第八章　方差分析
(analysis of variance, ANOVA)

张爱莲

新疆大学生命科学与技术学院　新疆生物资源基因工程重点实验室
zalxju@gmail. com
20101209

本章内容

8.1 方差分析的基本思路

8.2 单因素方差分析的方法

8.3 独立性检验

本章要点

- 拟合优度检验
- 独立性检验

课前提问

1. 统计假设检验的基本原理是什么？

对总体参数做出假设，包括无效假设和备择假设；研究样本的抽样分布，确定检验的统计量；根据小概率原理，规定一个显著水平，建立无效假设的拒绝域，从而做出统计推断。

2. 无效假设的含义是什么？以单个样本的假设检验为例，说明怎样通过无效假设以抽样分布为基础建立统计推断的方法。

无效假设的含义是：实验所得样本是对照总体的一个随机样本。在无效假设的前提下，就可以根据抽样分布计算出这个随机样本在总体中出现的概率。而不同的样本统计量有不同的抽样分布类型，从而建立起不同的检验方法。例如对于一个样本，假设……

3. 截至目前，我们已经学过几种假设检验的方法？分别列出各种方法的应用条件和计算统计量.

目前为止共学了8个假设检验方法。…….

⇓

3个或3个以上总体平均数如何比较呢？

统计方法的名称		应用条件	检验的统计量
单个样本	U检验	σ 已知	$u=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
	T检验	σ 未知	$t_{n-1}=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$
	χ^2检验	变异性检验	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$
			$u=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$
两个样本	U检验	σ 已知	
	成组数据t检验	σ 未知但相等	$t_{n_1+n_2-2}=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_{\bar{x}_1-\bar{x}_2}}=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-2)S_2^2}{(n_1-1)+(n_2-1)}(\frac{1}{n_1}+\frac{1}{n_2})}}$
			$t=\frac{\bar{d}}{S_{\bar{d}}}=\frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$
	配对数据t检验	配对数据	
	F检验	方差齐性检验	$F=\frac{S_{大}^2}{S_{小}^2}$

4、次数资料的χ^2检验的基本原理?

5、什么拟合优度检验？什么独立性检验？

它们的与 χ^2 的关系？它们的理论数怎样计算？

2、χ^2是度量实际观察次数与理论次数偏离程度

3、拟合优度检验即观察值与理论值是否吻合。

判断两类因子彼此相关或相互独立的假设检验就是独立性检验，独立性检验实际上是基于次数资料对因子间相关性的研究

8.1 方差分析的基本思路

8.1.1 方差分析的用途

8.1.2方差分析的基本思路

8.1.3 几个常用术语

8.1.1 方差分析的用途

例：以A、B、C、D 四种药剂处理水稻种子，其中A为对照，研究不同药剂对水稻苗高的影响。

株号	A	B	C	D	
1	19	21	28	22	
2	23	24	18	25	
3	21	27	19	27	
4	13	20	15	22	
平均	19	23	18	24	21

根据前面学过的方法，怎样比较四种处理间有无差异？

- 如果采用成组数据 t 检验.

H_{01}:μ_1= μ_2 , H_{02}:μ_1= μ_3 , H_{03}:μ_1= μ_4 , H_{04}:μ_2= μ_3, H_{05}:μ_2= μ_4, H_{06}:μ_3= μ_4，共检验6次，即 C_k^2 次。

- 进行3个以上平均数间的差异显著性检验时，*t* 检验法不适宜, 这是因为：

1、检验过程繁琐

若实验含k个处理，用*t* 检验要进行 C_k^2 次检验；

2、 无统一的实验误差

对同一实验的多个处理进行比较，应有一个统一的实验误差估计值。若用 t 检验法作两两比较，由于每次比较需计算一个 $S_{\bar{x}_1-\bar{x}_2}$，故使得各次比较误差的估计不统一。

3、误差估计的精确性和检验的灵敏性低

没有充分利用资料所提供的信息而使误差估计的精确性降低，从而降低检验的灵敏性，容易掩盖差异的显著性。

4、推断的可靠性低，检验的 I 型错误率大

用t 检验法进行多个处理平均数间的差异显著性检验，由于没有考虑相互比较的两个平均数的秩次问题，因而会增大犯I型错误的概率，降低推断的可靠性。

由于上述原因，多个平均数的差异显著性检验不宜用 t 检验，须采用方差分析法。

8.1.2方差分析的基本思路

- 方差分析 是由英国统计学家R.A.Fisher于1923年提出的.
- 方差分析是一种在若干能相互比较的资料组中，把产生变异的原因加以区分开来的方法与技术.
- 方差分析的实质：

