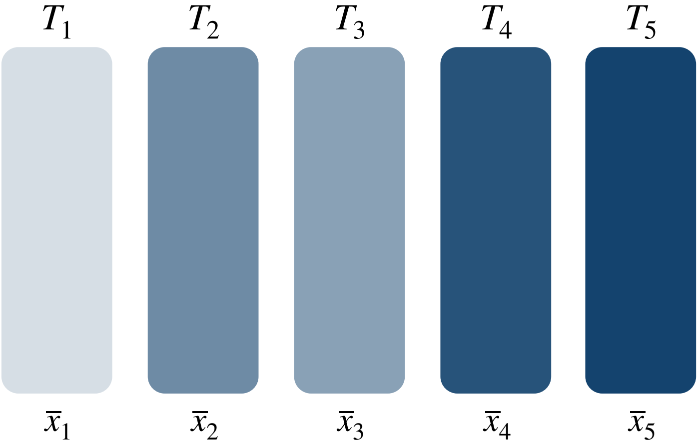
生物统计学

第九章 方差分析

云南大学 生命科学学院

會澤百家 至公天下


T_1
T_2
T_3
T_4
T_5
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
$\bar{x}_4$
$\bar{x}_5$

多次使用t 检验，理论上存在以下三个问题:

1. 检验的过程会变得烦琐，n 个样本就需要执行 C_n^2 次检验。
2. 误差只能在每次比较的两个样本范围内进行估计，精确性相对较低，进而造成检验的灵敏性不达预期。
3. 多次使用两两 t 检验，犯第一类错误的概率会增加，推断的可靠性会降低。

多次使用t 检验，理论上存在以下三个问题：

1. 检验的过程会变得烦琐，n 个样本就需要执行 C_n^2 次检验。
2. 误差只能在每次比较的两个样本范围内进行估计，精确性相对较低，进而造成检验的灵敏性不达预期。
3. 多次使用两两 t 检验，犯第一类错误的概率会增加，推断的可靠性会降低。

为了解决多个样本的平均数比较问题，Fisher 于 1918 年提出了方差分析(analysis of variance, ANOVA) 的方法。

1. 方差分析的基本原理
   - 方差分析的数学模型
   - 平方和与自由度分解
   - 不同效应方差的数学期望
   - 显著性检验——F 检验
   - 功效分析
   - 多重比较
   - 方差分析的基本条件
2. 单因素方差分析
3. 双因素方差分析

任何一个试验产生的观测值，造成它们取值差异的主要原因有二:

- 试验因素带来的处理效应(treatment effect, TE)
- 试验过程中偶然性因素带来的误差效应(error effect, EE)

任何一个试验产生的观测值，造成它们取值差异的主要原因有二:

- 试验因素带来的处理效应(treatment effect, TE)
- 试验过程中偶然性因素带来的误差效应(error effect, EE)

$$x_i - \overline{x} = \mathrm{TE} + \mathrm{EE} \tag{9.1}$$

表 **9.1** 单因素 k 水平（每组 n 个观测值）的数据资料表

处理	A_1	A_2	$\cdots$	A_i	$\cdots$	A_k	
	x_{11}	x_{21}	$\cdots$	x_{i1}	$\cdots$	x_{k1}	
	x_{12}	x_{22}	$\cdots$	x_{i2}	$\cdots$	x_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	x_{1j}	x_{2j}	$\cdots$	x_{ij}	$\cdots$	x_{kj}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	x_{1n}	x_{2n}	$\cdots$	x_{in}	$\cdots$	x_{kn}	
总和	$\sum_{j=1}^{n} x_{1j}$	$\sum_{j=1}^{n} x_{2j}$	$\cdots$	$\sum_{j=1}^{n} x_{ij}$	$\cdots$	$\sum_{j=1}^{n} x_{kj}$	$\sum\sum x_{ij}$
平均数	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\cdots$	$\bar{x}_{i\cdot}$	$\cdots$	$\bar{x}_{k\cdot}$	$\bar{x}_{\cdot\cdot}$

