **目的**：了解不同变量在不同水平下的数据分布
  - 例：等级制成绩与性别有关联吗？或男女同学的等级制成绩有无显著差异？

- **基本任务**：
  - 产生交叉列联表
  - 分析列联表中变量间的关系

1.2.2 列联表的主要内容



- **交叉列联表**是两个或两个以上的变量交叉分组后形成的频数分布表。

列变量

年龄* 血压 交叉制表

			血压			
			低血压	正常	高血压	合计
年龄	30岁以下	计数	27	48	23	98
		年龄 中的 %	27.6%	49.0%	23.5%	100.0%
		血压 中的 %	28.4%	20.7%	15.6%	20.7%
		总数的 %	5.7%	10.1%	4.9%	20.7%
	30-49岁	计数	37	91	51	179
		年龄 中的 %	20.7%	50.8%	28.5%	100.0%
		血压 中的 %	38.9%	39.2%	34.7%	37.8%
		总数的 %	7.8%	19.2%	10.8%	37.8%
	50岁以上	计数	31	93	73	197
		年龄 中的 %	15.7%	47.2%	37.1%	100.0%
		血压 中的 %	32.6%	40.1%	49.7%	41.6%
		总数的 %	6.5%	19.6%	15.4%	41.6%
合计		计数	95	232	147	474
		年龄 中的 %	20.0%	48.9%	31.0%	100.0%
		血压 中的 %	100.0%	100.0%	100.0%	100.0%
		总数的 %	20.0%	48.9%	31.0%	100.0%

行变量

观测频数

列百分比

行百分比

1.2.3 列联表行列变量间关系的分析

（1）列联表的卡方检验

（2）列联表卡方检验的说明

（1）列联表的卡方检验

① 提出原假设

② 计算检验统计量

③ 确定显著性水平和临界值

④ 得出结论和决策

① 提出原假设

- **原假设H_0**：行变量与列变量独立。

② 计算检验统计量

- **Pearson卡方统计量**

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}^{o} - f_{ij}^{e})^2}{f_{ij}^{e}}$$

r为列联表的行数

c为列联表的列数

f_{ij}^{o}为观测频数

f_{ij}^{e}为期望频数

$$f^e = \frac{RT}{n} \times \frac{CT}{n} \times n = \frac{RT \times CT}{n}$$

期望频数

RT – 单元格所在行的观测频数合计

CT – 单元格所在列的观测频数合计

年龄* 血压 交叉制表

			血压			
			低血压	正常	高血压	合计
年龄	30岁以下	计数	27	48	23	98
		期望的计数	19.6	48.0	30.4	98.0
		年龄 中的 %	27.6%	49.0%	23.5%	100.0%
		血压 中的 %	28.4%	20.7%	15.6%	20.7%
		总数的 %	5.7%	10.1%	4.9%	20.7%
	30-49岁	计数	37	91	51	179
		期望的计数	35.9	87.6	55.5	179.0
		年龄 中的 %	20.7%	50.8%	28.5%	100.0%
		血压 中的 %	38.9%	39.2%	34.7%	37.8%
		总数的 %	7.8%	19.2%	10.8%	37.8%
	50岁以上	计数	31	93	73	197
		期望的计数	39.5	96.4	61.1	197.0
		年龄 中的 %	15.7%	47.2%	37.1%	100.0%
		血压 中的 %	32.6%	40.1%	49.7%	41.6%
		总数的 %	6.5%	19.6%	15.4%	41.6%
合计		计数	95	232	147	474
		期望的计数	95.0	232.0	147.0	474.0
		年龄 中的 %	20.0%	48.9%	31.0%	100.0%
		血压 中的 %	100.0%	100.0%	100.0%	100.0%
		总数的 %	20.0%	48.9%	31.0%	100.0%

98×95÷474=19.6

95×20.7%=19.6

98×20.0%=19.6

期望频数的分布反映的是行列变量独立下的分布！

- **卡方统计量观测值**的大小取决于两个因素：
  - 列联表的单元格数目
  - 观测频数与期望频数的总差值

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}^{\,o} - f_{ij}^{\,e})^2}{f_{ij}^{\,e}}$$

- 在列联表确定的情况下，卡方统计量观测值的大小仅取决于**观测频数与期望频数的总差值**。总差值越大，卡方值也越大，实际分布与期望分布的差距越大，表明行列变量之间越不可能独立，反之同理。

③ 确定显著性水平和临界值

- 在卡方检验中，由于卡方统计量服从“**（行数-1）×（列数-1）**”个自由度的卡方分布，因此，在行列数目和显著性水平α确定时，卡方临界值是唯一确定的。

④ 得出结论和决策

- 第一，根据**统计量观测值**和临界值比较的结果进行决策。（观测值大于临界值，拒绝原假设，行列变量间不独立，存在相关关系）

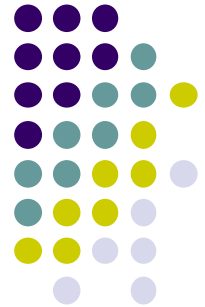
- 第二，根据**统计量观测值的概率P值**和显著性水平α比较的结果进行决策。（概率P值小于等于α，拒绝原假设，行列变量间不独立，存在相关关系）

（2）列联表卡方检验的说明

① 列联表各单元格中期望频数的大小

② 样本量的大小

① 列联表各单元格中期望频数的大小



- 列联表中不应有**期望频数小于1**的单元格，或不应有**大量的期望频数小于5**的单元格。
- 如果期望频数偏小的单元格大量存在，Pearson卡方统计量无疑会存在偏大趋势，容易导致拒绝原假设。
- 可以采用似然比卡方检验（**见Logistic回归模型部分内容**）等方法进行修正。

② 样本量的大小

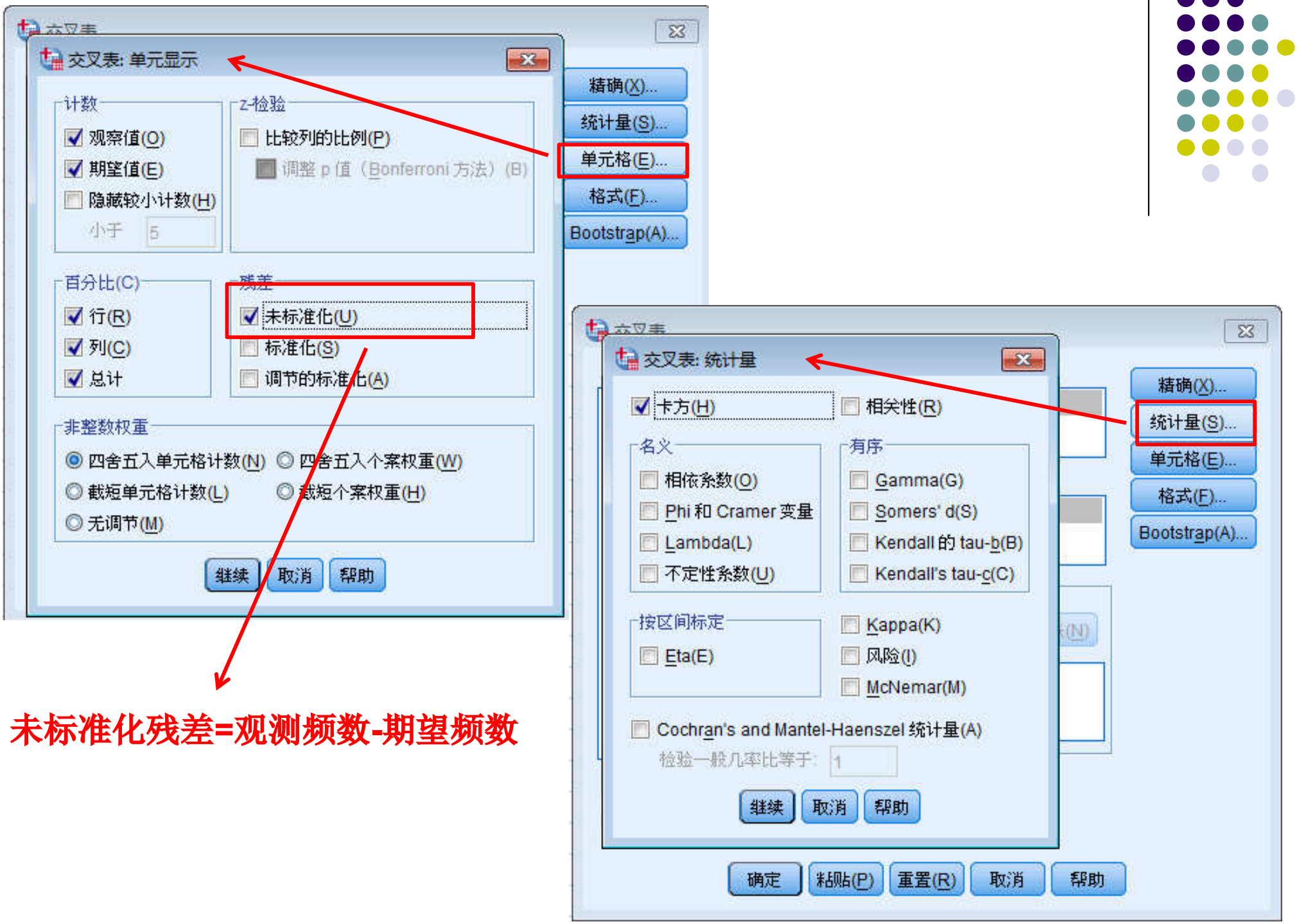
- 在某列联表中，如果各单元格中的样本量均扩大10倍，卡方值也会随之扩大10倍。但由于自由度和显著性水平并没有改变，卡方临界值不变，进而使拒绝原假设的可能性增加。
- 为此，有必要对Pearson卡方值进行必要的修正，以**剔除样本量的影响**。

1.2.4 卡方分析应用举例

- **案例**：分析不同性别学生在填报高考志愿时所考虑的因素是否存在差异，即**影响高考志愿填报的因素**与**性别**是否有关。

交叉列联表分析窗口


交叉表
专业分类 [专业分类]
年级 [年级]
志愿谁决定 [Q1]
规划课程指导 [Q3]
职业规划过程中最...
毕业后意向 [Q5]
对所学专业非常感...
很清楚所学专业的...
有明确的职业目标...
对自己未来的职业...
很了解职业生涯规...
认真阅读过一些职...
有自己明确的职业...
参加过许多次（5...
行(s):
性别 [性别]
列(C):
志愿决定因素 [Q2]
层1的1
上一张(V)
下一张(N)
在表层中显示层变量
显示复式条形图(B)
取消表格(T)
精确(X)...
统计量(S)...
单元格(E)...
格式(F)...
Bootstrap(A)...
确定
粘贴(P)
重置(R)
取消
帮助



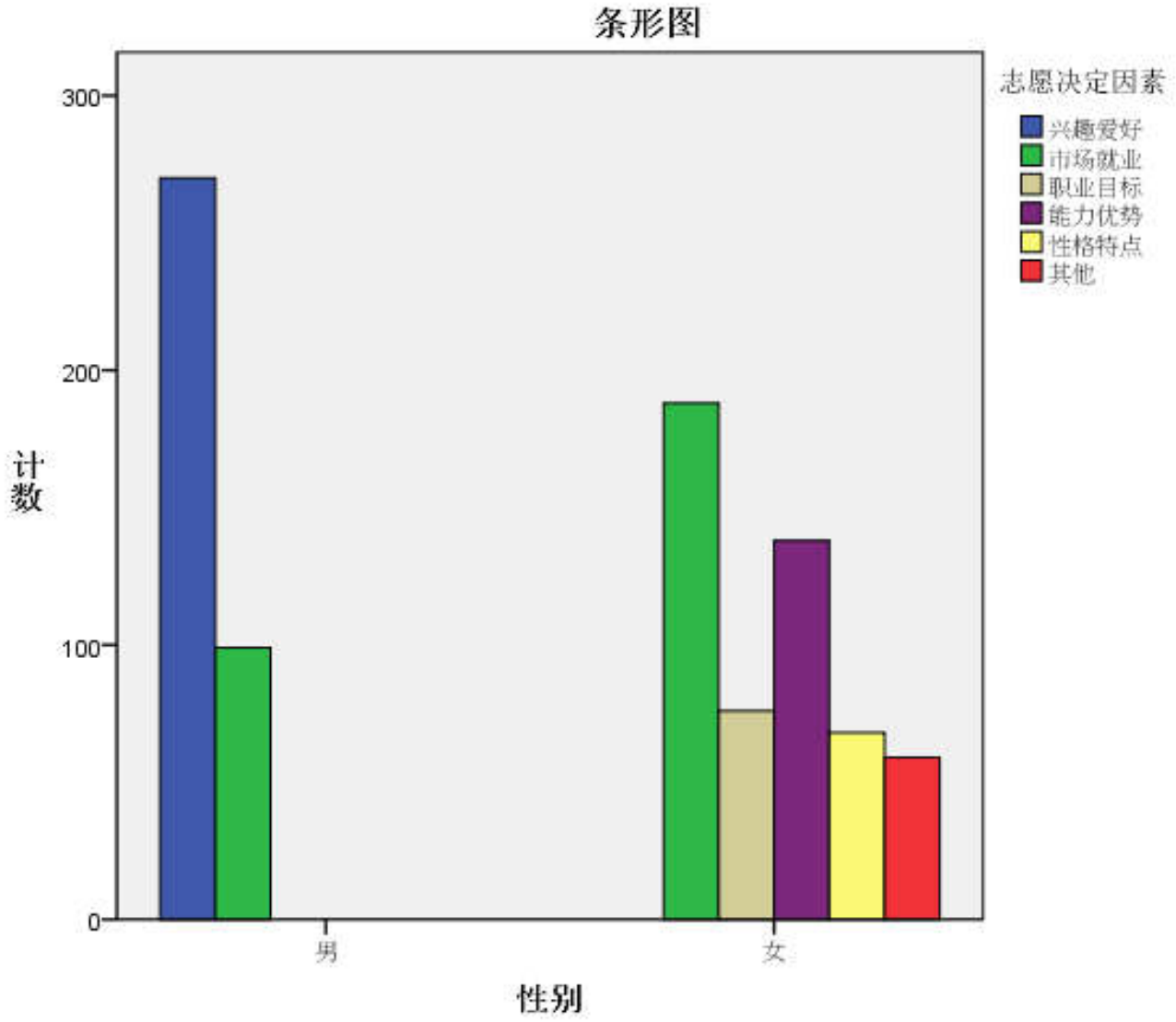
未标准化残差=观测频数-期望频数

性别和高考志愿决定因素的列联表

性别* 志愿决定因素 交叉制表

			志愿决定因素						
			兴趣爱好	市场就业	职业目标	能力优势	性格特点	其他	合计
性别	男	计数	270	99	0	0	0	0	369
		期望的计数	110.9	117.9	31.2	56.7	27.9	24.2	369.0
		性别 中的 %	73.2%	26.8%	0.0%	0.0%	0.0%	0.0%	100.0%
		志愿决定因素 中的 %	100.0%	34.5%	0.0%	0.0%	0.0%	0.0%	41.1%
		总数的 %	30.1%	11.0%	0.0%	0.0%	0.0%	0.0%	41.1%
		残差	159.1	-18.9	-31.2	-56.7	-27.9	-24.2	
	女	计数	0	188	76	138	68	59	529
		期望的计数	159.1	169.1	44.8	81.3	40.1	34.8	529.0
		性别 中的 %	0.0%	35.5%	14.4%	26.1%	12.9%	11.2%	100.0%
		志愿决定因素 中的 %	0.0%	65.5%	100.0%	100.0%	100.0%	100.0%	58.9%
		总数的 %	0.0%	20.9%	8.5%	15.4%	7.6%	6.6%	58.9%
		残差	-159.1	18.9	31.2	56.7	27.9	24.2	
合计		计数	270	287	76	138	68	59	898
		期望的计数	270.0	287.0	76.0	138.0	68.0	59.0	898.0
		性别 中的 %	30.1%	32.0%	8.5%	15.4%	7.6%	6.6%	100.0%
		志愿决定因素 中的 %	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		总数的 %	30.1%	32.0%	8.5%	15.4%	7.6%	6.6%	100.0%

不同性别高考志愿决定因素的条形图



BTW，条形图（柱状图）和直方图的区别？

BTW，条形图（柱状图）和直方图的区别？

- 条形图
  - 分类型变量
  - 比大小
- 直方图
  - 连续型变量
  - 数多少

举例：10个不同重量的苹果：
1）展示0-10g，10-20g，20-30g重量的苹果有多少个
2）展示每个苹果的具体重量

不同性别对高考志愿决定因素的一致性检验结果

样本量较大时，与Pearson卡方非常接近，检验结论通常也是一致的。

只适用于定序型变量，不能用于定类型变量。

卡方检验

	值	df	渐进 Sig. (双侧)
Pearson 卡方	630.094[a]	5	.000
似然比	846.425	5	.000
线性和线性组合	450.418	1	.000
有效案例中的 N	898		

a. 0 单元格(0.0%) 的期望计数少于 5。最小期望计数为 24.24。

拒绝原假设

说明适合做卡方检验

1.3 方差分析

1.3.1 常用术语

1.3.2 模型入门

1.3.3 适用条件

1.3.4 单因素方差分析模型案例

1.3.5 多因素方差分析模型案例

1.3.1 常用术语

（1）因素与水平

（2）单元

（3）元素

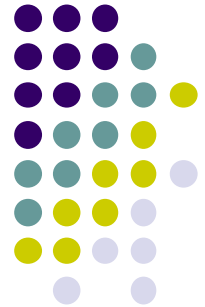
（4）均衡

（5）协变量

（6）交互作用

（7）固定因子与随机因子

（1）因素与水平



- **因素（factor）**也称为因子，是指可能对因变量有影响的分类变量。
- 分类变量的不同取值等级（类别）就称为**水平（level）**。

（2）单元

- **单元（cell）**又称为水平组合，是指各因素各个水平的组合。（**在研究性别（2水平）、血型（4水平）对成年人身高的影响时，最多可以有2*4=8个单元**）

（3）元素

- **元素（element）**是指用于测量因变量值的最小单位。根据具体的实验设计，一个单元格内可以有一个或多个元素，甚至也可以没有元素（**比较不同职业的月收入，元素就是每种职业的每一位受访者**）

（4）均衡

- 如果在一个实验设计中任一因素各水平在所有单元格中出现的次数相同，且每个单元格内的元素数相同，则该实验是**均衡（balance）**的，否则就是不均衡的。

（5）协变量

- **协变量（covariates）**是指对因变量可能产生影响，而且需要在分析时对其影响加以控制的连续变量。
- 实际上，可以简单地把因素和协变量分别理解为分类自变量和连续自变量。

（6）交互作用

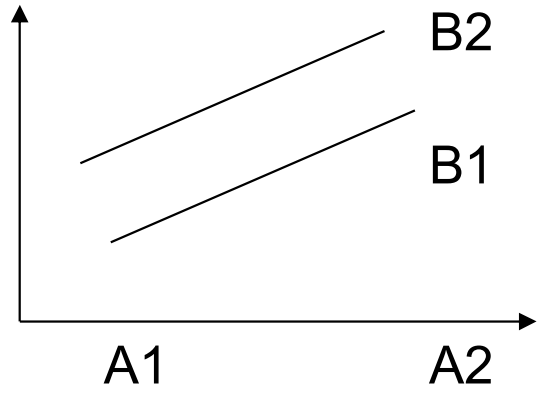
- 如果一个因素的效应大小在另一个因素不同水平下明显不同，则称为两因素间存在**交互作用（interaction）**。
- 当存在交互作用时，单纯研究某个因素的影响没有意义，必须在另一个因素的不同水平下研究该因素的影响大小。

注：如果所有单元格内都至多只有一个元素，则无法分析交互作用，只能不考虑它，如随机区组设计（配伍设计）方差分析

对交互作用的理解

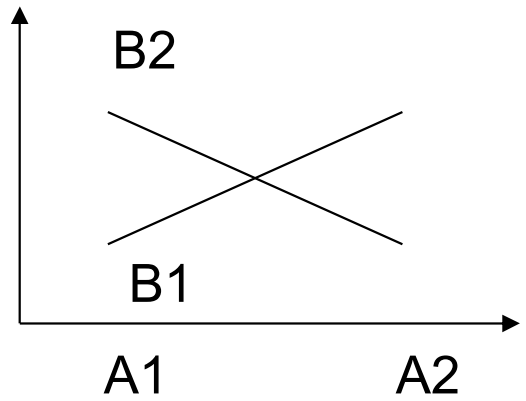
因素A和因素B无交互作用

	A1	A2
B1	2	5
B2	7	10



因素A和因素B有交互作用

	A1	A2
B1	2	5
B2	7	3



（7）固定因子与随机因子

- **固定因子（fixed factor）**指的是该因素在样本中所有可能水平都出现了。换言之，该因素的所有可能水平仅有此几种，对现有样本进行推断就可以得知该因素所有水平的状况，无需进行未知水平的外推。
- **随机因子（random factor）**指的是该因素所有可能的水平在样本中没有都出现，或不可能都出现。换言之，目前在样本中的这些水平是从总体中随机抽样出来的，如果重复本研究，则可能出现的因素水平会与现在完全不同。

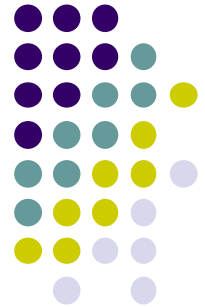
1.3.2 模型入门

（1）单因素方差分析模型的结构

（2）双因素方差分析模型的结构

（3）模型中效应的检验

（1）单因素方差分析模型的结构



- 假设因素A有k个水平，每个水平均有r个观测数据（r次试验），则在水平A_i下的第j次试验的观测值x_{ij}可以定义为：

$$x_{ij}=\mu+a_i+\varepsilon_{ij}, i=1,2,\cdots,k; j=1,2,\cdots,r$$

μ：观测变量总的平均值

a_i：因素水平A_i对试验结果产生的附加影响，称为水平A_i对观测变量产生的效应，且$\sum_{i=1}^{k}a_i=0$

ε_{ij}：抽样误差，是服从正态分布$N(0, \sigma^2)$的独立随机变量

- 如果因素*A*对观测变量没有影响，则各水平的效应a_i应**全部为0**，否则应**不全为0**。

H_0：对任意的i取值，都有$a_i = 0$

H_1：至少有一个$a_i \neq 0$

（2）双因素方差分析模型的结构

- 设因素A有k个水平，B有r个水平，每个交叉水平下均有l个样本（l次试验），则在因素A的水平A_i和因素B的水平B_j下的第m个观测值x_{ijm}可以定义为：

$$x_{ijm} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijm}$$

$$i = 1, 2, \cdots, k; \; j = 1, 2, \cdots, r; m = 1, 2, \cdots, l$$

ε_{ijm}：抽样误差，是服从正态分布$N(0, \sigma^2)$的独立随机变量

- 如果因素*A*（或*B*）对观测变量没有影响，则各水平的效应a_i（或b_j）应**全部为0**，否则**不全为0**。

$$H_0: a_i = 0; H_1: \text{至少有一个} a_i \neq 0$$

$$H_0: b_j = 0; H_1: \text{至少有一个} b_j \neq 0$$

- 同理，如果因素*A*和*B*对观测变量没有交互影响，则各水平的效应$(ab)_{ij}$应**全部为0**，否则**不全为0**。

（3）模型中效应的检验

$$SS_{总} = SS_{因素1} + SS_{因素2} + \cdots + SS_{误差}$$

$$MS_{因素i} = SS_{因素i} / DF_{因素i}$$

$$MS_{误差} = SS_{误差} / DF_{误差}$$

$$F_{因素i} = MS_{因素i} / MS_{误差}$$

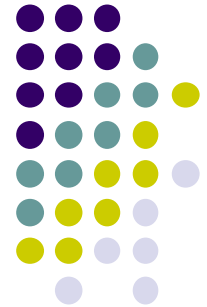
单因素方差分析：SST=SSA+SSE

双因素方差分析：SST=SSA+SSB+SSAB+SSE

三因素方差分析：

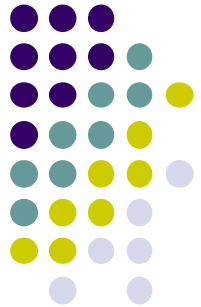
SST=SSA+SSB+SSC+SSAB+SSAC+SSBC+SSABC+SSE

方差分析的基本思想——变异分解



- 将样本的总变异分解为若干个部分，除有一部分代表**随机误差**的作用外，其余每个部分的变异分别代表了某个影响因素的作用（或交互作用）。
- 通过比较可能由某因素所致的变异与随机误差的大小，借助**F分布**做出推断，即可了解该因素对结果变量的影响是否存在。

检验步骤
——以单因素方差分析为例



① 提出原假设

② 选择检验统计量

③ 计算检验统计量的观测值和概率P值

④ 给定显著性水平α，并作出决策

① 提出原假设

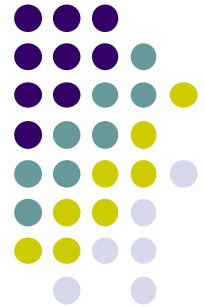
- **原假设H_0**：某因素A不同水平下观测变量各总体的均值无显著差异，该因素不同水平下的效应同时为0，记为$a_1=a_2=\ldots=a_k=0$，意味着该因素不同水平的变化没有对观测变量均值产生显著影响
- **备择假设H_1**：各效应值不同时为0

② 选择检验统计量

$$F=\frac{SSA/(k-1)}{SSE/(n-k)}=\frac{MSA}{MSE}\sim F(k-1,n-k)$$

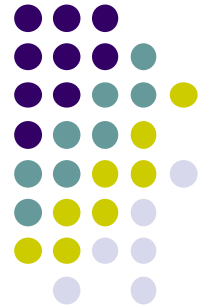
- **n**为总样本量；**k-1**和**n-k**分别是SSA和SSE的自由度；**MSA**是平均组间平方和，也称组间均方；**MSE**是平均组内平方和，也称组内均方。
- 除以自由度的目的是消除水平数和样本量对分析带来的影响。

③ 计算检验统计量的观测值和概率P值



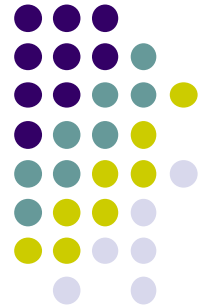
- 如果因素A对观测变量造成了显著影响，观测变量总的变差中该**因素**影响所占的比例相对于随机变量必然较大，**F值显著大于1**；
- 反之，如果因素A没有对观测变量造成显著影响，观测变量的变差应归结为由**随机变量**造成的，**F值接近1**。