  关于观测值变异原因的数量分析方法。

方差分析的具体做法

1.这种方法是将 k 个处理的观测值作为一个整体看待；

2.把观测值总变异的平方和及自由度分解为相应于不同变异来源的平方和及自由度，进而获得不同变异来源总体方差估计值；

3. 通过计算这些总体方差的估计值的适当比值，就能检验各样本所属总体平均数是否相等。

表中数据可以看出，组间苗高不同，同一组内苗高也不相等。

株号	A	B	C	D
1	19	21	28	22
2	23	24	18	25
3	21	27	19	27
4	13	20	15	22
平均	19	23	18	24

组内变异：误差引起

组间变异：处理效应+误差引起

如果处理效应很小或为零，则组间变异=组内变异，

如果处理效应较大，则组间变异＞组内变异。

Question：怎样比较两个变异的大小？

比较两个方差的齐性-----F检验

$\sigma^2_{组内}$是随机误差σ^2

$\sigma^2_{组间}$是随机误差+处理效应

$$F = \frac{S^2_{组间}}{\sigma^2_{组间}} \Big/ \frac{S^2_{组内}}{\sigma^2_{组内}}$$

H_O：处理效应为零 $\sigma^2_{组间} = \sigma^2_{组内} = \sigma^2$

则 $$F = \frac{S^2_{组间}}{S^2_{组内}}$$

当F> Fα时，P<0.05，拒绝H_0，处理效应显著。

当F< Fα时，P>0.05，接受H_0，处理效应不显著.

哪种处理效果最好？做平均数间的两两比较----多重比较。

- 方差分析的基本出发点在变异性上，故叫做方差分析。
- 基本思路是分析变异产生的来源，构建它们的方差，利用F检验来分析各组平均数之间的差异显著性。

8.1.3 几个常用术语

1、实验指标(experimental index)

在实验中具体测定的性状或观测的项目称为实验指标。

由于实验目的不同，选择的实验指标也不相同。

2、实验因素(experimental factor)

实验中所研究的影响实验指标的因素叫实验因素。

当实验中考察的因素只有一个时，称为单因素实验；

若同时研究两个或两个以上的因素对实验指标的影响时，则称为两因素或多因素实验。实验因素常用大写字母*A*、*B*、*C*、…等表示。

3、实验处理(treatment)

事先设计好的实施在实验单位上的具体项目叫实验处理，简称处理。

在单因素实验中，实施在实验单位上的具体项目就是实验因素的某一水平。进行单因素实验时，实验因素的一个水平就是一个处理。在多因素实验时，实验因素的一个水平组合就是一个处理。

5、重复(repetition)

在实验中，将一个处理实施在两个或两个以上的实验单位上，称为处理有重复；处理实施的实验单位数称为处理的重复数。

8.2 单因素方差分析的方法

8.2.1观察值的线性模型

8.2.2 方差分析的基本方法

8.2.3 多重比较

8.2.1观察值的线性模型

一、数据结构

假设某单因素有*a*个处理，每个处理有*n*次重复，共有*na* 个观测值。这类实验资料的数据结构如表。

处理	观	测		值			合计 $x_{i.}$	平均 $\bar{x}_{i.}$
A_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}	$x_{1.}$	$\bar{x}_{1.}$
A_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}	$x_{2.}$	$\bar{x}_{2.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮	⋮
A_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}	$x_{i.}$	$\bar{x}_{i.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮	⋮
A_k	x_{k1}	x_{k2}	…	x_{kj}	…	x_{kn}	$x_{k.}$	$\bar{x}_{k.}$
合计							$x_{..}$	$\bar{x}_{..}$

表中x_{ij} 表示第i个处理的第j个观测值

$$x_{i\bullet} = \sum_{i=1}^{n} x_{ij}$$

表示第i个处理n个观测值的和；

（i=1, 2, …, k；j=1, 2, …, n）；

$$\bar{x}_{i\bullet} = \frac{1}{n} x_{i\bullet}$$

表示第i个处理的平均数；

$$x_{\bullet\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n} x_{ij}$$

表示全部观测值的总和；

$$\bar{x}_{\bullet\bullet} = \frac{1}{an} x_{\bullet\bullet}$$

表示全部观测值的总平均数；

二、数学模型

- 本章所介绍的数学模型为线性模型，指将观测值表示为影响观测值大小的各个因素的效应的线性组合。
- 对于单因素资料而言影响观测值大小的因素分为：

（1）处理：

对各组实施不同的处理，即它们来自不同的总体；另一方面，同组个体接受的处理是相同的。

（2）随机误差：

对每个个体的影响都不同。

单因素实验的线性可加模型（linear model）

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

其中 μ 表示全实验观测值总体的平均数；

α_i 是 第 i 个 处理效应 显然有 $\sum_{i=1}^{a} \alpha_i = 0$ ；

ε_{ij} 是随机误差，相互独立，且服从 N（0，σ^2）分布。

- 单因素实验的数学模型可归纳为：

  效应的可加性（additivity）

  分布的正态性（normality）

  方差的同质性（homogeneity）

这是进行其它类型方差分析的前提或基本假定。

- 将x_{ij}的数据结构（模型） $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

  用样本符号来表示，$x_{ij} = \bar{x}_{\bullet\bullet} + (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) + (x_{ij} - \bar{x}_{i\bullet})$