任一观测值可以表示为：

$$x_{ij} = \mu_i + \varepsilon_{ij} \tag{9.2}$$

k 个处理对应 k 个小总体，每个小总体都有其平均数 μ_i。假定误差服从正态分布，即 $\varepsilon_{ij} \sim N(0, \sigma_e^2)$。

任一观测值可以表示为:

$$x_{ij} = \mu_i + \varepsilon_{ij} \tag{9.2}$$

k 个处理对应 k 个小总体，每个小总体都有其平均数 μ_i。假定误差服从正态分布，即 $\varepsilon_{ij} \sim N(0, \sigma_e^2)$。

μ_i 与 μ 之间的偏差，也就是 A 因素第 i 个水平的处理效应，记作 α_i，即 $\alpha_i = \mu_i - \mu$。

任一观测值可以表示为:

$$x_{ij} = \mu_i + \varepsilon_{ij} \tag{9.2}$$

k 个处理对应 k 个小总体，每个小总体都有其平均数 μ_i。假定误差服从正态分布，即 $\varepsilon_{ij} \sim N(0, \sigma_e^2)$。

μ_i 与 μ 之间的偏差，也就是 A 因素第 i 个水平的处理效应，记作 α_i，即 $\alpha_i = \mu_i - \mu$。

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

任一观测值可以表示为：

$$x_{ij} = \mu_i + \varepsilon_{ij} \tag{9.2}$$

k 个处理对应 k 个小总体，每个小总体都有其平均数 μ_i。假定误差服从正态分布，即 $\varepsilon_{ij} \sim N(0, \sigma_e^2)$。

μ_i 与 μ 之间的偏差，也就是 A 因素第 i 个水平的处理效应，记作 α_i，即 $\alpha_i = \mu_i - \mu$。

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

单因素试验资料的数学模型（线性模型，linear model）。

根据处理效应的性质不同，观测值的线性模型可以分为以下三种类型：

根据处理效应的性质不同，观测值的线性模型可以分为以下三种类型：

- 固定效应模型 (fixed effect model) 是指模型所涉及的处理效应 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 都是固定的常量。

根据处理效应的性质不同，观测值的线性模型可以分为以下三种类型：

- 固定效应模型 (fixed effect model) 是指模型所涉及的处理效应 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 都是固定的常量。
- 随机效应模型 (random effect model) 是指模型中所涉及的处理效应 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 都是相互独立的随机变量，且服从正态分布 $N(0, \sigma_\alpha^2)$。

根据处理效应的性质不同，观测值的线性模型可以分为以下三种类型：

- 固定效应模型 (fixed effect model) 是指模型所涉及的处理效应 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 都是固定的常量。
- 随机效应模型 (random effect model) 是指模型中所涉及的处理效应 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 都是相互独立的随机变量，且服从正态分布 $N(0, \sigma_\alpha^2)$。
- 在双因素和多因素试验中，如果既有固定效应因素，又有随机效应因素，那么相应的线性模型被称为混合效应模型 (mixed effect model)。

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

样本的线性模型：

$$x_{ij} = \overline{x} + a_i + e_{ij} \tag{9.5}$$

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

样本的线性模型：

$$x_{ij} = \overline{x} + a_i + e_{ij} \tag{9.5}$$

$$x_{ij} - \overline{x} = a_i + e_{ij} \tag{9.6}$$

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{9.4}$$

样本的线性模型：

$$x_{ij} = \overline{x} + a_i + e_{ij} \tag{9.5}$$

$$x_{ij} - \overline{x} = a_i + e_{ij} \tag{9.6}$$

等式左边方差化得 kn 个观测值的总样本方差，

$$s^2 = \frac{\sum_{i=1}^{nk}(x_i - \overline{x})^2}{kn - 1} \tag{9.7}$$

1. 方差分析的基本原理
    - 方差分析的数学模型
    - 平方和与自由度分解
    - 不同效应方差的数学期望
    - 显著性检验——F 检验
    - 功效分析
    - 多重比较
    - 方差分析的基本条件
2. 单因素方差分析
3. 双因素方差分析

