第七章 方差分析

教学要求：

1、掌握方差分析的前提条件：各种设计方案（完全随机设计、随机区组设计）变异和自由度的分解方法；多个样本均数的两两比较。

2、熟悉方差分析的基本思想。

3、了解方差齐性检验和变量转换。

方差分析(analysis of variance，简写为ANOVA)无论是在试验研究还是在调查研究中，如研究因素为k个，当k＝2时，两组总体均数比较是否相等的假设可采用前面介绍的t检验或Z检验（当然也可采用今天所介绍的方差分析）；当k>2时，即检验两组以上的总体均数是否相等时，已不能满足要求，如果用t检验将增大第一类错误，故需采用本章介绍的方差分析(analysis of variance 简写为ANOVA) .

方差分析用于两组及多个样本均数间的比较

第一节　方差分析的基本思想

方差分析的基本思想是把全部观察值间的变异按设计和需要分解成两个或多个组成部分，然后将各部分的变异与随机误差进行比较，以判断各部分的变异是否具有统计学意义。

方差分析的应用条件

1、各样本是相互独立的随机样本，均服从正态分布。

2、各样本的总体方差相等即方差齐性。

例7－1（P129）为研究钙离子对体重的影响作用,某研究者将**36**只肥胖模型大白鼠随机分为三组,每组**12**只,分别给予高脂正常剂量钙**(0.5%)**、高脂中剂量钙**(1.0%)**和高脂高剂量钙**(1.5%)**三种不同的饲料,喂养**9**周,测其喂养前后体重的差值.问三组不同喂养方式下大白鼠体重改变是否相同?

表7-1中的36个数据X间变异的分解见图7-1(教材130页).其中,散点表示36个数据;平行于横轴的长线表示总均数 $\bar{X}$ **(** $\bar{X}$ **=252.55);平行于横轴的短横线表示各组均数** $\bar{X}_i$ **(** $\bar{X}_1$ **=293.37,** $\bar{X}_2$ **=239.49,** $\bar{X}_3$**=224.78).**

表7-1 三种不同喂养方式下大白鼠体重喂养前后差值(g)

	正常(0.5%)	中剂量钙(1.0%)	高剂量钙(1.5%)	合计
X_{ij}	332.96	253.21	232.55	
	297.64	235.87	217.71	
	312.57	269.30	216.15	
	295.47	258.90	220.72	
	284.25	254.39	219.46	
	307.97	200.87	247.47	
	292.12	227.79	280.75	
	244.61	237.05	196.01	
	261.46	216.85	208.24	
	286.46	238.03	198.41	
	322.49	238.19	240.35	
	282.42	243.49	219.56	
n_i	12	12	12	36(N)
$\overline{X}_i$	293.37	239.49	224.78	252.55($\overline{X}$)
S_i^2	600.15	350.51	540.31	1364.52(S^2)

1. 总变异： 全部实验数据围绕总均数的变异。包含了处理效应和随机误差。用SS$_{总}$和总均方MS$_{总}$来描述。

$$ss_{总}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}\right)^2=s_{总}^2\left(N-1\right) \qquad v_{总}=N-1$$

MS$_{总}$=SS$_{总}$/ v $_{总}$

2.组间变异:各处理组的样本均数 $\overline{X}_1,\overline{X}_2..\overline{X}_k$ 与总体均数的差别。包含了处理效应和随机误差。用SS$_{组间}$与组间均方MS$_{组间}$来描述。

$$ss_{组间}=\sum_{i=1}^{k}n_i\left(\overline{X}_i-\overline{X}\right)^2 \qquad v_{组间}=k-1$$

MS$_{组间}$=SS$_{组间}$/ v $_{组间}$（反映了处理因素的作用，同时也包括了随机误差）

3. 组内变异　各处理组内部的观察值围绕样本均数的变异，它反映了观察值的随机误差（包括个体差异以及测定误差)。用 $SS_{组内}$和组内均方$MS_{组内}$来描述。

$$SS_{组内}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2$$

$$\nu_{组内}=N-k$$

$$MS_{组内}=SS_{组内}/\nu_{组内}$$

$$SS_{总}=SS_{组内}+SS_{组间}，且\nu_{总}=\nu_{组间}+\nu_{组内}$$

$$F=MS_{组间}/MS_{组内}$$

实践中,往往将上述过程总结为表7-2所示的方差分析表:

表7-2　完全随机设计方差分析表

变异来源	SS	ν	MS	F
总	$SS_{总}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}\right)^2$ $=S_{总}^2(N-1)$	N-1		
组间	$SS_{组间}=\sum_{i=1}^{k}n_i\left(\overline{X}_i-\overline{X}\right)^2$	K-1	$SS_{组间}/(k-1)$	$MS_{组间}/MS_{组内}$
组内	$SS_{总}-SS_{组间}$	N-K	$SS_{组内}/(N-k)$	

现以例7-1资料说明其计算方法:

$SS_{总} = S^2_{总}$**(N-1)=1364.52x(36-1)=47758.20**

$$\mathbf{SS}_{组间} = 12\times(293.37-252.55)^2 + 12\times(239.49-252.55)^2$$

$$+12\times(224.78-252.55)^2 = \mathbf{31291.67}$$

$v_{组间}$ **=k-1=3-1=2**

$MS_{组间}$**=**$SS_{组间}$ **/** $v_{组间}$**=15645.83**

$SS_{组内}$**=**$SS_{总}$ **-** $SS_{组间}$ **=16466.65**

$v_{组内}$**=N-k=33**

$MS_{组内}$ **=**$SS_{组内}$**/** $v_{组内}$**=498.99**

F =$MS_{组间}$ **/** $MS_{组内}$ **=31.36**

通常,将计算结果列成如表7-3的方差分析表:

表7-3 例7-1资料的方差分析表

变异来源	SS	df	MS	F	P
总变异	47758.32	35			
组间(处理组间)	31291.67	2	15645.83	31.36	<0.001
组内(误差)	16466.65	33	498.99		

第二节 方差分析的步骤

在上述例7-1的基础上,现将一般步骤归纳如下:

(1)建立假设检验,确定检验水准

H_0:三组不同喂养方式下大白鼠体重改变的总体平均水平相同

H_1;三组不同喂养方式下大白鼠体重改变的总体平均水平不全相同

$\alpha=0.05$

完全随机设计（completely randomized design）是将同质的受试对象随机地分配到各处理组，再观察其实验效应。是最常见的研究单因素两水平或多水平的实验设计方法，各组样本含量可以相等，也可不等。

(2)计算检验统计量

可根据表7-2的公式来计算出离均差平方和SS:包括组间离均差平方和SS组间与组内离均差平方和SS组内；自由度V：组间自由度V组间和组内自由度V组内；均方MS；组间均方MS组间和组内均方MS组内；F值 见表7-2。

(3)确定P值并作出推断结论

根据$V_1=V_{组间}=2$，$V_2=V_{组内}=33$。因附录三附表 3-1中V无33，在保守的原则下取不大于33且与其最接近者V=32，得到$\alpha=0.05$和$\alpha=0.01$的F=3.29和F=5.34,由F=31.36知P<0.01.按$\alpha=0.05$水准，差别有统计学意义，可以认为三组不同喂养方式下大白鼠体重改变的总体平均水平不全相同，即三个总体均数中至少有两个不等。至于多个总体均数中哪些均数不同，可用第三节的方法进行多个均数间的两两比较。

随机区组设计资料的方差分析

- 随机区组设计又称配伍组设计,通常是将受试对象按性质相同或相近者组成b个区组(配伍组)，每个区组中的受试对象分别随机分配到k个处理组中去。

- 例7-2 为探索丹参对肢体缺血再灌注损伤的影响，将30只纯种新西兰实验用大白兔，按窝别相同、体重相近划分为10个区组。每个区组的3只大白兔随机接受三种不同的处理,即在松止血带前分别给予丹参2ml/kg、 1ml/kg、生理盐水2ml/kg，并分别测定松止血带前及松后1小时血中白蛋白含量（g/L)，算出白蛋白减少量如 表7-4 所示，问三种处理效果是否不同？