④ 给定显著性水平α，并作出决策



- **概率P值小于显著性水平α**，则应拒绝原假设，认为因素A不同水平下观测变量各总体的均值存在显著差异，因素A的各个效应不同时为0，因素A的不同水平对观测变量均值产生了显著影响；
- **概率P值大于显著性水平α**，则不应拒绝原假设，认为因素A不同水平下观测变量各总体的均值无显著差异，因素A的各个效应同时为0，因素A的不同水平对观测变量均值没有产生显著影响。

检验步骤
——以双因素方差分析为例



① 提出原假设

② 选择检验统计量

③ 计算检验统计量的观测值和概率P值

④ 给定显著性水平α，并作出决策

① 提出原假设

- **原假设H_0**：各因素不同水平下观测变量各总体的均值无显著差异，各因素各效应和交互作用效应同时为0，这意味着各因素和它们的交互作用没有对观测变量产生显著影响。

② 选择检验统计量

- **固定效应模型**：

$$F_A = \frac{SSA/(k-1)}{SSE/[kr(l-1)]} = \frac{MSA}{MSE}$$

$$F_B = \frac{SSB/(r-1)}{SSE/[kr(l-1)]} = \frac{MSB}{MSE}$$

$$F_{AB} = \frac{SSAB/[(k-1)(r-1)]}{SSE/[kr(l-1)]} = \frac{MSAB}{MSE}$$

- **随机效应模型**：

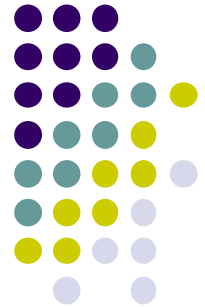
$$F_A = \frac{SSA/(k-1)}{SSAB/[(k-1)(r-1)]} = \frac{MSA}{MSAB}$$

$$F_B = \frac{SSB/(r-1)}{SSAB/[(k-1)(r-1)]} = \frac{MSB}{MSAB}$$

$$F_{AB} = \frac{SSAB/[(k-1)(r-1)]}{SSE/[kr(l-1)]} = \frac{MSAB}{MSE}$$

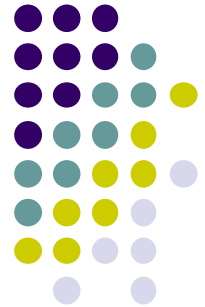
因素A有k个水平
因素B有r个水平
每个交叉水平下均有l个样本

③ 计算检验统计量的观测值和概率P值



- SPSS自动把相关数据代入各式，计算出各个F统计量的观测值和对应的概率P值。

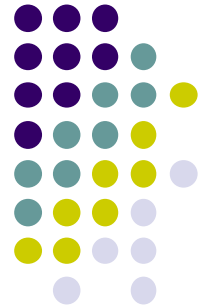
④ 给定显著性水平α，并作出决策



- **固定效应模型**：如果F_A的概率P值小于α，则应拒绝原假设，认为因素A不同水平下观测变量各总体的均值存在显著差异，因素A的各个效应不同时为0，因素A的不同水平对观测变量产生了显著影响。对因素B以及A，B交互作用的推断同理。
- **随机效应模型**：应首先对A，B的交互作用是否显著进行推断，再分别依次对A，B的效应进行检验。

注：若无交互作用，随机效应和固定效应设定的区别不大

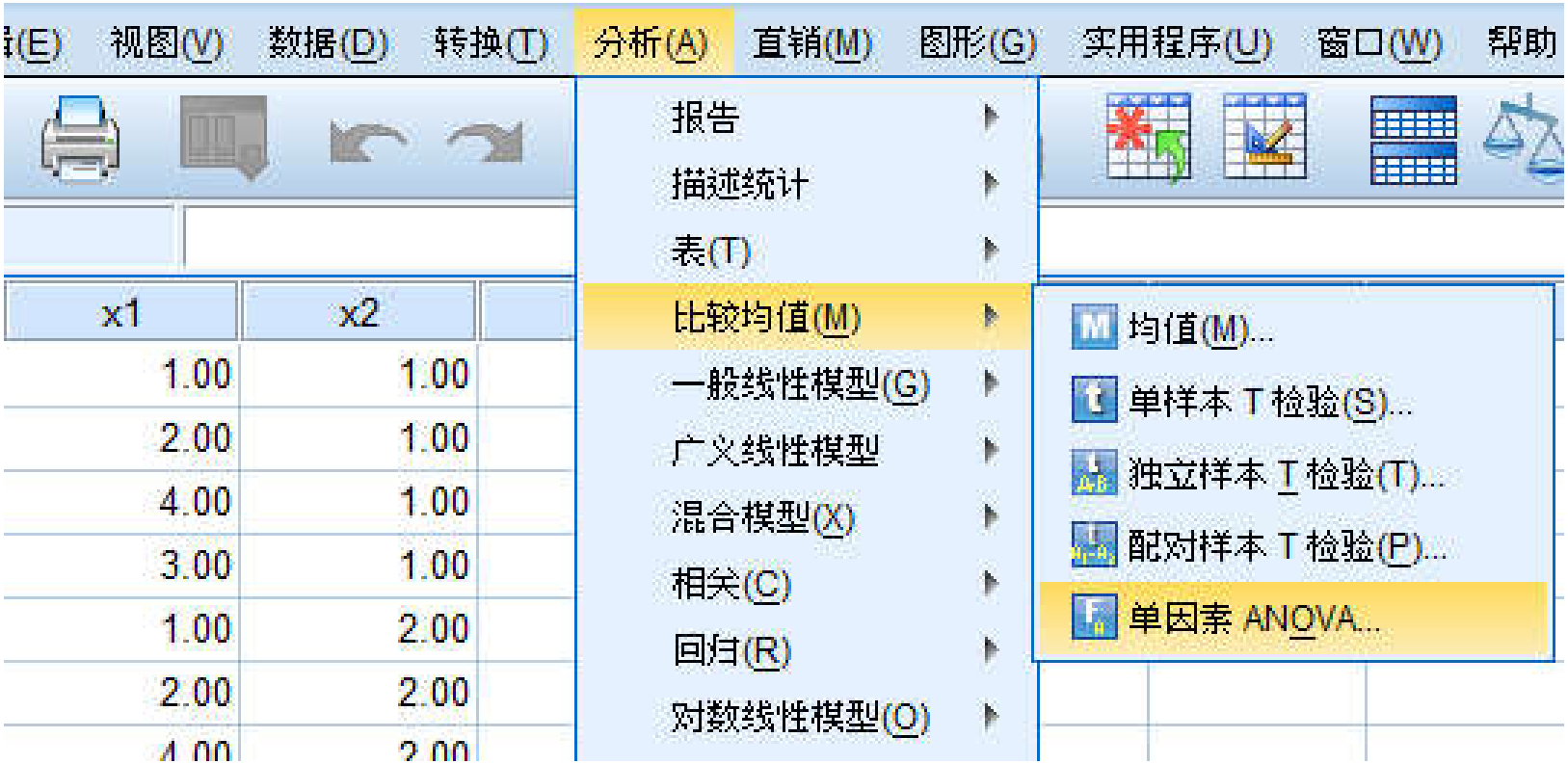
1.3.3 适用条件



- **独立性**：观察对象是来自于所研究因素的各个水平之下的独立随机抽样（Independence）
- **正态性**：每个水平下的因变量应当服从正态分布（Normality）
- **方差齐**：各水平下的总体具有相同的方差（Homoscedasticity）

1.3.4 单因素方差分析模型案例

- **案例**：广告形式或地区是否对商品销售额产生影响？



ANOVA分析窗口


单因素方差分析
地区 [x2]
因变量列表(E):
销售额 [x3]
因子(F):
广告形式 [x1]
对比(N)...
两两比较(H)...
选项(O)...
Bootstrap(B)...
确定
粘贴(P)
重置(R)
取消
帮助

ANOVA分析结果（广告形式VS地区）

单因素方差分析

不同广告形式对销售额的平均值产生显著影响

销售额

	平方和	df	均方	F	显著性
组间	5866.083 SSA	3	1955.361 MSA	13.483	.000
组内	20303.222 SSE	140	145.023 MSE		
总数	26169.306 SST	143			

单因素方差分析

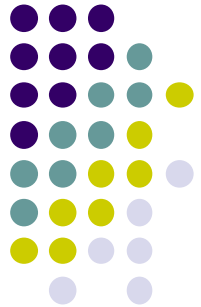
不同地区对销售额的平均值产生显著影响

销售额

	平方和	df	均方	F	显著性
组间	9265.306	17	545.018	4.062	.000
组内	16904.000	126	134.159		
总数	26169.306	143			

对比两张表，如果从单因素的角度考虑，广告形式对销售额的影响比地区大。

延伸思考：单因素方差分析的进一步分析



（1）方差齐性检验

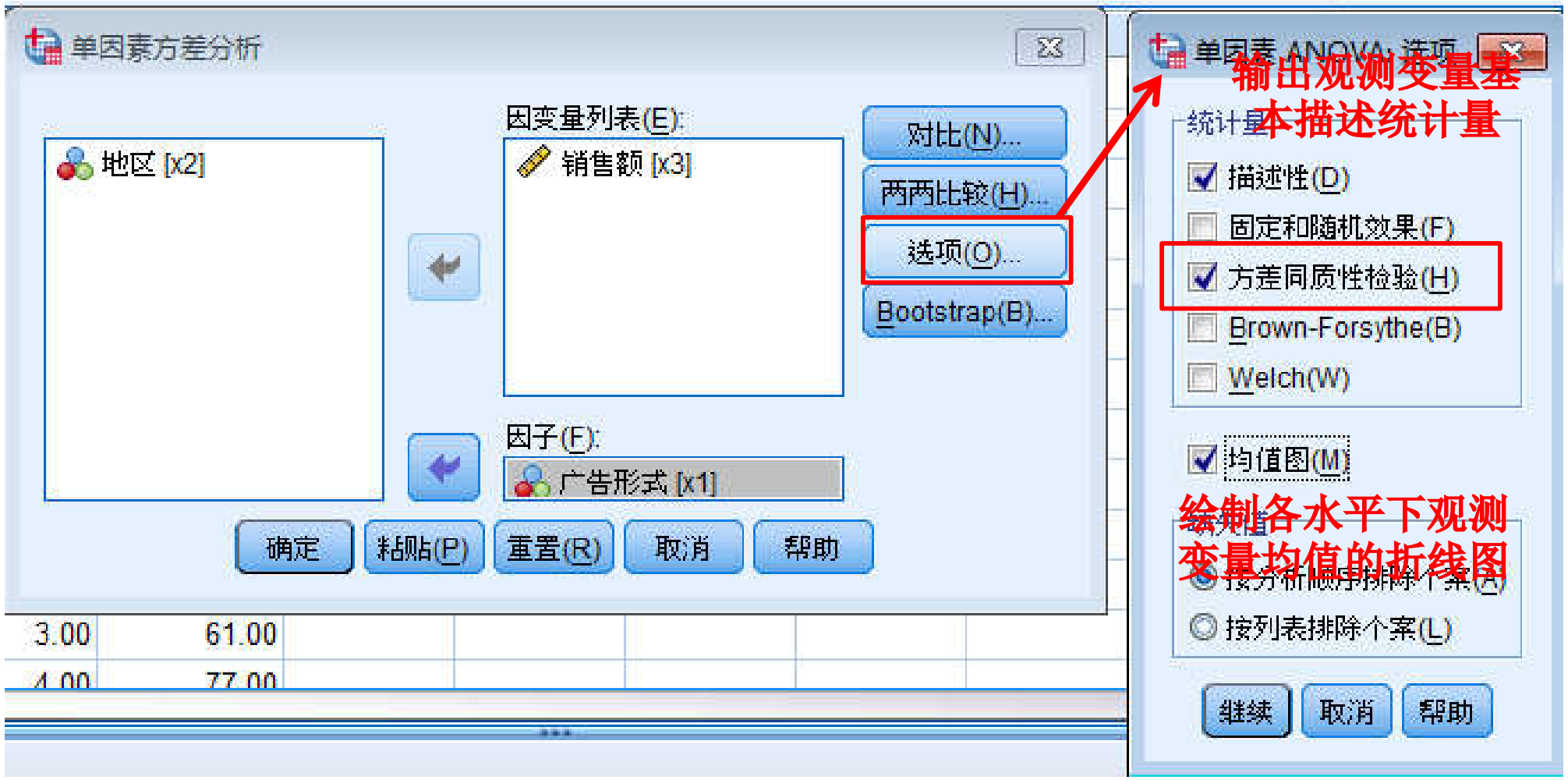
（2）多重比较检验

（3）对比检验

（4）趋势检验

（1）方差齐性检验

- **方差齐性检验**是对某因素不同水平下各观测变量总体方差是否相等进行分析。（**方差分析的前提要求**）
- 其**原假设H_0**是：各水平下观测变量总体方差无显著差异。实现思路同两独立样本t检验中的同方差检验。


单因素方差分析
因变量列表(E):
地区 [x2]
销售额 [x3]
对比(N)...
两两比较(H)...
选项(O)...
Bootstrap(B)...
因子(F):
广告形式 [x1]
确定
粘贴(P)
重置(R)
取消
帮助
3.00
61.00
单因素 ANOVA: 选项
统计量
描述性(D)
固定和随机效果(F)
方差同质性检验(H)
Brown-Forsythe(B)
Welch(W)
均值图(M)
按列表排除个案(L)
继续
取消
帮助
输出观测变量基本描述统计量
绘制各水平下观测变量均值的折线图

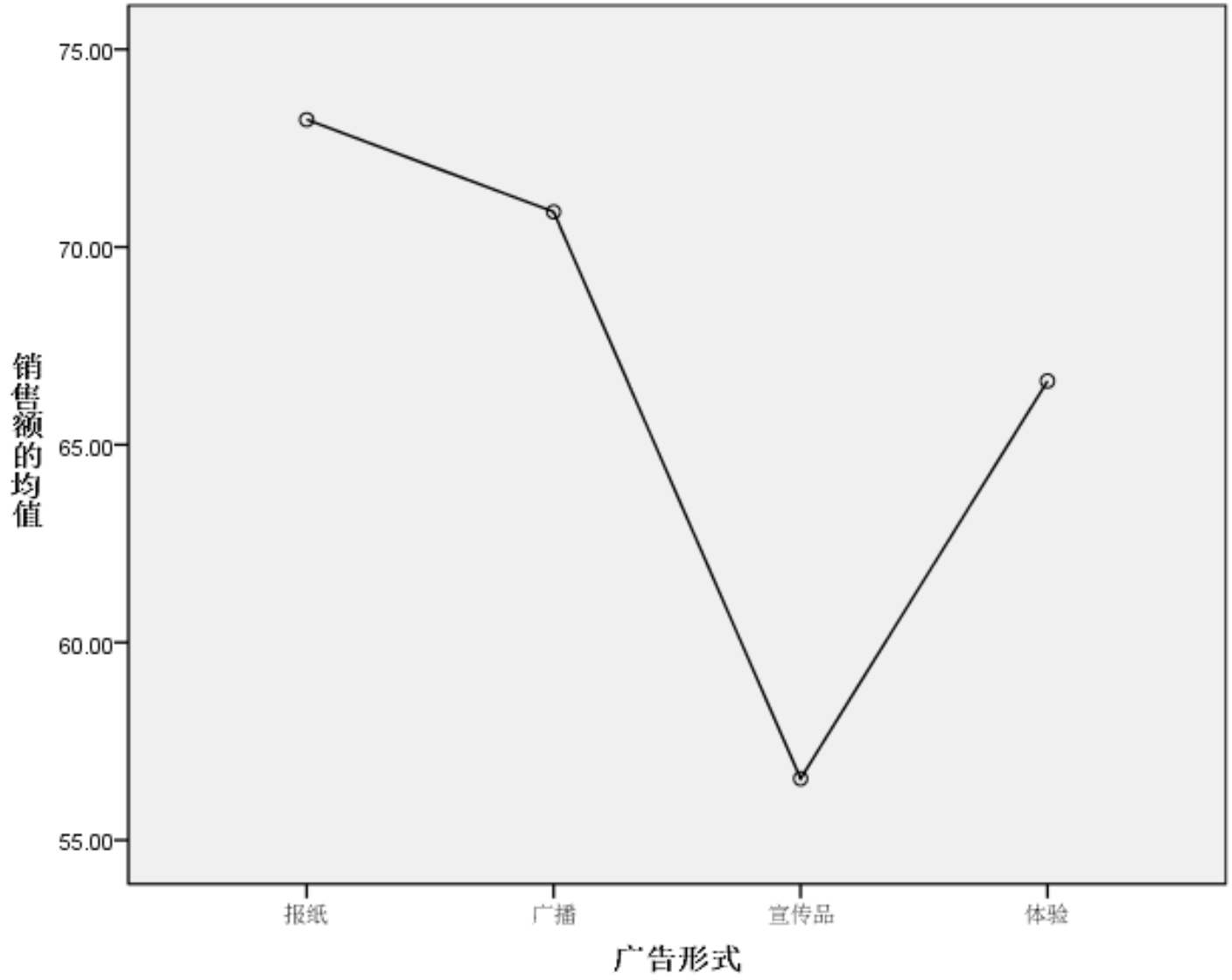
不同广告形式下销售额的基本描述统计量及95%置信区间

描述

销售额

	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
报纸	36	73.2222	9.73392	1.62232	69.9287	76.5157	54.00	94.00
广播	36	70.8889	12.96760	2.16127	66.5013	75.2765	33.00	100.00
宣传品	36	56.5556	11.61881	1.93647	52.6243	60.4868	33.00	86.00
体验	36	66.6111	13.49768	2.24961	62.0442	71.1781	37.00	87.00
总数	144	66.8194	13.52783	1.12732	64.5911	69.0478	33.00	100.00

不同广告形式下销售额均值折线图



不同广告形式下方差齐性检验结果

- 概率P值大于显著性水平，因此**不应拒绝原假设**，认为不同广告形式下销售额的总体方差无显著差异，满足方差分析的前提要求。

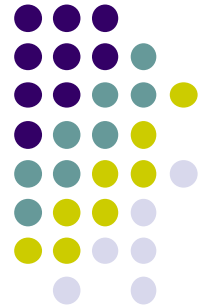
方差齐性检验

销售额

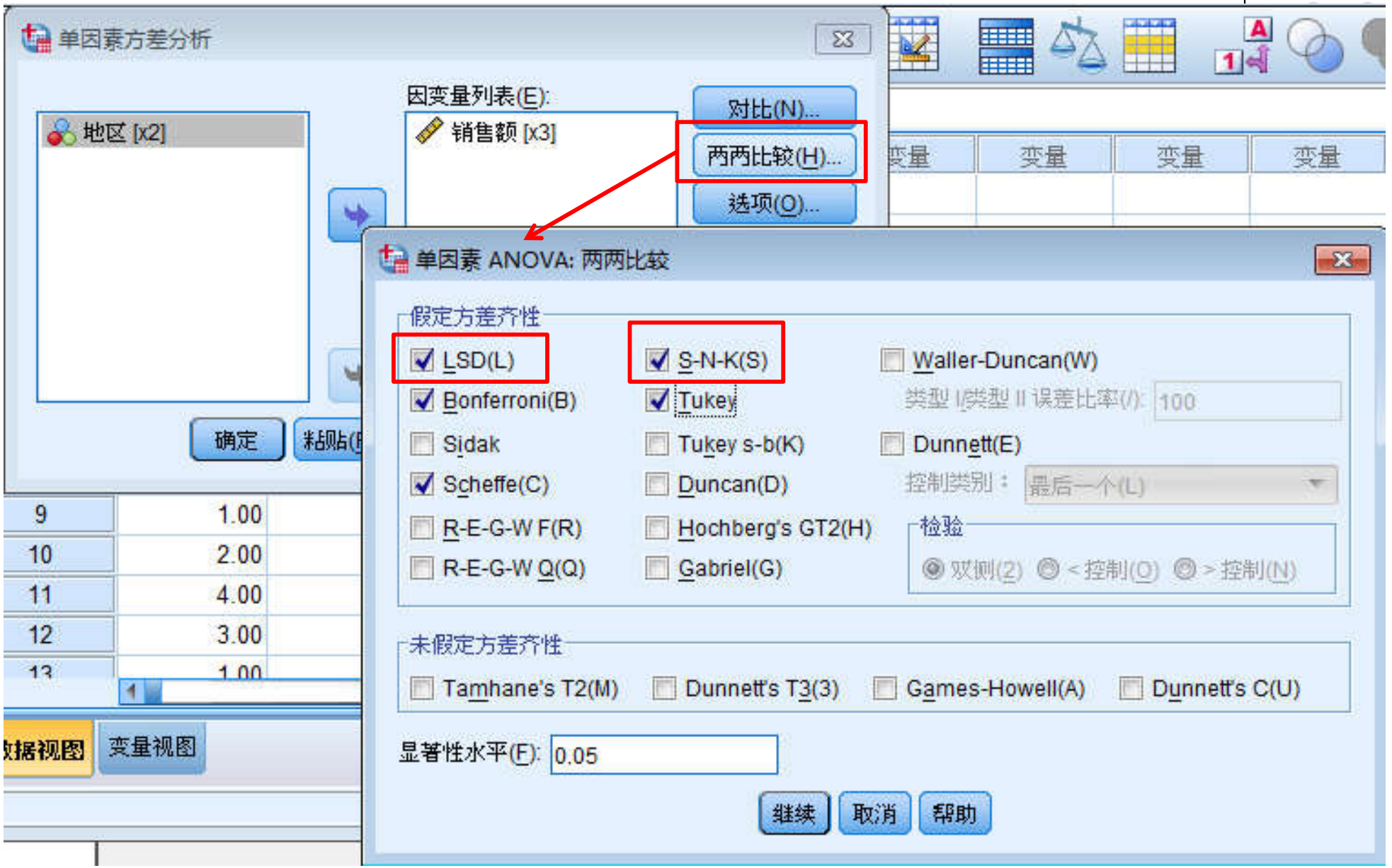
Levene 统计量	df1	df2	显著性
.765	3	140	.515

（2）多重比较检验

- 单因素方差分析的基本分析只能判断某因素是否对观测变量产生了显著影响。如果确实有影响，还应进一步确定该因素的不同水平对观测变量的影响程度如何，其中哪个水平的作用明显区别于其他水平，哪个水平的作用是不显著的，等等。
- **多重比较检验**利用了全部观测变量值，实现对各个水平下观测变量总体均值的逐对比较。



- 多重比较检验的**原假设H_0**是：第i和第j个水平下观测变量的总体均值μ_i和μ_j不存在显著差异
- 常用检验统计量的构造方法：
  - LSD方法（敏感度最高）
  - Bonferroni方法
  - Tukey方法
  - Scheffe方法
  - S-N-K方法


单因素方差分析
因变量列表(E):
地区 [x2]
销售额 [x3]
对比(N)...
两两比较(H)...
选项(O)...
确定
粘贴(P)
变量
变量
变量
变量
9 1.00
10 2.00
11 4.00
12 3.00
13 1.00
数据视图
变量视图
单因素 ANOVA: 两两比较
假定方差齐性
LSD(L)
S-N-K(S)
Waller-Duncan(W)
Bonferroni(B)
Tukey
类型 I/类型 II 误差比率(/): 100
Sidak
Tukey s-b(K)
Dunnett(E)
Scheffe(C)
Duncan(D)
控制类别: 最后一个(L)
R-E-G-W F(R)
Hochberg's GT2(H)
检验
R-E-G-W Q(Q)
Gabriel(G)
双侧(2)
< 控制(O)
> 控制(N)
未假定方差齐性
Tamhane's T2(M)
Dunnett's T3(3)
Games-Howell(A)
Dunnett's C(U)
显著性水平(F): 0.05
继续
取消
帮助

广告形式的多重比较检验（LSD方法）

多重比较

报纸和宣传品、报纸和体验、广播和宣传品、宣传品和体验两两显著差异

因变量: 销售额

	(I) 广告形式	(J) 广告形式	均值差 (I-J)	标准误	显著性	95% 置信区间	
						下限	上限
LSD	报纸	广播	2.33333	2.83846	.412	-3.2784	7.9451
		宣传品	16.66667*	2.83846	.000	11.0549	22.2784
		体验	6.61111*	2.83846	.021	.9993	12.2229
	广播	报纸	-2.33333	2.83846	.412	-7.9451	3.2784
		宣传品	14.33333*	2.83846	.000	8.7216	19.9451
		体验	4.27778	2.83846	.134	-1.3340	9.8896
	宣传品	报纸	-16.66667*	2.83846	.000	-22.2784	-11.0549
		广播	-14.33333*	2.83846	.000	-19.9451	-8.7216
		体验	-10.05556*	2.83846	.001	-15.6673	-4.4438
	体验	报纸	-6.61111*	2.83846	.021	-12.2229	-.9993
		广播	-4.27778	2.83846	.134	-9.8896	1.3340
		宣传品	10.05556*	2.83846	.001	4.4438	15.6673

*. 均值差的显著性水平为 0.05。

广告形式多重比较检验的相似性子集（S-N-K方法）

- 第一子集（宣传品组）组内相似（自身相似）的概率为1，第二子集组内相似的可能性大于0.05，为0.055。

销售额

	广告形式	N	alpha = 0.05 的子集 1	alpha = 0.05 的子集 2
Student-Newman-Keuls[a]	宣传品	36	56.5556	
	体验	36		66.6111
	广播	36		70.8889
	报纸	36		73.2222
	显著性		1.000	.055

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 36.000。

（3）对比检验

- 在多重比较检验中，如果发现某些水平与另外一些水平的均值差异显著，如有5个水平，其中，$\bar{x}_1$，$\bar{x}_2$，$\bar{x}_3$与$\bar{x}_4$，$\bar{x}_5$有显著差异，就可进一步比较这两组总的均值是否存在显著差异，即1/3($\bar{x}_1$+ $\bar{x}_2$+ $\bar{x}_3$)与1/2($\bar{x}_4$+ $\bar{x}_5$)是否有显著差异。
- 需事先指定各均值的系数c_i，如令c_1=1/3，c_2=1/3，c_3=1/3，c_4=-1/2，c_5=-1/2，且Σc_i=0，再对其线性组合进行检验的分析方法称为**对比检验**。