将总样本方差式 9.7 分子部分中，样本观测值的下标由一维变为两维，则有

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x})^2$$

称为总离均差平方和(total sum of squares)。

$$
\begin{aligned}
SS &= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot}+\overline{x}_{i\cdot}-\overline{x})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}\Big[(x_{ij}-\overline{x}_{i\cdot})+(\overline{x}_{i\cdot}-\overline{x})\Big]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}\Big[(x_{ij}-\overline{x}_{i\cdot})^2+(\overline{x}_{i\cdot}-\overline{x})^2+2(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x})\Big] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{x}_{i\cdot}-\overline{x})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x})
\end{aligned}
\tag{9.8}
$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x})=\sum_{i=1}^{k}(\overline{x}_{i\cdot}-\overline{x})\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot}) \tag{9.9}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x}) = \sum_{i=1}^{k}(\overline{x}_{i\cdot}-\overline{x})\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot}) \tag{9.9}$$

$$\begin{aligned} SS &= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{x}_{i\cdot}-\overline{x})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\overline{x}_{i\cdot})^2 + \sum_{i=1}^{k} n(\overline{x}_{i\cdot}-\overline{x})^2 \end{aligned} \tag{9.10}$$

$$SS = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x}_{i\cdot})^2 + \sum_{i=1}^{k} n(\overline{x}_{i\cdot} - \overline{x})^2 \tag{9.10}$$

总离均差平方和 SS 被成功分解为以下两部分：

- $\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x}_{i\cdot})^2$，称为组内离均差平方和 (sum of squares within groups)，记作 SS_e。
- $\sum_{i=1}^{k} n(\overline{x}_{i\cdot} - \overline{x})^2$，称为组间偏差平方和 (sum of squares between groups)，记作 SS_t。

- 观测值总样本方差的自由度为 $df = kn - 1$。

- 观测值总样本方差的自由度为 $df = kn - 1$。
- 误差方差的总自由度为 $df_e = k(n-1)$。

- 观测值总样本方差的自由度为 $df = kn - 1$。
- 误差方差的总自由度为 $df_e = k(n-1)$。
- 处理方差的自由度为 $df_t = k - 1$。

- 观测值总样本方差的自由度为 $df = kn - 1$。
- 误差方差的总自由度为 $df_e = k(n-1)$。
- 处理方差的自由度为 $df_t = k - 1$。

$$df = df_t + df_e \tag{9.19}$$

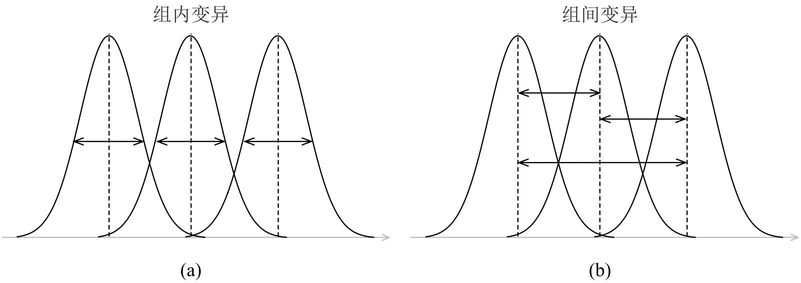


图 **9.1** 多组样本的变异类型

表 **9.2** 5 种抗生素的血浆蛋白结合率数据

重复观测值	A	B	C	D	E
obs1	81.2	75.1	82.3	79.8	75.6
obs2	79.8	77.3	83.7	80.5	74.9
obs3	81.6	76.8	81.9	78.9	76.3
obs4	81.1	75.5	82.5	81.2	75.8
obs5	81.9	78.1	83.2	79.4	74.0

