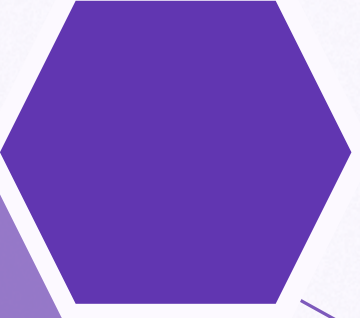


第四章 方差分析与秩和 Kruskal-Wallis检验

复旦大学附属肿瘤医院 周支瑞



4.1 单因素方差分析与 协方差分析

变异来源	SS	ν	MS	F	P
总	SS_{Total}	$n-1$			
组间	$SS_{Between}$	$k-1$	$MS_{Between}$	$MS_{Between}/MS_{Within}$	$\times.\times\times\times\times$
组内(误差)	SS_{Within}	$n-k$	MS_{Within}		

表 5-1 不同大鼠模型血清 IL-2 水平 (ng/ml) 的比较

$$SS_{\text{总}} = \sum_i \sum_j (X_{ij} - \bar{X})^2 \quad V_{\text{总}} = N-1$$

$$SS_{\text{组间}} = \sum n_i (\bar{X}_i - \bar{X})^2$$

各组均数也不相等 (组间变异) $V_{\text{组间}} = K-1$

$$MS_{\text{组间}} = SS_{\text{组间}} / (K-1)$$

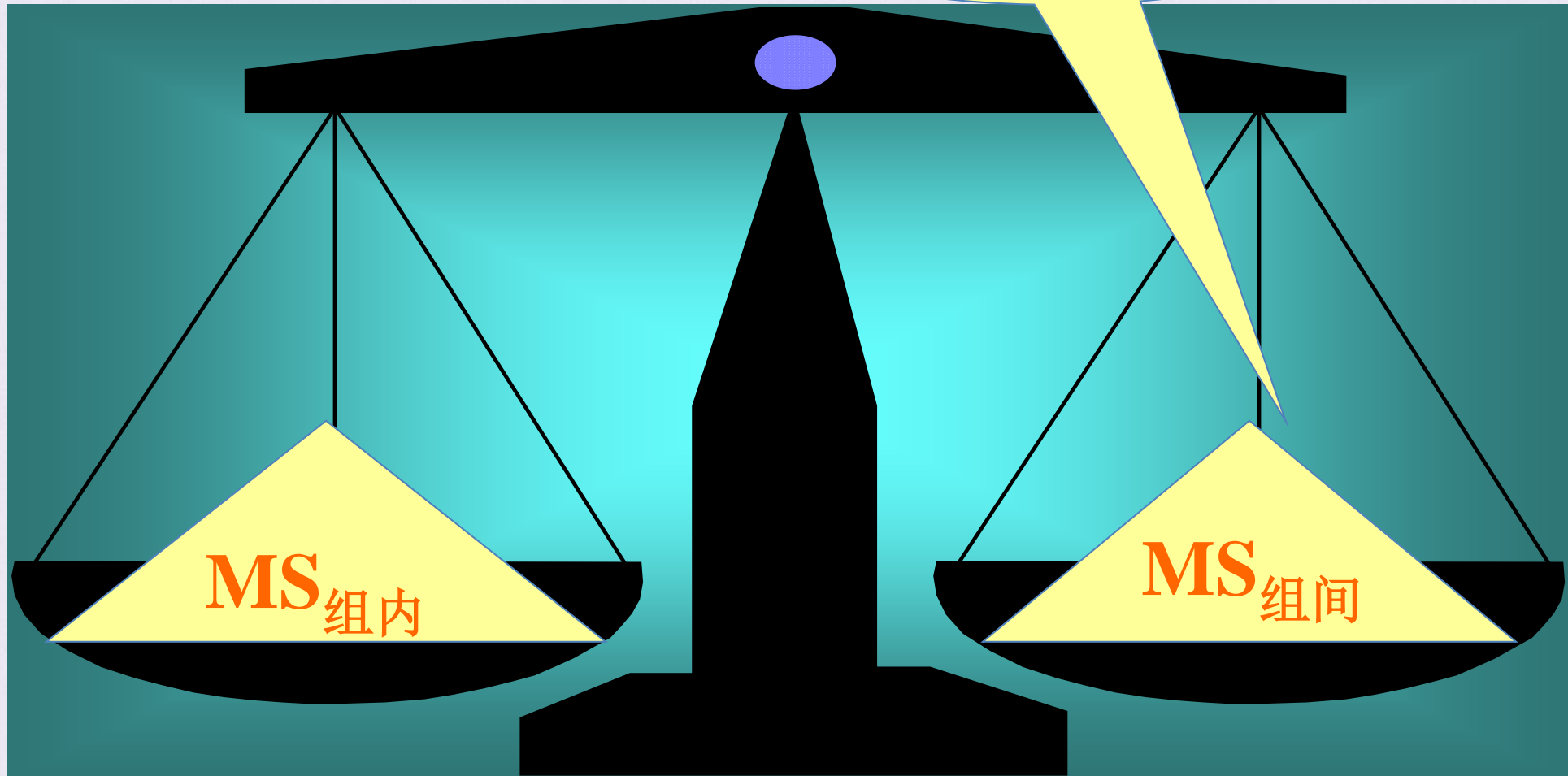
$$SS_{\text{组内}} = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2$$

组内各观察值之间各不相同 (组内变异) $V_{\text{组内}} = N-K$

$$MS_{\text{组内}} = SS_{\text{组内}} / (N-K)$$

血清 IL-2 水平					合计
甲组	乙组	丙组	丁组		
0.08	0.83	1.16	2.65	$SS_{\text{组内}} = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2$	
0.38	1.16	2.65	2.07		
0.40	1.55	2.07	2.13		
0.50	0.56	2.13	2.92		
0.60	0.36	2.92	2.12		
0.15	1.57	2.12	1.47		
0.10	1.04	2.09	2.09		
0.12	1.09	2.05	2.67		
n_i 8	8	8	8		(n) 32
$\sum_{j=1}^{n_i} x_{ij}$ 2.33	8.16	17.19	18.12		($\sum x$) 45.80
\bar{x}_i 0.2913	1.0200	2.1488	2.2650		(\bar{x}) 1.4313
$\sum_{j=1}^{n_i} x_{ij}^2$ 0.9677	9.6148	38.7813	42.5230		($\sum x^2$) 91.8868
S_i 0.2032	0.4296	0.5133	0.4600		(S) 0.92170

处理因素无效



$$MS_{\text{组内}} = MS_{\text{组间}}$$



Ronald Aylmer Fisher
(1890-1962)