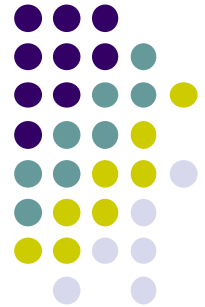
- 通过多重比较检验已知，宣传品效果最差，而其余三种略有差异。需进一步对**报纸的效果**与**广播和体验的整体效果**进行对比分析。

各种广告对比检验的系数说明

对比系数

对比	广告形式			
	报纸	广播	宣传品	体验
1	1	-.5	0	-.5

报纸效果与广播、体验整体效果的对比检验结果



t检验概率P值大于0.05，不应拒绝原假设

对比检验

报纸下的平均销售额比广播和体验下的平均销售额多4.47

看第一行结果

		对比	对比值	标准误	t	df	显著性（双侧）
销售额	假设方差相等	1	4.4722	2.45818	1.819	140	.071
	不假设等方差	1	4.4722	2.25053	1.987	90.771	.050

（4）趋势检验

- 当某因素为**定序型变量**时，趋势检验能够分析随着该因素水平的变化，观测变量值变化的总体趋势是怎样的（线性变化趋势，还是二次、三次等多项式变化趋势）。

- 假定不同地区的差异主要表现在人口密度方面（地区编号越大，人口密度越低），需进一步分析**不同地区的销售额总体上是否随着地区人口密度的降低而呈现出线性趋势**。

地区的趋势检验结果

单因素方差分析

销售额

		平方和	df	均方	F	显著性
组间	(组合)	9265.306	17	545.018	4.062	.000
	线性项 对比	543.938	1	543.938	4.054	.046
	偏差	8721.367	16	545.085	4.063	.000
组内		16904.000	126	134.159		
总数		26169.306	143			

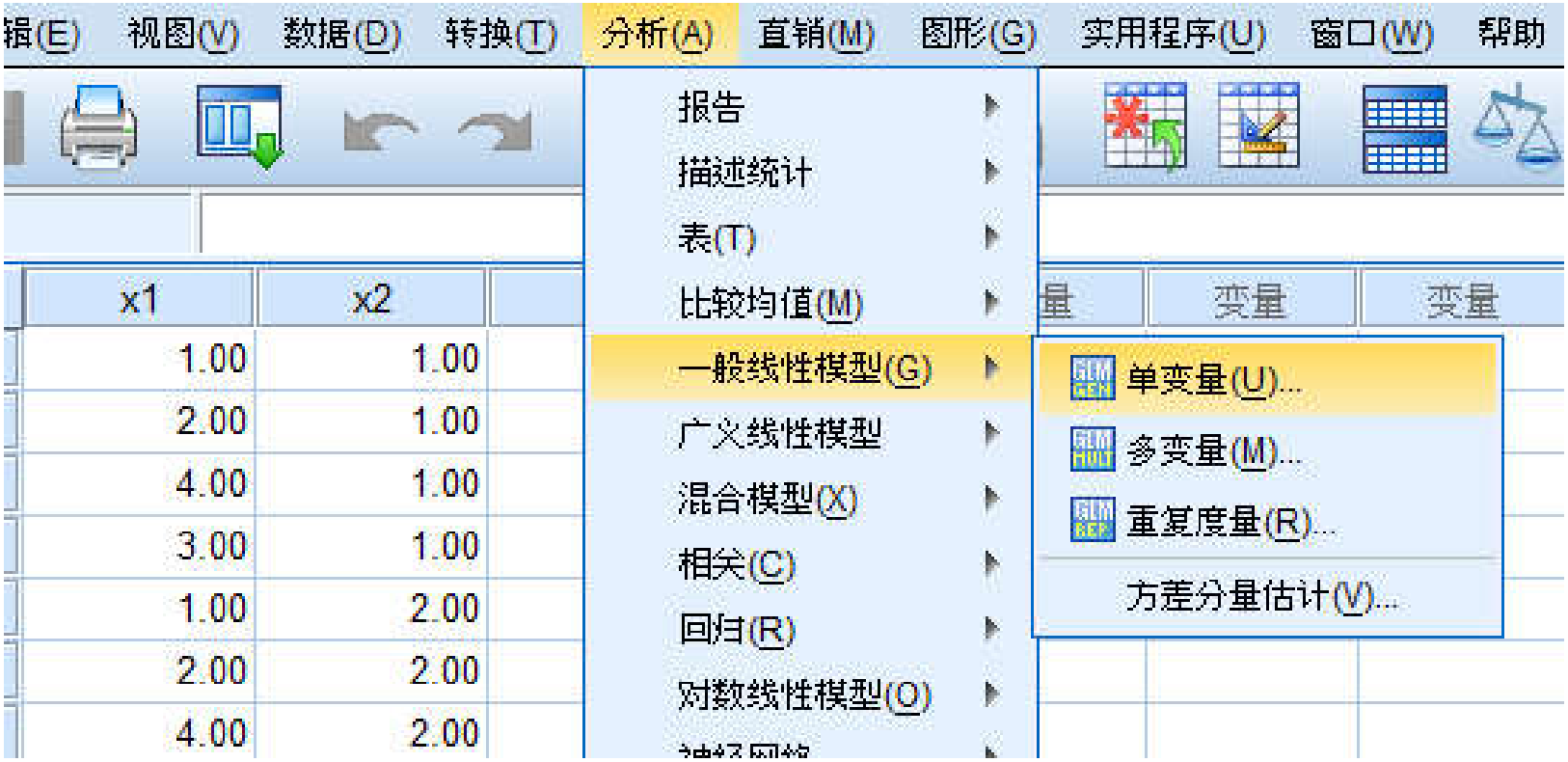
可被地区线性解释的变差，也即一元线性回归分析中的回归平方和

剩余的不可被地区线性解释的变差

概率P值小于0.05，拒绝原假设，认为地区和销售额之间不是零线性相关。但概率P值接近0.05，表明拒绝原假设，认为存在非零相关性的把握程度不高

1.3.5 多因素方差分析模型案例

- **案例**：某企业对广告形式、地区以及广告形式和地区的交互作用是否对商品销售额产生影响进行分析。



多因素方差分析窗口


单变量
因变量(D):
固定效应的因素
固定因子(F):
广告形式 [x1]
地区 [x2]
随机因子(A):
随机效应的因素
协变量(C):
WLS 权重(W):
模型(M)...
对比(N)...
绘制(T)...
两两比较(H)...
保存(S)...
选项(O)...
Bootstrap(B)...
确定
粘贴(P)
重置(R)
取消
帮助

销售额多因素方差分析结果

主体间效应的检验

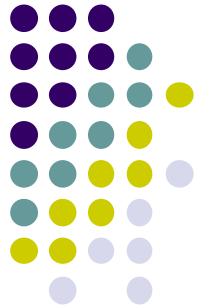
因变量: 销售额

源	III 型平方和	df	均方	F	Sig.
校正模型	20094.306[a]	71	283.018	3.354	.000
截距	642936.694	1	642936.694	7619.990	.000
x1	5866.083	SSA 3	1955.361	23.175	.000
x2	9265.306	SSB 17	545.018	6.459	.000
x1 * x2	4962.917	SSAB 51	97.312	1.153	.286
误差	6075.000	SSE 72	84.375		
总计	669106.000	144			
校正的总计	26169.306	SST 143			

a. R 方 = .768（调整 R 方 = .539）

不同广告形式和地区没有对销售额产生显著的交互作用，不同地区采用哪种形式的广告都不会对销售额产生显著影响

延伸思考：多因素方差分析的进一步分析



（1）建立非饱和模型

（2）均值比较分析

（3）控制变量交互作用的图形分析

（1）建立非饱和模型

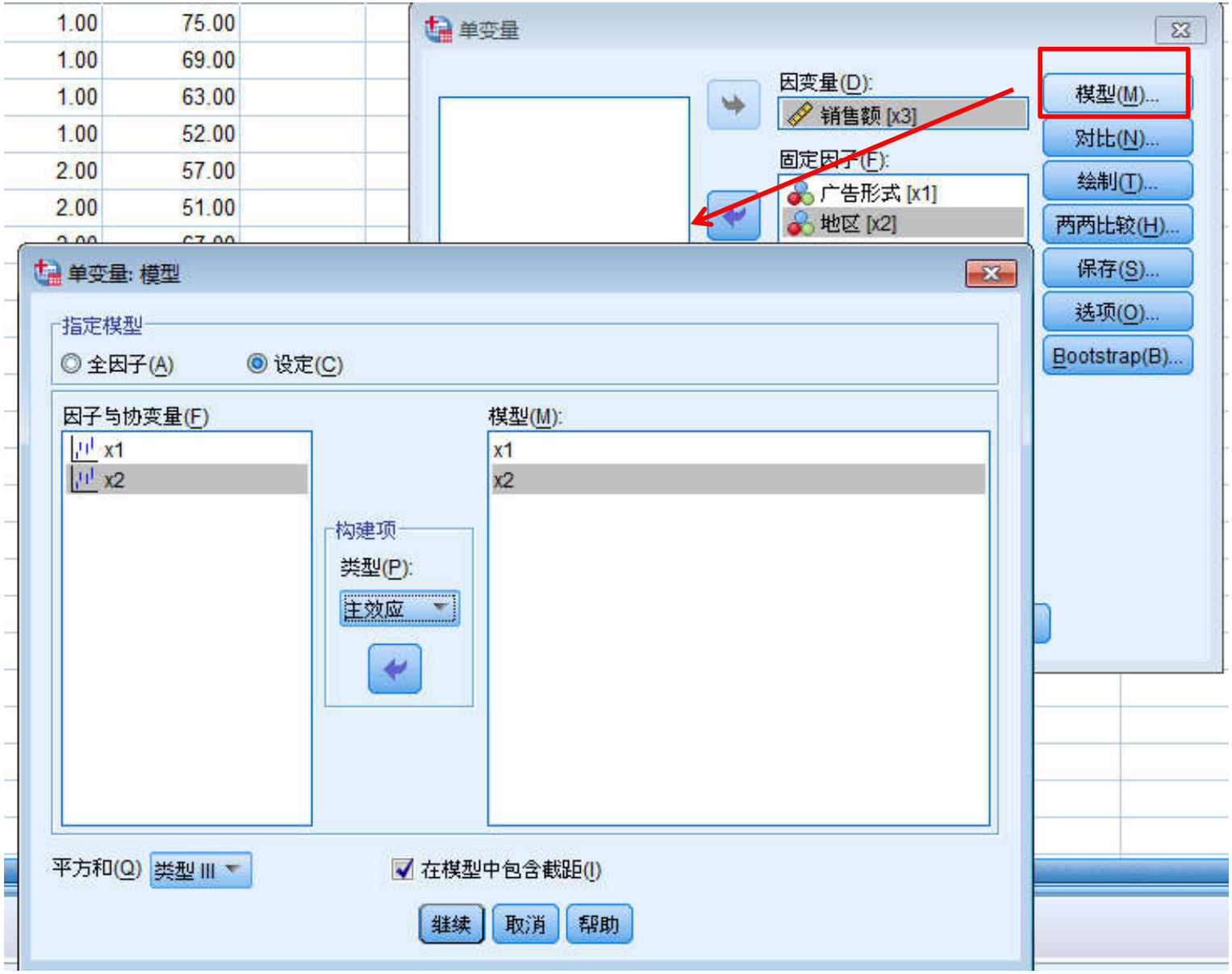
- 非饱和模型是针对饱和模型而言的。
  - **饱和模型**中的观测变量总变差被分解为各因素独立作用、各因素交互作用（包括二阶、三阶或更高阶的交互）以及抽样误差三大部分。
  - **非饱和模型**将其中某些对观测变量变差解释作用不显著的部分合并到SSE中。

两因素非饱和模型：

SST=SSA+SSB+SSE

三因素二阶非饱和模型：

SST=SSA+SSB+SSC+SSAB+SSAC+SSBC+SSE


1.00 75.00
1.00 69.00
1.00 63.00
1.00 52.00
2.00 57.00
2.00 51.00
单变量
因变量(D):
销售额 [x3]
固定因子(F):
广告形式 [x1]
地区 [x2]
模型(M)...
对比(N)...
绘制(T)...
两两比较(H)...
保存(S)...
选项(O)...
Bootstrap(B)...
单变量: 模型
指定模型
全因子(A)
设定(C)
因子与协变量(F)
x1
x2
模型(M):
x1
x2
构建项
类型(P):
主效应
平方和(Q)
类型 III
在模型中包含截距(I)
继续
取消
帮助

销售额多因素方差分析的非饱和模型

主体间效应的检验

因变量: 销售额

源	III 型平方和	df	均方	F	Sig.
校正模型	15131.389[a]	20	756.569	8.431	.000
截距	642936.694	1	642936.694	7164.505	.000
x1	5866.083	SSA 3	1955.361	21.789	.000
x2	9265.306	SSB 17	545.018	6.073	.000
误差	11037.917	123	89.739		
总计	669106.000	144			
校正的总计	26169.306	SST 143			

a. R 方 = .578（调整 R 方 = .510）

新的SSE即为原来的SSE+SSAB

各因素所能解释的变差比例相对于随机因素来说减少，导致各个F检验统计量的值变小，对应的概率P值变大，不易得到各因素不同水平对观测变量有显著影响的结论。

（2）均值比较分析

（2）对比检验

（1）多重比较检验，同单因素方差分析

对比检验的检验值指定为观测变量的均值

对比结果（K 矩阵）

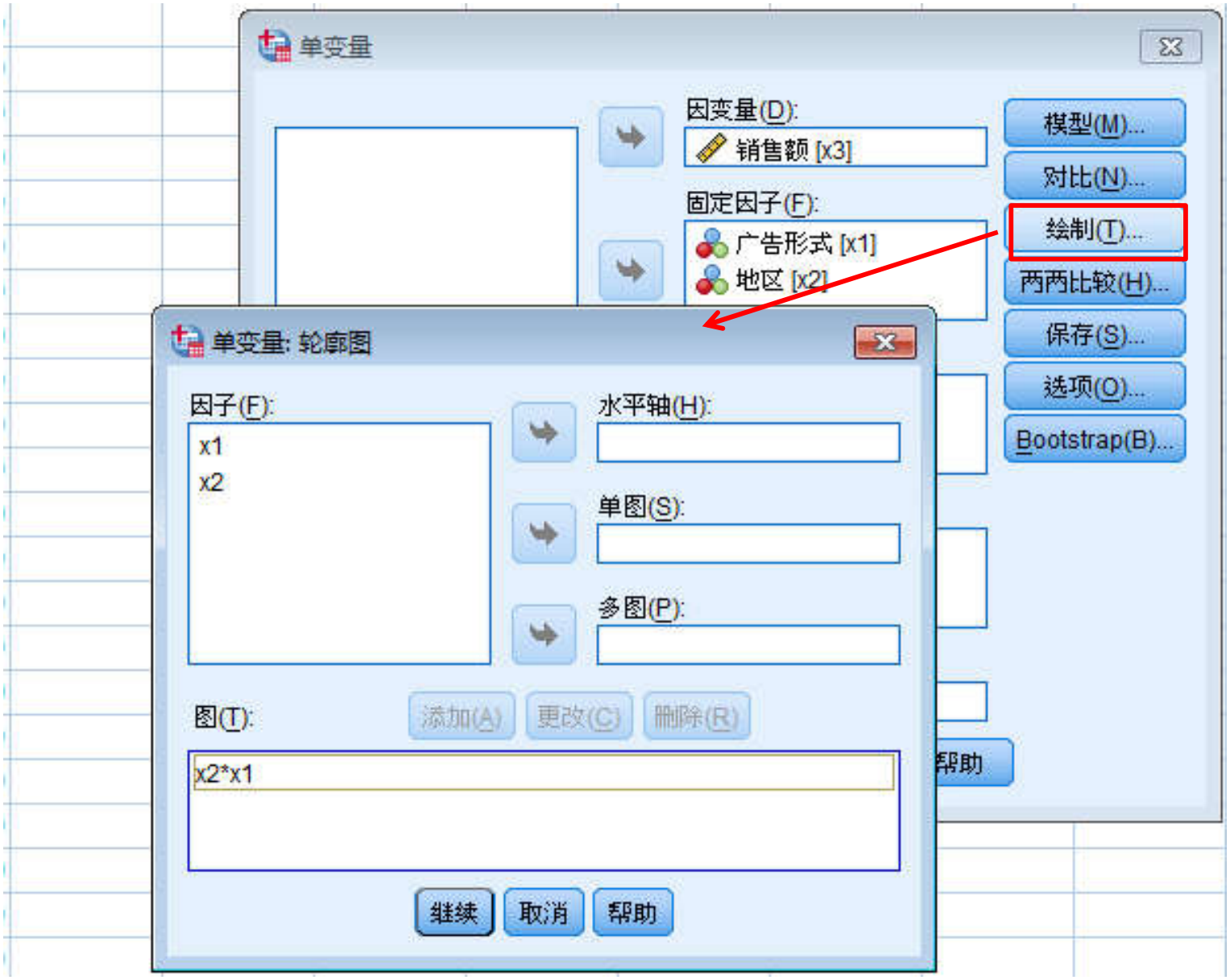
广告形式 偏差对比[a]			因变量 销售额
级别 1 和均值	对比估算值		6.403
	假设值		0
	差分（估计 - 假设）		6.403
	标准 误差		1.326
	Sig.		.000
	差分的 95% 置信区间	下限	3.760
		上限	9.046
级别 2 和均值	对比估算值		4.069
	假设值		0
	差分（估计 - 假设）		4.069
	标准 误差		1.326
	Sig.		.003
	差分的 95% 置信区间	下限	1.426
		上限	6.712
级别 3 和均值	对比估算值		-10.264
	假设值		0
	差分（估计 - 假设）		-10.264
	标准 误差		1.326
	Sig.		.000
	差分的 95% 置信区间	下限	-12.907
		上限	-7.621

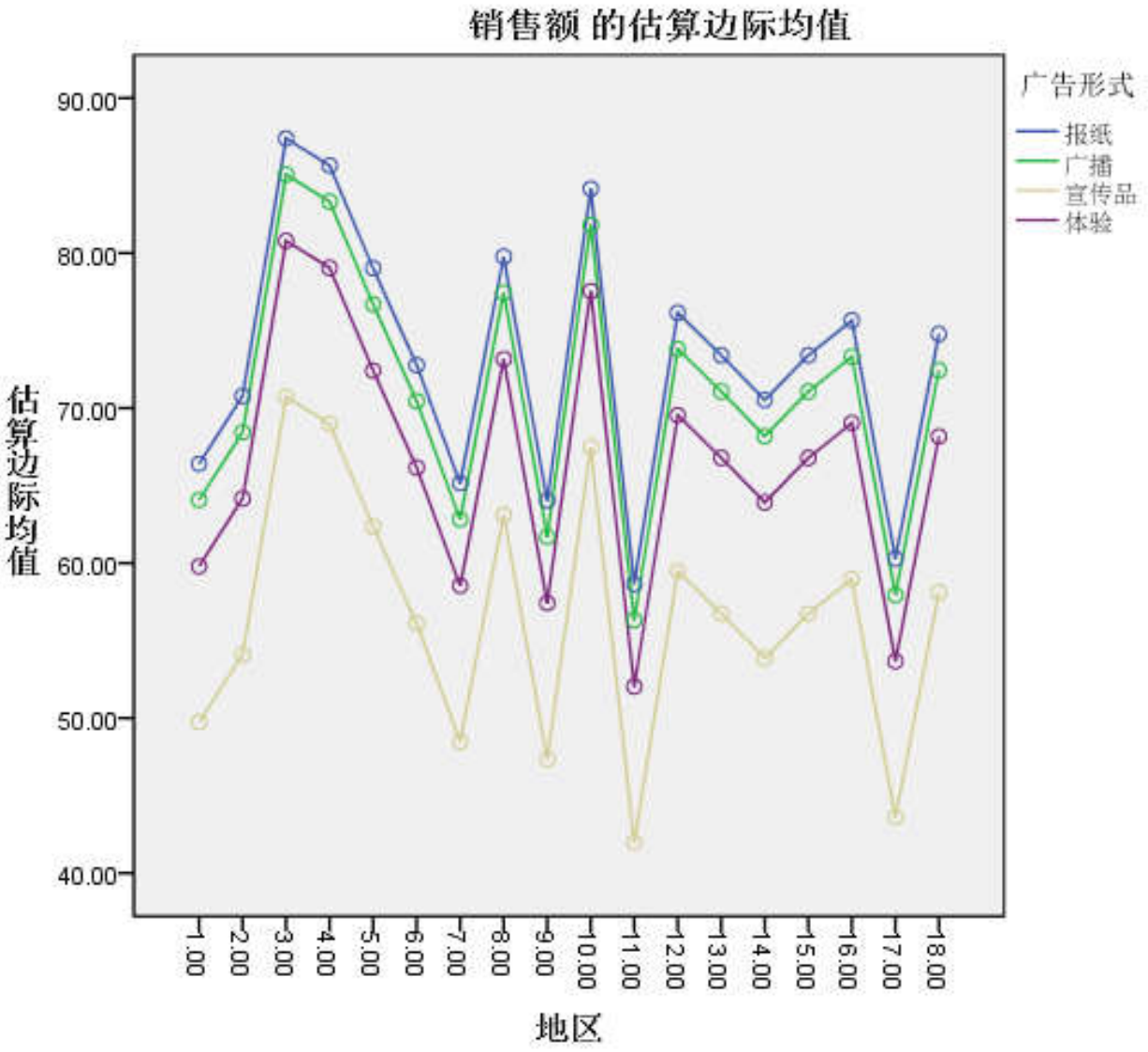
a. 省略的类别 = 4

t检验统计量概率P值为0，第一种广告形式下销售额均值与检验值（各水平下的整体均值）间存在显著差异，其明显高于整体水平

同理，第二种广告形式下销售额也明显高于总体水平，而第三种广告形式下销售额明显低于整体水平。
三种广告形式产生的效果有显著差异。

（3）控制变量交互作用的图形分析




销售额 的估算边际均值
估算边际均值
90.00
80.00
70.00
60.00
50.00
40.00
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00
17.00
18.00
地区
广告形式
报纸
广播
宣传品
体验

1.4 常用统计模型串讲

1.4.1 一般线性模型——方差分析与线性回归的统一

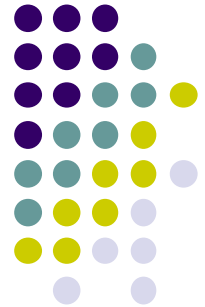
1.4.2 ★广义线性模型——线性回归与Logistic回归的统一

1.4.3 广义可加模型——脱离“线性”束缚

1.4.4 多水平模型——打破“独立”条件

1.4.5 结构方程模型——从单因单果到多因多果

1.4.1 一般线性模型
——方差分析与线性回归的统一



- **一般线性模型**（General Linear Model）的基本形式：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

自变量个数与类型	一般线性模型的具体方法
1 个二分类变量	t 检验
1 个多分类变量	方差分析
2 个（或多个）分类变量	多因素方差分析（不是多元方差分析）
1 个连续变量	单因素线性回归
多个连续变量	多因素线性回归（不是多元线性回归）
1 个连续变量、1 个分类变量	协方差分析

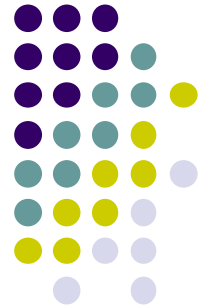
ANOVA

MANOVA

ANCOVA

方法名称	中文全称	因变量数量	自变量中的因素类型	核心目的
ANOVA	方差分析	**一个**	只有**类别型**自变量	比较不同组别在一个连续因变量上的均值差异。
ANCOVA	协方差分析	**一个**	**类别型**自变量 + **连续型**协变量	在**控制**了协变量的影响后，比较类别自变量对因变量的效应。
MANOVA	多变量/多元方差分析	**多个**	只有**类别型**自变量	同时比较不同组别在**多个相关**的连续因变量上的均值差异。
MANCOVA	多变量/多元协方差分析	**多个**	**类别型**自变量 + **连续型**协变量	在**控制**了协变量的影响后，比较类别自变量对**多个相关**因变量的效应。

1.4.2 广义线性模型
——线性回归与Logistic回归的统一



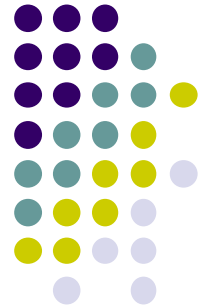
- **广义线性模型**（Generalized Linear Model）的基本形式：

连接函数 $g(\mu)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\varepsilon$

资料类型	分布	$g(\mu)$的具体形式	广义线性模型的具体方法
连续资料	正态分布	μ	线性回归模型
分类资料	二项分布或多项分布	$\ln\frac{\mu}{1-\mu}$ 或 $\text{logit}(\mu)$	Logistic 回归模型
计数资料	Poisson 分布	$\ln(\mu)$	Poisson 回归模型
计数资料	负二项分布	$\ln(1-\mu)$	负二项回归模型

共同性：参数估计一般采用最大似然估计，回归系数检验一般是Wald检验、似然比检验、score检验（得分检验），等等。

1.4.3 广义可加模型
——脱离“线性”束缚



- 广义可加模型（Generalized Additive Model）的基本形式：

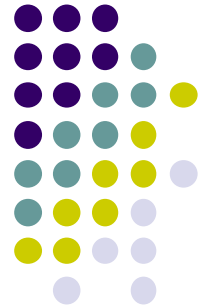
可线性也可非线性的函数关系

$$g(\mu)=\beta_0+f_1x_1+f_2x_2+\ldots+f_px_p+\varepsilon$$

- 广义可加模型没有回归系数可估计，只是在寻找一条拟合效果相对更好（也即偏差-方差权衡）的曲线（**非参数回归，以探索**（初步探索自变量与因变量的恰当关系）**和预测**（只是预测，无须给出参数模型的具体形式）**为主**）