  与 $\bar{x}_{\bullet\bullet}$、$(\bar{x}_{i.} - \bar{x}_{\bullet\bullet})$、$(x_{ij} - \bar{x}_{i\bullet})$ 比较可知，

  分 别是 μ、α_i、ε_{ij} 的估计值。

- 故an个观测值的总变异可分解为：

  处理间的变异

  处理内的变异两部分。

8.2.2 方差分析的基本方法

一、平方和与自由度的分解

- 在方差分析中用样本方差即均方（mean squares，MS）来度 量资料的变异程度。观测值的总变异以总均方来度量。
- 将总变异分解为处理间变异和处理内变异，即总均方分解为处理间均方和处理内均方。

总均方=总离均差平方和（总平方和）/总自由度

- 总平方和 SS_T
  - 处理间平方和：SS_A
  - 处理内平方和：SS_e
- 总自由度 df_t
  - 处理间自由度：df_A
  - 处理内自由度df_e

1. 总平方和的分解

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) + (x_{ij} - \bar{x}_{i\bullet})\right]^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(x_{ij} - \bar{x}_{i\bullet}) + (x_{ij} - \bar{x}_{i\bullet})^2\right]$$

$$= n\sum_{i=1}^{a}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2 + 2\sum_{i=1}^{a}[(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})] + \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2$$

其中 $\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet}) = 0$

所以 $\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{\bullet\bullet})^2 = n\sum_{i=1}^{a}(\bar{x}_{i\bullet}-\bar{x}_{\bullet\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\bullet})^2$

- $n\sum_{i=1}^{a}(\bar{x}_{i\bullet}-\bar{x}_{\bullet\bullet})^2$ 为各处理平均数与总平均数的离均差平方和与重复数n的乘积，反映了重复n次的处理间变异，称为处理间平方和，记为SS_A，即 $SS_A = n\sum_{i=1}^{a}(\bar{x}_{i\bullet}-\bar{x}_{\bullet\bullet})^2$

- $\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\bullet})^2$为 各处 理内离均差平方和之和，反映了各处理内的变异即误差，称为处理内平方和或误差平方和，记为SS_e，即 $SS_e = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\bullet})^2$

于是： $SS_T = SS_A + SS_e$

这个关系式中三种平方和的简便计算公式如下：

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - C$$

$$SS_A = \frac{1}{n}\sum_{i=1}^{a} x_{i.}^2 - C$$

$$SS_e = SS_T - SS_A$$

其中， $C = \dfrac{x_{\bullet\bullet}^2}{an}$ 称为矫正数。

2. 总自由度的分解

- 在计算总平方和时，资料中的各个观测值要受这一条件的约束: $\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{\cdot\cdot})=0$, $df_T=an-1$.

- 处理间平方和时，各处理均数 $\bar{x}_{i\cdot}$ 要受 $\sum_{i=1}^{a}(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})=0$ 的约束，$df_A=a-1$.

- 在计算处理内平方和时，要受a个条件 $\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i\cdot})=0$ 的约束， $dfe=an-a=a(n-1)$。

$$df_T = df_A + df_e$$

$$df_T = an - 1$$

$$df_A = a - 1$$

$$df_e = df_T - df_A$$

二、计算各部分均方

各部分平方和除以各自的自由度得到均方， 分别记为 MS_A 和MS_e

即： $$MS_A = SS_A / df_A$$

$$MS_e = SS_e / df_e$$

三、F检验

$H_0: \sigma_\alpha^2 = 0$,　　$H_A: \sigma_\alpha^2 \neq 0$;

或 $H_0: \mu_1 = \mu_2 = \ldots = \mu_a$;　　H_A: 各 μ_i 不全相等,

或 $H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$;　H_A: 各 $\alpha_i \neq 0$

$F = MS_A / MS_e$

判断处理间均方是否显著大于处理内(误差)均方。

如果接受H_0，那么MS_A与MS_e都是总体误差 σ_2的估计值，理论上讲F值等于1;