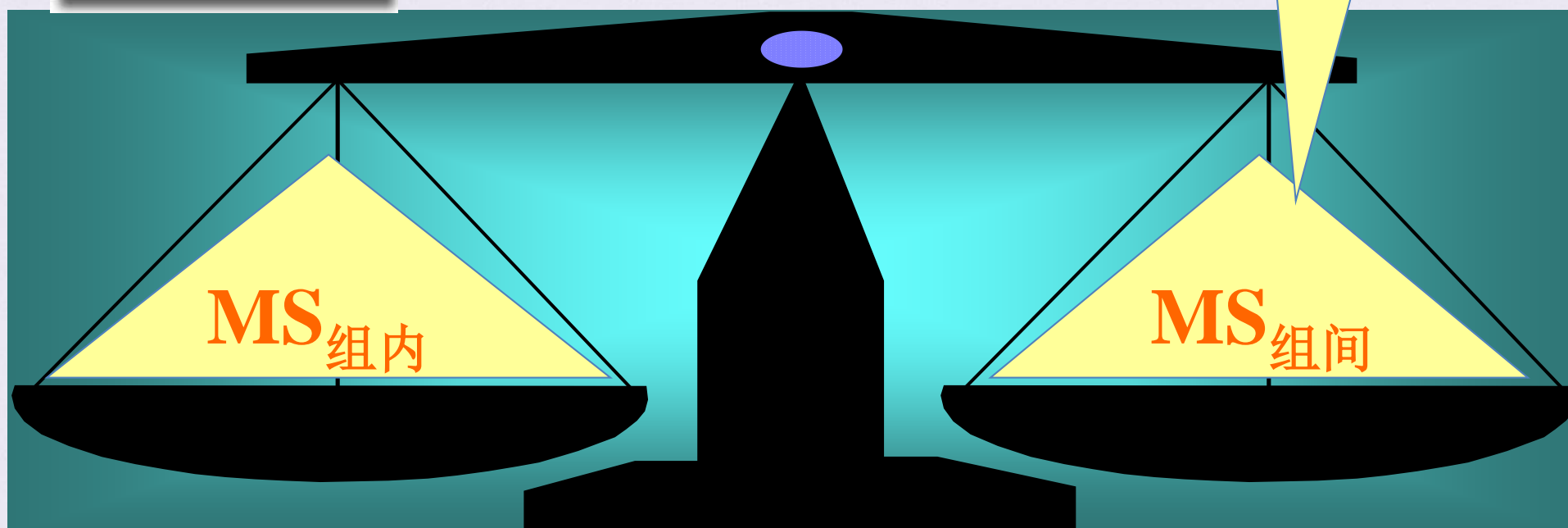
Fisher

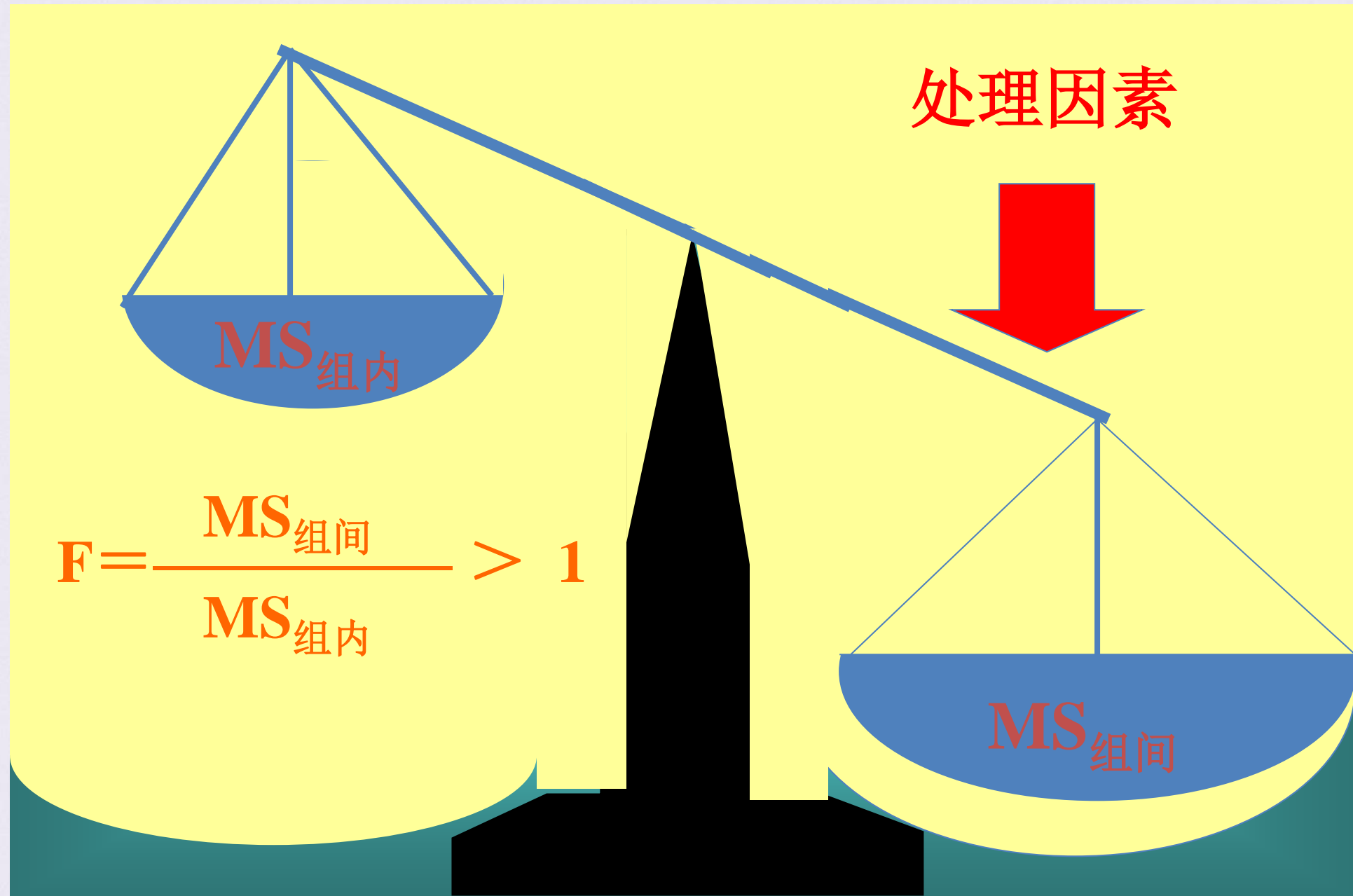
$MS_{\text{组间}}$

$$F = \frac{\quad}{\quad} = 1$$

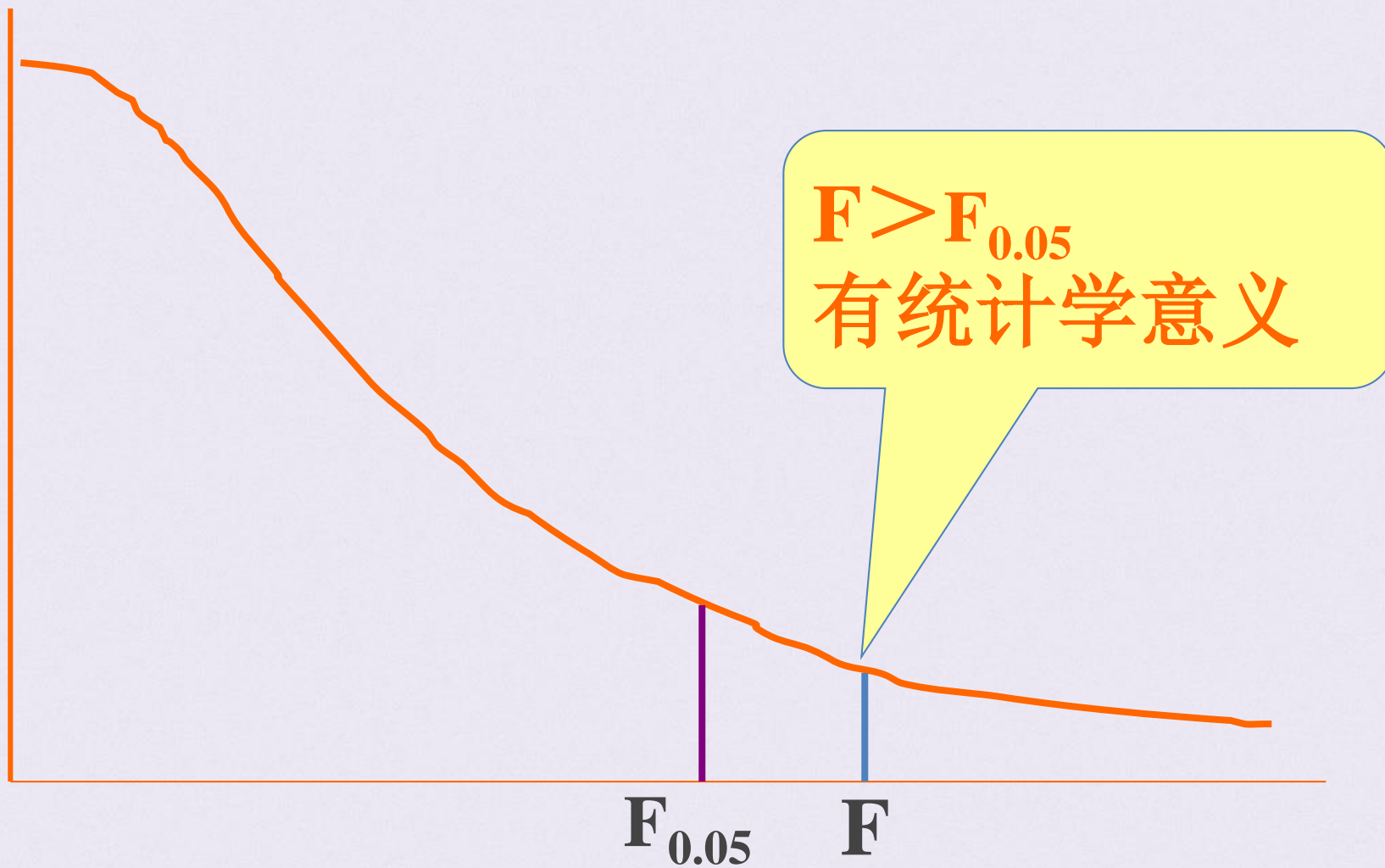
$MS_{\text{组内}}$

处理因素无效





如何判断是否有统计学意义呢?



将计算得到的F值与F分布的界值相比较