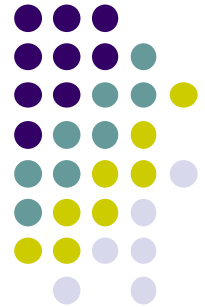
注：非线性回归通过一定的变换后仍可能满足线性关系，仍属于参数回归范畴

1.4.4 多水平模型
——打破“独立”条件



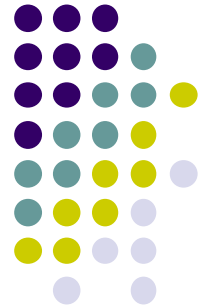
- 广义线性模型不满足“线性”条件可考虑广义可加模型，不满足“独立性”条件可考虑**多水平模型**（Multilevel Model），也称分层线性模型（Hierarchical Linear Model）、混合效应模型（Mixed Effect Model）等。
- 多水平模型同时包含了多个水平的数据，从而在多个水平上都存在残差。多水平模型的基本思想是**把高水平上的差异估计出来**，这就使得残差变小，估计的结果更为可靠。

1.4.5 结构方程模型
——从单因单果到多因多果



- **结构方程模型**（Structural Equation Modeling，SEM）可看做路径分析（Path Analysis）和验证性因子分析（Confirmatory Factor Analysis，非EFA）的组合（路径分析可以分析变量间的直接和间接关系，验证性因子分析可以分析潜变量与显变量之间的关系，路径分析中可以含有潜变量）。
  - 测量模型：相当于验证性因子分析，潜变量和显变量
  - 结构模型：相当于路径分析，内生变量和外生变量

2. Logistic回归模型



2.1 二项Logistic回归模型（**Logit模型**）

2.2 多项Logistic回归模型（**广义Logit模型**）

2.3 多项有序Logistic回归模型（**累积Logit模型**）

2.4 Probit回归模型

附：Logistic回归
——统计学和机器学习不同视角（1）

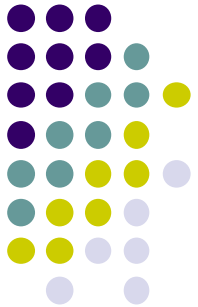
“解释性模型”　　“预测性工具”

维度	统计学侧重点	机器学习侧重点
核心目标	**理解与解释**。关注自变量(X)与因变量(Y)之间关系的强度和方向。模型是用于检验理论和假设的工具。	**预测与泛化**。关注模型对未知新数据做出准确预测的能力。模型本身是一个黑箱或工具，预测精度是首要目标。
问题范式	“这些因素中，哪个对结果有显著影响？其影响有多大？”	“给定一些特征，这个样本最可能属于哪个类别？”

附：Logistic回归
——统计学和机器学习不同视角（2）

维度	统计学视角	机器学习视角
目标	解释变量关系	优化预测性能
评估	模型拟合度、p值	测试集准确率、AUC
正则化	较少使用	常规使用（L1/L2）
假设检验	核心关注	较少关注
过拟合	通过理论限制	通过正则化/验证集
样本量	关注最小样本量	尽可能多收集数据
软件工具	R, SAS, Stata	Python (scikit-learn), TensorFlow
输出焦点	参数估计、置信区间	预测概率、分类结果

2.1 二项Logistic回归模型



2.1.1 Logistic回归分析概述

2.1.2 二项Logistic回归方程

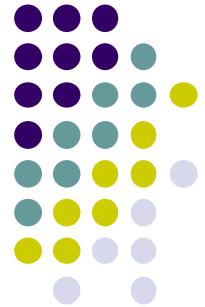
2.1.3 二项Logistic回归方程系数的含义

2.1.4 二项Logistic回归方程的检验

2.1.5 二项Logistic回归分析中的虚拟变量

2.1.6 案例分析

2.1.1 Logistic回归分析概述



- 研究二分类变量与其他变量之间的关系
  - 例如：研究吸烟对是否得肺癌的影响，并以年龄和性别作为控制变量，特点：
    - 因变量是二分类变量
    - 自变量有分类变量和数值变量
    - 吸烟与肺癌概率之间并非一种线性关系
- 对二分类的因变量可否直接采用一般多元线性回归分析方法？
  - 结论：不可以

- 当因变量为0/1二分类变量时，如何从一般线性回归模型中得到启示：
  - 一般回归模型是对因变量的均值的预测，可将其转化为对因变量取值为1的概率P的预测

$$E(y_i)=\beta_0+\sum_{i=1}^{k}\beta_i x_i \Longrightarrow P_{y=1}=\beta_0+\sum_{i=1}^{k}\beta_i x_i$$

  - 一般回归模型的因变量的取值范围是－∞～＋∞，而这里因变量的取值范围是0～1，可对概率P做合理转换处理，使其取值范围与一般回归模型吻合
  - 一般回归模型的自变量与概率P间的关系只能是线性的，实际应用中往往是非线性关系，对概率P的转换处理应采用非线性转化

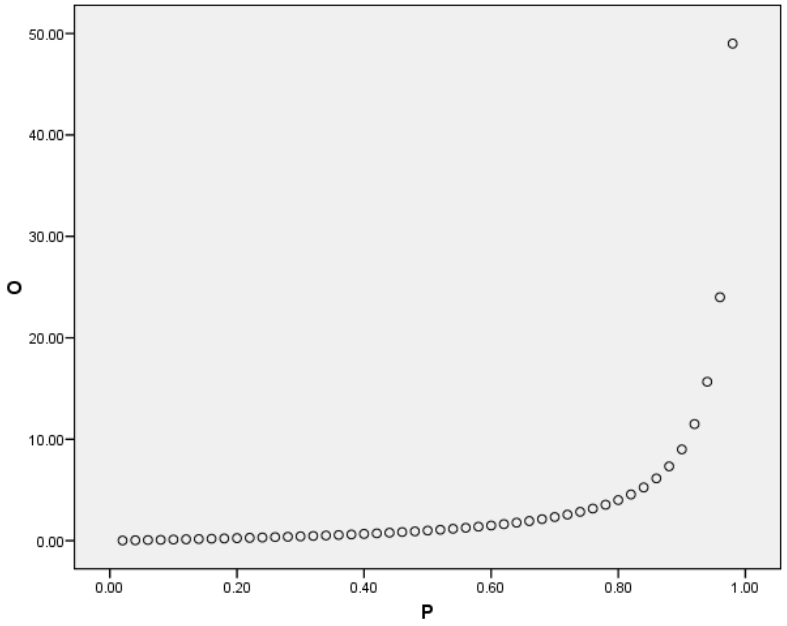
- **解决问题的方向：**
  - 对概率P进行转换处理后，应使其取值范围与一般线性回归模型吻合
  - 对概率P应采用非线性转化处理
  - 所有的转换都不应改变自变量和因变量之间关系的方向（函数单调性一致）

二项Logistic回归：理论上的处理

- 进行**两步转换处理**：
  - 第一步，将P转换成Ω
    - Ω称为优势（Odds）
    - 对P的转化是非线性的
    - Ω是P的单调增函数
    - 优势的取值范围：0～＋∞

某事件发生概率与不发生概率之比

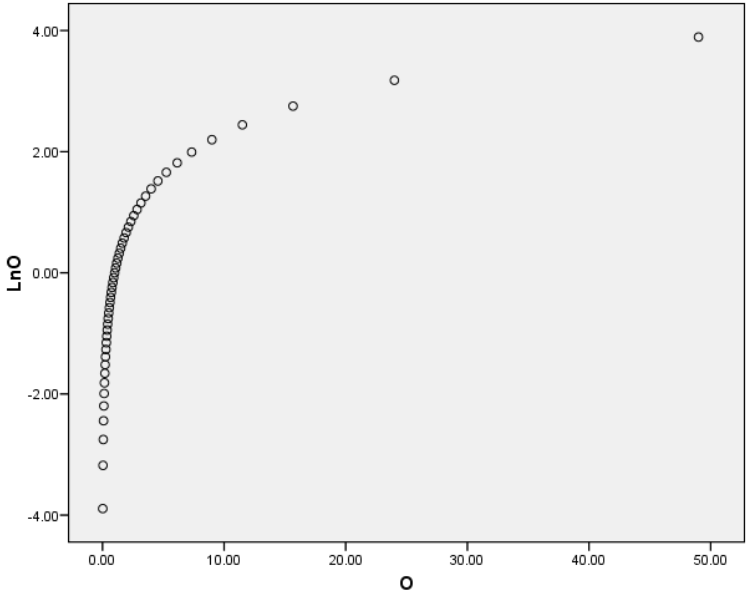
$$\Omega = \frac{P}{1 - P}$$



二项Logistic回归：理论上的处理

- 进行**两步转换处理**：
  - 第二步，将Ω转换成lnΩ
    - lnΩ称为Logit P
    - Logit P与Ω仍呈增长（或下降）的一致性关系
    - Logit P的取值于－∞～＋∞

$$\ln(\Omega) = \ln(\frac{P}{1-P})$$



2.1.2 二项Logistic回归方程

- Logit P与自变量间为**线性**关系（两回归方程等同）：

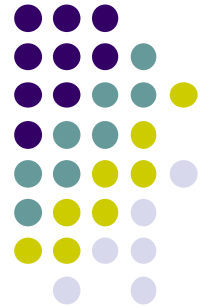
$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

$$\text{Logit } P = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

- P与自变量间为**非线性**关系（**Sigmoid函数**）：

$$P = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}}$$

2.1.3 二项Logistic回归方程系数的含义



- **回归系数**表示当其他自变量保持不变时，自变量x_i每增加一个单位，将引起**Logit P**（也即**ln Ω**）平均增加（或减少）β_i个单位。

- 在实际应用中人们关心的是自变量变化引起事件发生概率**P**变化的程度（非线性变化），人们通常更关心自变量给**优势Ω**带来的变化。

- **优势Ω=P/(1-P)**，即某事件发生的概率与不发生的概率之比，利用**优势比（OR，Odds Ratio）**可进行不同组之间相对风险的近似对比分析
  - 例如，如果吸烟组得肺癌的概率是0.25，不吸烟组得肺癌的概率是0.10，则两组的优势比为：

$$OR_{\text{A VS..B}} = \frac{P(D_A)}{1-P(D_A)} \Big/ \frac{P(D_B)}{1-P(D_B)} = \frac{1}{3} \Big/ \frac{1}{9} = 3$$

  - 吸烟组的相对风险近似是不吸烟组的3倍，吸烟患肺癌的风险高于不吸烟

- 设因变量Y(1=得肺癌/0=没得肺癌)，自变量X_1(1=吸烟/0=不吸烟)，则**Logistic方程**为：

$$\text{Logit}[P(Y=1)]=\beta_0+\beta_1 X_1$$

  - 吸烟与不吸烟组的方程分别是：

$$\text{Logit}[P(Y=1)]=\ln(\Omega_1)=\beta_0+\beta_1\times 1=\beta_0+\beta_1$$

$$\text{Logit}[P(Y=1)]=\ln(\Omega_2)=\beta_0+\beta_1\times 0=\beta_0$$

  - 两组**优势比**为：

$$\text{OR}_{\text{S VS. NS}}=\frac{\Omega_1}{\Omega_2}=\frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}}=e^{\beta_1}$$

- 设因变量Y(1=得肺癌/0=没得肺癌)，自变量X (x_1吸烟/x_2年龄/x_3性别)，则**Logistic方程**为：

$$\mathrm{Logit}[P(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

  - 控制年龄（45）和性别（1）后，吸烟和不吸烟组的方程分别是：

$$\mathrm{Logit}[P(Y=1)] = \ln(\Omega_1) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 45 + \beta_3 \times 1$$

$$\mathrm{Logit}[P(Y=1)] = \ln(\Omega_2) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 45 + \beta_3 \times 1$$

  - 两组**优势比**为：

$$\mathrm{OR}_{\mathrm{S\ VS.\ NS}} = \frac{\Omega_1}{\Omega_2} = e^{(1-0)\beta_1 + (45-45)\beta_2 + (1-1)\beta_3} = e^{\beta_1}$$

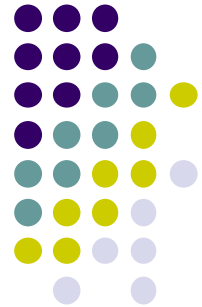
- **一般化处理**：将x_1变化一个单位后的优势设为Ω^*

$$\Omega = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

$$\Omega^* = \exp(\beta_1 + \beta_0 + \sum_{i=1}^{k} \beta_i x_i) = \Omega \exp(\beta_1)$$

$$\frac{\Omega^*}{\Omega} = \exp(\beta_1)$$

- 由此可知，x_1增加一个单位所导致的优势是原来的$\exp(\beta_1)$倍，即相对风险近似为$\exp(\beta_1)$



- **再一般化处理**：

$$\frac{\Omega^{*}}{\Omega} = \exp(\beta_i)$$

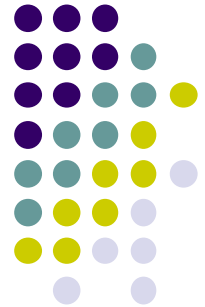
- 上述公式表明，当其他自变量保持不变时，x_i每增加一个单位所导致的优势是原来的 $\exp(\beta_i)$ 倍，即**优势比**为 $\exp(\beta_i)$，相对风险近似为 $\exp(\beta_i)$

2.1.4 二项Logistic回归方程的检验

- 采用**极大似然估计法（Maximum Likelihood Estimate，MLE）**进行参数估计：
  - 以**似然函数值L**达到最大时的参数值作为总体参数的估计值，该值在0～1之间，反映了在所估计参数的总体中抽到特定样本的可能性，越接近1越好
  - 为便于处理，通常将似然函数取自然对数，得到**对数似然函数值LL**，当似然函数值最大为1，对数似然函数值取到最大值0
  - LL越大，模型较好地拟合样本数据的可能性越大，所得模型的拟合优度越高，否则反之（**注：LL为负数！**）

“利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值”

二项Logistic回归的检验——回归方程显著性检验（1）

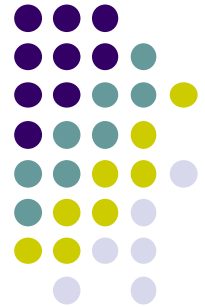


- **回归方程的显著性检验**：自变量全体与Logit P的线性关系是否显著
  - 原假设H_0：回归系数同时为0
  - 采用对数似然比测度拟合程度是否显著提高
  - 设x_i未引入回归方程前的对数似然函数值为 LL ，x_i引入回归方程后的对数似然函数值为 LL_{x_i}，则对数似然比为：

$$\frac{LL}{LL_{x_i}}$$

  - 如果对数似然比与1无显著差异，则说明引入x_i后，自变量全体对Logit P的线性解释无显著改善；如果对数似然比远远大于1，与1有显著差异，则说明引入x_i后，自变量全体与Logit P的线性关系随x_i的进入得到显著提升

二项Logistic回归的检验 ——回归方程显著性检验（2）



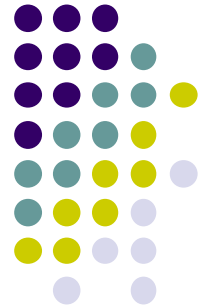
- **回归方程的显著性检验**：自变量全体与Logit P的线性关系是否显著

$$\frac{LL}{LL_{x_i}}$$

- 由于对数似然比 的分布未知，通常采用 $-\ln(\frac{L}{L_{xi}})^2$，它在原假设成立条件下近似服从卡方分布，也称为**似然比卡方**，它反映了x_i引入回归方程前后对数似然值的变化幅度，该值越大表明x_i的引入越有意义

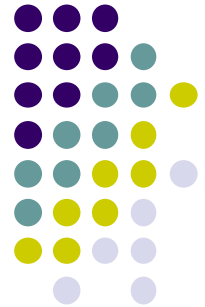
$$-\ln(\frac{L}{L_{xi}})^2 = -2\ln(\frac{L}{L_{xi}}) = -2\ln(L) - (-2\ln(L_{xi})) = -2LL - (-2LL_{xi})$$

二项Logistic回归的检验——回归方程显著性检验（3）



- **回归方程的显著性检验**：自变量全体与Logit P的线性关系是否显著
  - 进一步，如果似然比卡方的观测值对应概率P值小于给定显著性水平，则应拒绝原假设，认为目前方程中所有回归系数不同时为0，自变量全体与LogitP之间的线性关系显著；反之则不应拒绝原假设，认为目前方程中所有的回归系数同时为0，自变量全体与LogitP之间的线性关系不显著

二项Logistic回归的检验——回归系数显著性检验

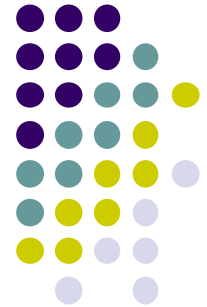


- **回归系数的显著性检验**：各个自变量与Logit P的线性关系是否显著
  - 原假设H_0: $\beta_i = 0$，即某回归系数与零无显著差异，相应的自变量与Logit P之间的线性关系不显著
  - 采用Wald检验统计量，近似服从卡方分布

$$Wald_i = (\frac{\beta_i}{S_{\beta i}})^2$$

  - 若某自变量Wald观测值对应概率P值小于给定的显著性水平，则应拒绝原假设，该自变量应保留在方程中；反之则不应拒绝原假设，该自变量不应保留在方程中

二项Logistic回归的检验——回归方程拟合优度检验（1）



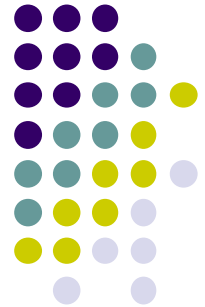
- **回归方程的拟合优度检验：(1)回归方程能够解释因变量变差的程度**；(2)由回归方程计算出的预测值与实际值之间吻合的程度，即方程总体错判率:
  - Cox & Snell R^2统计量，其中LL_0为方程中只包含常数项时的对数似然值，LL_p为当前方程的对数似然值（取值范围不易确定）：

$$1-\left(\frac{LL_0}{LL_p}\right)^{\frac{2}{n}}$$

  - Nagelkerke R^2统计量（取值范围0-1之间）：

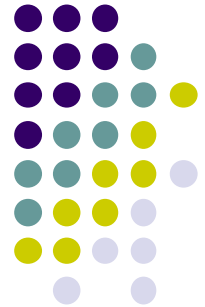
$$\frac{1-\left(\frac{LL_0}{LL_p}\right)^{\frac{2}{n}}}{1-(LL_0)^{\frac{2}{n}}}$$

二项Logistic回归的检验——回归方程拟合优度检验（2）



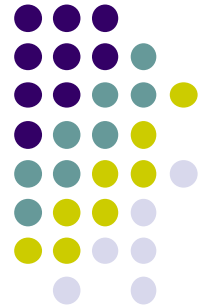
- **回归方程的拟合优度检验**：(1)回归方程能够解释被解释变量变差的程度；**(2)由回归方程计算出的预测值与实际值之间吻合的程度，即方程总体错判率**:
  - 分类表（通过矩阵表格展示模型预测值与实际观测值的吻和程度）
  - Hosmer-Lemeshow检验（生成交叉列联表，原假设H_0是观测频数/实际值的分布与期望频数/预测值的分布无显著差异，H-L统计量的概率P值**大于**给定的显著性水平，则不应拒绝原假设，即观测频数的分布与期望频数的分布没有显著差异，表明样本实际值与预测值的整体差异较小，模型拟合效果越好，否则反之）

2.1.5 二项Logistic回归分析中的虚拟变量



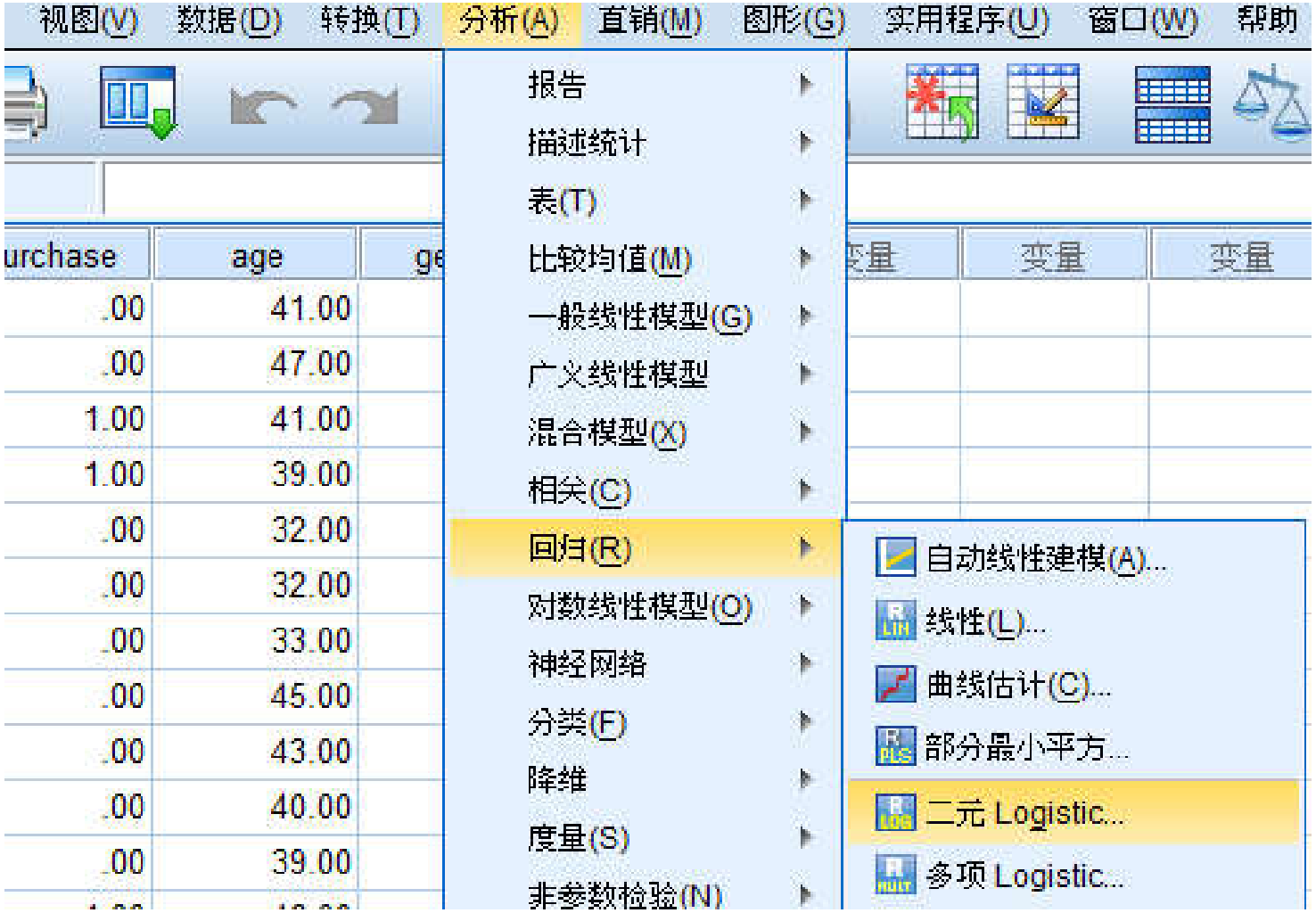
- 分类型变量参与回归分析的主要目的是研究各类对因变量影响的差异性
- 对于具有k个类别的分类型变量，当确定参照类别后，只需设置k-1个虚拟变量即可，k-1个虚拟变量参与回归分析
- 虚拟自变量回归系数的含义是：**相对于参照类别，各个类别对因变量平均贡献的差**。进而可以研究各类别间对因变量平均贡献的差异。

2.1.6 案例分析

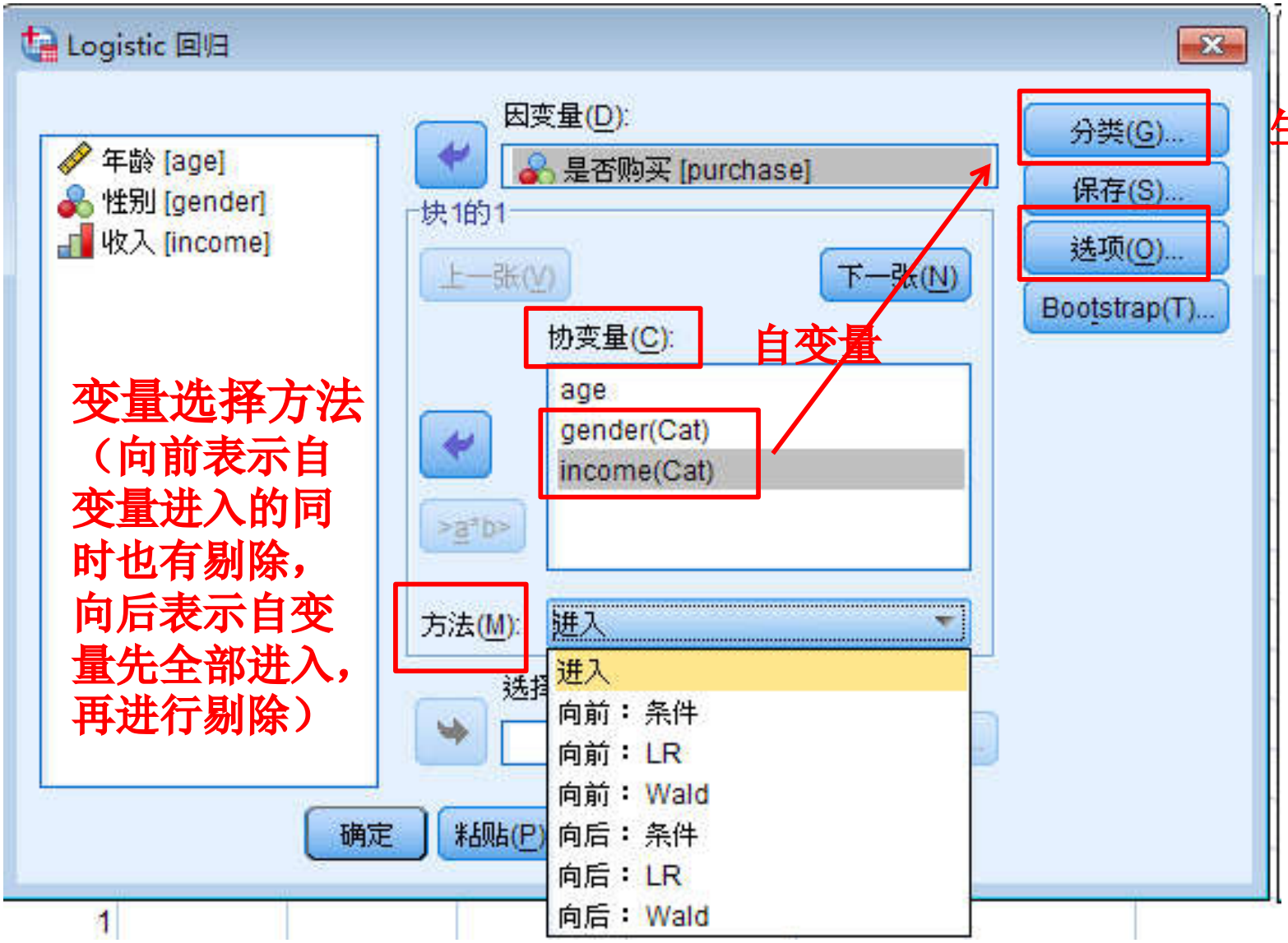


- 431个随机样本数据，变量有是否购买（Purchase，1为购买，0为不购买）、年龄（Age）、性别（Gender，1为男，2为女）和收入水平（Income，1为低收入，2为中收入，3为高收入）
- 建立客户购买的预测模型，分析影响因素，其中是否购买为因变量，其余变量为自变量