如果拒绝H_0，那么 MS_A之期望均方中的α就不等于零，理论上讲F 值就必大于1。

- 若$F < F_{0.05(df_1,df_2)}$，即$P > 0.05$，不能否定H_0，表述为：各处理间差异不显著，在F值的右上方标记“*ns*”，或不标记符号；
- 若$F_{0.05(df_1,df_2)} \leqslant F < F_{0.01(df_1,df_2)}$，即$0.01 < P \leqslant 0.05$，否定$H_0$，接受$H_A$，表述为：各处理间差异显著，在$F$值的右上方标记“*”；
- 若$F \geqslant F_{0.01(df_1,df_2)}$，即$P \leqslant 0.01$，否定$H_0$，接受$H_A$，结果表述为：各处理间差异极显著，在$F$值的右上方标记“**”。

【例8.1】以A、B、C、D 四种药剂处理水稻种子，其中A为对照，研究不同药剂对水稻苗高的影响。

株号	A	B	C	D	
1	19	21	28	22	
2	23	24	18	25	
3	21	27	19	27	
4	13	20	15	22	
平均	19	23	18	24	21

1.自由度与平方和分解

$$df_T = an - 1 = 4 \times 4 - 1 = 15$$

$$df_{处理} = a - 1 = 4 - 1 = 3$$

$$df_{误差} = an - a = 4 \times 4 - 4 = 12$$

$$C = \frac{x_{..}^2}{an} = \frac{336^2}{4 \times 4} = 7056$$

$$SS_T = \sum x_{ij}^2 - C = 19^2 + 23^2 + \ldots + 22^2 - 7056 = 222$$

$$SS_{处理} = \frac{1}{n}\sum_1^a x_{i\cdot}^2 - C = \frac{1}{4}\left(76^2 + 92^2 + 72^2 + 96^2\right) - 7056 = 104$$

$$SS_e = SS_T - SS_A = 222 - 104 = 118$$

2. F检验

$\alpha=0.05$, H_0: $\alpha_i=0$ H_A: $\alpha_i>0$

$$MS_{处理}=\frac{SS_{处理}}{df_{处理}}=\frac{104}{3}=34.67$$

$$MS_e=\frac{SS_e}{df_e}=\frac{118}{12}=9.83$$

$$F=\frac{MS_{处理}}{MS_e}=\frac{34.67}{9.83}=3.53$$

$$F_{0.05,3,12}=3.49$$

$F>_{F0.05}$,拒绝H_0 处理效应显著。

方差分析表

变异来源	平方和	自由度	均方	F	$F_{0.05}$
药剂间	104	3	34.67	3.53*	$F_{3,12,0.05}=3.49$
误差	118	12	9.83		
总和	222	15			

单因素方差分析表

变异来源	平方和	自由度	均方	F	$F_{0.05}$
处理间	SS_A	a-1	MS_A	MS_A/MSe	
误差	SSe	an-a	MSe		
总和	SS_T	an-1			

【例8.2】 某水产研究所为了比较四种不同配合饲料对鱼的饲喂效果，选取了条件基本相同的鱼20尾，随机分成四组，投喂不同饲料，经一个月实验以后，各组鱼的增重结果列于下表。

表8-2 饲喂不同饲料的鱼的增重 （单位：10g）

饲料	鱼的增重（x_{ij}）					合计 $x_{i.}$	平均 $\bar{x}_{i.}$
A_1	31.9	27.9	31.8	28.4	35.9	155.9	31.18
A_2	24.8	25.7	26.8	27.9	26.2	131.4	26.28
A_3	22.1	23.6	27.3	24.9	25.8	123.7	24.74
A_4	27.0	30.8	29.0	24.5	28.5	139.8	27.96
合计						$x_{..}$=550.8	