$$F \geq F_{\alpha, v1, v2},$$

$$P \leq \alpha$$

$$F < F_{\alpha, v1, v2},$$

$$P > \alpha$$

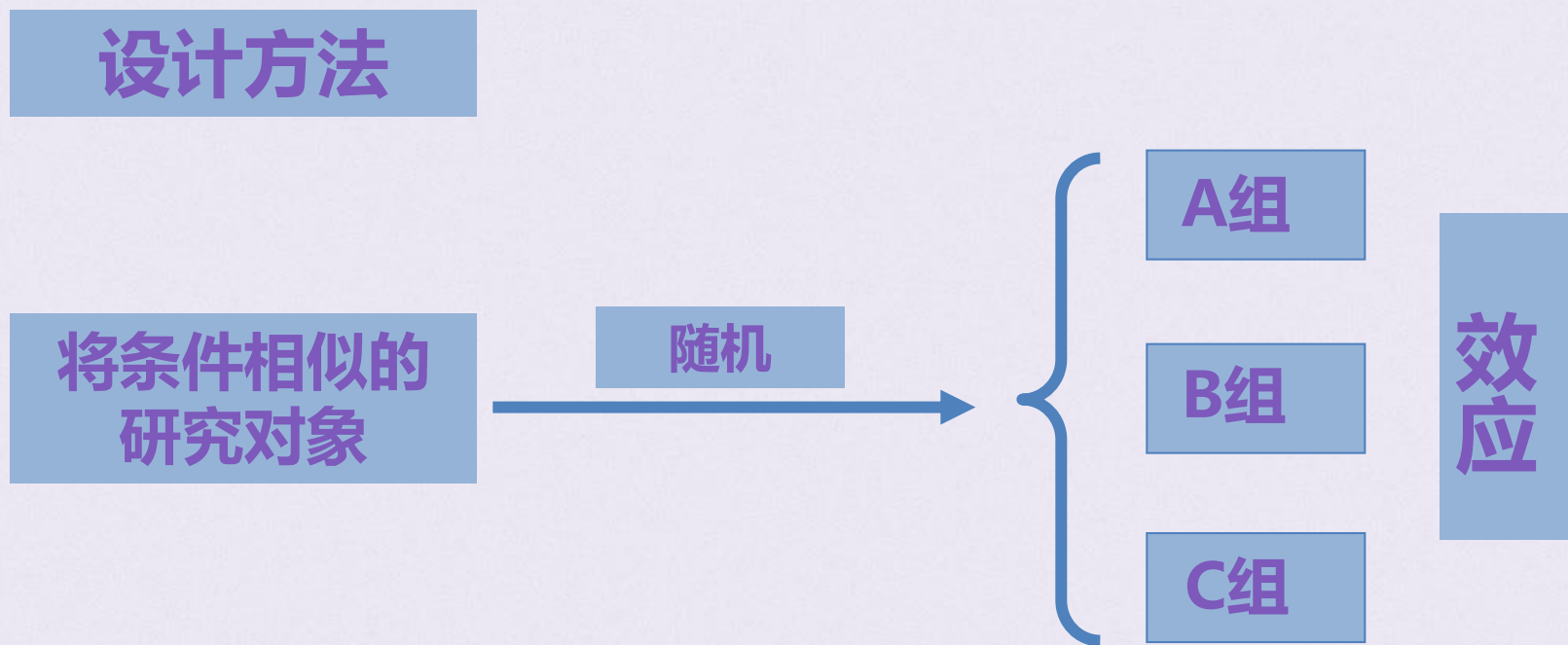
- 检验两个或多个样本均数间的差异有无统计学意义；注意：两个样本均数的比较可以采用t检验或 F检验，两个以上样本均数的比较只能用F检验。
- 回归方程的线性假设检验；
- 检验两个或多个因素间有无交互作用。