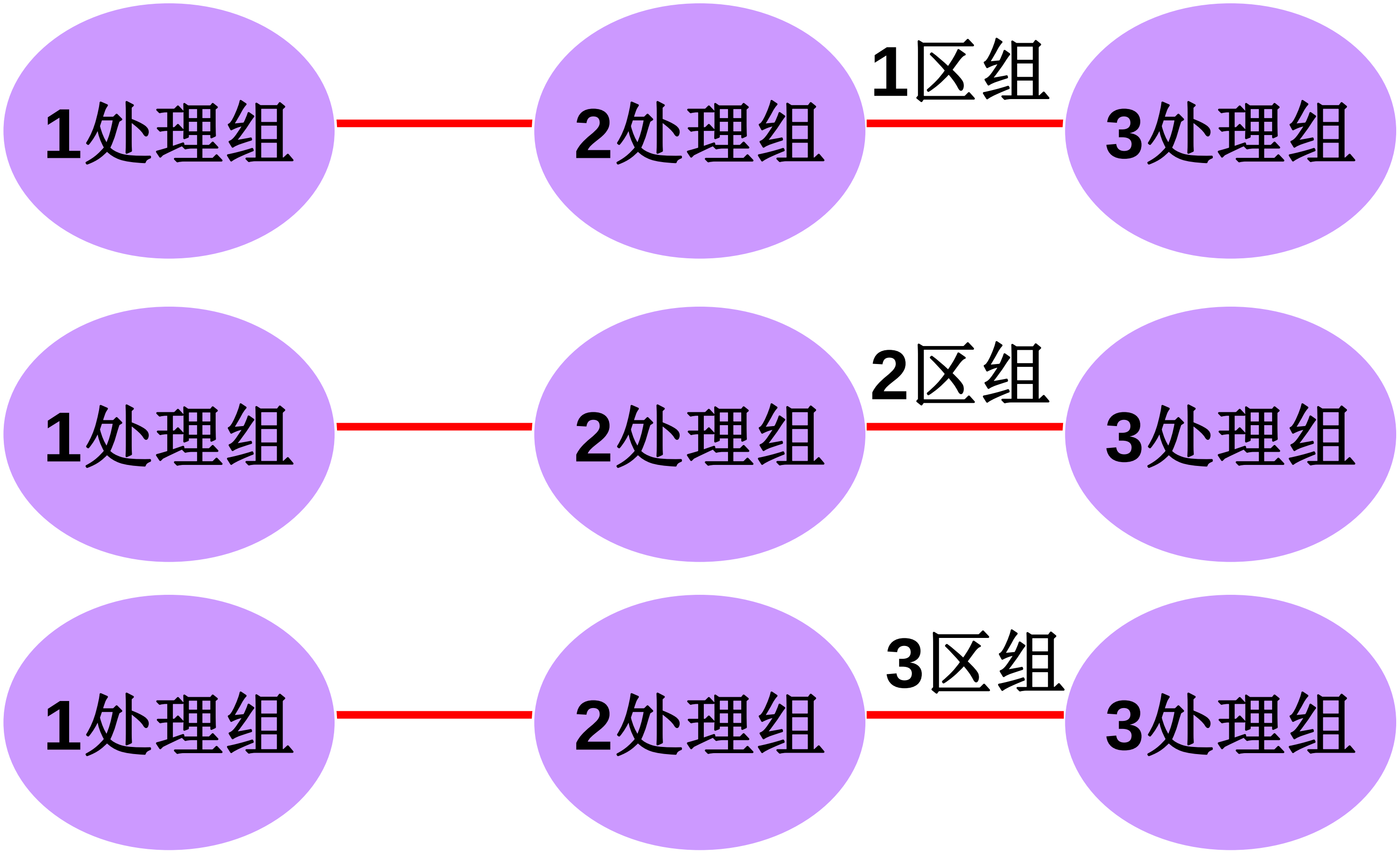

1处理组
2处理组
1区组
3处理组
1处理组
2处理组
2区组
3处理组
1处理组
2处理组
3区组
3处理组

表7-4　三种方案处理后大白兔血中白蛋白减少量(g/L)