各项平方和及自由度计算如下：

矫正数 $C = x_{..}^2 / na = 550.8^2 /(4\times 5) = 15169.03$

$$SS_T = \sum\sum x_{ij}^2 - C$$

$$= 31.9^2 + 27.9^2 + \cdots + 28.5^2 - C$$

$$= 15368.7 - 15169.03 = 199.67$$

$$SS_A = \frac{1}{n}\sum x_{i.}^2 - C$$

$$= \frac{1}{5}(155.9^2 + 131.4^2 + 123.7^2 + 139.8^2) - C$$

$$= 15283.3 - 15169.03 = 114.27$$

$$SS_e = SS_T - SS_A$$

$$= 199.67 - 114.27 = 85.40$$

总自由度 $df_T = na - 1 = 5 \times 4 - 1 = 19$

处理间自由度 $df_A = a - 1 = 4 - 1 = 3$

处理内自由度 $df_e = df_T - df_A = 19 - 3 = 16$

用SS_A、SS_e分别除以df_A和df_e得到MS_A及MS_e。

$$MS_A = SS_A / df_A = 114.27 / 3 = 38.09$$

$$MS_e = SS_e / df_e = 85.40 / 16 = 5.34$$

$$F = \frac{MS_{处理}}{MS_e} = \frac{38.09}{5.34} = 7.13$$

$$F_{0.01,3,16} = 5.417$$

方差分析表

变异来源	平方和	自由度	均方	F值
处理间	114.27	3	38.09	7.13**
处理内	85.40	16	5.34	
总变异	199.67	19		

F_{α}

$F_{0.01}$

5.417

$F > F_{0.01}$，拒绝H_0，$P<0.01$，四种不同配合饲料对鱼的饲喂效果达到了极显著的水平。

复 习

1. 方差分析的用途

2. 方差分析的基本思路

3. 方差分析的出发点

4. 方差分析的步骤

5. 单因素方差分析中 SS_T， SS_A， SS_e 的含义及三者的关系。

1、3个或3个以上总体平均数的差异显著性检验

2、基本思路是分析变异产生的来源，构建它们的方差，利用F检验来分析各组平均数之间的差异显著性

3、方差分析的基本出发点在变异性上，故叫做方差分析。

8.2.3　多重比较

(multiple comparisons)

一、多重比较的定义

- 多个平均数两两间的相互比较称为多重比较

  F值显著或极显著，否定H_0，实验的总变异主要来源于处理间的变异，实验中各处理平均数间存在显著或极显著差异，但并不意味着每两个处理平均数间的差异都显著或极显著，也不能具体说明哪些处理平均数间有显著差异，哪些差异不显著。

- 多重比较的方法很多，介绍最小显著差数法(LSD法)和最小显著极差法(LSR法) 。

二、最小显著差数法

(*LSD*法，least significant difference)

LSD法实际上是成组 t 检验.

在*F*检验显著的 前提下，先计算出显著水平为 α 的最小显著差数 LSDα，然后将任意两个处理平均数的差数的绝对值 $|\bar{x}_{1\bullet}-\bar{x}_{2\bullet}|$ 与其比较。若 $|\bar{x}_{1\bullet}-\bar{x}_{2\bullet}|>LSD\alpha$时，则 与$\bar{x}_{1\bullet}$ 在$\bar{x}_{2\bullet}$α水平上差异显著；反之，则在 α 水平上差异不显著。

LSD计算方法简述如下：

对于任意两组数据的平均数，差数（$\bar{x}_1 - \bar{x}_2$）的差异显著性检验，可以用成组数据 t 检验，

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \Longrightarrow t = \frac{\bar{x}_1 - \bar{x}_2}{S_{(\bar{x}_1 - \bar{x}_2)}}$$

$$S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{MS_e\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

每一组的组内方差为误差均方，所以 $s_1^2 = s_2^2 = MSe$,

当$n_1 = n_2$时

$$S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{2MS_e}{n}}$$

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{2MS_e}{n}}}$$ ，df=$an-a$

其中MS_e为误差均方，n为每一处理的观察次数，于是

当 $t>t_{0.05}$ 时差异显著，

当 $t>t_{0.01}$时差异极显著。

因此，当差异显著时 $\frac{|\bar{x}_1-\bar{x}_2|}{\sqrt{\frac{2MS_e}{n}}}>t_{0.05}$

并可得到，当 $\left|\bar{x}_1-\bar{x}_2\right|>t_{0.05}\sqrt{\dfrac{2MS_e}{n}}$ 时差异显著。

$t_{0.05}\sqrt{2MS_e}/n$ 称为最小显著差数，记为 LSD。每一对平均数的差与LSD比较，当 $\left|\bar{x}_1-\bar{x}_2\right|$ >LSD时，差异显著；否则差异不显著。

- LSD是一种很有用的检验方法，计算起来很方便，也容易比较。但是它有难以克服的缺点，即这种比较方法将会加大Ⅰ型错误的概率。

关于*LSD* 法的几点说明

- *LSD* 法实质上是 *t* 检验。
- 但利用*F*检验中的误差自由度 df_e 查临界t_α值，利用误差均方 MS_e 计算均数差异标准误， 不同于每次利用两组数据进行多个平均数两两比较的t 检验 。
- *LSD* 法克服了t 检验法程繁琐， 无统 一 实验误差且估计误差的精确性和检验的灵敏性低。
- 但未考虑平均数间相距的远近，并未解决推断的可靠性降低、犯I型错误概率变大的问题。

二、 最小显著极差法

(LSR法 , Least significant ranges)

考虑平均数间相距的远近，把平均数的差作为极差，根据极差范围内所包含的处理数(称为秩次距) k 的不同而采用不同的检验尺度。这些在显著水平 α 上依秩次距 k 的不同而采用的不同检验尺度叫做最小显著极差*LSR*。

因此，有k个平均数相互比较，有 k-1 种秩次距 (k，k-1，k-2，…，2)，因而需求得k-1个最小显著极差(LSR_{α}，k)，分别作为判断具有相应秩次距的平均数的极差是否显著的标准。