- 各个样本是相互独立的随机样本；
- 各个样本来自正态总体；
- 各个处理组的总体方差相等，即方差齐。

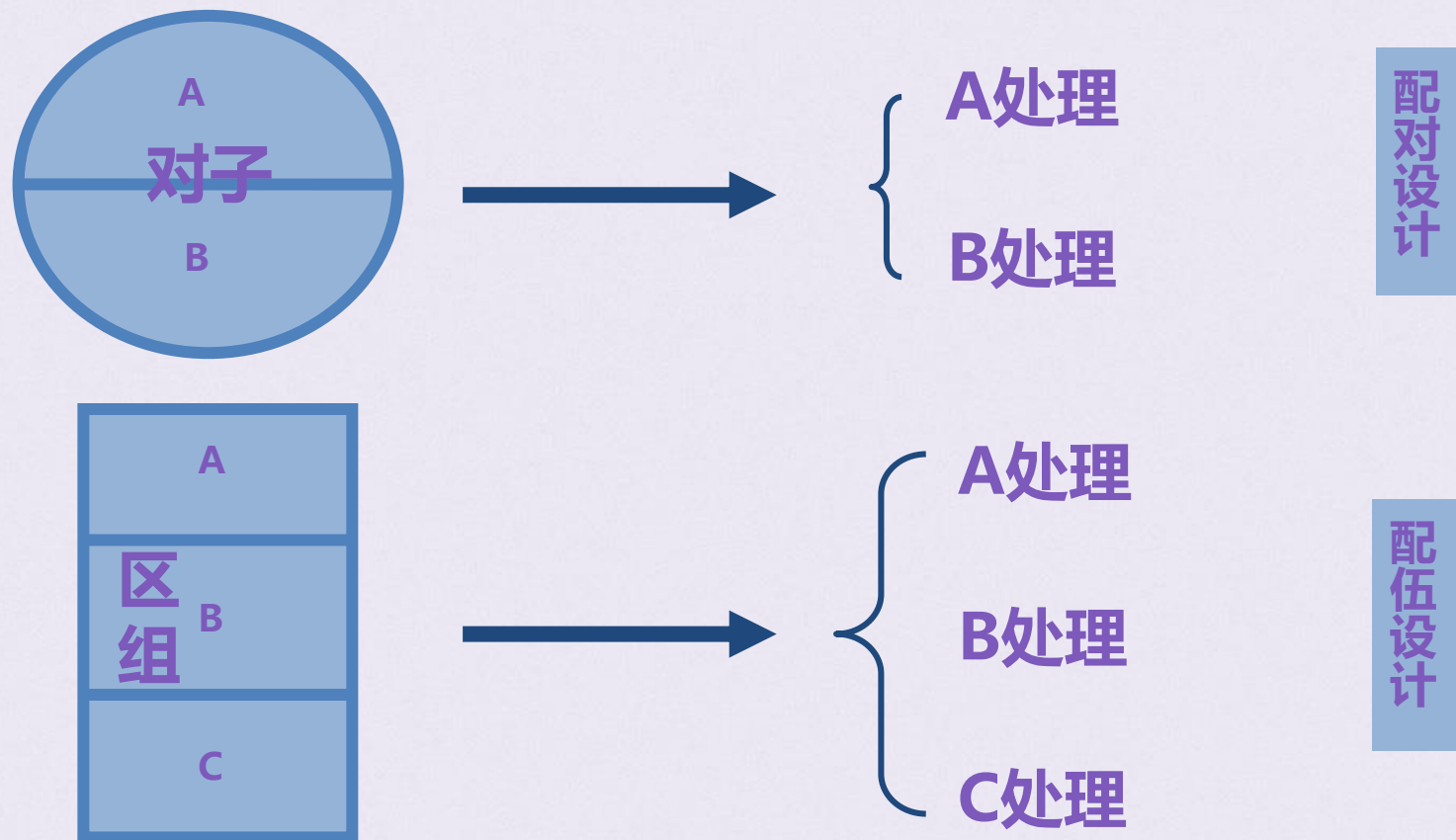
不满足方差分析应用条件时处理方法：

- 进行变量变换，以达到方差齐或正态的要求
- **采用非参数法（秩和检验）**
- 使用近似F检验

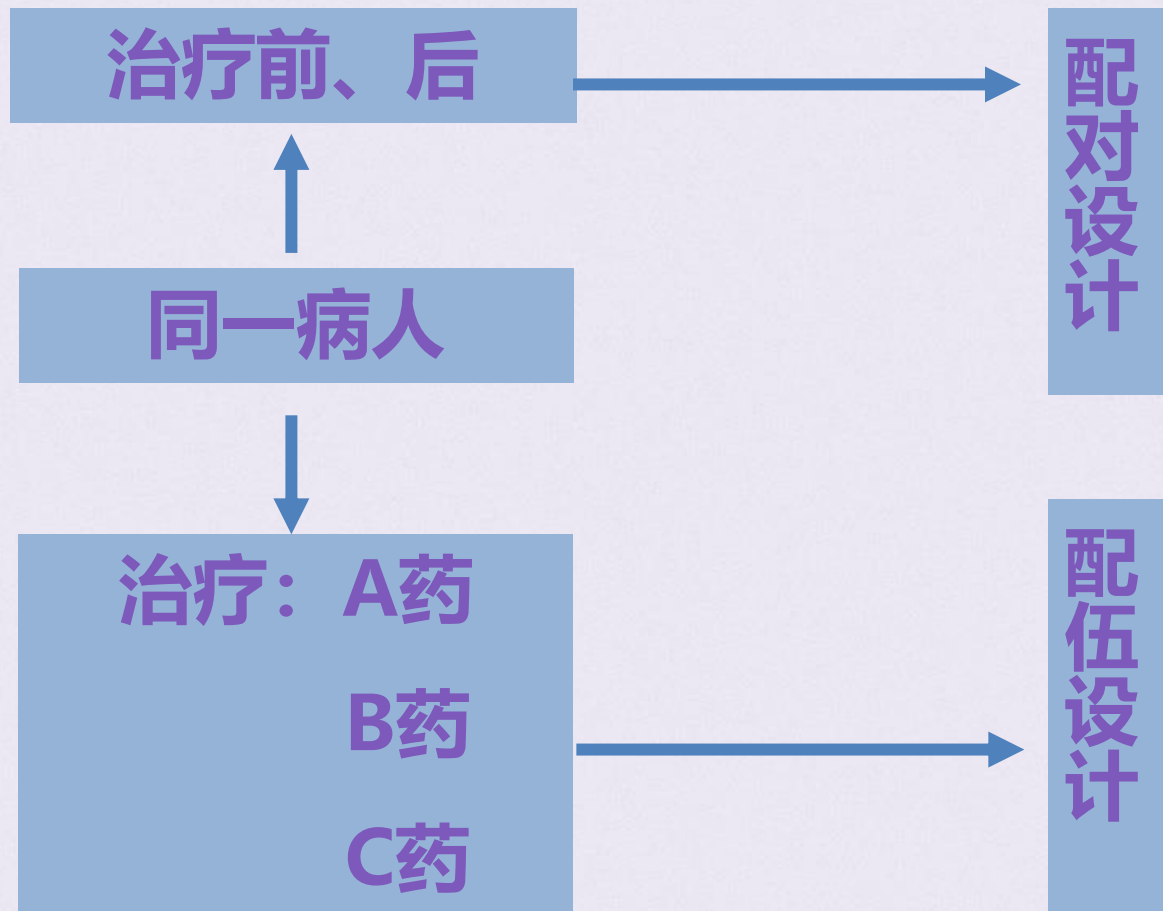
完全随机设计的方差分析



配对设计与配伍（区组）设计



配对设计与配伍（区组）设计 举例



- 由于涉及的对比组数大于2，就不能应用前面介绍的t检验。若仍用前述前述的t检验方法，对每两个对比组作比较，会使犯第I类错误(拒绝了实际上成立的H0所犯的误差)的概率 α 增大，即可能把本来无差别的两个总体均数判为有差别。