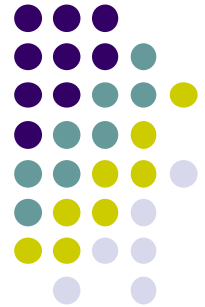
（1）操作说明


视图(V)
数据(D)
转换(T)
分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W)
帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
自动线性建模(A)...
线性(L)...
曲线估计(C)...
部分最小平方...
二元 Logistic...
多项 Logistic...
urchase
age
变量
变量
变量
.00 41.00
.00 47.00
1.00 41.00
1.00 39.00
.00 32.00
.00 32.00
.00 33.00
.00 45.00
.00 43.00
.00 40.00
.00 39.00

（1）操作说明


Logistic 回归
年龄 [age]
性别 [gender]
收入 [income]
因变量(D):
是否购买 [purchase]
块1的1
上一张(V)
下一张(N)
协变量(C):
age
gender(Cat)
income(Cat)
方法(M):
进入
进入
向前：条件
向前：LR
向前：Wald
向后：条件
向后：LR
向后：Wald
分类(G)...
保存(S)...
选项(O)...
Bootstrap(T)...
确定
粘贴(P)
变量选择方法（向前表示自变量进入的同时也有剔除，向后表示自变量先全部进入，再进行剔除）
自变量
生成虚拟变量

附：变量选择/模型选择的评判标准



- 选择不同变量组合会得到不同模型，**AIC（赤池信息准则）**越小，模型越好；**BIC（贝叶斯信息准则）**越小，模型越好。

$$AIC = -2\ln(L) + 2k$$

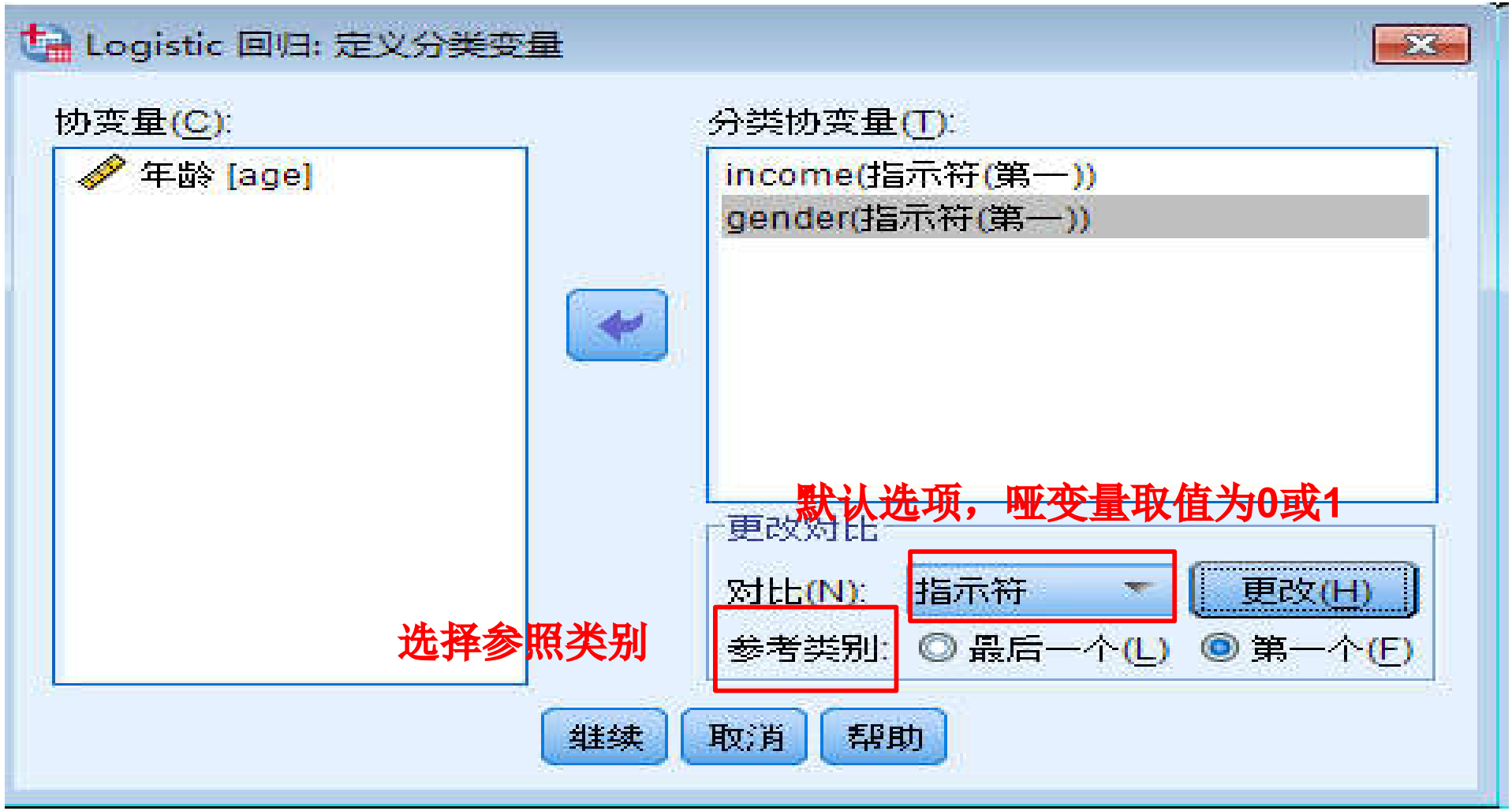
$$BIC = -2\ln(L) + k\ln(n)$$

惩罚项

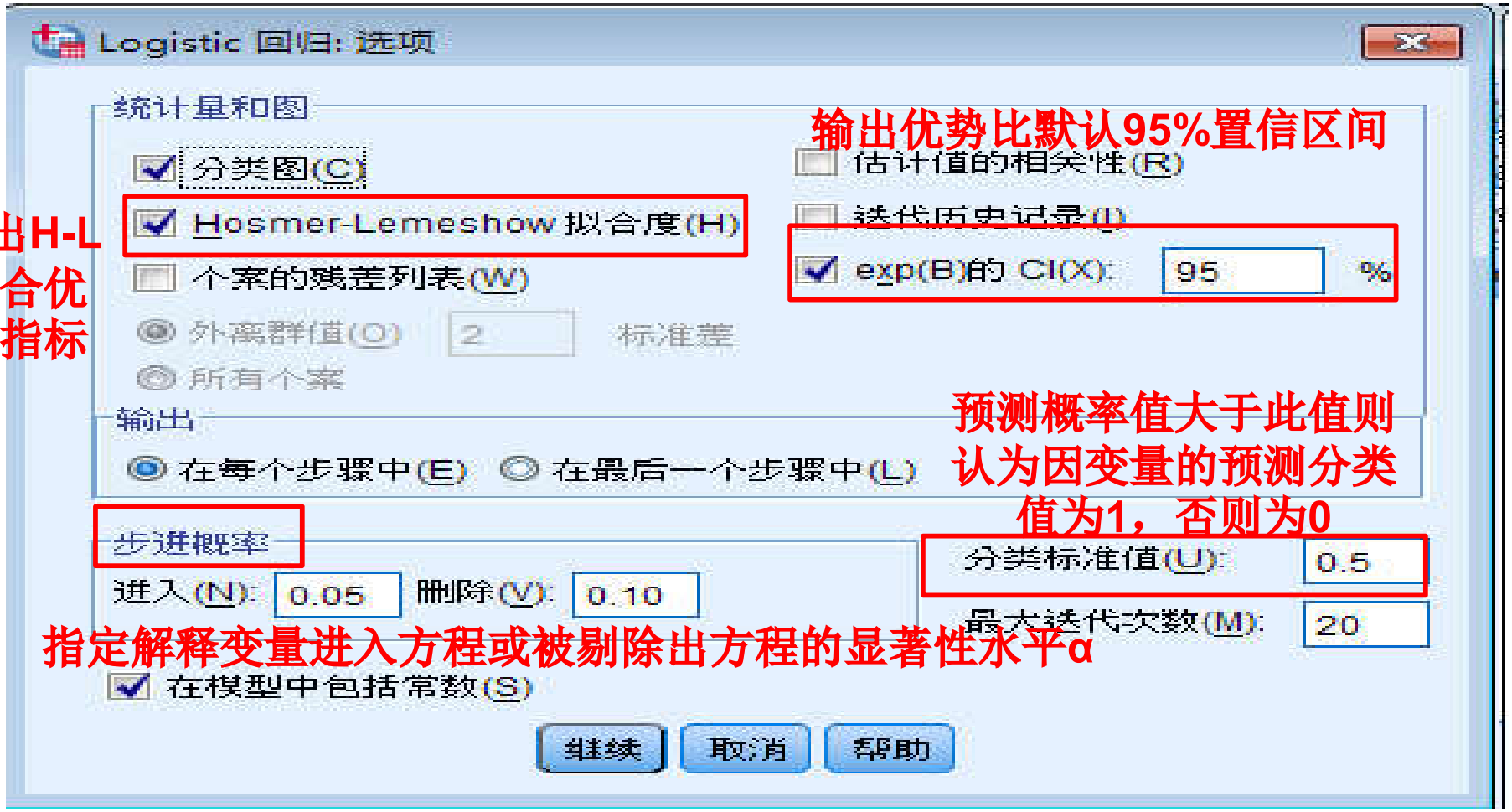
L是模型似然函数值，k是模型参数个数，n是样本量

- BIC选出来的变量个数往往少于AIC选出的变量个数。
- 被两种方法都选中的变量更可能是重要变量，被两种方法都抛弃的变量可能暂时不用考虑，被AIC选中而被BIC抛弃的变量则需更审慎的分析。

（1）操作说明——分类窗口


Logistic 回归: 定义分类变量
协变量(C):
年龄 [age]
分类协变量(T):
income(指示符(第一))
gender(指示符(第一))
默认选项，哑变量取值为0或1
更改对比
对比(N):
指示符
更改(H)
参考类别:
最后一个(L)
第一个(F)
选择参照类别
继续
取消
帮助

（1）操作说明——选项窗口



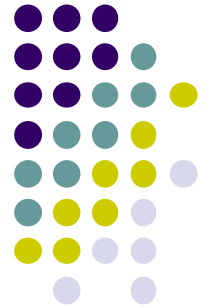
输出H-L拟合优度指标

输出优势比默认95%置信区间

预测概率值大于此值则认为因变量的预测分类值为1，否则为0

指定解释变量进入方程或被剔除出方程的显著性水平α

（2）结果解释Ⅰ（注：采用“进入”策略，强行令所有变量进入回归方程）



分类变量编码

		频率	参数编码 (1)	参数编码 (2)	
收入	低收入	132	.000	.000	参照类
	中收入	144	1.000	.000	
	高收入	155	.000	1.000	
性别	男	191	.000		参照类
	女	240	1.000		

（2）结果解释Ⅰ——回归方程显著性检验

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	18.441	4	.001
	块	18.441	4	.001
	模型	18.441	4	.001

P<0.05，拒绝原假设，认为所有回归系数不同时为0，自变量全体与LogitP之间的线性关系显著，采用该模型合理

（2）结果解释Ⅰ——回归方程拟合优度检验之Cox&SnellR²和NagelkerkeR²

越小代表模型拟合优度越高

模型汇总

越接近1越好

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	552.208[a]	.042	.057

a. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 4 处终止。

（2）结果解释Ⅰ——回归方程拟合优度检验之Hosmer和Lemeshow检验

- H-L统计量的观测值为15.353，概率P值为0.053，大于显著性水平α=0.05，因此不应拒绝原假设，认为因变量实际类别值的分布与预测类别值的分布无显著差异，模型拟合优度较好

= Hosmer 和 Lemeshow 检验 =

步骤	卡方	df	Sig.
1	15.353	8	.053

（2）结果解释Ⅰ——回归方程拟合优度检验之分类表（混淆矩阵）

分类表[a]

			已预测		
			是否购买		
	已观测		不购买	购买	百分比校正
步骤 1	是否购买	不购买	TN 236	FP 33	87.7
		购买	FN 131	TP 31	19.1
	总计百分比				61.9

负类被正确预测为负类的数量

准确率

a. 切割值为 .500

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$ 准确率

$$Precision = \frac{TP}{TP + FP}$$ 精确率

$$Recall = \frac{TP}{TP + FN}$$ 召回率

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$ F1分数

（2）结果解释Ⅰ——回归系数显著性检验

不应拒绝原假设，认为该回归系数与0无显著差异，它与Logit P的线性关系不显著，不应保留在方程中，该模型不可用，重新建模

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	age	.025	.018	1.974	1	.160	1.026	.990	1.062
	gender(1)	.511	.209	5.954	1	.015	1.667	1.106	2.513
	income			12.305	2	.002			
	income(1)	.101	.263	.146	1	.703	1.106	.660	1.853
	income(2)	.787	.253	9.676	1	.002	2.196	1.338	3.606
	常量	-2.112	.754	7.843	1	.005	.121		

a. 在步骤 1 中输入的变量: age, gender, income.

（3）结果解释II（注：采用“向前：LR”策略）

分类变量编码

		频率	参数编码	
			(1)	(2)
收入	低收入	132	.000	.000
	中收入	144	1.000	.000
	高收入	155	.000	1.000
性别	男	191	.000	
	女	240	1.000	

（3）结果解释II——回归方程显著性检验

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	10.543	2	.005
	块	10.543	2	.005
	模型	10.543	2	.005
步骤 2	步骤	5.917	1	.015
	块	16.459	3	.001
	模型	16.459	3	.001

若剔除income，则-2LL将增大10.543（似然比卡方值），-285.325即为第0步模型的LL值（剔除income后），-280.053即为第1步模型的LL值（剔除gender后）。

-2*(-285.325)-(-2*(-280.053))

如果移去项则建模

变量		模型对数似然性	在 -2 对数似然中的更改	df	更改的显著性
步骤 1	income	-285.325	10.543	2	.005
步骤 2	gender	-280.053	5.917	1	.015
	income	-282.976	11.761	2	.003

（3）结果解释Ⅱ——回归方程拟合优度检验之Cox&SnellR²和NagelkerkeR²

越小越好　　模型汇总　　越接近1越好

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	560.107[a]	.024	.033
2	554.190[b]	.037	.051

a. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 3 处终止。

b. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 4 处终止。

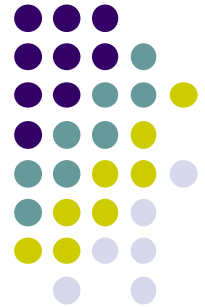
（3）结果解释Ⅱ ——回归方程拟合优度检验之Hosmer和Lemeshow检验

= Hosmer 和 Lemeshow 检验 =

步骤	卡方	df	Sig.
1	.000	1	1.000
2	8.943	4	.063

不应拒绝原假设，认为因变量实际类别值的分布与预测类别值的分布无显著差异，模型拟合优度（统计检验角度得到的，而非一般的描述性指标）较好

（3）结果解释Ⅱ——回归方程拟合优度检验之分类表（混淆矩阵）



分类表[a]

已观测			已预测		
			是否购买		
			不购买	购买	百分比校正
步骤 1	是否购买	不购买	269	0	100.0
		购买	162	0	.0
	总计百分比				62.4
步骤 2	是否购买	不购买	225	44	83.6
		购买	126	36	22.2
	总计百分比				60.6

a. 切割值为 .500

虽然总体预测正确率下降了一些，但大大提高了对购买人群预测的正确率

（3）结果解释Ⅱ——回归系数显著性检验

Income（低收入）、gender（男）为参照类别

相应自变量变化一个单位导致的优势比

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	income			10.512	2	.005			
	income(1)	.006	.259	.001	1	.982	1.006	.606	1.670
	income(2)	.672	.247	7.424	1	.006	1.958	1.208	3.174
	常量	-.762	.187	16.634	1	.000	.467		
步骤 2[b]	gender(1)	.504	.209	5.824	1	.016	1.656	1.099	2.493
	income			11.669	2	.003			
	income(1)	.096	.263	.134	1	.714	1.101	.658	1.843
	income(2)	.761	.251	9.147	1	.002	2.139	1.307	3.502
	常量	-1.113	.240	21.432	1	.000	.329		

a. 在步骤 1 中输入的变量: income.

b. 在步骤 2 中输入的变量: gender.

① 低收入顾客群的Logit回归方程

- Logit P=-1.11+0.504gender(1)

  0.504反映了相同收入群体的不同性别在购买上的差异。这里，由于参照水平为男，因此表示女性相对男性使LogitP平均增长0.504个单位。结合优势可知，女性的优势是男性的1.656倍，两者的优势比为1.656，女性更倾向于购买该产品

方程中的变量

		B	S.E.	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	income			10.512	2	.005			
	income(1)	.006	.259	.001	1	.982	1.006	.606	1.670
	income(2)	.672	.247	7.424	1	.006	1.958	1.208	3.174
	常量	-.762	.187	16.634	1	.000	.467		
步骤 2[b]	gender(1)	.504	.209	5.824	1	.016	1.656	1.099	2.493
	income			11.669	2	.003			
	income(1)	.096	.263	.134	1	.714	1.101	.658	1.843
	income(2)	.761	.251	9.147	1	.002	2.139	1.307	3.502
	常量	-1.113	.240	21.432	1	.000	.329		

a. 在步骤 1 中输入的变量: income.

b. 在步骤 2 中输入的变量: gender.

② 中收入顾客群的Logit回归方程

- Logit P=-1.11+0.504gender(1)+0.096income(1)

  0.096反映了相同性别的顾客群中中收入与低收入在购买上的差异。相同性别的顾客，中收入相对低收入使LogitP平均增长0.096个单位。结合优势可知，中收入的优势是低收入的1.101倍，即两者的优势比为1.101，且总体优势比有95%的把握在0.658-1.843之间，统计上并不显著。

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	income			10.512	2	.005			
	income(1)	.006	.259	.001	1	.982	1.006	.606	1.670
	income(2)	.672	.247	7.424	1	.006	1.958	1.208	3.174
	常量	-.762	.187	16.634	1	.000	.467		
步骤 2[b]	gender(1)	.504	.209	5.824	1	.016	1.656	1.099	2.493
	income			11.669	2	.003			
	income(1)	.096	.263	.134	1	.714	1.101	.658	1.843
	income(2)	.761	.251	9.147	1	.002	2.139	1.307	3.502
	常量	-1.113	.240	21.432	1	.000	.329		

a. 在步骤 1 中输入的变量: income.

b. 在步骤 2 中输入的变量: gender.

③ 高收入顾客群的Logit回归方程

- Logit P=-1.11+0.504gender(1)+0.761income(2)

0.761反映了相同性别的顾客群中高收入与低收入在购买上的差异。相同性别的顾客，高收入相对低收入使LogitP平均增长0.761个单位。结合优势可知，高收入的优势是低收入的2.139倍，显然高出较多，且总体优势比有95%的把握在1.307-3.502之间，具有统计显著性。

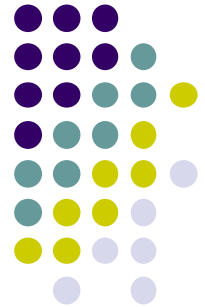
方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	income			10.512	2	.005			
	income(1)	.006	.259	.001	1	.982	1.006	.606	1.670
	income(2)	.672	.247	7.424	1	.006	1.958	1.208	3.174
	常量	-.762	.187	16.634	1	.000	.467		
步骤 2[b]	gender(1)	.504	.209	5.824	1	.016	1.656	1.099	2.493
	income			11.669	2	.003			
	income(1)	.096	.263	.134	1	.714	1.101	.658	1.843
	income(2)	.761	.251	9.147	1	.002	2.139	1.307	3.502
	常量	-1.113	.240	21.432	1	.000	.329		

a. 在步骤 1 中输入的变量: income.

b. 在步骤 2 中输入的变量: gender.

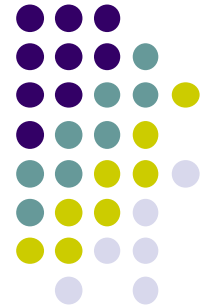
2.2 多项Logistic回归模型



2.2.1 多项Logistic回归分析概述

2.2.2 案例分析

2.2.1 多项Logistic回归分析概述



- 因变量为**多分类变量**，研究目的是分析因变量各类别与参照类别的对比情况

$$\ln(\frac{P_j}{P_J}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- 其中Pj为因变量为第j类的概率，PJ为因变量为第J（J≠j）类的概率，且第J类为参照类，$\ln(\frac{P_j}{P_J})$ 称为**广义Logit P**，是两概率比率的自然对数，该模型称为**广义Logit模型**
- 如果因变量有k个类别，则需建立k-1个模型

- 例如，如果因变量有A，B，C三个类别，且以C类别作为参照类别，则应建立两个**广义Logit模型**：

$$LogitP_A = \ln\left[\frac{P(y=A|X)}{P(y=C|X)}\right] = \beta_0{}^A + \sum_{i=1}^{p}\beta_i{}^A x_i$$

$$LogitP_B = \ln\left[\frac{P(y=B|X)}{P(y=C|X)}\right] = \beta_0{}^B + \sum_{i=1}^{p}\beta_i{}^B x_i$$

$$LogitP_C = \ln\left[\frac{P(y=C|X)}{P(y=C|X)}\right] = \ln 1 = 0$$

$$P_A + P_B + P_C = 1$$

2.2.2 案例分析

案例：研究和预测顾客的品牌选择取向

profession	gender	brand	变量	变量	变量	变量	变量	变量	变量	变量
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	1.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								
1.00	1.00	2.00								


多项 Logistic 回归: 模型
独立效应
独立效应以及交互效应
自行指定、筛选策略
指定模型
主效应
全因子(A)
设定/步进式(C)
因子与协变量(F)
gender
profession
建立项
交互
交互
强制输入项(O):
步进项(P):
步进法(W):
向前进入
在模型中包含截距(N)
继续
取消
帮助
多项 Logistic 回归: 统计量
输出模型的拟合优度指标
个案处理摘要(S)
模型
伪 R 方(P)
步骤摘要(M)
模型拟合度信息(D)
信息标准
单元格可能性
分类表(T)
拟合度(G)
单调性测量(O)
输出回归方程显著性检验结果
参数
估计(E)
似然比检验(L)
渐进相关(A)
渐进协方差(C)
输出回归系数的估计值和默认95%置信区间
置信区间(C):
定义子总体
由因子和协变量定义的协变量模式(F)
由下面的变量列表定义的协变量模式(V)
子群体(U):
性别 [gender]
职业 [profession]
继续
取消
帮助

结果解释——个案处理摘要

案例处理摘要

		N	边际百分比
购买品牌	A	79	23.4%
	B	85	25.1%
	C	174	51.5%
职业	职业一	120	35.5%
	职业二	128	37.9%
	职业三	90	26.6%
性别	男	163	48.2%
	女	175	51.8%
有效		338	100.0%
缺失		0	
总计		338	
子总体		6	

结果解释——模型拟合信息

输出零模型与当前模型的回归方程显著性检验结果

模型拟合信息

模型	模型拟合标准	似然比检验		
	-2 倍对数似然值	卡方	df	显著水平
仅截距	78.915			
最终	50.445	28.470	6	.000

拒绝回归方程显著性检验的原假设，说明自变量全体与广义Logit P之间的线性关系显著，模型选择正确

结果解释——似然比检验

输出模型引入（或剔除）各自变量后的似然比卡方值

似然比检验

效应	模型拟合标准	似然比检验		
	简化后的模型的 -2 倍对数似然值	卡方	df	显著水平
截距	50.445[a]	.000	0	.
profession	66.830	16.385	4	.003
gender	61.539	11.094	2	.004

卡方统计量是最终模型与简化后模型之间在 -2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0。

a. 因为省略效应不会增加自由度，所以此简化后的模型等同于最终模型。

职业和性别的卡方检验的概率P值都小于显著性水平0.05，则应拒绝回归系数为0的假设，即它们对广义Logit P的线性贡献均是显著的

结果解释——伪R方

伪 R 方

Cox 和 Snell	.081
Nagelkerke	.093
McFadden	.041

$$\rho^2 = 1 - \frac{LL_p}{LL_0}$$

LL_0为零模型的对数似然值，LL_p为当前模型的对数似然值

McFadden伪R方可直观视为相对零模型而言，当前模型解释信息的比率

结果解释——分类

分类

观察值	预测值			
	A	B	C	百分比校正
A	15	0	64	19.0%
B	15	0	70	0.0%
C	16	0	158	90.8%
总百分比	13.6%	0.0%	86.4%	51.2%

模型对C品牌的预测准确率较高，这与样本在品牌上的分布有一定关系，样本中C品牌的占比远高于其他两个品牌

结果解释——参数估计

- 分析**职业**、**性别**在**品牌选择倾向**（A，B，C三种品牌，且C为参照类别）时的影响

参数估计

购买品牌[a]		B	标准误	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95% 下限	Exp(B) 的置信区间 95% 上限
A	截距	-.656	.296	4.924	1	.026			
	[profession=1.00]	-1.315	.384	11.727	1	.001	.269	.127	.570
	[profession=2.00]	-.232	.333	.486	1	.486	.793	.413	1.522
	[profession=3.00]	0[b]	.	.	0	.	.	.	.
	[gender=1.00]	.747	.282	7.027	1	.008	2.112	1.215	3.670
	[gender=2.00]	0[b]	.	.	0	.	.	.	.
B	截距	-.653	.293	4.986	1	.026			
	[profession=1.00]	-.656	.339	3.730	1	.053	.519	.267	1.010
	[profession=2.00]	-.475	.344	1.915	1	.166	.622	.317	1.219
	[profession=3.00]	0[b]	.	.	0	.	.	.	.
	[gender=1.00]	.743	.271	7.533	1	.006	2.101	1.237	3.571
	[gender=2.00]	0[b]	.	.	0	.	.	.	.