- 常用的 LSR 法有 q 检验法和 Duncan法（新复极差法）两种。
- Duncan检验与LSD的一个明显不同是Duncan检验中，不同对平均数的差有不同的临界值R_k 。
- 最小显著极差计算公式为：

$$R_{a,k} = r_{a(k,df_e)} S_{\bar{x}}$$

*LSR*法的原理

- 比如，相互比较10个$\bar{x}$ ，先将10个 $\bar{x}$ 依数值大小排列， $\bar{x}_1$与$\bar{x}_{10}$ 的差异显著性，由其差数是否大于秩次距***k*=10**时的最小显著极差决定；
- $\bar{x}_1$与$\bar{x}_9$ 的差异显著性，则由极差是否大于***k*=9** 时的最小显著极差决定；
- 直到任何两个相邻平均数的差数由这些差数是否大于*k*=2 时的最小显著极差决定为止。

Duncan检验程序

①将需要比较的a个平均数依次排列好，使之：

$\bar{x}_1 > \bar{x}_2 > \cdots > \bar{x}_a$ 并将每一对平均数的差列成下表：

	a	$a-1$	$\cdots$	3	2
1	$\bar{x}_1 - \bar{x}_a$	$\bar{x}_1 - \bar{x}_{a-1}$	$\cdots$	$\bar{x}_1 - \bar{x}_3$	$\bar{x}_1 - \bar{x}_2$
2	$\bar{x}_2 - \bar{x}_a$	$\bar{x}_2 - \bar{x}_{a-1}$	$\cdots$	$\bar{x}_2 - \bar{x}_3$	
$a-2$	$\bar{x}_{a-2} - \bar{x}_a$	$\bar{x}_{a-2} - \bar{x}_{a-1}$			
$a-1$	$\bar{x}_{a-1} - \bar{x}_a$				

②查出不同秩次距k 的r_α 。

$r_a = r_a$（k，df）的值可以从附表“多重比较中的Duncan表”中查出：表的最左边一列是误差自由度df ＝a（n－1），最上一列为k值，表体为r_a (k, df)。表中的 k 值是相比较的两个平均数之间所包含的平均数的个数。如两个要比较的平均数相邻时k＝2，两个要比较的平均数中间隔一个平均数时k＝3，依此类椎。因为平均数共有a个，所以需查出a—1个r_a .

③计算不同秩次距 k的临界值R_k

$$R_k = r_\alpha s_{\bar{x}}\ ,\ \ k = 2,3,\cdots,a$$

其中 $s_{\bar{x}} = \sqrt{\dfrac{MS_e}{n}}$

α =0.05 和 α =0.01 水平下的最小显著极差为：

$$R_{0.05,k} = r_{0.05(df_e,k)} S_{\bar{x}}$$

$$R_{0.01,k} = r_{0.01(df_e,k)} S_{\bar{x}}$$

④每两个平均数的差与相应的临界值比较

显著的打上一个星花“*”，极显著的打上两个星花“**”。

先从表的第一行最左边的一个差x_1-x_a开始比较。若$x_1-x_a>R_a$，则x_1与x_a的差异显著；否则差异不显著，然后比较下一个。若$x_1-x_{a-1}>R_{a-1}$，则x_1与x_{a-1}差异显著，否则差异不显著，…。第一行比较完之后用同样的方法比较第二行。先从第二行最左边的一个差x_2-x_a开始，在x_2到x_a这个范围内共包含a－1个平均数，因此x_2-x_a应与R_{a-1}比较，若$x_2-x_a>R_{a-1}$，则差异显著，否则不显著，…。第二行比较完再比较第三行，第四行，…。直到所有平均数的差均与其相应的R_k比较完为止。对于显著的标上“*”，极显著的标上“* *”。

例　题

药剂	D	B	A	C
平均数	24	23	19	18
顺序号	1	2	3	4

列出差数半矩阵表:

	D	B	A
C	$\bar{x}_1-\bar{x}_4=6^*$	$\bar{x}_1-\bar{x}_3=5$	$\bar{x}_1-\bar{x}_2=1$
A	$\bar{x}_2-\bar{x}_4=5$	$\bar{x}_2-\bar{x}_3=4$	
B	$\bar{x}_3-\bar{x}_4=1$		

已知算得 **MSe=9.83, n=4,** $S_{\bar{x}} = \sqrt{\frac{MS_e}{n}} = \sqrt{\frac{9.83}{4}} = 1.57(cm)$

对于k = 2, 3, 4 分别查出$r_{0.05}$, $R_{a,k} = r_{a(k,df_e)} S_{\bar{x}}$

算出$R_{0.05}$

df_e	k	$r_{0.05}$	$R_{0.05}$
12	**2**	**3.08**	**4.84**
	3	**3.23**	**5.07**
	4	**3.33**	**5.23**