区 组	丹参2ml/kg	丹参1ml/kg	生理盐水	$\overline{X}_j$
1	2.21	2.91	4.25	3.1233
2	2.32	2.64	4.56	3.1733
3	3.15	3.67	4.33	3.7167
4	1.86	3.29	3.89	3.0133
5	2.56	2.45	3.78	2.9300
6	1.98	2.74	4.62	3.1133
7	2.37	3.15	4.71	3.4100
8	2.88	3.44	3.56	3.2933
9	3.05	2.61	3.77	3.1433
10	3.42	2.86	4.23	3.5033
b	10	10	10	30(N)
$\overline{X}_i$	2.5800	2.9760	4.1700	3.2420($\overline{X}$)
S_i^2	0.2743	0.1581	0.1605	0.6565 (S^2)

一、离均差平方和与自由度的分解

- 从表7-4中可以看出，随机区组设计资料的变异除了总变异、处理的变异和随机误差外，还分离出区组的变异，这是由于大白兔遗传特征所致。其大小可用各区组的均数的离均差平方和表示，即：

$$SS_{区组}=\sum_{j} k\left(\overline{X}_{j}-\overline{X}\right)^{2} \quad \nu_{区组}=b-1$$

$$MS_{区组}=SS_{区组}/\nu_{区组}$$

$$SS_{区组}=1.5577 \qquad MS_{区组}=\frac{1.5577}{9}=0.1731$$

总变异和处理组变异的计算同于完全随机设计资料的方差分析。即：

$$SS_{总}=S_{总}^{2}(N-1)=19.0385$$

$$SS_{处理}=\sum_{i} b\left(\bar{X}_{i}-\bar{X}\right)^{2} \qquad \nu_{处理}=k-1=2$$

$$SS_{处理}=13.7018 \qquad MS_{处理}=\frac{13.7018}{2}=6.8509$$

数理统计证明：在随机区组设计的方差分析中，总变异分为三部分，即：

$$SS_{总}=SS_{处理组间}+SS_{区组间}+SS_{误差}$$

$$SS_{误差}=SS_{总}-SS_{处理组间}-SS_{区组间} \qquad \nu_{误差}=\nu_{总}-\nu_{处理组间}-\nu_{区组间}$$

$$=N-1-(k-1)-(b-1)$$

进一步计算出处理和区组的F值。

总 变 异

处理组间变异：**A**、**B**、**C**不同方案的影响及随机测量误差。

区组间变异：既反映了十个区组不同的影响同时又包括了随机测量误差。

误差变异： 个体差异及血白蛋白的随机测定误差。

表7-5　随机区组设计方差分析的计算公式

变异来源	SS	Df	MS	F
处理组	$\sum_{j} k\left(\bar{x}_j - \bar{x}\right)^2$	k−1	$SS_{处理}/(k-1)$	$MS_{处理}/MS_{误差}$
区组	$\sum_{i} b\left(\bar{x}_i - \bar{x}\right)^2$	b−1	$SS_{区组}/(b-1)$	$MS_{区组}/MS_{误差}$
误差	SS总-SS处理-SS区组	N−k−b+1	$SS_{误差}/(N-k-b+1)$	
总	$S^2{}_{总}(N-1)$	N−1		

表7－6　例7－2的方差分析表

变异来源	SS	df	MS	F	P
总变异	19.0385	29			
处理组	13.7018	2	6.8509	32.639	<0.01
区组	1.5577	9	0.1731	0.825	>0.05
误差	3.7790	18	0.2099		