表 **9.2** 5 种抗生素的血浆蛋白结合率数据

重复观测值	A	B	C	D	E
obs1	81.2	75.1	82.3	79.8	75.6
obs2	79.8	77.3	83.7	80.5	74.9
obs3	81.6	76.8	81.9	78.9	76.3
obs4	81.1	75.5	82.5	81.2	75.8
obs5	81.9	78.1	83.2	79.4	74.0

$$s^2 = 8.778233, \quad s_t^2 = 48.3224, \quad s_e^2 = 0.8694$$

当误差同分布的条件满足时，误差效应的样本方差 s_e^2 的数学期望等于误差的总体方差 σ_e^2。即

$$E(s_e^2) = \sigma_e^2$$

对于处理效应的样本方差 s_t^2 的数学期望，有

$$E(s_t^2) = \frac{n}{k-1} E\left(\sum_{i=1}^{k} (\alpha_i - \overline{\alpha})^2 \right) + \sigma_e^2 \tag{9.30}$$

对于处理效应的样本方差 s_t^2 的数学期望，有

$$E(s_t^2) = \frac{n}{k-1} E\left(\sum_{i=1}^{k} (\alpha_i - \overline{\alpha})^2 \right) + \sigma_e^2 \tag{9.30}$$

如果试验因素符合固定效应模型，那么 $\overline{\alpha} = 0$，而各个水平的效应值 α_i 都为常量

$$E(s_t^2) = \frac{n}{k-1} \sum_{i=1}^{k} \alpha_i^2 + \sigma_e^2 \tag{9.31}$$

对于处理效应的样本方差 s_t^2 的数学期望，有

$$E(s_t^2) = \frac{n}{k-1} E\left(\sum_{i=1}^{k} (\alpha_i - \overline{\alpha})^2\right) + \sigma_e^2 \tag{9.30}$$

对于处理效应的样本方差 s_t^2 的数学期望，有

$$E(s_t^2) = \frac{n}{k-1} E\left(\sum_{i=1}^{k} (\alpha_i - \overline{\alpha})^2 \right) + \sigma_e^2 \tag{9.30}$$

如果试验因素符合随机效应模型，那么各个水平的效应值 α_i 不再是常量，而是服从 $N(0, \sigma_\alpha^2)$ 的随机变量。

$$E(s_t^2) = n\sigma_\alpha^2 + \sigma_e^2 \tag{9.33}$$

F 检验应用到方差分析的场景中，则有

$$F = \frac{s_t^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_t^2} \tag{1}$$

此时 F 分布的自由度分别为 $k-1$ 和 $k(n-1)$。

F 检验应用到方差分析的场景中，则有

$$F = \frac{s_t^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_t^2} \tag{1}$$

此时 F 分布的自由度分别为 $k-1$ 和 $k(n-1)$。

在检验时，假定零假设 $\sigma_t^2 = \sigma_e^2$ 成立。

F 检验应用到方差分析的场景中，则有

$$F = \frac{s_t^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_t^2} \tag{1}$$