结论：上述实验四个处理的苗高，只有处理D和C差异显著，其它处理间差异皆不显著。

例8.2　调查5个不同小麦品系株高，结果见下表：

	品系				
	I	II	III	IV	V
1	64.6	64.5	67.8	71.8	69.2
2	65.3	65.3	66.3	72.1	68.2
3	64.8	64.6	67.1	70.0	69.8
4	66.0	63.7	66.8	69.1	68.3
5	65.8	63.9	68.5	71.0	67.5
和	326.5	322.0	336.5	354.0	343.0
平均数	65.3	64.4	67.3	70.8	68.6

	品		系			
	I	II	III	IV	V	
1	−0.4	−0.5	2.8	6.8	4.2	
2	0.3	0.3	1.3	7.1	3.2	
3	−0.2	−0.4	2.1	5.0	4.8	
4	1.0	−1.3	1.8	4.1	3.3	
5	0.8	−1.1	3.5	6.0	2.5	总和
$x_{i.}$	1.5	−3.0	11.5	29.0	18.0	57.0
$x_{i.}^2$	2.25	9.00	132.25	841.00	324.00	1308.50
Σx_{ij}^2	1.93	3.40	29.43	174.46	68.06	277.28

$$C = \frac{x_{..}^2}{na} = \frac{57^2}{(5)(5)} = 129.96$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - C = 277.28 - 129.96 = 147.32$$

$$SS_A = \frac{1}{n}\sum_{i=1}^{a} x_{i.}^2 - C = \frac{1308.50}{5} - 129.96 = 131.74$$

$$SS_e = SS_T - SS_A = 147.32 - 131.74 = 15.58$$

将以上结果列成方差分析表：

变差来源	平方和	自由度	均方	F
品系间	131.74	4	32.72	41.95**
误 差	15.58	20	0.78	
总 和	147.32	24		

** $\alpha=0.01$

$F_{4,\ 20,\ 0.05}=2.87$，$F_{4,\ 20,\ 0.01}=4.43$。$F>F_{0.01}$。$P<0.01$

因此，上述5个不同小麦品系株高差异极显著。

习惯上以“ * ”表示在 $\alpha=0.05$水平上差异显著，

以“ ** ”表示在 $\alpha=0.01$水平上差异显著。

对5个小麦品系株高平均数做Duncan多重比较.

首先将平均数按从高到低顺序排列好。

品系号	IV	V	III	I	II
平均数	70.8	68.6	67.3	65.3	64.4
顺序号	1	2	3	4	5

再列成下表

	5	4	3	2
1	$x_1-x_5=6.4^{**}$	$x_1-x_4=5.5^{**}$	$x_1-x_3=3.5^{**}$	$x_1-x_2=2.2^{**}$
2	$x_2-x_5=4.2^{**}$	$x_2-x_4=3.3^{**}$	$x_2-x_3=1.3^{*}$	
3	$x_3-x_5=2.9^{**}$	$x_3-x_4=2.0^{**}$		
4	$x_4-x_5=0.9$			

将临界值也列成表。根据MS_e＝0.78，n＝5，

$$S_{\bar{x}} = \sqrt{\frac{0.78}{5}} = 0.395$$

对于k＝2，3，4，5，df＝a（n—1）＝20，

分别求出$r_{0.05}$和$r_{0.01}$的R_k值并列成表：

df	k	$r_{0.05}$	Rk	$r_{0.01}$	Rk
20	2	2.95	1.165	4.02	1.588
	3	3.10	1.225	4.22	1.667
	4	3.18	1.256	4.33	1.710
	5	3.25	1.284	4.40	1.738

再以 $\alpha=0.05$和 $\alpha=0.01$水平上的临界值与下表中的平均数的差做比较，差异显著的打上“*”，差异极显著的打上“**”。

	5	4	3	2
1	$x_1-x_5=6.4^{**}$	$x_1-x_4=5.5^{**}$	$x_1-x_3=3.5^{**}$	$x_1-x_2=2.2^{**}$
2	$x_2-x_5=4.2^{**}$	$x_2-x_4=3.3^{**}$	$x_2-x_3=1.3^{*}$	
3	$x_3-x_5=2.9^{**}$	$x_3-x_4=2.0^{**}$		
4	$x_4-x_5=0.9$			

df	k	$r_{0.05}$	Rk	$r_{0.01}$	Rk
20	2	2.95	1.165	4.02	1.588
	3	3.10	1.225	4.22	1.667
	4	3.18	1.256	4.33	1.710
	5	3.25	1.284	4.40	1.738

8.3 方差分析应具备的条件

1. 可加性

每个处理效应与误差效应是可加的，$x_{ij} = \mu + a_i + \varepsilon_{ij}$。由于有这一假定，不同的效应才能被分解，才能最终判断处理效应是否比误差更显著。