二、分析计算步骤

（1） 建立假设并建立检验水准

对于处理组：

H_0：三个处理组总体均数相等

H_1：三个处理组总体均数不全相等

对于区组：

H_0：10个区组总体均数全相等

H_1：10个区组总体均数不全相等

$\alpha=0.05$

（2） 计算检验统计量F值

3) 确定P值并作出推断结论

以分子的自由度$v_{处理}$ =2为v_1，分母的自由度$v_{误差}$ =18为v_2，查附表3，方差分析用F界值表，$F_{0.05(2,18)}$=3.55，$F_{处理}$=32.639> $F_{0.05(2,18)}$=3.55，P <0.05。在α=0.05水准上,拒绝H_0，接受H_1，可以认为3个总体均数不全相同。

以分子的自由度$v_{区组}$ = 9 为v_1，分母的自由度$v_{误差}$ =18为v_2，查附表3，方差分析用F界值表，$F_{0.05(9，18)}$=2.46，$F_{区组}$=0.825< $F_{0.05(9,18)}$=2.46，P >0.05。在α=0.05水准上,不拒绝H_0，还不能认为10个区组总体均数不全相同。

第三节 多个样本均数间的两两比较

在检验多组均数差别的无效假设H_0时，常见的有以下两种情况：

1. 检验某几个特定的总体均数是否相等，其无效假设称为部分无效假设即某几个组所对应的总体均数相等，H_0：$\mu_i=\mu_j$（i≠j），如研究者对实验结果有一个大致的设想，在设计阶段就根据研究目的或专业知识决定了某些均数间的两两比较，常见于事先有明确假设的证实性实验研究。如：多个实验组与一个对照组的比较。可采用**Dunnett-t**检验、**LSD-t**检验、**Bonferroni**法等。

2. 检验全部k个总体均数是否相等，其无效假设称为完全无效假设即所有各组所对应的总体均数都相等，H_0：$\mu1=\mu2=\mu3=\ldots=\mu k$。例如在研究设计阶段对实验结果知之不多的探索性研究，或经数据结果的提示后，才决定的多个均数间的两两比较，这类情况往往涉及到每两个均数的两两比较。可采用SNK法、Bonferroni法等。

下面我们主要介绍：SNK法、Dunnett-t检验、Bonferroni法

一、多个样本均数间每两个均数的比较：即SNK法，也称多重极差检验或q检验。

$$q = \frac{\overline{X}_A - \overline{X}_B}{S_{\bar{d}}}$$

$$S_{\bar{d}} = \sqrt{\frac{MS_{误差}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$\nu = \nu$误差

$\overline{X}_A, \overline{X}_B$ 为两个对比组的样本均数。

$S_{\bar{d}}$ 为比较两组均数差值的标准误。

$MS_{误差}$为方差分析中算得的误差均方（组内均方），n_A和n_B分别为两对比组的样本例数。

计算的检验统计量为q值，q的分布与两比较组之间的组间跨度a及自由度ν有关。组间跨度a是指 $\overline{X}_A, \overline{X}_B$ 之间涵盖的均数个数（包括 $\overline{X}_A, \overline{X}_B$ 自身在内）

每个对比组所包含的组数（组间跨度） a =2，3，…K。

根据检验统计量q值，组间跨度a，误差自由度 $\nu_{误差}$及检验水准α，查q界值表，确定P值。

例7-3　对例7-1资料经 方差分析得出3组不同喂养方式下大白鼠体重改变的总体均数不全相同，而哪些均数不同需做两两比较

1) 建立假设检验,确定检验水准

H_0：两 对比组的总体均数相等，即$\mu_A = \mu_B$

H_1：两 对比组的总体均数不等，即$\mu_A \neq \mu_B$

α=0.05

2）计算检验统计量　　将三个样本均数从大到小排列，并编组次：

组次	1	2	3
均数	293.37	239.49	224.78
组别	正常钙组	中剂量钙	高剂量钙

列出两两计算表(见表7-10)