X1职业

X2性别

参照类别

a. 参考类别是: C。

b. 因为此参数冗余，所以将其设为零。

①广义Logit方程（1）

- 选择A品牌与选择C品牌的概率比的自然对数模型

$$LogitP_A = \ln\left[\frac{P(y=A|X)}{P(y=C|X)}\right] = -0.656 - 1.315X_1(1) - 0.232X_1(2) + 0.747X_2(1)$$

  - 当性别相同时，职业1的ln(Pa/Pc)比职业3(参照水平)平均减少1.315，职业1的(Pa/Pc)是职业3的0.269倍。如果以Pc为基准，则职业1选择A品牌的倾向不如职业3，且统计上显著，即职业1选择A品牌的倾向性与职业3有显著差异。
  - 当职业相同时，男性的ln(Pa/Pc)比女性(参照水平)平均多0.747，男性的(Pa/Pc)是女性的2.112倍。如果以Pc为基准，则男性较女性更倾向选择A品牌，且统计上显著，即男性选择A品牌的倾向性与女性有显著差异。

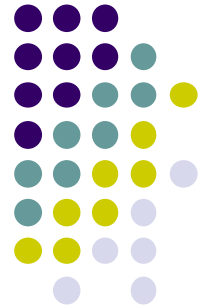
② 广义Logit方程（2）

- 选择B品牌与选择C品牌的概率比的自然对数模型

$$LogitP_B = \ln\left[\frac{P(y=B|X)}{P(y=C|X)}\right] = -0.653 - 0.656X_1(1) - 0.475X_1(2) + 0.743X_2(1)$$

  - 当性别相同时，职业1的ln(Pb/Pc)比职业3(参照水平)平均减少0.656，职业1的(Pb/Pc)是职业3的**0.519**倍。如果以Pc为基准，职业1选择B品牌的倾向不如职业3，但统计上不显著，即职业1选择B品牌的倾向与职业3并无显著差异。
  - 当职业相同时，男性的ln(Pb/Pc)比女性(参照水平)平均多0.743，男性的(Pb/Pc)是女性的**2.101**倍。如果以Pc为基准，则男性较女性更倾向选择B品牌，且统计上显著，即男性选择B品牌的倾向性与女性有显著差异。

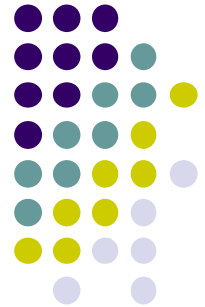
2.3 多项有序Logistic回归模型



2.3.1 多项有序回归分析概述

2.3.2 案例分析

2.3.1 多项有序回归分析概述



- 有序多分类变量——
  - 病情：一级、二级、三级和四级
  - 症状感觉：不痛、微痛、较痛和剧痛
  - 幸福感：很不幸福、不幸福、一般、幸福和很幸福
  - 满意度：很不满意、不满意、一般、满意和很满意

  。 。 。 。 。 。

多项有序回归分析：两个分析角度

（1）与多项Logistic回归类似，建立**k-1个广义优势模型**：

$$\mathrm{Logit}_1 = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0^1 + \sum_{i=1}^{p}\beta_i x_i$$

$$\mathrm{Logit}_2 = \ln\left(\frac{\pi_1+\pi_2}{1-\pi_1-\pi_2}\right) = \beta_0^2 + \sum_{i=1}^{p}\beta_i x_i$$

$$\vdots$$

$$\mathrm{Logit}_{k-1} = \ln\left(\frac{\pi_1+\pi_2+\cdots+\pi_{k-1}}{1-\pi_1-\pi_2-\cdots-\pi_{k-1}}\right) = \beta_0^{k-1} + \sum_{i=1}^{p}\beta_i x_i$$

回归系数默认相同，k-1个模型对应的回归线（或面）平行，只是截距不同（平行性假设/比例优势假设，也即一个自变量在所有有序类别的累积分割点上的效应强度（系数或优势比）是恒定不变的）

各分类的概率，因为类别有序，所以累积概率才有意义

（2）建立**k-1个补充对数-对数模型**：

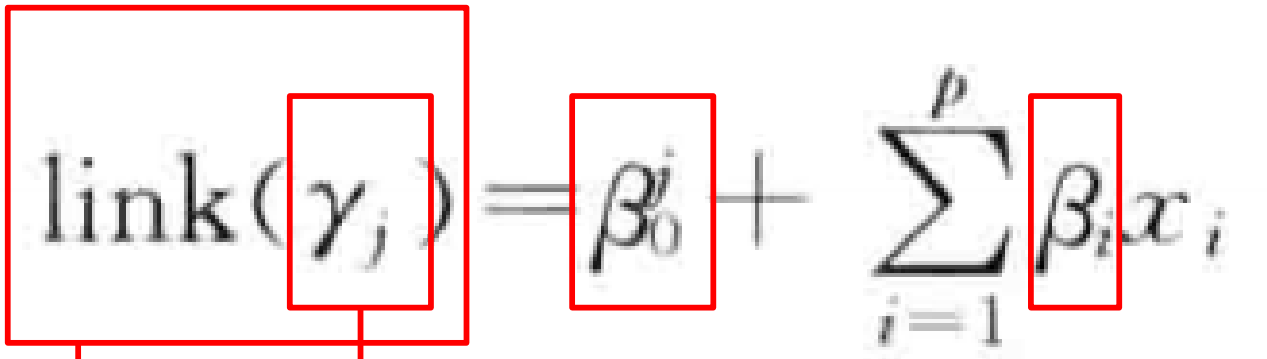
$$\ln[-\ln(1-\pi_1)]=\beta_0^1+\sum_{i=1}^{p}\beta_i x_i$$

$$\ln[-\ln(1-\pi_1-\pi_2)]=\beta_0^2+\sum_{i=1}^{p}\beta_i x_i$$

$$\vdots$$

$$\ln[-\ln(1-\pi_1-\pi_2-\cdots-\pi_{k-1})]=\beta_0^{k-1}+\sum_{i=1}^{p}\beta_i x_i$$

多项有序回归分析：位置（Location）模型



阈值

位置参数

前j类的累积概率

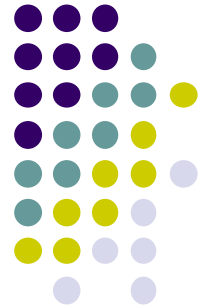
连接函数，可以是Logit形式，也可以是补充对数-对数形式，还可以是其他形式

表 10—6　　常用的连接函数

连接函数	函数形式	一般应用场合
Logit	$\ln\left(\frac{\gamma_j}{1-\gamma_j}\right)$	各类别的概率分布大致均匀
补充 Log-Log	$\ln[-\ln(1-\gamma_j)]$	高类别的概率较高
负 Log-Log	$-\ln[-\ln(\gamma_j)]$	低类别的概率较高
Probit	$\Phi^{-1}(\gamma_j)$（Φ 为标准正态分布的分布函数）	潜变量服从正态分布
Cauchit	$\tan[\pi(\gamma_j-0.5)]$ 或 $0.5+\arctan(\gamma_j)/\pi$	存在极端值

Probit回归，见后

2.3.2 案例分析



- **案例**：分析住户特征是如何影响其打算购买的房屋类型的。**自变量**包括文化程度、户口状况、年龄和家庭收入；**因变量**为购买类型（1为二手房，2为多层商品房、3为高层商品房，4为别墅）

（1）操作说明


转换(T)
分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W)
帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
自动线性建模(A)...
线性(L)...
曲线估计(C)...
部分最小平方...
二元 Logistic...
多项 Logistic...
有序...
Probit...
化程度
从业状况
婚姻
1.00 3.00
3.00 5.00
1.00 2.00
1.00 2.00

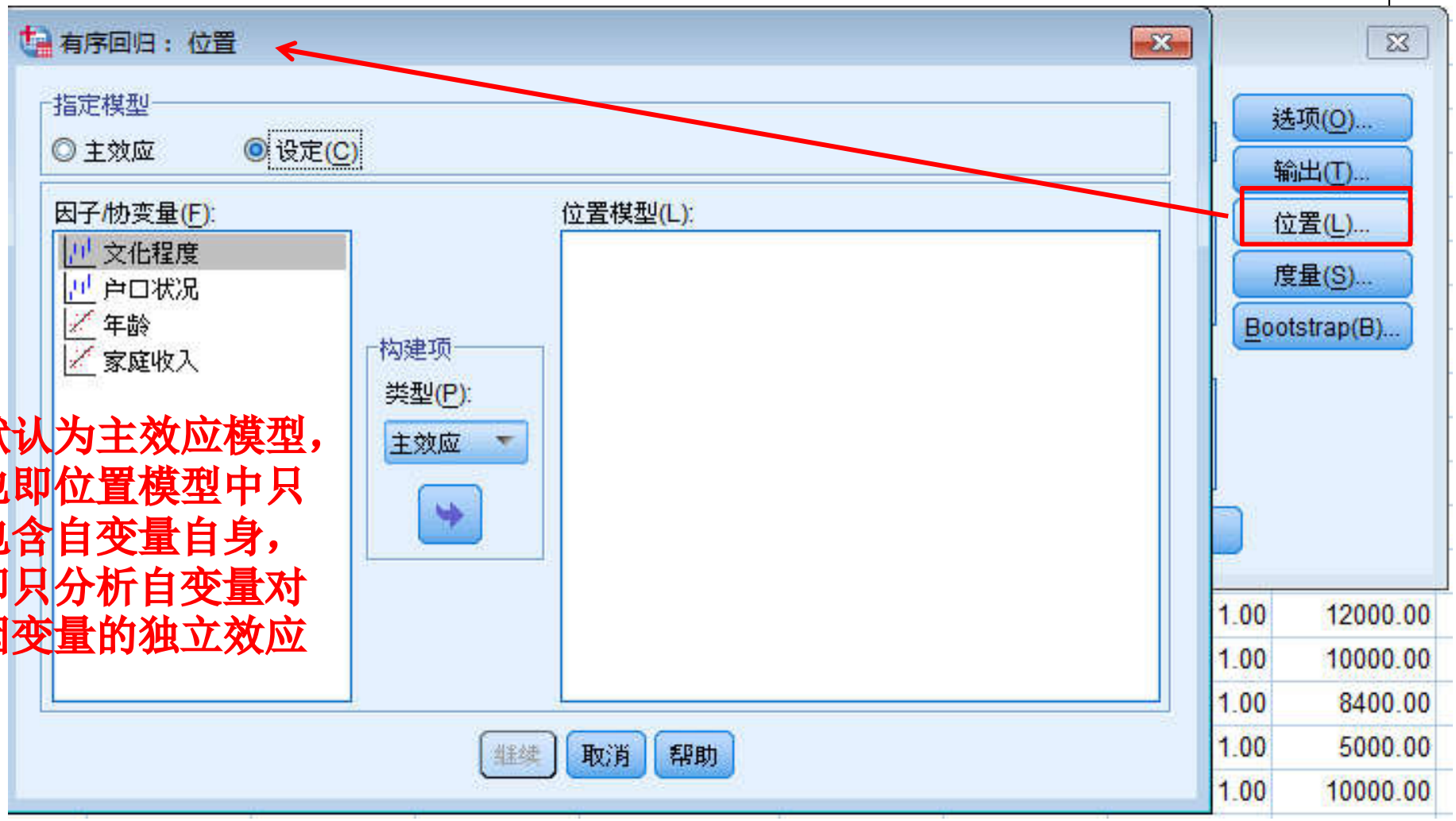
（1）操作说明


Ordinal 回归
编码
所在区
性别
从业状况
婚姻
常住人口
现住面积
人均面积
房屋产权
住房满意
因变量(D):
购买类型
因子(F)
分类型自变量
文化程度
户口状况
协变量(C)
数值型自变量
年龄
家庭收入
选项(O)...
输出(T)...
位置(L)...
度量(S)...
Bootstrap(B)...
确定
粘贴(P)
重置(R)
取消
帮助

（1）操作说明——选项窗口


Ordinal 回归
Ordinal 回归: 选项
迭代
最大迭代(M): 100
最大步骤对分(X): 5
对数似然性收敛性(L): 0
参数收敛性(P): 0.000001
置信区间(C): 95
Delta(D): 0
奇异性容许误差(S): 0.00000001
链接(K):
Logit
Cauchit
补充对数-对数
Logit
负对数-对数
概率
继续
取消
选项(O)...
输出(T)...
位置(L)...
度量(S)...
Bootstrap(B)...
编码
所在区
性别
从业状况
婚姻
常住人口
现住面积
人均面积
房屋产权
住房满意
选择连接函数，
默认为Logit函数

（1）操作说明——位置窗口



默认为主效应模型，也即位置模型中只包含自变量自身，即只分析自变量对因变量的独立效应

（1）操作说明——输出窗口

表示输出回归线（面）平行检验结果，以判断选择的连接函数是否恰当

（2）结果解释——个案处理摘要

案例处理摘要

		N	边际百分比
购买类型	二手房	100	13.9%
	多层商品房	497	69.1%
	高层商品房	113	15.7%
	别墅	9	1.3%
文化程度	初中及以下	141	19.6%
	高中（中专）	309	43.0%
	大学（专、本科）	255	35.5%
	研究生及以上	14	1.9%
户口状况	本市户口	660	91.8%
	外地户口	59	8.2%
有效		719	100.0%
缺失		0	
合计		719	

（2）结果解释——伪R方

伪 R 方

Cox 和 Snell	.095
Nagelkerke	.115
McFadden	.057

联接函数：Logit。

该模型整体拟合优度不理想

（2）结果解释——模型拟合信息

模型拟合信息

模型	-2 对数似然值	卡方	df	显著性
仅截距	1131.732			
最终	1059.715	72.017	6	.000

联接函数：Logit。

拒绝回归方程显著性检验的原假设，说明自变量全体与连接函数（这里为Logit）之间的线性关系显著，模型选择正确

（2）结果解释——平行线检验之

回归线（面）平行是位置模型的基本假设。如果违背该假设，则说明连接函数选择不正确。
原假设H0：模型的位置参数即斜率，在因变量的不同类别上无显著差异

平行线检验[a]

模型	-2 对数似然值	卡方	df	显著性
零假设	1059.715			
广义	1038.215[b]	21.499[c]	12	.044

零假设规定位置参数（斜率系数）在各响应类别中都是相同的。

a. 联接函数：Logit。

b. 在达到最大步骤对分次数后，无法进一步增加对数似然值。

c. 卡方统计量的计算基于广义模型最后一次迭代得到的对数似然值。检验的有效性是不确定的。

拒绝原假设，表明各模型的斜率存在显著差异，选择Logit连接函数不恰当

注：若各种连接函数都无法满足平行性假设，则建议使用多项无序Logistic回归模型

（2）结果解释——平行线检验之二

平行线检验[a]

模型	-2 对数似然值	卡方	df	显著性
零假设	1058.922			
广义	1040.424	18.498	12	.101

零假设规定位置参数（斜率系数）在各响应类别中都是相同的。

a. 联接函数：负对数-对数。

低类别的概率比较高，重新选择负Log-Log模型

（2）结果解释——参数估计值

参数估计值

		估计	标准误	Wald	df	显著性	95% 置信区间	
							下限	上限
阈值	[购买类型 = 1.00]	-1.400	.438	10.206	1	.001	-2.258	-.541
	[购买类型 = 2.00]	1.121	.437	6.573	1	.010	.264	1.978
	[购买类型 = 3.00]	3.853	.543	50.290	1	.000	2.788	4.918
位置	年龄	-.012	.005	4.860	1	.027	-.022	-.001
	家庭收入	1.273E-005	2.603E-006	23.939	1	.000	7.634E-006	1.784E-005
	[文化程度=1.00]	-1.045	.398	6.885	1	.009	-1.825	-.264
	[文化程度=2.00]	-.788	.390	4.091	1	.043	-1.552	-.024
	[文化程度=3.00]	-.433	.390	1.234	1	.267	-1.197	.331
	[文化程度=4.00]	0[a]	.	.	0	.	.	.
	[户口状况=1.00]	.244	.175	1.952	1	.162	-.098	.586
	[户口状况=2.00]	0[a]	.	.	0	.	.	.

研究生及以上

参考类

外地户

联接函数：负对数-对数。

a. 因为该参数为冗余的，所以将其置为零。

初步分析结论：

1. **随着年龄的增加，打算购买的房屋类型会降低；**
2. **随着家庭收入的增加，打算购买的房屋类型越高级；**
3. **文化程度的上升会带来更高级房屋类型的购买意向，但大学水平和研究生及以上水平的房屋类型购买意向可能无差异，可以考虑对这两个类别进行合并**

最终的三个方程

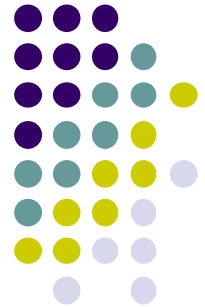
只在阈值上有差别

$$-\ln[-\ln(\gamma_1)] = -1.4 + 0.244\ \text{户口}(1) - 1.045\ \text{文化程度}(1) - 0.788\ \text{文化程度}(2) - 0.433\ \text{文化程度}(3) - 0.012\ \text{年龄} + 0.000\,01\ \text{家庭收入}$$

$$-\ln[-\ln(\gamma_2)] = 1.121 + 0.244\ \text{户口}(1) - 1.045\ \text{文化程度}(1) - 0.788\ \text{文化程度}(2) - 0.433\ \text{文化程度}(3) - 0.012\ \text{年龄} + 0.000\,01\ \text{家庭收入}$$

$$-\ln[-\ln(\gamma_3)] = 3.853 + 0.244\ \text{户口}(1) - 1.045\ \text{文化程度}(1) - 0.788\ \text{文化程度}(2) - 0.433\ \text{文化程度}(3) - 0.012\ \text{年龄} + 0.000\,01\ \text{家庭收入}$$

对方程回归系数的解读（1）



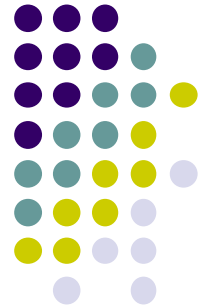
- 在控制了年龄和家庭收入的条件下，**外地户口研究生及以上文化程度人群**（**均为参考类**）：
  - 购买二手房可能性的负Log-Log值为-1.4，则 $\pi_1 = 0.017$
  - 购买二手房可能性与购买多层商品房可能性之和的负Log-Log值为1.121，则 $\pi_1 + \pi_2 = 0.722, \pi_2 = 0.705$
  - 购买二手房可能性与购买多层商品房可能性及购买高层商品房可能性之和的负Log-Log值为3.853，则 $\pi_1 + \pi_2 + \pi_3 = 0.979, \pi_3 = 0.257$
  - 购买别墅可能性 $\pi_4 = 0.021$

对方程回归系数的解读（2）

- 负Log-Log值与年龄成反比（**-0.012**），与家庭收入成正比（**0.00001**）。
- 在文化程度、家庭收入、年龄相同的条件下，本地户口的购房可能性的负Log-Log值比外地户口（参考类）平均高出**0.244**，即购房可能性平均高出**45.7%**。其他分析同理。

$-\ln[-\ln(r_1)]=0.244, r_1=45.7\%$

2.4 Probit回归模型



- Probit即概率单元（Probability Unit），用于因变量为分类变量的数据统计分析。
  - 二分类：分析 → 回归 → Probit
  - 有序和无序多分类：分析 → 回归 → Logistic，将连接函数改为Probit
- Probit回归模型中的β_i代表其他自变量取值保持不变时，自变量每改变一个单位，出现阳性结果的概率单元的改变量。β_0代表自变量全部取值为0时的概率单元值。

2.4.1 模型简介

$$\Pr obit(P) = \Phi^{-1}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n$$

Probit回归模型

对应的累积概率函数，即标准正态分布的累积概率函数:

$$P = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n) = \int_{-\infty}^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n} \phi(z)dz$$

$$\log it(P) = \ln(\frac{P}{1-P}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n$$

Logistic回归模型

变换可得等价形式，即logit累积积概率数 $P = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n)}$

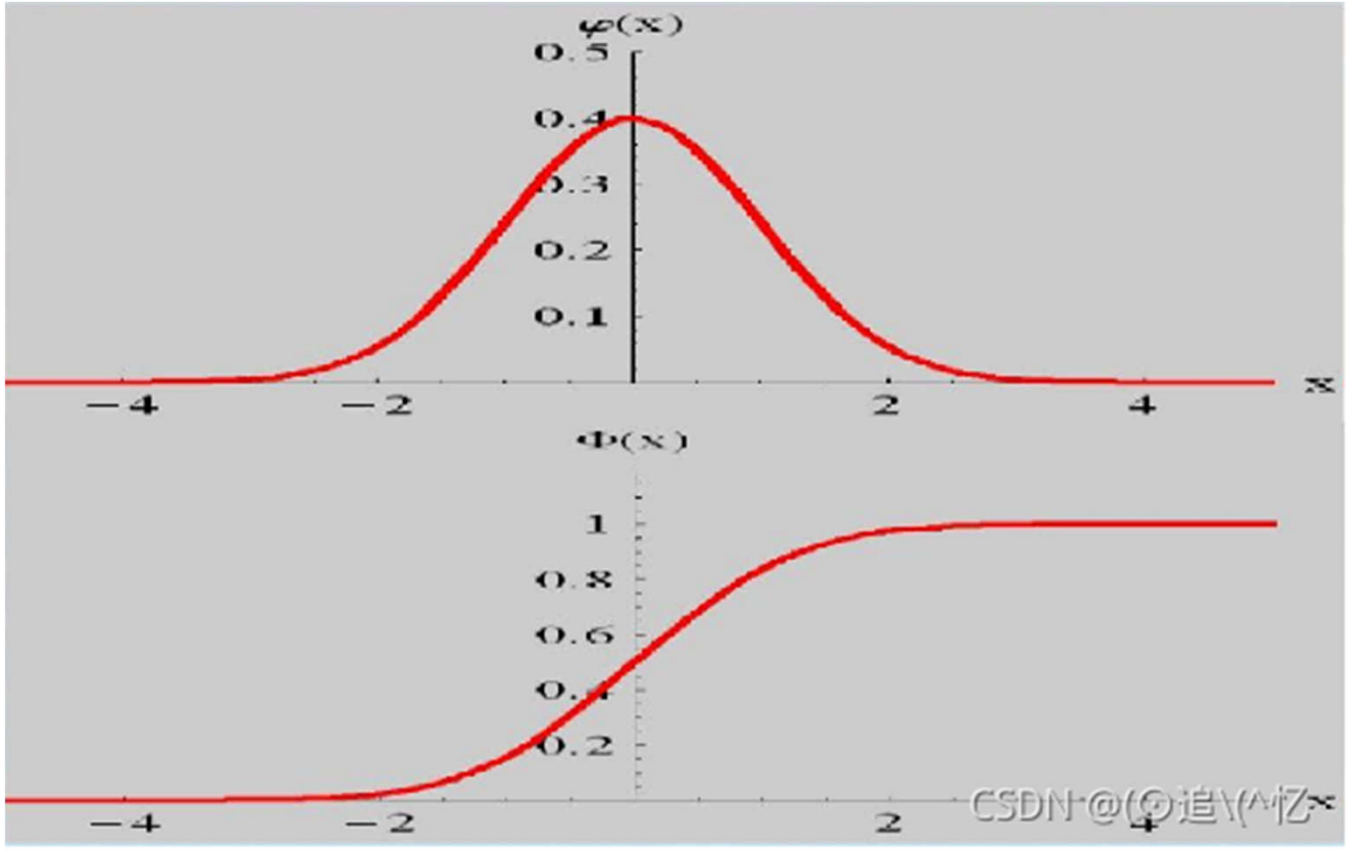
$$f(x) = F'(x)$$

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt, -\infty < x < +\infty$$

- ***Φ(x)*：标准正态分布的概率分布函数、累积概率函数**或**累积分布函数**（CDF,Cumulative Distribution Function），可用于计算随机变量小于或等于x的概率，是已知横轴（某一事件）求纵轴（概率）的过程。将概率密度函数在定义域上进行**积分**可获得。
- ***φ(z)*：标准正态分布的概率密度函数**（PDF,Probability Density Function），就是概率的密度，反映的是概率在x点处的“密集程度”，可以表示随机变量每个取值有多大的可能性。对累积概率函数**求导**可获得。
- ***$\Phi^{-1}(.)$*：Φ(.)的反函数，即Probit函数，百分点函数**（PPF,Percent Point Function）或**逆累积分布函数**（ICDF）。给定概率p求相应累积分布的随机变量x，是已知纵轴（概率）求横轴（某一事件）的过程。

附：标准正态分布的概率密度函数和分布函数

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad \Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$



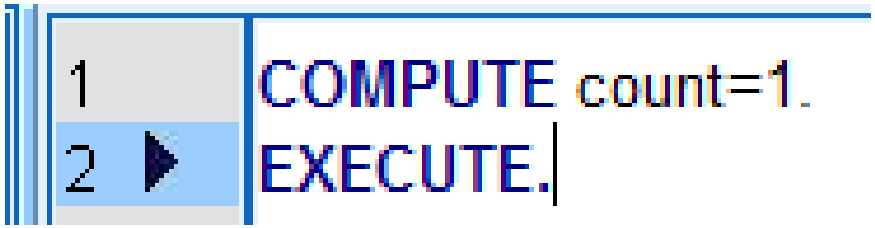
2.4.2 案例：与Logistic回归模型比较

- 案例：分析产妇在妊娠期间吸烟与否(smoke)是否导致分娩低出生体重儿(low)?