表9-3 含组间和组内因子的双因素方差分析

		患 者	时 间	
			5周	6个月
疗 法	CBT	s1		
		s2		
		s3		
		s4		
		s5		
	EMDR	s6		
		s7		
		s8		
		s9		
		s10		

- ANOVA和回归方法都是独立发展，但从函数形式上，它们都是广义线性模型的特例。回归中的lm()函数也能分析ANOVA模型。两个函数的结果是等同的。
- aov()函数的语法为：aov(formula, data=dataframe)，表9-4列举了表达式中可以使用的特殊符号。表9-4中的y是因变量，字母A、B、C代表因子。

表9-4 R表达式中的特殊符号	
符 号	用 法
~	分隔符号，左边为响应变量，右边为解释变量。例如，用 A、B 和 C 预测 y，代码为 <code>y ~ A + B + C</code>
:	表示变量的交互项。例如，用 A、B 和 A 与 B 的交互项来预测 y，代码为 <code>y ~ A + B + A:B</code>
*	表示所有可能交互项。代码 <code>y ~ A * B * C</code> 可展开为 <code>y ~ A + B + C + A:B + A:C + B:C + A:B:C</code>
^	表示交互项达到某个次数。代码 <code>y ~ (A + B + C)^2</code> 可展开为 <code>y ~ A + B + C + A:B + A:C + B:C</code>
.	表示包含除因变量外的所有变量。例如，若一个数据框包含变量 y、A、B 和 C，代码 <code>y ~ .</code> 可展开为 <code>y ~ A + B + C</code>

➤ 表9-5列举了一些常见的研究设计表达式。在表9-5中，小写字母表示定量变量，大写字母表示组别因子，Subject是对被试者独有的标识变量。

表9-5 常见研究设计的表达式	
设 计	表 达 式
单因素 ANOVA	$y \sim A$
含单个协变量的单因素 ANCOVA	$y \sim x + A$
双因素 ANOVA	$y \sim A * B$
含两个协变量的双因素 ANCOVA	$y \sim x1 + x2 + A*B$
随机化区组	$y \sim B + A$ (B 是区组因子)
单因素组内 ANOVA	$y \sim A + \text{Error}(\text{Subject}/A)$
含单个组内因子 (W) 和单个组间因子 (B) 的重复测量 ANOVA	$y \sim B * W + \text{Error}(\text{Subject}/W)$

- 表达式中效应的顺序在两种情况下会造成影响：(a)因子不止一个，并且是非平衡设计；(b)存在协变量。出现任意一种情况时，等式右边的变量都与其他每个变量相关。此时，我们无法清晰地划分它们对因变量的影响。例如，对于双因素方差分析，若不同处理方式中的观测数不同，那么模型 $y \sim A*B$ 与模型 $y \sim B*A$ 的结果不同

- 样本大小越不平衡，效应项的顺序对结果的影响越大。一般来说，越基础性的效应越需要放在表达式前面。具体来讲，首先是协变量，然后是主效应，接着是双因素的交互项，再接着是三因素的交互项，以此类推。对于主效应，越基础性的变量越应放在表达式前面，因此性别要放在处理方式之前。有一个基本的准则：若研究设计不是正交的（也就是说，因子和/或协变量相关），一定要谨慎设置效应的顺序。

含因子A、B和因变量y的双因素不平衡因子设计，有三种效应：A和B的主效应，A和B的交互效应。假设你正使用如下表达式对数据进行建模： $Y \sim A + B + A:B$ 有三种类型的方法可以分解等式右边各效应对y所解释的方差。

- 类型 I（序贯型）效应根据表达式中先出现的效应做调整。A不做调整，B根据A调整，A:B交互项根据A和B调整。
- 类型 II（分层型）效应根据同水平或低水平的效应做调整。A根据B调整，B依据A调整，A:B交互项同时根据A和B调整。
- 类型 III（边界型）每个效应根据模型其他各效应做相应调整。A根据B和A:B做调整，A:B交互项根据A和B调整。

R默认调用类型I方法，其他软件（比如SAS和SPSS）默认调用类型III方法。

- 请注意car包中的Anova()函数（不要与标准anova()函数混淆）提供了使用类型II或类型III方法的选项，而aov()函数使用的是类型I方法。若想使结果与其他软件（如SAS和SPSS）提供的结果保持一致，可以使用Anova()函数，细节可参考help(Anova,package="car")。

- 单因素方差分析用于比较分类因子定义的两个或多个组别中的因变量均值。以 multcomp 包中的 cholesterol 数据集为例（取自Westfall、Tobia、Rom、Hochberg, 1999），50个患者均接受降低胆固醇药物治疗（trt）五种疗法中的一种疗法。其中三种治疗条件使用药物相同，分别是20mg一天一次（1time）、10mg一天两次（2times）和5mg一天四次（4times）。剩下的两种方式（drugD和drugE）代表候选药物。哪种药物疗法降低胆固醇（响应变量）最多呢？分析过程见如下代码。


```
> library(multcomp)
> attach(cholesterol)
> table(trt)
```

① 各组样本大小

```
trt
  1time 2times 4times  drugD  drugE
      10      10      10      10      10
```

```
> aggregate(response, by=list(trt), FUN=mean)
```

② 各组均值

```
  Group.1      x
1   1time  5.78
2   2times  9.22
3   4times 12.37
4   drugD 15.36
5   drugE 20.95
```

```
> aggregate(response, by=list(trt), FUN=sd)
```

③ 各组标准差

```
  Group.1      x
1   1time  2.88
2   2times  3.48
3   4times  2.92
4   drugD  3.45
5   drugE  3.35
```

```
> fit <- aov(response ~ trt)
```

```
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	4	1351	338	32.4	9.8e-13 ***
Residuals	45	469	10		

④ 检验组间差异 (ANOVA)

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> library(gplots)
```

```
> plotmeans(response ~ trt, xlab="Treatment", ylab="Response",
  main="Mean Plot\nwith 95% CI")
```

```
> detach(cholesterol)
```

⑤ 绘制各组均值及其置信区间的图形

- 虽然ANOVA对各疗法的F检验表明五种药物治疗效果不同，但是并没有告诉你哪种疗法与其他疗法不同。多重比较可以解决这个问题。例如，TukeyHSD()函数提供了对各组均值差异的成对检验（见如下代码）。

代码清单9-2 Tukey HSD的成对组间比较

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = response ~ trt)