此时 F 分布的自由度分别为 $k-1$ 和 $k(n-1)$。

在检验时，假定零假设 $\sigma_t^2 = \sigma_e^2$ 成立。

$$\sigma_t^2 = E(s_t^2) = n\eta_\alpha^2 + \sigma_e^2$$

F 检验应用到方差分析的场景中，则有

$$F = \frac{s_t^2}{s_e^2} \times \frac{\sigma_e^2}{\sigma_t^2} \tag{1}$$

此时 F 分布的自由度分别为 $k-1$ 和 $k(n-1)$。

在检验时，假定零假设 $\sigma_t^2 = \sigma_e^2$ 成立。

$$\sigma_t^2 = E(s_t^2) = n\eta_\alpha^2 + \sigma_e^2$$

$$\eta_\alpha^2 = \frac{1}{k-1}\sum_{i=1}^{k}\alpha_i^2$$

根据假设检验的一般操作流程，做 F 检验如下。

1. 按照单尾检验问题的提法，设定零假设 $H_0: \sigma_t^2 = \sigma_e^2$，备择假设 $H_1: \sigma_t^2 > \sigma_e^2$。
2. 选取显著性水平 $\alpha = 0.05$。
3. 计算检验统计量和 P 值。
   - 检验统计量 $F_c = \frac{s_t^2}{s_e^2} = \frac{48.3224}{0.8694} \approx 55.581$;
   - 单尾 F 检验的 P 值：$P(F \geq F_c \approx 55.581 | H_0) = 1.49222e - 10$。
4. 做出统计推断。

   不同抗生素的血浆蛋白结合率具有统计学意义上的显著差异。

表 **9.3** 5 种抗生素的血浆蛋白结合率的方差分析表

变异来源	平方和 SS	自由度 df	方差 s^2	F 值	P 值
抗生素间 (处理间)	193.2896	4	48.3224	55.58132	1.49222e-10
重复间 (处理内)	17.388	20	0.8694		
总变异	210.6776	24			

1. 方差分析的基本原理
   - 方差分析的数学模型
   - 平方和与自由度分解
   - 不同效应方差的数学期望
   - 显著性检验——F 检验
   - 功效分析
   - 多重比较
   - 方差分析的基本条件
2. 单因素方差分析
3. 双因素方差分析

对于固定效应，F 检验得出有显著差异的结论，仅仅意味着 $\eta_\alpha^2 \neq 0$。

对于固定效应，F 检验得出有显著差异的结论，仅仅意味着 $\eta_\alpha^2 \neq 0$。

常用的多重比较法（又称多重比较检验）包括:

- 最小显著差数法 (the method of least significant difference, LSD)
- 最小显著极差法 (the method of least significant range，LSR)

比较两个样本平均数时，标准化统计量为 $t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_{\bar{x}_1-\bar{x}_2}}$。

当 H_0 成立时 $\mu_1 = \mu_2$，有检验统计量 $t_c = \frac{\bar{x}_1-\bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}}$。

比较两个样本平均数时，标准化统计量为 $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_{\overline{x}_1 - \overline{x}_2}}$。

当 H_0 成立时 $\mu_1 = \mu_2$，有检验统计量 $t_c = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}}$。

显著性水平 $\alpha = 0.05$ 的双尾检验需要 $|t_c| > t_{0.025,df}$。

比较两个样本平均数时，标准化统计量为 $t = \frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{s_{\overline{x}_1-\overline{x}_2}}$。

当 H_0 成立时 $\mu_1=\mu_2$，有检验统计量 $t_c = \frac{\overline{x}_1-\overline{x}_2}{s_{\overline{x}_1-\overline{x}_2}}$。

显著性水平 $\alpha = 0.05$ 的双尾检验需要 $|t_c| > t_{0.025,df}$。

$$|\overline{x}_1 - \overline{x}_2| > t_{0.025,df} \times s_{\overline{x}_1-\overline{x}_2}$$

比较两个样本平均数时，标准化统计量为 $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_{\overline{x}_1 - \overline{x}_2}}$。
当 H_0 成立时 $\mu_1 = \mu_2$，有检验统计量 $t_c = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}}$。

显著性水平 $\alpha = 0.05$ 的双尾检验需要 $|t_c| > t_{0.025,df}$。

$$|\overline{x}_1 - \overline{x}_2| > t_{0.025,df} \times s_{\overline{x}_1 - \overline{x}_2}$$