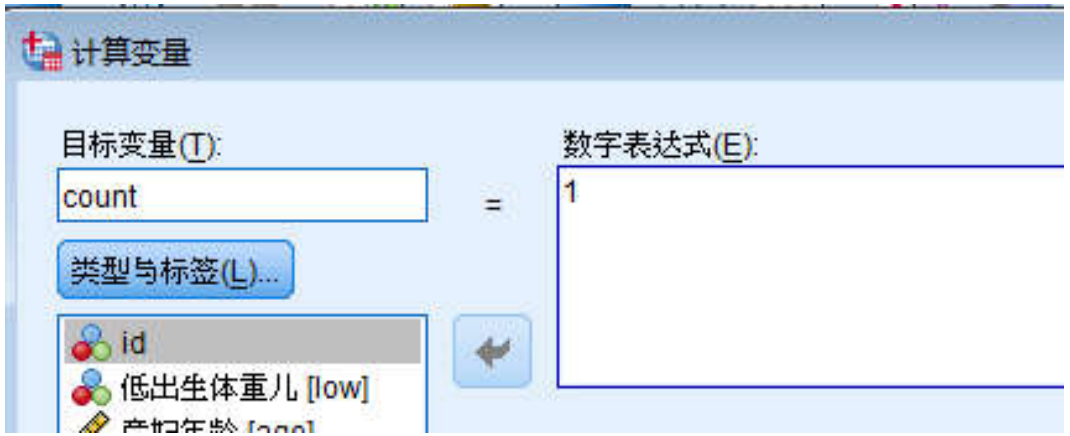
	正常体重儿	低体重儿
吸烟	44	30
不吸烟	86	29

（1）操作说明

- SPSS默认对**频数表资料**进行Probit回归分析，而本数据并非各自变量不同取值水平组合的频数表资料，每一个个案表示一个观察对象，因此需指定一个**频数变量count=1**。



方法1：文件 → 新建 → 语法



方法2：转换 → 计算变量

注：因变量的赋值必须为0和1，且1为阳性，SPSS默认取值为1表示出现阳性结果（低出生体重儿）

（2）结果解释——卡方检验

卡方检验

		卡方	df[b]	Sig.
PROBIT	Pearson 拟合度检验	189.000	187	.445[a]

a. 由于显著性水平大于 .150，因此在置信限度的计算中未使用异质因子。

b. 基于单个个案的统计量与基于分类汇总个案的统计量不同。

H_0假设：模型能够拟合数据，P=0.445，说明当前模型对数据拟合良好

（2）结果解释——参数估计值

参数估计值

	参数	估计	标准误	z	Sig.	95% 置信区间	
						下限	上限
PROBIT[a]	产妇在妊娠期间是否吸烟	.428	.194	2.204	.028	.047	.809
	截距	-.668	.127	-5.263	.000	-.795	-.541

a. PROBIT 模型: PROBIT(p) = 截距 + BX

Probit回归模型：Probit(p)=-0.668+0.428*smoke

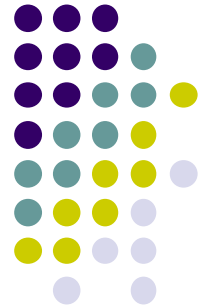
方程中的变量

OR值

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	smoke	.704	.320	4.852	1	.028	2.022
	常量	-1.087	.215	25.627	1	.000	.337

a. 在步骤 1 中输入的变量: smoke.

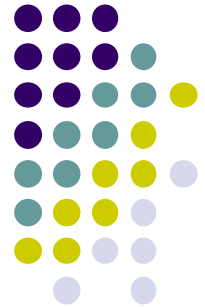
Logistic回归模型：Logit(p)=-1.087+0.704*smoke



$Probit(p)= \Phi^{-1}(p) =-0.668+0.428*smoke$

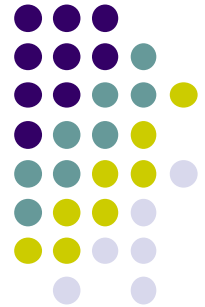
- β_0=-0.668，表示smoke=0（不吸烟组，即对照组，基线状态）的概率单元值
- β_1=0.428，表示smoke=1（吸烟组）与smoke=0（不吸烟组）的概率单元值差值，相比不吸烟组，吸烟组的概率单元值增加（β_1=0.428>0），即孕母孕期吸烟会增加低出生体重的儿童概率，结果具有统计学意义(P=0.028<0.05)。

延伸思考：与Logistic回归分析的结论是一致的吗？



- 也即检验OR值是否为**2.02**？
  - 方法1：**Probit回归模型**中求OR值
  - 方法2：**交叉列联表**中求OR值

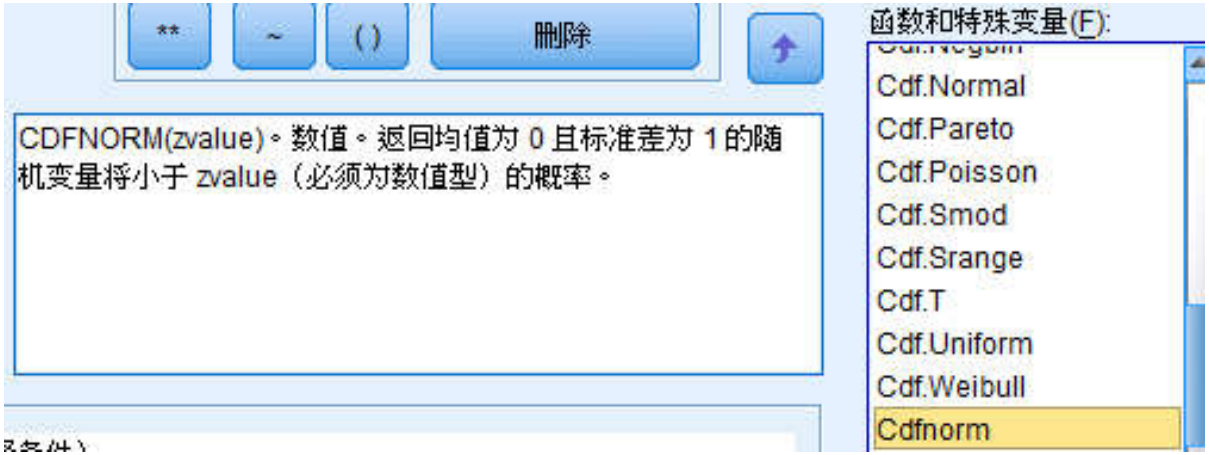
方法1：Probit回归模型中求OR值



- 不吸烟组：$p=\Phi(-0.668)=0.2521$ 29/(29+86)
- 吸烟组：$p=\Phi(-0.668+0.428)=0.4052$

- **OR值**：OR=[0.4052/(1-0.4052)] /[0.2521/(1-0.2521)] =**2.02**

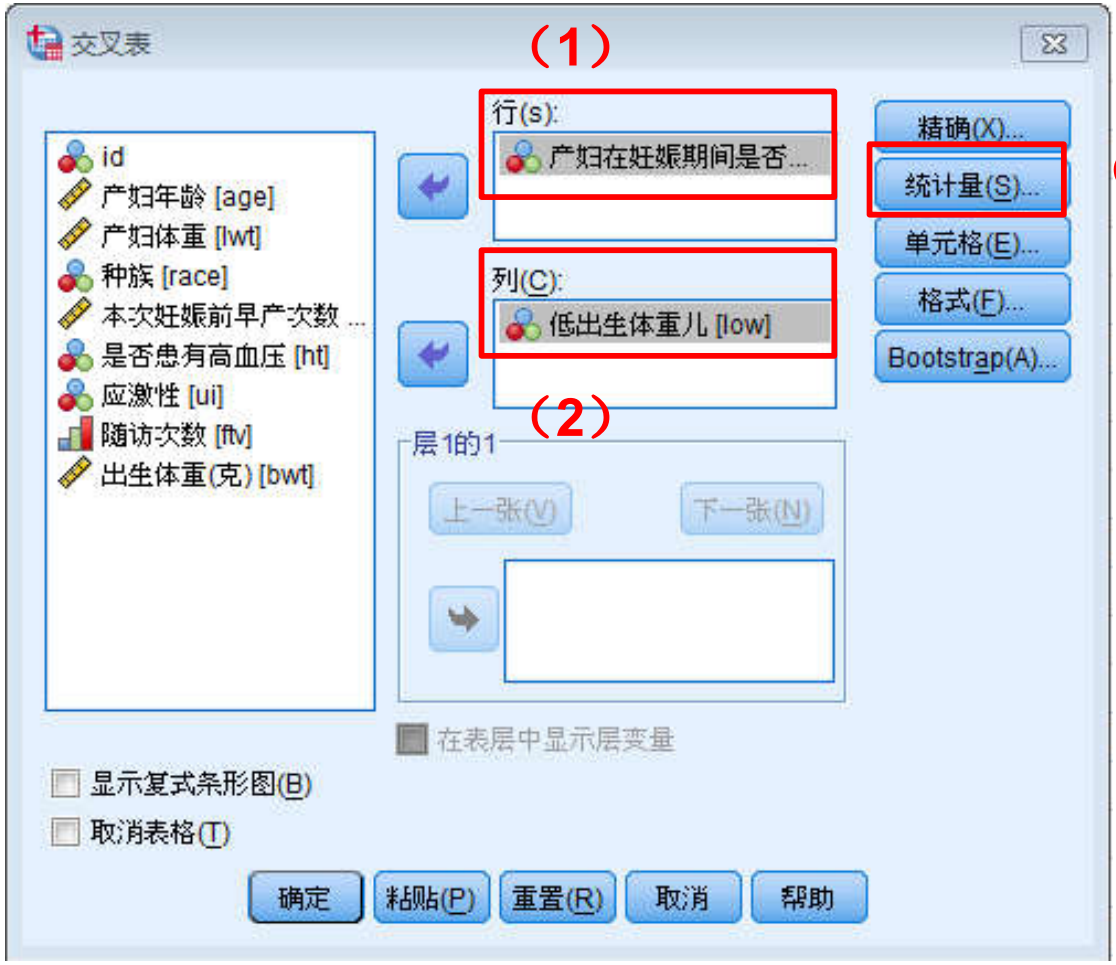
- **SPSS操作**：转换→ 计算变量 → CDFNORM函数



- **Excel操作**：公式→ 插入函数 → NORMSDIST函数

方法2：交叉列联表中求OR值

- **SPSS操作**：分析 → 描述统计 → 交叉表



（3）

卡方检验

	值	df	渐进 Sig. (双侧)	精确 Sig.(双侧)	精确 Sig.(单侧)
Pearson 卡方	4.924[a]	1	.026		
连续校正[b]	4.236	1	.040		
似然比	4.867	1	.027		
Fisher 的精确检验				.036	.020
线性和线性组合	4.898	1	.027		
有效案例中的 N	189				

a. 0 单元格(0.0%) 的期望计数少于 5。最小期望计数为 23.10。

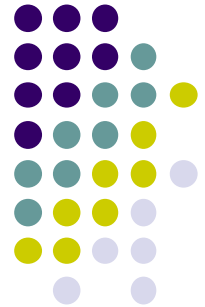
b. 仅对 2x2 表计算

风险估计

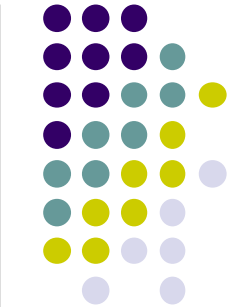
	值	95% 置信区间	
		下限	上限
产妇在妊娠期间是否吸烟 (不吸烟 / 吸烟) 的几率比	2.022	1.081	3.783
用于 cohort 低出生体重儿 = 正常	1.258	1.013	1.561
用于 cohort 低出生体重儿 = 低出生体重	.622	.409	.945
有效案例中的 N	189		

OR值

总结



- **Logistic回归模型**强调的是随着解释变量的变化，结局变量的阳性结果是发生还是不发生。（**二项分布**）
- **Probit回归模型**则倾向于研究解释变量阳性结果发生概率的变化情况。（**正态分布**）



3. 对数线性模型

3.1 模型简介

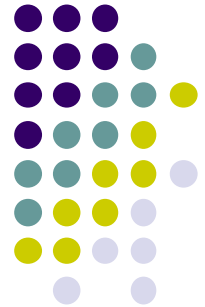
3.2 一般对数线性模型

3.3 Poisson对数线性模型

3.4 Logit对数线性模型

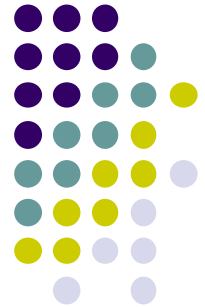
3.5 分层对数线性模型

3.1 模型简介



- **分类数据**的常用分析方法：
  - **卡方检验**：二维列联表
  - **方差分析**：因变量是连续数据
  - **对数线性模型和Logistic回归模型**：可对多个分类变量间的关系给出一个综合评价，可在控制其他因素作用的同时对变量的效应做出评估

3.1.1 模型入门：从两因素方差分析模型到二维列联表的对数线性模型



$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijk}$$

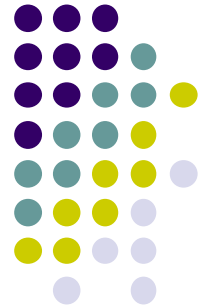
- y_{ijk}是A因素i水平和B因素j水平下的第k个观测值（**假定服从正态分布**），α_i、β_j分别表示A因素i水平和B因素j水平的**主效应**，$\alpha_i\beta_j$则为A因素与B因素的**交互效应**，ε_{ijk}是随机误差，服从正态分布$N(0,\sigma^2)$

$$\ln(f_{ij}) = \ln(\text{常数}) + \ln(\text{A的主效应}) + \ln(\text{B的主效应}) + \ln(\text{A与B的交互作用})$$

$$\ln(f_{ij}) = \mu.. + \alpha_A + \beta_B + (\alpha\beta)_{AB}$$

- $\ln(f_{ij})$是对单元格中的频数取自然对数（**假定频数服从多项分布**），α_A、β_B、$(\alpha\beta)_{AB}$同上

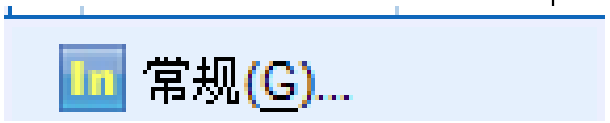
注：对数线性模型之——饱和模型和简约模型



- **饱和模型**中包含了所有主效应和交互作用项，**不饱和模型**或**简约模型**将某些无统计意义的交互作用项从饱和模型中去除。
- 拟合饱和模型必定得到实际频数完全等于理论频数，拟合优度卡方值等于0的结果。这是因为**饱和模型中独立参数的个数等于列联表的单元格数**，各单元格中的频数并无自由度可用于变化。

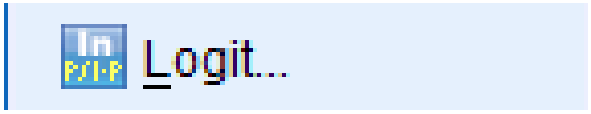
3.1.2 SPSS软件实现

- **1.一般/Poisson对数线性模型**



适用于对某些特定效应进行分析，只考虑因素之间是否相关，不考虑因果。

- **2.Logit对数线性模型**



适用于已明确区分出因变量与自变量，并且因变量为二分类变量，分析的目的是因变量和自变量的关系，结果和Logistic回归等价。

- **3.分层对数线性模型**

适用于探索性分析，没有具体分出因变量和自变量，也没有预先对某些效应感兴趣，只是设想某些变量可能存在联系，并无明确假设，该模型输出的结果最为详细且繁杂。

3.2 一般对数线性模型

- 案例：分析产妇在妊娠期间吸烟与否(smoke)以及是否导致分娩低出生体重儿(low) 这两个因素对单元格中的频数是否存在交互作用？

	正常体重儿	低体重儿
吸烟	44	30
不吸烟	86	29

3.2.1 操作说明

模型会将所有单元格的频数加上该值进行参数估计，以避免某些单元格频数为0时引起的计算问题。如果数据不存在空单元格可以将其改为0

3.2.2 结果解释——拟合优度检验

注1：模型的拟合优度检验即检验当前模型与饱和模型的预测效果之差是否有统计意义，就可以明确当前模型是否需要做进一步改善。

注2：饱和模型是对当前数据能够拟合的最完美的模型，不可能再做进一步的改进。

拟合度检验[a,b]

	值	df	Sig.
似然比	.000	0	.
Pearson 卡方检验	.000	0	.

a. 模型：多项式

b. 设计:常量 + low + smoke + low * smoke

该模型包含了所有主效应和交互作用项，也即拟合的是饱和模型，拟合结果必定最优，拟合优度卡方值等于0

3.2.2 结果解释——单元计数和残差

单元计数和残差[a,b]

低出生体重儿		产妇在妊娠期间是否吸烟		观测 计数	观测 %	期望的 计数	期望的 %	残差	标准化残差	调整残差	偏差
正常	0	不吸烟	0	86	45.5%	86.000	45.5%	.000	.000	.	.000
		吸烟	1	44	23.3%	44.000	23.3%	.000	.000	.000	.000
低出生体重	1	不吸烟	0	29	15.3%	29.000	15.3%	.000	.000	.000	.000
		吸烟	1	30	15.9%	30.000	15.9%	.000	.000	.	.000

a. 模型：多项式

b. 设计:常量 + low + smoke + low * smoke

由于拟合的是饱和模型，因此各单元格的实际频数和理论频数完全相同，各单元格拟合的残差、校正残差与偏差均为0

3.2.2 结果解释——参数估计

low，smoke取值均为1时单元格中频数的自然对数值，即ln30=3.4

参数估计[c,d]

参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
常量	3.401[a]					
[low = 0]	.383	.237	1.618	.106	-.081	.847
[low = 1]	0[b]	.	.	.	.	.
[smoke = 0]	-.034	.260	-.130	.896	-.544	.477
[smoke = 1]	0[b]	.	.	.	.	.
[low = 0] * [smoke = 0]	.704	.320	2.203	.028	.078	1.331
[low = 0] * [smoke = 1]	0[b]	.	.	.	.	.
[low = 1] * [smoke = 0]	0[b]	.	.	.	.	.
[low = 1] * [smoke = 1]	0[b]	.	.	.	.	.

a. 在多项式假设中常量不作为参数使用，因此不计算它们的标准误差

b. 此参数为冗余参数，因此将被设为零

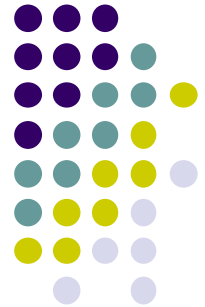
c. 模型：多项式

d. 设计:常量 + low + smoke + low * smoke

同Logistic回归分析中的回归系数

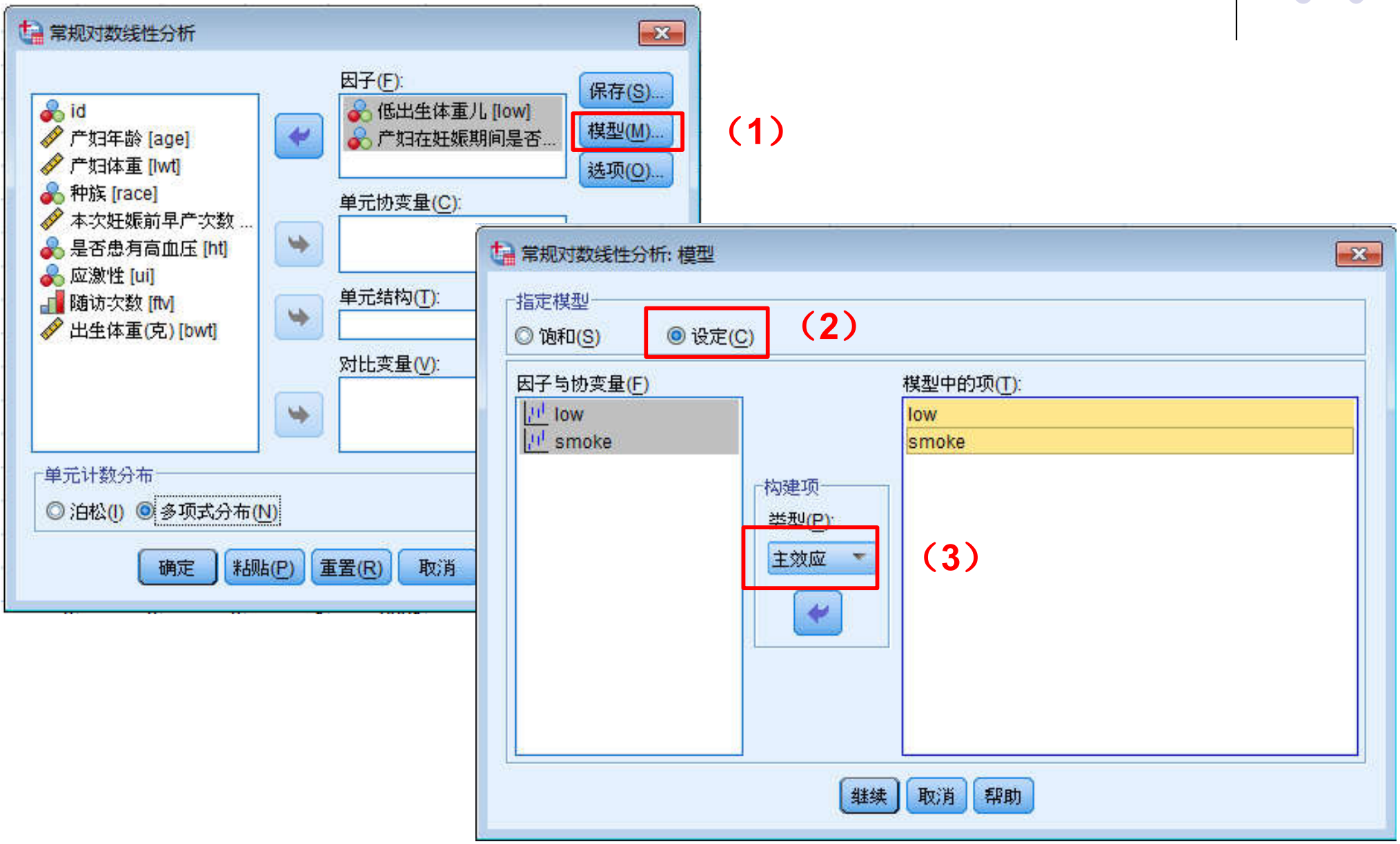
P<0.05，low和smoke的交互作用有统计意义

3.2.3 延伸思考：对交互作用检验的另一种方法

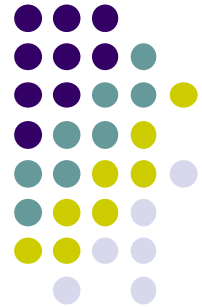


- 二维列联表的饱和模型中包含了交互作用项，在饱和模型中将该项去掉，检验**此简约模型与饱和模型的拟合优度有无统计学差异**，如果无差异，则说明该交互作用实际上不存在。

（1）操作说明


常规对数线性分析
因子(F):
低出生体重儿 [low]
产妇在妊娠期间是否...
保存(S)...
模型(M)...
选项(O)...
单元协变量(C):
单元结构(T):
对比变量(V):
id
产妇年龄 [age]
产妇体重 [lwt]
种族 [race]
本次妊娠前早产次数 ...
是否患有高血压 [ht]
应激性 [ui]
随访次数 [ftv]
出生体重(克) [bwt]
单元计数分布
泊松(I)
多项式分布(N)
确定
粘贴(P)
重置(R)
取消
（1）
常规对数线性分析: 模型
指定模型
饱和(S)
设定(C)
（2）
因子与协变量(F):
low
smoke
构建项
类型(P):
主效应
模型中的项(T):
low
smoke
（3）
继续
取消
帮助

（2）结果解释——拟合优度检验



拟合度检验[a,b]

	值	df	Sig.
似然比	4.867	1	.027
Pearson 卡方检验	4.924	1	.026

同Logistic回归分析结果

同交叉表分析结果

a. 模型：多项式

b. 设计:常量 + low + smoke

P值都小于0.05，说明该模型和饱和模型存在差异，即两因素交互作用有意义

3.3 Poisson对数线性模型

- 在拟合对数线性模型时，如果假设单元格中的频数服从Poisson分布，则相应的模型就被称为**Poisson对数线性模型**。

- 该模型可用于描述服从Poisson分布的事件发生数与各影响因素间的关系，它和**Poisson回归模型**完全等价。
  - **Poisson对数线性模型**：分析 → 对数线性模型 → 常规，“单元计数分布”框组选中“Poisson”
  - **Poisson回归模型**：分析 → 广义线性模型 → 广义线性模型，“模型类型”选项卡选中“计数”中的“Poisson对数线性”


分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W)
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
变量
变量
广义线性模型(G)...
广义估计方程(E)...


