统计学与 R 语言

第 18 讲　方差分析

张敬信

2022 年 5 月 3 日

哈尔滨商业大学

方差分析（ANOVA)，是针对连续变量的参数检验，检验多个分组的均值有无差异，其中分组是按影响因素的不同水平值组合进行划分的。它是对总变异进行分解，看总变异是由哪些部分组成的，这些部分间的关系如何。

方差分析可用于：

- 完全随机设计（单因素)、随机区组设计（双因素)、析因设计、拉丁方设计和正交设计等资料；
- 可对两因素间交互作用差异进行显著性检验；
- 进行方差齐性检验。

一. 相关术语

- 案例 1：现有 4 个行业 23 家企业一年内投诉次数的数据

```
library(tidyverse)
complaints = tibble(id = 1:7,
  零售业 = c(57,66,49,40,34,53,44),
  旅游业 = c(68,39,29,45,56,51,NA),
  航空公司 = c(31,49,21,34,40,NA,NA),
  家电制造业 = c(44,51,65,77,58,NA,NA))
```

```
complaints
#> # A tibble: 7 x 5
#>      id 零售业 旅游业 航空公司 家电制造业
#>   <int>  <dbl>  <dbl>    <dbl>       <dbl>
#> 1     1     57     68       31          44
#> 2     2     66     39       49          51
#> 3     3     49     29       21          65
#> 4     4     40     45       34          77
#> 5     5     34     56       40          58
#> 6     6     53     51       NA          NA
#> 7     7     44     NA       NA          NA
```

分析 4 个行业之间的服务质量是否有显著差异，也就是要判断 “行业” 对 “投诉次数” 是否有显著影响

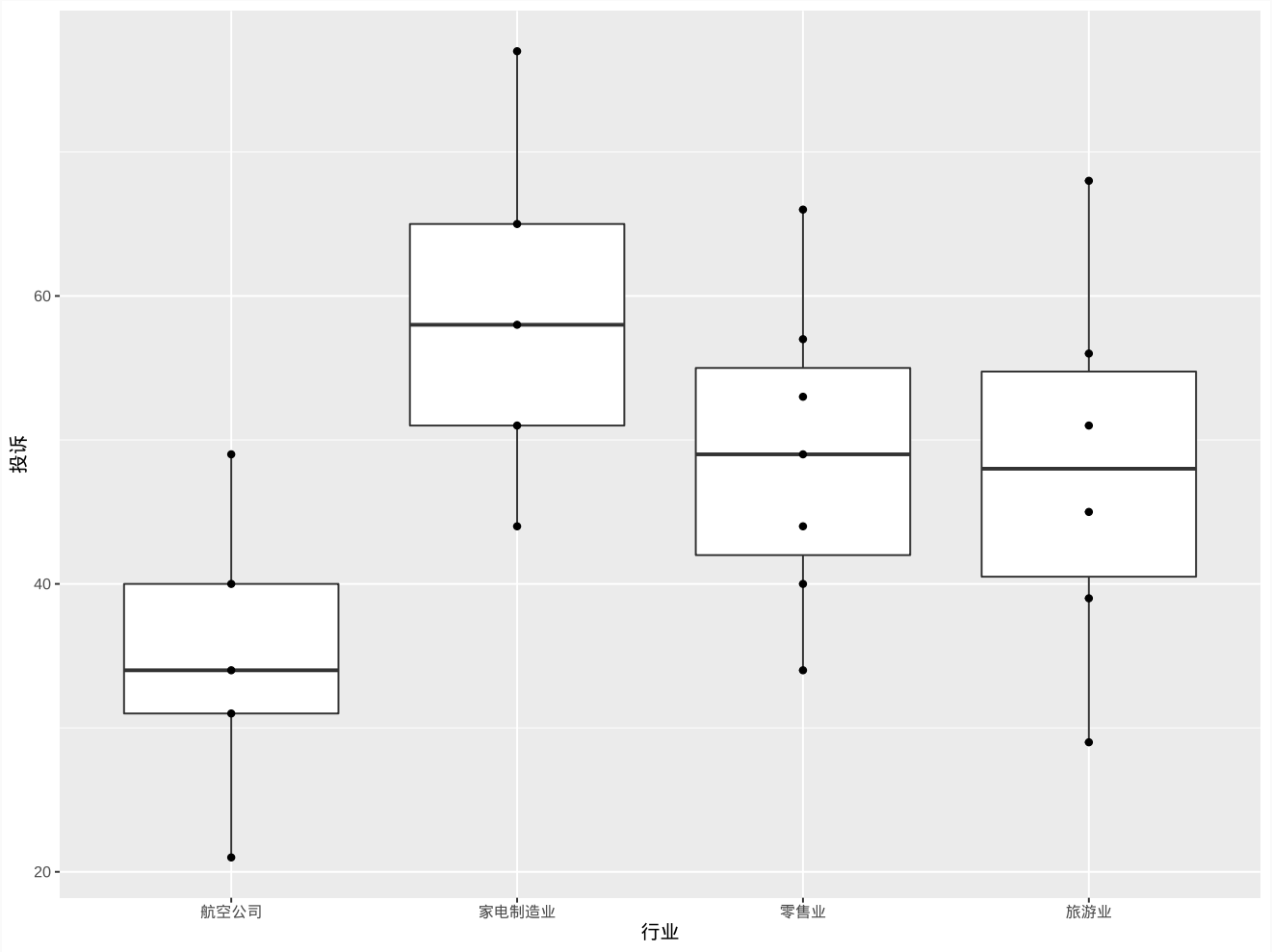
即检验这 4 个行业被投诉次数的均值是否相等：

- 若均值相等，则意味着 “行业” 对投诉次数是没有影响的，即它们之间的服务质量没有显著差异
- 若均值不全相等，则意味着 “行业” 对投诉次数是有影响的，它们之间的服务质量有显著差异