平均数的差异达到显著水平的最小差数为

$$\mathrm{LSD}_{0.05} = t_{0.025,df} \times s_{\overline{x}_1 - \overline{x}_2} \tag{9.38}$$

$$
\begin{aligned}
\mathrm{LSD}_{0.05} &= t_{0.025,20} \times s_{\overline{x}_1-\overline{x}_2} = 2.085963 \times 0.5897118 = 1.230117 \\
\mathrm{LSD}_{0.01} &= t_{0.005,20} \times s_{\overline{x}_1-\overline{x}_2} = 2.84534 \times 0.5897118 = 1.67793
\end{aligned} \tag{9.39}
$$

$$\begin{aligned} \mathrm{LSD}_{0.05} &= t_{0.025,20} \times s_{\overline{x}_1-\overline{x}_2} = 2.085963 \times 0.5897118 = 1.230117 \\ \mathrm{LSD}_{0.01} &= t_{0.005,20} \times s_{\overline{x}_1-\overline{x}_2} = 2.84534 \times 0.5897118 = 1.67793 \end{aligned} \tag{9.39}$$

9.4 5 种抗生素的血浆蛋白结合率的多重比较（LSD 法）

抗生素	平均数	平均数差值	$\alpha = 0.05$	$\alpha = 0.01$
C	82.72		a	A
A	81.12	1.60	b	AB
D	79.96	1.16	b	B
B	76.56	3.40	c	C
E	75.32	1.24	d	C

LSD 法虽然摆脱了两两 t 检验在操作上复杂性，但是犯第一类错误可能性高的问题并没有解决。

LSD 法虽然摆脱了两两 t 检验在操作上复杂性，但是犯第一类错误可能性高的问题并没有解决。

显著性水平膨胀 (the inflation of the significance level)

LSD 法虽然摆脱了两两 t 检验在操作上复杂性，但是犯第一类错误可能性高的问题并没有解决。

显著性水平膨胀 (the inflation of the significance level)

连续使用 C 次 t 检验，犯第一类错误的概率 $\mathcal{A} = 1 - (1 - \alpha)^C$，其中 α 是一次检验中的显著性水平。

LSD 法虽然摆脱了两两 t 检验在操作上复杂性，但是犯第一类错误可能性高的问题并没有解决。

显著性水平膨胀 (the inflation of the significance level)

连续使用 C 次 t 检验，犯第一类错误的概率 $\mathcal{A} = 1 - (1-\alpha)^C$，其中 α 是一次检验中的显著性水平。

$$\alpha = 1 - (1 - \mathcal{A})^{\frac{1}{C}}$$

LSD 法虽然摆脱了两两 t 检验在操作上复杂性，但是犯第一类错误可能性高的问题并没有解决。

显著性水平膨胀 (the inflation of the significance level)

连续使用 C 次 t 检验，犯第一类错误的概率 $\mathcal{A} = 1 - (1-\alpha)^C$，其中 α 是一次检验中的显著性水平。

$$\alpha = 1 - (1 - \mathcal{A})^{\frac{1}{C}}$$

$$\mathcal{A} = 0.05, \quad \alpha \approx 0.00512$$

Sidak 法中涉及到分数幂，在没有计算机辅助的情况下难以计算，所以瑞典生物统计学家 Sture Holm 在 1979 年提出了一种基于 Bonferroni 不等式（称为 Bonferroni 法）的近似计算公式

$$\alpha \approx \frac{\mathcal{A}}{C}$$

连续 10 次的 t 检验，用 Bonferroni 法矫正后的 $\alpha \approx 0.005$。

Sidak 法中涉及到分数幂，在没有计算机辅助的情况下难以计算，所以瑞典生物统计学家 Sture Holm 在 1979 年提出了一种基于 Bonferroni 不等式（称为 Bonferroni 法）的近似计算公式

$$\alpha \approx \frac{\mathcal{A}}{C}$$

连续 10 次的 t 检验，用 Bonferroni 法矫正后的 $\alpha \approx 0.005$。

Bonferroni 法是迄今为止使用频次最高的显著性水平矫正方法之一。

最小显著极差法 (the method of least significant range，LSR) 是在一定的显著性水平 α 上，通过确定达到显著差异的最小极差 LSR；然后与各组平均数之差进行比较的方法。