$trt
      diff      lwr      upr p adj
2times-1time  3.44 -0.658  7.54 0.138
4times-1time  6.59  2.492 10.69 0.000
drugD-1time   9.58  5.478 13.68 0.000
drugE-1time  15.17 11.064 19.27 0.000
4times-2times  3.15 -0.951  7.25 0.205
drugD-2times  6.14  2.035 10.24 0.001
drugE-2times 11.72  7.621 15.82 0.000
drugD-4times  2.99 -1.115  7.09 0.251
drugE-4times  8.57  4.471 12.67 0.000
drugE-drugD   5.59  1.485  9.69 0.003

> par(las=2)
> par(mar=c(5,8,4,2))
> plot(TukeyHSD(fit))
```

- 单因素方差分析中，我们假设因变量服从正态分布，各组方差相等。可使用Q-Q图来检验正态性假设：
 `>library(car)`
 `>qqPlot(lm(response ~ trt, data=cholesterol),simulate=TRUE, main="Q-Q Plot", labels=FALSE)`
 注意qqPlot()要求用lm()拟合，落在95%的置信区间内，说明满足正态性假设。
- R提供了一些可用来做方差齐性检验的函数。例如，可以通过如下代码来做bartlett检验：
 `>bartlett.test(response ~ trt, data=cholesterol)`
- 方差齐性分析对离群点非常敏感。可利用car包中的outlierTest()函数来检测离群点：
 `>library(car)`
 `>outlierTest(fit)`

- 单因素协方差分析 (ANCOVA) 扩展了单因素方差分析 (ANOVA)，包含一个或多个定量协变量。下面的例子来自于multcomp包中的litter数据集（见Westfall et al., 1999）。怀孕小鼠被分为四个小组，每个小组接受不同剂量（0、5、50或500）的药物处理。产下幼崽的体重均值为因变量，怀孕时间为协变量。分析代码如下。

代码清单9-3 单因素ANCOVA

```
> data(litter, package="multcomp")
> attach(litter)
> table(dose)
dose
 0   5  50 500
20  19  18  17
> aggregate(weight, by=list(dose), FUN=mean)
  Group.1      x
1        0 32.3
2         5 29.3
3        50 29.9
4       500 29.6
> fit <- aov(weight ~ gesttime + dose)
> summary(fit)
```


- 由于使用了协变量，你可能想要获取调整的组均值，即去除协变量效应后的组均值。可使用effects包中的effects()函数来计算调整的均值：
 > library(effects)
 > effect("dose", fit)

- > library(multcomp)
- > contrast <- rbind("no drug vs. drug" = c(3, -1, -1, -1))
- > summary(glht(fit, linfct=mcp(dose=contrast)))
- 对照c(3, -1, -1, -1)设定第一组和其他三组的均值进行比较。假设检验的t统计量(2.581) 在 $p < 0.05$ 水平下显著, 因此, 可以得出未用药组比其他用药条件下的出生体重高的结论。其他对照可用rbind()函数添加

- ANCOVA与ANOVA相同，都需要正态性和同方差性假设，可以用上节中相同的步骤检验这些假设条件。另外，ANCOVA还假定回归斜率相同。本例中，假定四个处理组通过怀孕时间来预测出生体重的回归斜率都相同。ANCOVA模型包含怀孕时间×剂量的交互项时，可对回归斜率的同质性进行检验。交互效应若显著，则意味着时间和幼崽出生体重间的关系依赖于药物剂量的水平。代码如下：

```
> library(multcomp)
> fit2 <- aov(weight ~ gesttime*dose, data=litter)
> summary(fit2)
```




4.2 双因素方差分析

- 双因素方差分析中，受试者被分配到两因子交叉类别组中。以基础安装包中的 ToothGrowth 数据集为例，随机分配60只豚鼠，分别采用两种喂食方法（橙汁或维生素C），各喂食方法中抗坏血酸含量有三种水平（0.5mg、1mg或2mg），每种处理方式组合都被分配10只豚鼠。牙齿长度为因变量，代码如下
- ```
> attach(ToothGrowth)
> table(supp, dose)
> aggregate(len, by=list(supp, dose), FUN=mean)
> aggregate(len, by=list(supp, dose), FUN=sd)
> dose <- factor(dose)
> fit <- aov(len ~ supp*dose)
> summary(fit)
```

- table语句的预处理表明该设计是均衡设计（各设计单元中样本大小都相同），aggregate语句处理可获得各单元的均值和标准差。dose变量被转换为因子变量，这样aov()函数就会将它当做一个分组变量，而不是一个数值型协变量。用summary()函数得到方差分析表，可以看到主效应（supp和dose）和交互效应都非常显著。

- 有多种方式对结果进行可视化处理。此处可用interaction.plot()函数来展示双因素方差分析的交互效应。
- `interaction.plot(dose, supp, len, type="b",col=c("red","blue"), pch=c(16, 18),main = "Interaction between Dose and Supplement Type")`

一般来说，研究设计中考虑以下问题时应采用重复测量设计：

- 研究主要目的之一是考察某指标在不同时间的变化情况。
- 研究个体间变异很大，应用普通研究设计的方差分析时，方差分析表中的误差项值将很大，即计算F值时的分母很大，对反应变量有作用的因素常难以识别。
- 有的研究中研究对象很难征募到足够多的数量，此时可考虑对所征募到的对象在不同条件下的反应进行测量。



- 基本思想：仍然应用方差分析的基本思想，将反应变量的变异分解成以下四个部分：研究对象内的变异（即测量时间点或测量条件下的效应）、研究对象间的变异（即处理因素效应）、上述两者的交互作用、随机误差变异。
- 因素：
  - 受试者内因素----用于区分重复测量次数的变量
  - 受试者间因素----在重复测量时保持恒定的因素
- 分析目的：一是分析受试者间因素的作用；二是考察随着测量次数的增加，测量指标是如何发生变化的，以及分组因素的作用是否会随时间发生，即是否和时间存在交互作用。

- 反应变量之间存在相关关系。
- 反应变量的均数向量服从多元正态分布。
- 对于自变量的各取值水平组合而言，反应变量的方差、协方差矩阵相等。

- 因变量是二氧化碳吸收量 (uptake) , 单位为ml/L, 自变量是植物类型Type (魁北克VS.密西西比) 和七种水平 (95~1000  $\mu\text{mol}/\text{m}^2 \text{ sec}$ ) 的二氧化碳浓度 (conc) 。另外, Type是组间因子, conc是组内因子。Type已经被存储为一个因子变量, 但你还需要先将conc转换为因子变量。