广义线性模型
模型类型
响应(E)
预测
模型
估计
统计量
EM 均值
保存
导出
选择下列一种模型类型或指定分布和链接函数的自定义组合。
尺度响应
线性
具有对数链接的 Gamma
计数
Poisson 对数线性
具有对数链接的负二项式
混合
具有对数链接的 Tweedie
具有恒等式链接的 Tweedie
有序响应
有序 logistic
有序 probit
二元响应或事件/试验数据
二元 logistic
二元 probit
间隔检查生存

补充：常见的广义线性模型的概率分布和连接函数

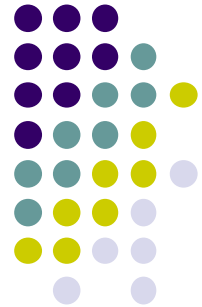
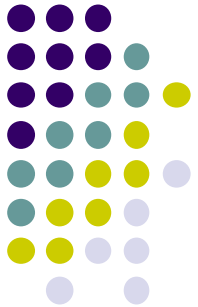


表 5.1　常见的广义线性模型的概率分布和连接函数

因变量类型	分布	连接函数	模型
连续变量	正态分布	恒等连接	直线回归
分类变量	二项分布	Logit 函数	Logistic 回归
分类变量	二项分布	$\phi^{-1}(\pi)$	Probit 回归
分类变量 离散变量	Poisson 分布	对数	Poisson 回归
分类变量	Gamma 分布		Gamma
分类变量	逆正态分布	μ^{-2}	逆正态
分类变量	负二项分布		负二项回归

注：R语言中的glm()函数可以根据概率分布和连接函数选择不同的模型



- glm()函数

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

二项Logistic回归模型

mymodel1 <- glm (Y ~ X1+X2+X3 , family = binomial (link=logit) , data = mydata)

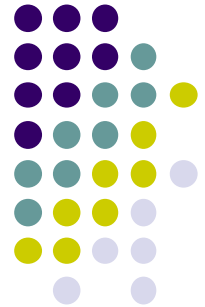
二项Probit回归模型

mymodel2 <- glm (Y ~ X1+X2+X3 , family = binomial (link=probit) , data = mydata)

Poisson回归模型

mymodel3 <- glm (Y ~ X1+X2+X3 , family = poisson ( ) , data = mydata)

3.3.1 模型简介



- **Poisson分布**用来描述单位时间/单位面积/单位空间内某事件发生次数的规律，其概率函数为：

$$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!},x=0,1,2,\cdots$$

- 在拟合对数线性模型时，如果假设单元格中的频数服从Poisson分布，则相应模型被称为**Poisson对数线性模型**。

- 设每个单元格（观察单位）内事件的发生数为λ_{ij}，则模型为

$$\ln(\lambda_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- 当单元格发生事件的观察单位数n_{ij}不同时，需要将发生数化为相同基数进行分析

$$\ln(P_{ij}) = \ln(\lambda_{ij} / n_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\ln(\lambda_{ij}) = \ln(n_{ij}) + \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

事件发生率（IR，incidence-rate）

偏移量（offset），用于去除观察单位数不相等的影响

- β_0——参照水平的事件发生率的自然对数
- $exp(\beta_0)$——参照水平的事件发生率
- β_i——其他自变量取值不变时，自变量x_i每改变一个单位，所引起的事件发生率自然对数值的改变量
- $exp(\beta_i)$——自变量x_i每改变一个单位，事件发生率是原来的$exp(\beta_i)$倍

$$IRR = \frac{IR_1}{IR_0} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_i(x_i+1)+\ldots\beta_m x_m}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_i x_i+\ldots\beta_m x_m}} = e^{\beta_i}$$

事件发生率的比值（incidence-rate ratios）

延伸思考：IRR与OR

	发生	不发生
组A	a	c
组B	b	d

$$IRR = \frac{a/(a+c)}{b/(b+d)}$$

事件发生率的比值(Incidence-rate Ratio)

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

优势的比值(Odds Ratio)，其中优势即某事件发生概率与不发生概率之比

3.3.2 案例：冠心病死亡与吸烟的关系

- 现收集某一年代英国男性医生冠心病死亡数(died)与抽烟(smoke)关系的年龄分组(agecls)数据。请推断英国男性医生冠心病死亡数与抽烟、年龄是否有关。

smoke	agecls	died	obsnum
1	1	32	52307
1	2	104	43248
1	3	206	28612
1	4	186	12663
0	1	2	18790
0	2	12	10673
0	3	28	5710
0	4	28	2585

观察人数

（1）将死亡数(died) 指定为权重变量

- 数据 → 加权个案

激活加权变量

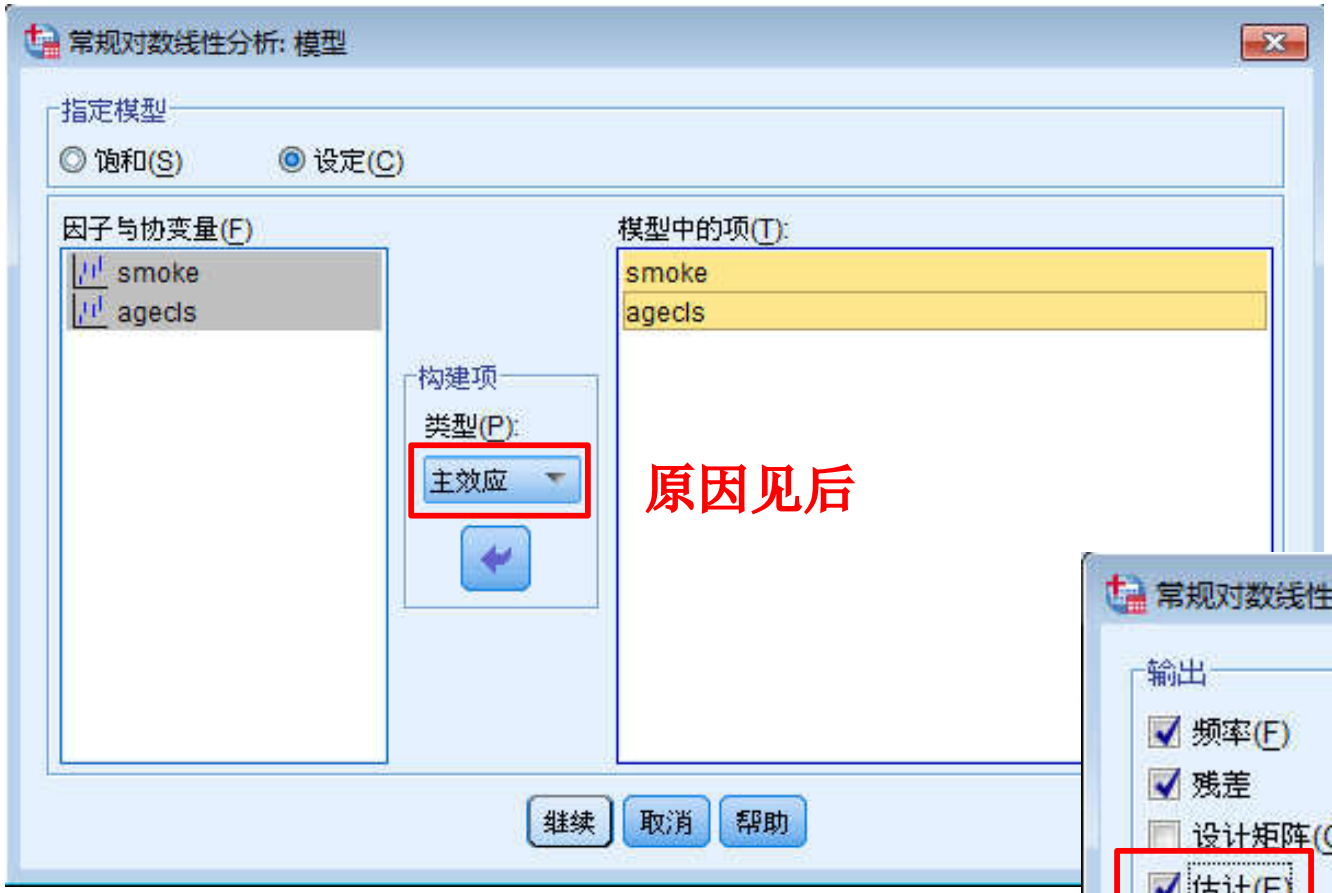
（2）Poisson对数线性模型操作说明

- 分析 → 对数线性模型 → 常规

设置
见后

单元格结构选入单元格的观察单位数，用于计算偏移量，去除基数的不同对模型造成的影响

死亡人数
服从泊松
分布



原因见后

模型会将所有单元格的频数加上该值进行参数估计，以避免某些单元格频数为0时引起的计算问题。如果数据不存在空单元格可以将其改为0

（3）结果解释——拟合优度检验

拟合度检验[a,b]

	值	df	Sig.
似然比	6.274	3	.099
Pearson 卡方检验	5.336	3	.149

a. 模型：泊松

b. 设计:常量 + smoke + agecls

P>0.05，当前模型与饱和模型相比没有统计学差异，无需再纳入两个变量的交互作用项

（3）结果解释——参数估计

参数估计[b,c]

参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
常量	-4.197	.070	-60.145	.000	-4.334	-4.060
[smoke = 0]	-.500	.127	-3.929	.000	-.750	-.251
[smoke = 1]	0[a]	.	.	.	.	.
[agecls = 1]	-3.338	.185	-18.065	.000	-3.701	-2.976
[agecls = 2]	-1.863	.115	-16.158	.000	-2.089	-1.637
[agecls = 3]	-.723	.095	-7.647	.000	-.909	-.538
[agecls = 4]	0[a]	.	.	.	.	.

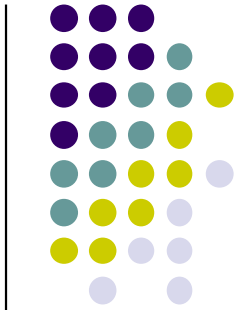
不抽烟
抽烟
35-44岁
45-54岁
55-64岁
65-74岁

a. 此参数为冗余参数，因此将被设为零。

b. 模型：泊松

c. 设计:常量 + smoke + agecls

（3）结果解释——参数估计



- 与抽烟相比，不抽烟的死亡风险较低，不抽烟是抽烟死亡风险的**0.61**倍(**IRR=exp(-0.5)=0.61**)，或者说抽烟是不抽烟死亡风险的**1.65**倍(**IRR=exp(0.5)=1.65，或IRR=1/exp(-0.5)**)。
- 随着年龄的增加，死亡风险也在逐渐上升。65-74岁死亡风险分别是35-44岁、45-54岁、55-64岁的**28.16**倍、**6.44**倍和**2.06**倍。

（3）结果解释——参数估计

- 45-54岁相比35-44岁的死亡**IRR=exp(-1.863+3.338)=4.37**，或者**IRR=28.16/6.44=4.37**，两者95%CI没有重叠部分，表明具有统计学意义。（注：其他任意两水平的IRR可用exp(两者系数相减值)计算，统计学检验可通过95%CI是否有重叠进行大体评估）

3.3.3 延伸思考：如何通过广义线性模型拟合Poisson回归？

smoke	agecls	died	obsnum
1	1	32	52307
1	2	104	43248
1	3	206	28612
1	4	186	12663
0	1	2	18790
0	2	12	10673
0	3	28	5710
0	4	28	2585

请留意数据录入格式的区别！

52307

smoke	agecls	outcome	count
1	1	死亡1	32
1	1	0	52275
1	2	1	104
1	2	0	43144
1	3	1	206
1	3	0	28406
1	4	1	186
1	4	0	12477
0	1	1	2
0	1	0	18788
0	2	1	12
0	2	0	10661
0	3	1	28
0	3	0	5682
0	4	1	28
0	4	0	2557

关键步骤

1）对**count变量**加权

2）模型类型选择**Poisson对数线性**

3）响应选择**outcome变量**

4）预测选择**smoke变量**和**agecls变量**

5）模型选择**主效应**

6）统计量选择**包括指数参数估计**

广义线性模型拟合Poisson回归结果

与Poisson对数线性模型分析结果完全一致！

参数估计

参数	B	标准 误差	95% Wald 置信区间		假设检验			Exp (B)	95% Exp(B) 的 Wald 置信区间	
			下限	上限	Wald 卡方	df	Sig.		下限	上限
（截距）	-4.197	.0698	-4.334	-4.060	3617.471	1	.000	.015	.013	.017
[smoke=0]	-.500	.1274	-.750	-.251	15.435	1	.000	.606	.472	.778
[smoke=1]	0[a]	.	.	.	.	.	.	1	.	.
[agecls=1]	-3.338	.1848	-3.701	-2.976	326.247	1	.000	.035	.025	.051
[agecls=2]	-1.863	.1153	-2.089	-1.637	261.070	1	.000	.155	.124	.194
[agecls=3]	-.723	.0946	-.909	-.538	58.481	1	.000	.485	.403	.584
[agecls=4]	0[a]	.	.	.	.	.	.	1	.	.
(刻度)	1[b]									

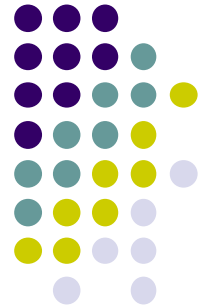
因变量: outcome

模型: （截距）, smoke, agecls

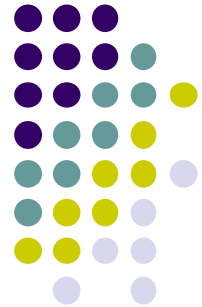
a. 此参数是冗余的，因此设置为零。

b. 固定在显示值。

3.4 Logit对数线性模型



- 该模型明确分出**因变量**和**自变量**，分析因变量和自变量之间的因果关系。
- 该模型自动引入自变量与因变量的**交互作用项**。
- 在拟合结果上该模型与**Logistic回归模型**等价。



- **案例（同一般对数线性模型）**：分析产妇在妊娠期间吸烟与否(smoke)是否导致分娩低出生体重儿(low)？也即分析这两个因素对单元格中的频数是否存在交互作用？

3.4.1 操作说明


增加的因变量框
Logit 对数线性分析
因变量(D):
低出生体重儿 [low]
因子(F):
产妇在妊娠期间是否...
单元协变量(C):
单元结构(T):
对比变量(V):
id
产妇年龄 [age]
产妇体重 [lwt]
种族 [race]
本次妊娠前早产次数 ...
是否患有高血压 [ht]
应激性 [ui]
随访次数 [ftv]
出生体重(克) [bwt]
保存(S)...
模型(M)...
选项(O)...
确定
粘贴(P)
重置(R)
取消
帮助
(1)
(2)


Logit 对数线性分析: 选项
输出
频率(F)
残差
设计矩阵(G)
估计(E)
迭代历史记录(H)
图
调节的残差值(S)
调节残差值的正态概率(N)
偏差残差(D)
偏差的正态概率(P)
置信区间(I) 95 %
标准
最大迭代(M): 20
收敛性(C): 0.001
Delta(T): 0
继续
取消
帮助
(3)
(4)

3.4.2 结果解释——离散分析和相关性测量

模型的解释度，类似回归模型的决定系数

离散分析[a,b]

	熵	集中度	df
模型	2.434	2.114	1
残差	114.902	79.050	187
总计	117.336	81.164	188

a. 模型：多项 Logit

b. 设计:常量 + low + low * smoke

2.434 / 117.336 = 0.021

相关性度量[a,b]

熵	.021
集中度	.026

a. 模型：多项 Logit

b. 设计:常量 + low + low * smoke

3.4.2 结果解释——参数估计

此为新增输出，也即smoke为0，low为1时单元格中频数的自然对数值，也即ln29=3.367

参数估计[c,d]

参数		估计	标准误	Z	Sig.	95% 置信区间	
						下限	上限
常量	[smoke = 0]	3.367[a]					
	[smoke = 1]	3.401[a]					
[low = 0]		.383	.237	1.618	.106	-.081	.847
[low = 1]		0[b]	.	.	.	.	.
[low = 0] * [smoke = 0]		.704	.320	2.203	.028	.078	1.331
[low = 0] * [smoke = 1]		0[b]	.	.	.	.	.
[low = 1] * [smoke = 0]		0[b]	.	.	.	.	.
[low = 1] * [smoke = 1]		0[b]	.	.	.	.	.

a. 在多项式假设中常量不作为参数使用。因此不计算它们的标准误差。

b. 此参数为冗余参数，因此将被设为零。

c. 模型：多项 Logit

d. 设计:常量 + low + low * smoke

对比一般对数线性模型的表格输出（结果解释见3.2.2）

参数估计[c,d]

参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
常量	3.401[a]					
[low = 0]	.383	.237	1.618	.106	-.081	.847
[low = 1]	0[b]	.	.	.	.	.
[smoke = 0]	-.034	.260	-.130	.896	-.544	.477
[smoke = 1]	0[b]	.	.	.	.	.
[low = 0] * [smoke = 0]	.704	.320	2.203	.028	.078	1.331
[low = 0] * [smoke = 1]	0[b]	.	.	.	.	.
[low = 1] * [smoke = 0]	0[b]	.	.	.	.	.
[low = 1] * [smoke = 1]	0[b]	.	.	.	.	.

a. 在多项式假设中常量不作为参数使用。因此不计算它们的标准误差。

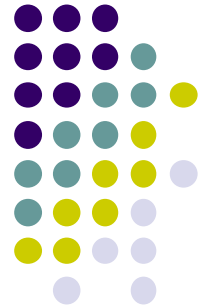
b. 此参数为冗余参数，因此将被设为零。

c. 模型：多项式

d. 设计:常量 + low + smoke + low * smoke

所有参数估计值和检验结果均与一般对数线性模型的结果相同

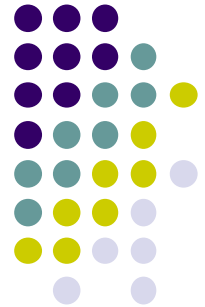
3.5 分层对数线性模型



- 对数线性模型要筛选出**最佳的不饱和模型**：
  - 模型尽量简单
  - 不含无意义的高阶交互作用

- 对数线性模型筛选的**约束条件**：
  - 一旦一个低阶的交互效应为0，相应的其他高阶交互效应就全部为0。
  - 当模型中高阶交互作用有统计学意义时，即使低阶的各项无统计学意义，也将其保留在模型中。

3.5.1 模型的选择策略

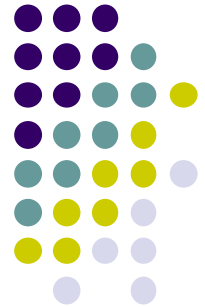


（1）建立饱和模型，然后检查每个系数的统计量或置信区间，消去无意义的效应。

（2）把所有效应一起包含到模型中，逐步从检验概率大于标准值的效应中，淘汰拟合优度变化最小的效应。**（自后淘汰法）（注：SPSS默认选择）**

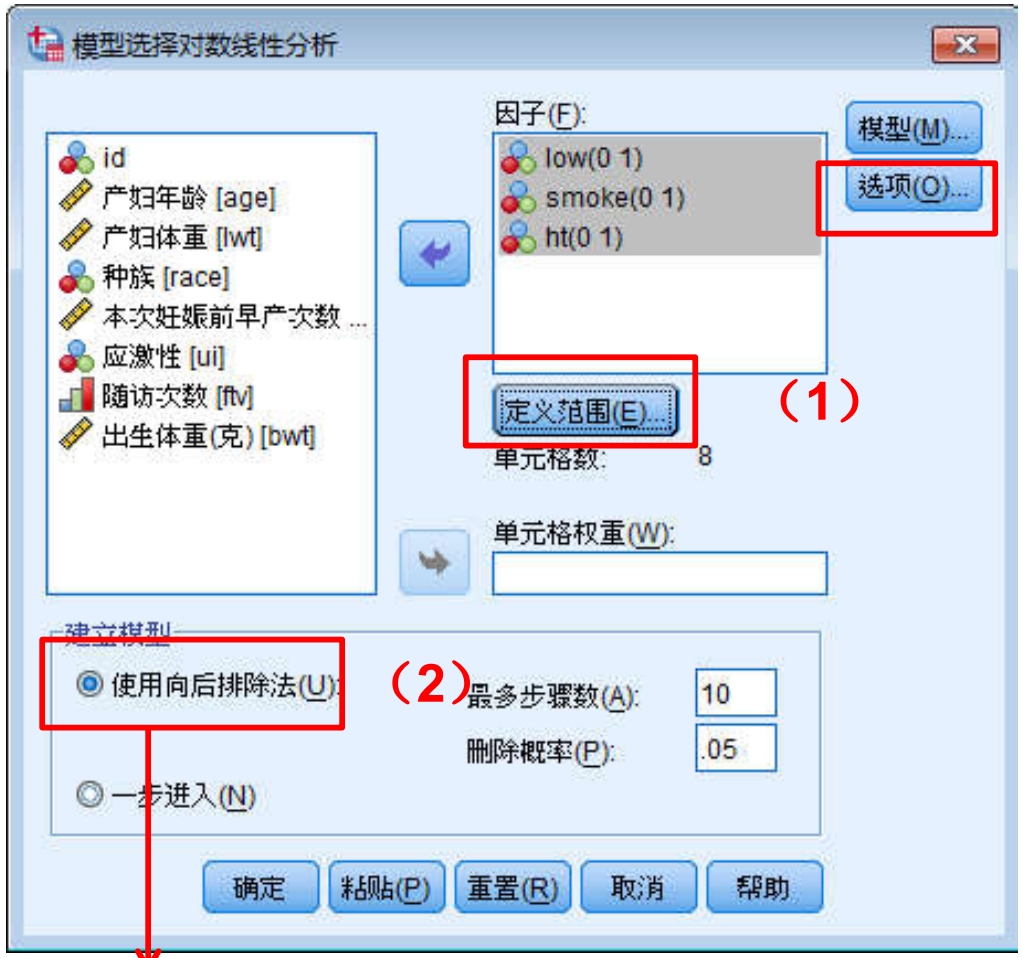
（3）系统地检查每次项的效应对模型的贡献，例如先建立二次项交互效应模型，然后建立只有主效应的模型，两种模型的似然值之差，就是交互效应对模型的贡献，通过检验拟合优度有无差异，可以得知交互效应能否被去除。**（逐一加入法）**

3.5.2 分析案例



- **案例（在一般对数线性模型案例基础上增加了一个分类自变量）：** 分析产妇在妊娠期间吸烟与否(smoke)以及高压血压与否(ht)是否导致分娩低出生体重儿(low)?

（1）操作说明



也即自后淘汰法的模型选择策略

（2）结果解释——K维和高阶效果

K-Way 和高阶效果

	K	df	似然比		Pearson		迭代数
			卡方	Sig.	卡方	Sig.	
K-Way 和高阶效果[a]	1	7	218.106	.000	236.185	.000	0
	2	4	9.179	.057	9.475	.050	2
	3	1	.265	.607	.270	.603	2
K-way 效果[b]	1	3	208.927	.000	226.710	.000	0
	2	3	8.914	.030	9.205	.027	0
	3	1	.265	.607	.270	.603	0

a. 检验 k-way 和高阶效果是否为零。

b. 检验 k-way 效果是否为零。

上表用于检验模型中K 维交互作用及K维以上交互作用是否有统计学意义，以及K维交互作用自身是否有统计学意义。
上表显示三维交互作用无统计学意义，但二维交互和主效应均有统计学意义。

（2）结果解释——步骤摘要

步骤摘要

步骤[a]			效果	卡方[c]	df	Sig.	迭代数
0	生成类[b]		low*smoke*ht	.000	0	.	
	已删除的效果	1	low*smoke*ht	.265	1	.607	2
1	生成类[b]		low*smoke, low*ht, smoke*ht	.265	1	.607	
	已删除的效果	1	low*smoke	4.858	1	.028	2
		2	low*ht	4.013	1	.045	2
		3	smoke*ht	.024	1	.876	2
2	生成类[b]		low*smoke, low*ht	.289	2	.865	
	已删除的效果	1	low*smoke	4.867	1	.027	2
		2	low*ht	4.022	1	.045	2
3	生成类[b]		low*smoke, low*ht	.289	2	.865	

a. 在每一步骤中，如果最大显著性水平大于 .050，则删除含有"似然比更改"的最大显著性水平的效果。

b. 在步骤 0 之后，将在每一步骤显示最佳模型的统计量。

c. 对于"已删除的效果"，从模型中删除该效果之后，这是卡方中的更改。

去除low*smoke*ht三阶交互作用项后拟合优度的改变无统计学意义，可考虑去除

当前模型拟合优度与饱和模型相比无统计学差异

（2）结果解释——最终模型的拟合优度检验

拟合优度检验

	卡方	df	Sig.
似然比	.289	2	.865
Pearson	.291	2	.865

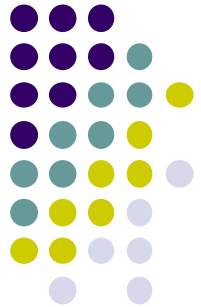
P值都大于0.05，最终模型拟合优度与饱和模型不存在差异，模型拟合良好，已得到最佳简约模型

（2）结果解释——参数估计

参数估计值

效果	参数	估计	标准误	Z	Sig.	95% 置信区间	
						下限	上限
low*smoke*ht	1	.080	.155	.518	.605	-.223	.383
low*smoke	1	.109	.155	.708	.479	-.194	.412
low*ht	1	.292	.155	1.888	.059	-.011	.595
smoke*ht	1	-.011	.155	-.072	.943	-.314	.292
low	1	.119	.155	.767	.443	-.184	.422
smoke	1	.162	.155	1.049	.294	-.141	.465
ht	1	1.299	.155	8.405	.000	.996	1.602

注意：此处只提供饱和模型的参数估计，不能输出简约模型的参数估计！！！在用它得到最佳简约模型后，还应当采用一般对数线性模型来得到具体的参数估计值和检验结果！！！



4. 模型间关系小结

4.1 对数线性模型 **VS** 方差分析模型

4.2 对数线性模型 **VS** Logistic回归模型

4.3 SPSS操作中的关联性

4.4 R语言操作中的关联性

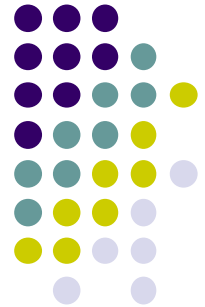
4.1 对数线性模型 VS 方差分析模型

- **相似**：**对数线性模型**的分析逻辑是以**方差分析模型**为基础，它们的作用类似，都能够分析各变量的主效应及变量间的交互效应。
- **差异**：
  - **方差分析模型**的因变量是连续变量，对数据的分布有特定要求（正态性、方差齐性等）
  - **对数线性模型**主要研究多个分类变量间的独立性和相关性，一般不分因变量和自变量，只分析各分类变量对交叉单元格内频数的影响，通常假设频数服从多项式分布

4.2 对数线性模型 VS Logistic回归模型

- **相似**：**对数线性模型**主要研究多个分类变量间的独立性与相关性，而**Logistic回归模型**的因变量也是分类变量，如果自变量也是分类变量，便与对数线性模型等价。
- **差异**：**对数线性模型**不用区分因变量和自变量，而**Logistic回归模型**则需要明确因变量和自变量；此外，当考虑的变量太多时，**对数线性模型**过于复杂（要考虑的不饱和模型太多）。

4.3 SPSS操作中的关联性



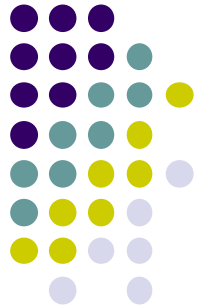
- **常规对数线性模型操作界面**——单元计数分布选“泊松”，即为Poisson对数线性模型。

- **多项有序Logistic回归模型操作界面**——将连接函数改为“概率”，即为多项Probit回归模型。

- **二项Probit回归模型操作界面**——模型选“Logit”，即为二项Logistic回归模型。

统一在“广义线性模型”大框架之下

4.4 R语言操作中的关联性——glm函数



```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

Fitting Generalized Linear Models

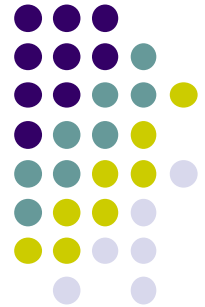
```
family(object, ...)

binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

Family Objects for Models

link character; one of "logit", "probit", "cauchit", "cloglog", "identity", "log", "sqrt", "1/mu^2", "inverse".

Create a Link for GLM Families

4.4 R语言操作中的关联性——MASS::polr函数



```
polr(formula, data, weights, start, ..., subset, na.action,
     contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "loglog", "cloglog", "cauchit"))
```

Ordered Logistic or Probit Regression