- Tukey's HSD (honestly significant difference) 法
- SNK (Student-Newman-Keuls) 法
- 新复极差检验法（又称 Duncan 法）

Tukey's HSD 法需要计算出每两组之间的平均数差数和标准误，然后根据学生化极差分布 (studentized range distribution) 来确定哪些组之间存在显著差异。

$$q_{k,df_{\overline{x}}} = \frac{\overline{x}_{\max} - \overline{x}_{\min}}{s_{\overline{x}}} = \frac{\overline{x}_{\max} - \overline{x}_{\min}}{s_e / \sqrt{n}} \tag{9.40}$$

与 LSD 法类似，当 $q_{k,df_{\overline{x}}} > q_{\alpha,k,df_{\overline{x}}}$ 时，$\overline{x}_{\max} - \overline{x}_{\min}$ 在 α 水平上存在显著差异。所以 α 水平上的最小显著极差为

$$LSR_{\alpha} = q_{\alpha,k,df_{\overline{x}}} \times \sqrt{\frac{s_e^2}{n}} \tag{9.41}$$

表 **9.9** 5 种抗生素的血浆蛋白结合率的多重比较（$\alpha = 0.05$）

药物	平均数	LSD	LSD+bonferroni	Tukey	SNK	Duncan
C	82.72	a	a	a	a	a
A	81.12	b	ab	ab	b	b
D	79.96	b	b	b	b	b
B	76.56	c	c	c	c	c
E	75.96	d	c	c	d	d

1. 方差分析的基本原理
   - 方差分析的数学模型
   - 平方和与自由度分解
   - 不同效应方差的数学期望
   - 显著性检验——F 检验
   - 功效分析
   - 多重比较
   - 方差分析的基本条件
2. 单因素方差分析
3. 双因素方差分析

1. 效应的可加性
2. 误差分布的正态性
3. 误差方差的同质性

1. 方差分析的基本原理

2. 单因素方差分析
   - 组内观测值数量相等
   - 组内观测值数量不相等
   - F 检验与 t 检验的关系

3. 双因素方差分析

例 (9.1)

5 头公牛分别与一组母牛交配，记录五组母牛所产的牛犊中 8 头雄性牛犊（来自不同母畜）的出生体重 (见 `calvesWeight` 数据集)。试用方差分析检验不同的父系对雄性牛犊的出生体重是否有显著的影响。

方差齐性的 Bartlett 检验:

```
> data(calvesWeight)
> bartlett.test(weight ~ sire, data = calvesWeight)
    Bartlett test of homogeneity of variances

data:  weight by sire
Bartlett's K-squared = 6.8291, df = 4, p-value = 0.1452
```

误差正态性检验:

```
> aov.fit <- aov(weight ~ sire, data = calvesWeight)
> resid <- residuals(aov.fit)
> shapiro.test(resid)
    Shapiro-Wilk normality test

data:  residuals(aov.fit)
W = 0.97839, p-value = 0.6301
```

方差分析:

```
> fit.lm <- lm(weight ~ sire, data = calvesWeight)
> anova.table <- anova(fit.lm); anova.table
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
sire       4  5591.2 1397.79  3.0138 0.03087 *
Residuals 35 16232.8  463.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

1. 方差分析的基本原理

2. 单因素方差分析
   - 组内观测值数量相等
   - 组内观测值数量不相等
   - F 检验与 t 检验的关系

3. 双因素方差分析

1. 方差分析的基本原理

2. 单因素方差分析
   - 组内观测值数量相等
   - 组内观测值数量不相等
   - F 检验与 t 检验的关系

3. 双因素方差分析

例 (9.3)

对例题 8.8 中的睡眠数据（`sleep` 数据集），试作方差分析。

例 (9.3)

对例题 8.8 中的睡眠数据（`sleep` 数据集），试作方差分析。

例 (8.8)

为比较两种安眠药 (编号 1 和 2) 的效果，分别随机选择 10 名失眠者试药，每名受试者服用一次后记录睡眠的延长时间 (R 自带数据包 `datasets` 的 `sleep` 数据集)。在显著水平 $\alpha = 0.05$ 下，试问两种安眠药的疗效有无显著差异?

方差分析:

```
> data(sleep)
> anova(lm(formula = extra ~ group, data = sleep))
Analysis of Variance Table

Response: extra
          Df Sum Sq Mean Sq F value  Pr(>F)
group      1 12.482 12.4820  3.4626 0.07919 .
Residuals 18 64.886  3.6048
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

t 检验:

```
> t.test(extra ~ group, data = sleep, var.equal = T)
    Two Sample t-test


data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means between
 group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

- 无重复观测值的双因素方差分析
- 有重复观测值的双因素方差分析
- 重复观测值不等的双因素方差分析

双因素方差分析有数学模型：

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{9.46}$$

双因素方差分析有数学模型：

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{9.46}$$

当因素的主效和因素间的互作效应符合固定模型时，

$$F_A = \frac{s_A^2}{s_e^2}, \quad F_B = \frac{s_B^2}{s_e^2}, \quad F_{AB} = \frac{s_{AB}^2}{s_e^2} \tag{9.47}$$

双因素方差分析有数学模型:

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{9.46}$$

双因素方差分析有数学模型：

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{9.46}$$

当因素的主效 A 和 B 同为随机模型时，

$$F_A = \frac{s_A^2}{s_{AB}^2}, \quad F_B = \frac{s_B^2}{s_{AB}^2}, \quad F_{AB} = \frac{s_{AB}^2}{s_e^2} \tag{9.48}$$

双因素方差分析有数学模型：

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{9.46}$$

当因素的主效 A 和 B 同为随机模型时，

$$F_A = \frac{s_A^2}{s_{AB}^2}, \quad F_B = \frac{s_B^2}{s_{AB}^2}, \quad F_{AB} = \frac{s_{AB}^2}{s_e^2} \tag{9.48}$$

当 A 为固定因素，B 为随机因素时，

$$F_A = \frac{s_A^2}{s_{AB}^2}, \quad F_B = \frac{s_B^2}{s_e^2}, \quad F_{AB} = \frac{s_{AB}^2}{s_e^2} \tag{9.49}$$

表 **9.10** 不同效应模型的期望方差

	样本方差	期望方差		
		固定效应模型	随机效应模型	混合模型 (A 固定，B 随机)
A 因素	s_A^2	$bn\eta_\alpha^2+\sigma_e^2$	$n\sigma_{\alpha\beta}^2+bn\sigma_\alpha^2+\sigma_e^2$	$n\sigma_{\alpha\beta}^2+bn\eta_\alpha^2+\sigma_e^2$
B 因素	s_B^2	$an\eta_\beta^2+\sigma_e^2$	$n\sigma_{\alpha\beta}^2+an\sigma_\beta^2+\sigma_e^2$	$an\sigma_\beta^2+\sigma_e^2$
$A\times B$	s_{AB}^2	$n\eta_{\alpha\beta}^2+\sigma_e^2$	$n\sigma_{\alpha\beta}^2+\sigma_e^2$	$n\sigma_{\alpha\beta}^2+\sigma_e^2$
误差	s_e^2	σ_e^2	σ_e^2	σ_e^2

本章小结

1. 方差分析的基本原理
    - 方差分析的数学模型
    - 平方和与自由度分解
    - 不同效应方差的数学期望
    - 显著性检验——F 检验
    - 功效分析
    - 多重比较
    - 方差分析的基本条件
2. 单因素方差分析
    - 组内观测值数量相等
    - 组内观测值数量不相等
    - F 检验与 t 检验的关系
3. 双因素方差分析