3）确定P值并作出推断结论：

表7-10 例7-1资料的SNK检验计算表P_{136}

对比组	两均数之差	差值的	q	对比组内	q临界值		P
A与B		标准误		包含组数a	0.05	0.01	
(1)	(2)	(3)	（4）=(2)/(3)	(5)	(6)	(7)	(8)
1与3	68.59	9.12	7.52	3	3.49	4.45	<0.01
1与2	53.87	9.12	5.91	2	2.89	3.89	<0.01
2与3	14.71	9.12	1.61	2	2.89	3.89	>0.05

二、 Dunnett-t检验：它适用于多个实验组与一个对照组的比较

$$t_D = q' = \frac{\bar{X}_T - \bar{X}_C}{S_{\bar{X}_T - \bar{X}_C}} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{MS_{误差}(\frac{1}{n_T} + \frac{1}{n_C})}} \qquad \nu = \nu_{误差}$$

公式以P137为主。

根据算得的t值、误差的自由度、试验组数（K-1）及检验水准查Dunnett-t界值表，作出统计推断结论。

例7-4　对例7-2资料，问两种不同剂量丹参浓度分别与生理盐水对照组比较其总体均数是否不同？

1)建立假设检验，确定检验水准：

H_0: 试验组与对照组的总体均数相等

H_1：试验组与对照组的总体均数不等

α =0.05

2)计算检验统计量（MS误差=0.2099）

表7-11　例7-2资料的Dunnett-t检验计算表

对比组T与C	两均数之差	t_D	t_D临界值		P
(1)	(2)	（3）=（2）/0.2049	0.05	0.01	(4)
A与C	-1.5900	-7.760	2.40	3.17	<0.01
B与C	-1.1940	-5.827	2.40	3.17	<0.01

3)确定P值作出推断结论:查Dunnett-t界值表得P值，列于上表.按 α=0.05 水准，A方案与C方案及B方案与C方案均拒绝H_0，差别有统计学意义，可认为实验组与对照组相比较大白兔血中白蛋白的减少量不同。

三、Bonfferoni 法

该法又称Bonfferoni t检验(Bonfferoni t test). Bonfferoni提出，若每次检验水准为α’，共进行m次比较，当H_0为真时，犯第一类错误的累计概率不超过mα’,即有 Bonfferoni不等式(Bonfferoni inequality)α< mα’ 成立。

例：四个样本均数经ANOVA有统计学意义，需对任意两个均数进行比较，其比较的次数$m=k(k-1)/2=6$,若$\alpha=0.05$，则6次比较均不犯第一类错误的概率为$(1-0.05)^6=0.746$，犯第一类错误的累计概率$=1-0.746=0.254<6x0.05=0.3$。故要使多次比较后犯第一类错误的累计概率保持不变或至少不超过原有水准，则现有每次比较的检验水准α’可利用上述Bonfferoni不等式确定为α/m。从实质上讲Bonfferoni 法是对检验水准进行调整，故又称Bonfferoni 调整(Bonfferoni adjustment)法。该法的思想适用于所有的两两比较，无论是本章介绍的多个均数比较，还是前面的多个频率比较。

例7-5 请对例7-1资料三组总体均数进行两两比较。

1）建立假设检验,确定检验水准

H_0：两对比组的总体均数相等

H_1：两对比组的总体均数不等

$\alpha' = \alpha/m = 2\alpha/k(k-1) = 0.0167$

2)计算检验统计量

$$t = \frac{\overline{X}_A - \overline{X}_B}{S_{\overline{X}_A - \overline{X}_B}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{MS_{误差}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$\nu = \nu_{误差}$

表7-12　例7-1资料的Bonferroni法计算表

对比组A与B (1)	两均数之差 (2)	标准误差 (3)	t (4)	P (5)
1与2	53.873	9.1195	5.907	<0.0001
1与3	68.587	9.1195	7.521	<0.0001
2与3	14.713	9.1195	1.613	0.1163

3)确定P值，做出统计推断结论

需注意： Bonfferoni t检验是两两比较方法中最为保守的。当比较次数不多时，该法效果较好；当比较的次数较多时，由于其检验水准选择得较低，结论偏于保守。尤其是当m=20、α=0.05时mα就已经等于1，也就是说7个及以上的样本均数要进行任两个均数的比较时，由于m=21>20,故不能使用该法。