```
> CO2$conc <- factor(CO2$conc)
```

```
> w1b1 <- subset(CO2, Treatment == 'chilled')
```

```
> fit <- aov(uptake ~ (conc*Type) + Error(Plant/(conc)), w1b1)
```

```
> summary(fit)
```



例 8-12 在研究某药物对某癌细胞株增殖影响的研究中，分别于细胞培养后 24h、48h、72h、96h 检测实验组和对照组细胞株的细胞抑制率，实验数据见表 8-20。使用重复测量资料分析的方法，得出专业结论。

表 8-20 癌细胞生长不同时间的抑制作用

| 分 组 | 编 号 | 细胞生长抑制率 (h) |       |       |       |
|-----|-----|-------------|-------|-------|-------|
|     |     | 24          | 48    | 72    | 96    |
| 对照组 | 1   | 1.431       | 1.519 | 1.477 | 1.364 |
|     | 2   | 1.385       | 1.562 | 1.459 | 1.372 |
|     | 3   | 1.473       | 1.487 | 1.612 | 1.414 |
|     | 4   | 1.452       | 1.535 | 1.537 | 1.403 |
|     | 5   | 1.371       | 1.469 | 1.268 | 1.296 |
| 实验组 | 6   | 1.257       | 0.976 | 0.725 | 0.578 |
|     | 7   | 1.232       | 0.934 | 0.828 | 0.609 |
|     | 8   | 1.298       | 1.036 | 0.813 | 0.512 |
|     | 9   | 1.216       | 1.247 | 0.694 | 0.579 |
|     | 10  | 1.275       | 0.942 | 0.675 | 0.621 |



例 8-13 试验名称为“缺氧细胞放射增敏剂（代号 808）效应大小的试验研究”。将 Lewis 肺癌瘤株接种到 C57 小鼠的腿部，使腿的接种处于缺氧条件下，对随机分入甲组的小鼠直接给予 12Gy $\gamma$  射线照射，对随机分入乙组的小鼠先给予“808 增敏剂”，然后再给予 12Gy $\gamma$  射线照射，甲、乙两组各有 4 只小鼠，照射后分别在 2、4、6、8、10 天后观测肿瘤的大小（相对体积），相同的试验共做了两批。使用重复测量资料分析的方法，得出专业结论，如表 8-21 所示。

表 8-21 缺氧细胞放射增敏剂效应大小的试验结果

| A（批次） | B（808 增敏剂） | 鼠编号 | 肿瘤的相对体积 |       |       |        |        |
|-------|------------|-----|---------|-------|-------|--------|--------|
|       |            |     | T1(2)   | T2(4) | T3(6) | T4(8)  | T5(10) |
| A1    | B1         | 1   | 0.640   | 1.439 | 2.580 | 3.667  | 6.353  |
|       |            | 2   | 1.103   | 2.195 | 3.546 | 4.028  | 5.592  |
|       |            | 3   | 2.795   | 5.617 | 7.831 | 10.304 | 19.518 |
|       |            | 4   | 1.133   | 1.749 | 1.723 | 2.528  | 3.199  |
|       | B2         | 5   | 2.585   | 3.868 | 5.555 | 7.101  | 10.618 |
|       |            | 6   | 2.807   | 5.569 | 8.176 | 10.790 | 16.771 |
|       |            | 7   | 2.709   | 5.729 | 6.263 | 11.486 | 16.455 |
|       |            | 8   | 2.034   | 4.167 | 7.524 | 12.479 | 13.222 |
| A2    | B1         | 9   | 1.649   | 3.830 | 4.650 | 7.653  | 9.620  |
|       |            | 10  | 2.522   | 3.939 | 6.070 | 8.910  | 16.250 |
|       |            | 11  | 1.658   | 4.612 | 5.903 | 11.295 | 13.409 |
|       |            | 12  | 0.848   | 2.040 | 2.371 | 5.000  | 8.976  |

续表

| A（批次） | B（808 增敏剂） | 鼠编号 | 肿瘤的相对体积 |       |       |        |        |
|-------|------------|-----|---------|-------|-------|--------|--------|
|       |            |     | T1(2)   | T2(4) | T3(6) | T4(8)  | T5(10) |
| A2    | B2         | 13  | 3.001   | 4.587 | 4.911 | 11.293 | 15.138 |
|       |            | 14  | 3.383   | 5.438 | 6.636 | 16.650 | 22.396 |
|       |            | 15  | 1.621   | 3.625 | 7.712 | 11.164 | 17.063 |
|       |            | 16  | 1.586   | 1.673 | 2.291 | 4.937  | 8.429  |



## 4.3 多元方差分析



例 将某肝炎病人随机分成两组，一组施以新的疗法，另一组仍用传统的治疗方法，考察用二种方法治疗后对反映病人肝功能指标（如SGPT、AST、ALT、HCV等）的影响。

- 与一个反应变量的方差分析相似，都是将反应变量的变异分解成为两部分：一部分为两组间变异（组别因素的效应），一部分为组内变异（随机误差）。然后对这两部分变异进行比较，看是否组间变异大于组内变异。不同的是，后者都是对组间均方与组内均方进行比较，而前者是对组间方差协方差矩阵与组内方差协方差矩阵进行比较。



- 各因变量服从多元正态分布：只要一个反应变量不服从正态分布，则这几个反应变量的联合分布肯定不服从多元正态分布。
- 各观察对象之间相互独立。
- 各组观察对象反应变量的方差-协方差矩阵相等。
- 反应变量间的确存在一定的关系，这可以从专业或研究目的角度予以判断。

- 当因变量（结果变量）不止一个时，可用多元方差分析（MANOVA）对它们同时进行分析。以MASS包中的UScereal数据集为例（Venables, Ripley (1999)），我们将研究美国谷物中的卡路里、脂肪和糖含量是否会因为储存架位置的不同而发生变化；其中1代表底层货架，2代表中层货架，3代表顶层货架。卡路里、脂肪和糖含量是因变量，货架是三水平（1、2、3）的自变量。分析过程见代码清单如下：

```
> library(MASS)
> attach(UScereal)
> shelf <- factor(shelf)
> y <- cbind(calories, fat, sugars)
> aggregate(y, by=list(shelf), FUN=mean)
> cov(y)
> fit <- manova(y ~ shelf)
> summary(fit)
> summary.aov(fit)
```

- 首先，我们将shelf变量转换为因子变量，从而使它在后续分析中能作为分组变量。cbind() 函数将三个因变量（卡路里、脂肪和糖）合并成一个矩阵。  
aggregate() 函数可获取货架的各个均值，cov()则输出各谷物间的方差和协方差。manova()函数能对组间差异进行多元检验。上面F值显著，说明三个组的营养成分测量值不同。注意shelf变量已经转成了因子变量，因此它可以代表一个分组变量。由于多元检验是显著的，可以使用summary.aov()函数对每一个变量做单因素方差分析。

- 单因素多元方差分析有两个前提假设，一个是多元正态性，一个是方差 协方差矩阵同质性。第一个假设即指因变量组合成的向量服从一个多元正态分布。可以用Q-Q图来检验该假设条件

```
> center <- colMeans(y)
> n <- nrow(y)
> p <- ncol(y)
> cov <- cov(y)
> d <- mahalanobis(y, center, cov)
> coord <- qqplot(qchisq(ppoints(n), df=p),
 d, main="QQ Plot Assessing Multivariate Normality",
 ylab="Mahalanobis D2")
> abline(a=0, b=1)
> identify(coord$x, coord$y, labels=row.names(UScereal))
```



- 若有一个 $p \times 1$ 的多元正态随机向量 $x$ ，均值为 $\mu$ ，协方差矩阵为 $\Sigma$ ，那么 $x$ 与 $\mu$ 的马氏距离的平方服从自由度为 $p$ 的卡方分布。Q-Q图展示卡方分布的分位数，横纵坐标分别是样本量与马氏距离平方值。如果点全部落在斜率为1、截距项为0的直线上，则表明数据服从多元正态分布。

- 若数据服从多元正态分布，那么点将落在直线上。还可以使用mvoutlier包中的ap.plot()函数来检验多元离群点。