2. 正态性

实验误差应当是服从正态分布N（0，σ^2）的独立随机变量。

3. 方差齐性

说各处理的总体方差必须一致。一般来说，在方差分析之前，应先做方差齐性检验。

8.4　数据的变换　（data transformation）

在方差分析应具备的条件不能满足时做方差分析，往往导致错误的结论。这时，必须采取适当的变换，使变换后的数值能满足上述条件，用变换后的数值做方差分析。

数据变换的目的

主要是满足方差分析的要求，同时对正态性以及可加性的要求都可得到较好的满足。

常用的变换方法：

- 平方根变换（square root transforma-tion）
- 反正弦变换（inverse sine transformation）
- 对数变换（logarithmic transfor-mation）

8.4.1平方根变换

- 属于Poisson分布的数据，它们的平均数与方差等值，常常需采取这种变换方法。
- 例如，单位面积内菌落数、一定区域内某种昆虫或某种植物数、放射性物质在单位时间内放射的次数、单位数量的种子中混有的草籽数等等。都属于这种情况。
- 对于这类数据，应该将每个 x_{ij} 取其平方根 $\sqrt{x_{ij}}$，然后再计算。
- 当数值很小时，如有几个数小于10时，为了矫正，可以使用 $\sqrt{x_{ij}+1}$ 变换。

例 施用5种不同的除草剂后，稻田中混杂的杂草株数见下表，

表8－1　不同除草剂除草后的杂草株数

		处理				
		A	B	C	D	E
地块	一	438	538	77	17	18
	二	442	422	61	31	26
	三	319	377	157	87	77
	四	380	315	52	16	20

将以上数据做 $\sqrt{x_{ij}}$ 变换得下表，

表8－2　表8－1中数据经平方根变换后的结果

		处理				
		A	B	C	D	E
地块	一	20.9	23.2	8.8	4.1	4.2
	二	21.0	20.5	7.8	5.6	5.1
	三	17.9	19.4	12.5	9.3	8.8
	四	19.5	17.7	7.2	4.0	4.5

8.4.2 反正弦变换

- 属于二项分布的数据，常常需采取这种变换方法。
- 生物学中常见的一些百分数资料，如某些多发病的患病率、某些病原体的感染率、白细胞的分类计数（%）、淋巴细胞的转换率、玫瑰花的簇生形成率等等数据，都需要进行反正弦变换。
- 反正弦变换的方法是，取每个观察值的平方根的反正弦的值，使它变成一个角度。然后，用这个角度做方差分析。

- 二项分布的百分比数据变化范围很大时，要做反正弦变换。$\theta = \sin^{-1}\sqrt{p}$ 使其变换成一个角度。
- 当百分数的范围在0～20％或80～100％时，可以用平方根变换， 80～100％时在变换之前，应先用100减去各百分数。
- 若百分数的范围是在30～70％，则不一定非要做变换。

不同贮藏时间有活力花粉百分数（%）

处理	2小时	4小时	6小时	对照
1	85	80	55	95
2	80	70	65	90
3	91	75	49	91
4	82	65	52	85
5	90	60	60	80
6	78	50	50	93

不同贮藏时间有活力花粉百分数的反正弦值（%）

处理	2小时	4小时	6小时	对照
1	67.2	63.4	47.9	77.1
2	63.4	56.8	53.7	71.6
3	72.5	60.0	44.4	72.5
4	64.9	53.7	46.1	67.2
5	71.6	50.8	50.7	63.4
6	62.0	53.7	45.0	74.7
平均数	66.9	56.4	48.0	71.1

8.4.3 对数变换

如果已知资料中的效应成比例而不是可加的，或者标准差（或极差）与平均数大体成比例时，经对数变换后，则方差具备齐性。对数变换适用于大范围的正整数，对于一些小的数值，例如小于10时，每一c_{ij}都加上1再变换，即取lg（x_{ij}＋1）。

以下数据是在5次3个相继的夜里捕获昆虫的几何平均数，实验采用3种捕灯蛾诱捕。

捕获昆虫统计及捕获数的对数值

时期	捕灯蛾			对数值		
	I	II	III	I	II	III
1	19.1	50.1	123.0	1.28	1.70	2.09
2	23.4	166.0	407.4	1.37	2.22	2.61
3	39.5	233.9	398.1	1.47	2.45	2.60
4	23.4	58.9	229.1	1.37	1.77	2.36
5	16.6	64.6	251.2	1.22	1.81	2.40
平均数	22.4	112.7	281.3	1.34	1.97	2.41
极差	12.9	173.8	284.4	0.25	0.85	0.52

作　　业

P118　　8.5　8.6　　8.7