第四节　方差分析的前提条件和数据变换

方差分析的前提条件为各样本是相互独立的随机样本均服从正态分布且各样本的总体方差相等。

在进行方差分析时，实际资料有时不能完全满足任何观察值都独立地来自具有等方差正态总体的假定，此时进行方差分析时，可能导致F值偏大，从而有增大第一类错误的危险。在样本例数较多的情况下，数据可看成近似正态分布，此时，方差分析对总体的非正态性并不苛求；当每组样本例数相等时，方差分析对于方差的非齐性并不苛求，故在方差分析时，最好采用每组例数相等的平衡设计方案。

二、方差齐性检验

（一）***Bartlett*** χ^2检验

（二）***Levene*** 检验

三、考察前提条件的残差图法

方差齐性最为简单、直观和有效的判断方法是残差图。

不同设计方案资料的方差分析其残差计算的具体公式不同。

如果数据符合方差分析的要求，则残差图的散点不应该有任何特殊的结构，应均匀分布在残差为0线条的上下。

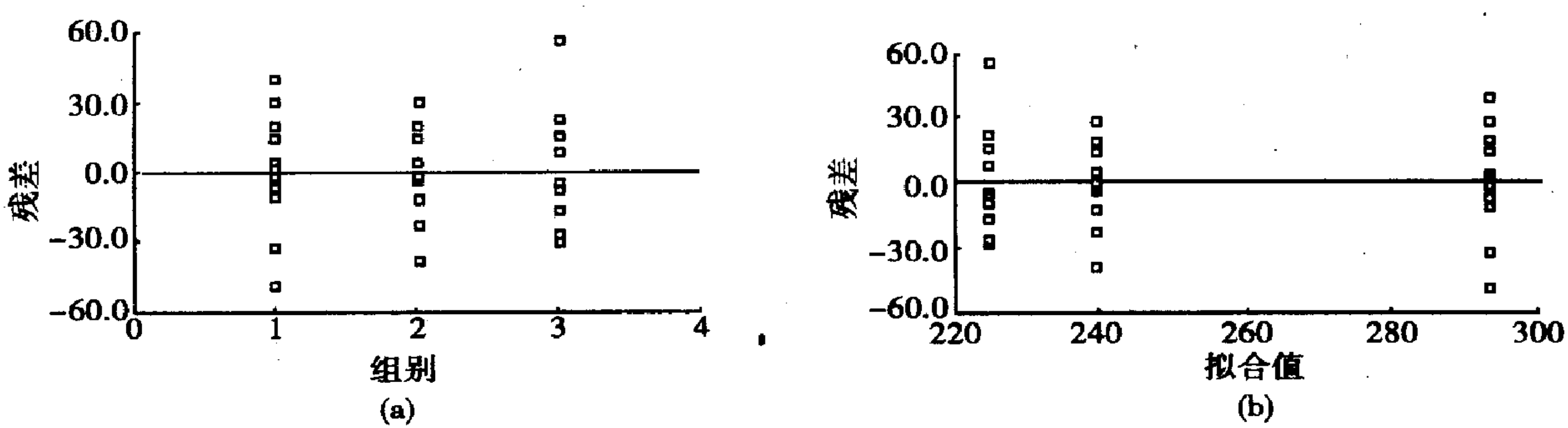
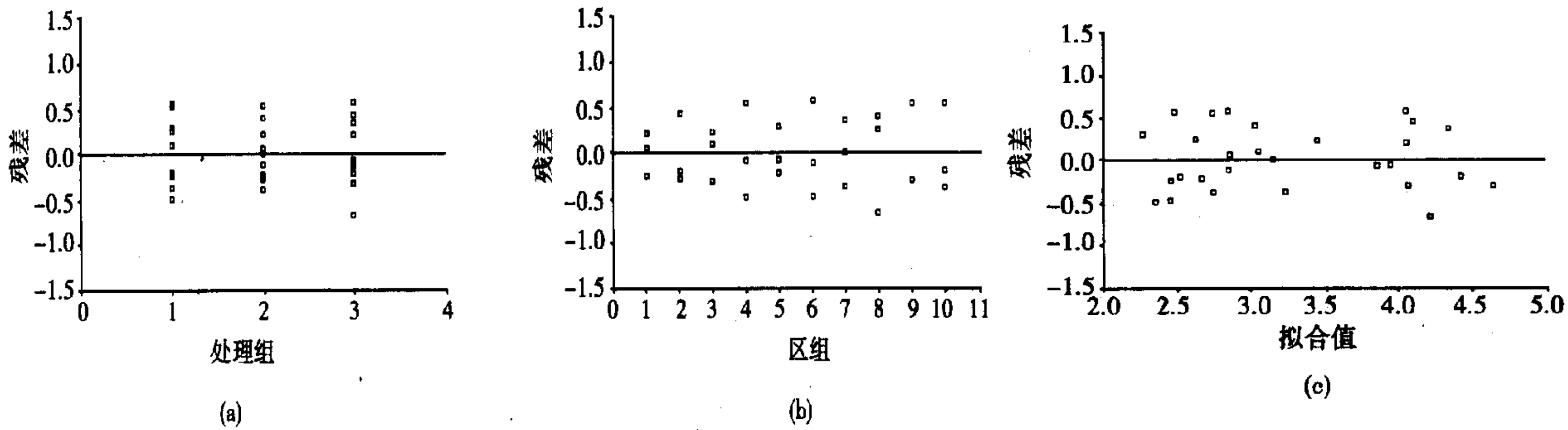