```
> library(mvoutlier)
```

```
> outliers <- aq.plot(y)
```

```
> outliers
```

- 如果多元正态性或者方差 协方差均值假设都不满足或者你担心多元离群点，那么可以考虑用稳健或非参数版本的MANOVA 检验。稳健单因素MANOVA 可通过rrcov 包中的Wilks.test()函数实现。vegan包中的adonis()函数则提供了非参数MANOVA的等同形式。

```
> library(rrcov)
```

```
> Wilks.test(y,shelf, method="mcd")
```

- ANOVA和回归都是广义线性模型的特例。因此，本章所有的设计都可以用 `lm()` 函数来分析。但是，为了更好地理解输出结果，需要弄明白在拟合模型时，R是如何处理类别型变量的。以单因素ANOVA问题为例，即比较五种降低胆固醇药物疗法（trt）的影响。

```
> library(multcomp)
```

```
> levels(cholesterol$trt)
```

```
> fit.aov <- aov(response ~ trt, data=cholesterol)
```

```
> summary(fit.aov)
```

```
> fit.lm <- lm(response ~ trt, data=cholesterol)
```

```
> summary(fit.lm)
```

```
> # fit.lm <- lm(response ~ trt, data=cholesterol, contrasts="contr.helmert") # wrong
```

```
> fit.lm <- lm(response ~ trt, data=cholesterol, contrasts=list(trt="contr.helmert "))
```

```
> summary(fit.lm)
```



表9-6 内置对照

| 对照变量创建方法                     | 描 述                                                                           |
|------------------------------|-------------------------------------------------------------------------------|
| <code>contr.helmert</code>   | 第二个水平对照第一个水平, 第三个水平对照前两个的均值, 第四个水平对照前三个的均值, 以此类推                              |
| <code>contr.poly</code>      | 基于正交多项式的对照, 用于趋势分析 (线性、二次、三次等) 和等距水平的有序因子                                     |
| <code>contr.sum</code>       | 对照变量之和限制为 0。也称作偏差找对, 对各水平的均值与所有水平的均值进行比较                                      |
| <code>contr.treatment</code> | 各水平对照基线水平 (默认第一个水平)。也称作虚拟编码                                                   |
| <code>contr.SAS</code>       | 类似于 <code>contr.treatment</code> , 只是基线水平变成了最后一个水平。生成的系数类似于大部分 SAS 过程中使用的对照变量 |



## 4.4 Kruskal-Wallis检验 Friedman检验

- 如果无法满足ANOVA设计的假设，那么可以使用非参数方法来评估组间的差异。如果各组独立，则Kruskal-Wallis检验将是一种实用的方法。如果各组不独立（如重复测量设计或随机区组设计），那么Friedman检验会更合适。



- Kruskal-Wallis检验的调用格式为: `kruskal.test(y ~ A, data)`
- 其中的y是一个数值型结果变量, A是一个拥有两个或更多水平的分组变量 (若有两个水平, 则它与Mann-Whitney U检验等价) 。



- Friedman检验的调用格式为：`friedman.test(y ~ A | B, data)`
- 其中的`y`是数值型结果变量，`A`是一个分组变量，而`B`是一个用以认定匹配观测的区组变量（blocking variable）。在以上两例中，`data`皆为可选参数，它指定了包含这些变量的矩阵或数据框。

- 考虑state.x77数据集。它包含了美国各州的人口、收入、文盲率、预期寿命、谋杀率和高中毕业率数据。如果你想比较美国四个地区（东北部、南部、中北部和西部）的文盲率，应该怎么做呢？

```
> states <- data.frame(state.region, state.x77)
```

```
> kruskal.test(illiteracy ~ state.region, data=states)
```

## kruskal.test 案例 2

| 表 14-8 |    | 3 种药物灭杀钉螺的死亡率 (%) |    |       |    |
|--------|----|-------------------|----|-------|----|
| 甲 药    |    | 乙 药               |    | 丙 药   |    |
| 死 亡 率  | 秩  | 死 亡 率             | 秩  | 死 亡 率 | 秩  |
| 32.5   | 10 | 16.0              | 4  | 6.5   | 1  |
| 35.5   | 11 | 20.5              | 6  | 9.0   | 2  |
| 40.5   | 13 | 22.5              | 7  | 12.5  | 3  |
| 46.0   | 14 | 29.0              | 9  | 18.0  | 5  |
| 49.0   | 15 | 36.0              | 12 | 24.0  | 8  |
| $R_i$  | 63 | —                 | 38 | —     | 19 |
| $n_i$  | 5  | —                 | 5  | —     | 5  |

本例为百分率资料，不符合正态分布，故采用 Kruskal-Wallis H 检验进行分析。



例 14-17 8 名受试对象在相同的试验条件下分别接受 4 种不同频率声音的刺激，他们的反应率 (%) 资料见表 14-12。问 4 种频率声音刺激的反应率是否有差别。

表 14-12 8 名受试对象对 4 种不同频率声音刺激的反应率比较

| 受试对象 | 反应率 (%) | 声音 | 反应率 (%) | 声音 | 反应率 (%) | 声音 | 反应率 (%) | 声音 |
|------|---------|----|---------|----|---------|----|---------|----|
| 1    | 8.4     | 1  | 9.6     | 2  | 9.8     | 3  | 11.7    | 4  |
| 2    | 11.6    | 1  | 12.7    | 4  | 11.8    | 2  | 12.0    | 3  |
| 3    | 9.4     | 2  | 9.1     | 1  | 10.4    | 4  | 9.8     | 3  |

264

续表

| 受试对象  | 反应率 (%) | 声音 | 反应率 (%) | 声音 | 反应率 (%) | 声音   | 反应率 (%) | 声音   |
|-------|---------|----|---------|----|---------|------|---------|------|
| 4     | 9.8     | 2  | 8.7     | 1  | 9.9     | 3    | 12.0    | 4    |
| 5     | 8.3     | 2  | 8.0     | 1  | 8.6     | 3.5  | 8.6     | 3.5  |
| 6     | 8.6     | 1  | 9.8     | 3  | 9.6     | 2    | 10.6    | 4    |
| 7     | 8.9     | 1  | 9.0     | 2  | 10.6    | 3    | 11.4    | 4    |
| 8     | 7.8     | 1  | 8.2     | 2  | 8.5     | 3    | 10.8    | 4    |
| $R_i$ | —       | 11 | —       | 16 | —       | 23.5 | —       | 29.5 |



- 【1】 Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016
- 【2】 Peter Dalgaard 著, 《R语言统计入门》 》(第2版), 人民邮电出版社, 2014
- 【3】 薛毅 陈立萍 著, 《R语言实用教程》, 清华大学出版社, 2014
- 【4】 张铁军 陈兴栋 刘振球 著, 《R语言与医学统计图形》, 人民卫生出版社, 2018

# Thanks!

感谢您的观看!