图 8-3　例 8-1 资料的残差图



四、数据变换

变量变换的目的:

1）使各组达到方差齐性;

2）使资料转换为正态分布，以满足方差分析和t检验的应用条件。通常情况下，一种适当的函数转换可使上述两个目的同时达到。

3）直线化 常用于曲线拟合。

1. 对数变换　即将原始数据**X**的对数值作为新的分析数据。

$$X' = \lg X$$

常用于：**1**）使服从对数正态分布的数据正态化。可用对数变换改善其正态性。**2**）使资料达到方差齐性的要求，特别 是各样本的标准差与均数成比例或变异系数**CV**接近一个常数时。

2. 平方根变换 即将原始数据X的平方根作为新的分析。 $X' = \sqrt{X}$

数据常用于：1）使服从Poission分布的计数资料或轻度偏态资料正态化，可用平方根变换使其正态化。2）当各样本的方差与均数呈正相关时，可使资料达到方差齐性。

3. 平方根反正旋变换 即将原始数据X的平方根反正旋值做为新的分析数据。

$$X = \sin^{-1}\sqrt{X}$$

常用于服从二项分布的率或百分比的资料。一般认为等总体率较小(如＜30%)时或较大(如＞70%时)，偏离正态较为明显，通过样本率的平方根正旋变换，可使资料接近正态分布，达到方差齐性的要求。

小结

1、方差分析的基本思想是把分部观察值总的离均差平方和分解为至少两部分，其自由度也分解为相应几个部分。每一部分有一定意义，其中至少有一部分表示各组均数间的变异，另一部分表示误差。离均差平方和除以自由度得均方，组间均方与误差均方之比为F值。F值远大于1，表示各组均数间有显著性，F值近于1，表示差别无显著性，其界点查F界值表（方差分析用）。

2、方差分析的用途很广，本章介绍了多个样本均数比较中的成组设计单因素方差分析、随机区组设计的两因素方差分析，其目的在于推断各总体均数是否相等。

3、若方差分析发现各总体均数有差别，必要时可进一步作两两比较。

4、作方差分析前要满足其应用条件，必须来自正态总体和方差齐。若不来自正态总体应采用非参数检验和变量变换；若方差不齐应采用变量变换、非参检验和近似F。

5、变量变换的目的是使方差齐，使资料正态化，可用于曲线直线化。应根据的性质选用适当的变量变换的方法。