- **因素（factor）**：要检验的对象或要考虑的影响因素，本例是“行业”；
- **水平或处理（treatment）**：因素的不同取值，本例是“零售业、旅游业、航空公司、家电制造业”，因素的每一个水平，可看作是一个总体；
- **观测值**：每个因素水平下的样本数据，本例是投诉次数，可以看作是从 4 个总体中抽取的样本数据。

```
complaints = complaints %>%
  pivot_longer(-id, names_to = " 行业", values_to = " 投诉",
               values_drop_na = TRUE) %>%
  mutate(行业 = factor(行业))
```

```
complaints %>%
  ggplot(aes(行业, 投诉)) +
  geom_boxplot() +
  geom_point()
```


60
40
20
投诉
航空公司
家电制造业
零售业
旅游业
行业

可见，同一行业，不同企业投诉次数有差异（**组内随机差异**）；不同行业之间也有差异（**组间因素差异**）。

方差分析对数据的要求：满足**正态性**（各组分别来自正态总体）和**方差齐性**(各组方差相等)，在这两个条件下，若各组有差异，则只可能是来自影响因素的不同水平（下均值不同）。

方差分析假定每一个观测值都由若干部分累加而成，即总的效应可分解为若干部分，每一部分都有特定含义，称为**效应的可加性。**

数据的总差异，用**总离均差平方和**表示。根据效应的可加性，它可以分为：

- 组内随机差异，用**组内离均差平方和**表示
- 组间因素差异，用**组间离均差平方和**表示

总自由度也相应地分成组内和组间自由度，组内、组间离均差平方和除以其自由度得到组内、组间的**均方**（Mean Square），两个均方之比服从 F 分布。

其它术语

以焦虑症的治疗疗效为例，一个因素是治疗方案，有 2 种治疗方案，即该因素有 2 个水平；（治疗方案称为**组间因素**，因为每个患者只能被分配到一个组别中，没有患者同时接受两种治疗）；再考虑一个因素治疗时间，也有两个水平：治疗 5 周和治疗 6 个月，同一患者在 5 周和 6 个月不止一次地被测量（两次），称为**重复测量**（治疗时间称为**组内因素**，因为每个患者在所有水平下都进行了测量）。

建立方差分析模型时，既要考虑两个因素治疗方案和治疗时间（主效应），又要考虑治疗方案和时间的交互影响（交互效应），称为**两因素混合模型方差分析。**

当某个因素的各个水平下的因变量的均值呈现统计显著性差异时，必要时可作两两水平间的比较，称为**均值间的两两比较。**

二. 单因素方差分析

1 个因变量，1 个影响因素，其模型可表示为：

$$\text{总差异 } Y_{ij} = \text{平均差异 } \mu + \text{因素 1 差异 } \alpha_i + \text{随机差异 } \varepsilon_{ij}$$

以行业投诉数据为例，提出假设检验：

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4, \qquad H_1 : \mu_1, \mu_2, \mu_3, \mu_4 \text{ 不全相等}$$

- 正态性检验

```
library(rstatix)
complaints %>%
  group_by(行业) %>%
  shapiro_test(投诉)
#> # A tibble: 4 x 4
#>   行业       variable statistic     p
#>   <fct>      <chr>        <dbl> <dbl>
#> 1 航空公司   投诉         0.995 0.994
#> 2 家电制造业 投诉         0.986 0.963
#> 3 零售业     投诉         0.993 0.997
#> # ... with 1 more row
```

- 方差齐性检验

```
complaints %>%
  levene_test(投诉 ~ 行业)
#> # A tibble: 1 x 4
#>     df1   df2 statistic     p
#>   <int> <int>     <dbl> <dbl>
#> 1     3    19     0.197 0.897
```

1. 手动计算 ANOVA 表

数据按行业分为 $k = 4$ 组，第 j 组的样本数记为 n_j.

计算分组均值和总均值：

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}, \qquad \bar{x} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j$$

```
xj = tapply(complaints$投诉, complaints$行业, mean)
xj
#>   航空公司 家电制造业     零售业     旅游业
#>         35         59         49         48
mu = mean(complaints$投诉)
mu
#> [1] 47.9
```

- 计算总离均差平方和、组间离均差平方和、组内离均差平方和：

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

$$SSA = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

三者满足：$SST = SSA + SSE.$

```r
SST = sum((complaints$投诉 - mu) ^ 2)
SST              # 总离均差平方和
#> [1] 4165
nj = table(complaints$行业) %>% as.vector()
SSA = sum(nj * (xj - mu) ^ 2)
SSA              # 组间离均差平方和
#> [1] 1457
complaints = arrange(complaints, 行业)
SSE = sum((complaints$投诉 - rep(xj, times = nj)) ^ 2)
SSE              # 组内离均差平方和
#> [1] 2708
```

自由度：

- SST 的自由度为 $n-1$，其中 n 为全部观测值的个数
- SSA 的自由度为 $k-1$，其中 k 为因素水平（总体）的个数
- SSE 的自由度为 $n-k$

计算均方：离差平方和除以相应自由度

```
n = 23
k = 4
(MSA = SSA / (k-1))     # 组间均方
#> [1] 486
(MSE = SSE / (n-k))     # 组内均方
#> [1] 143
```

- 计算 F 统计量：组间均方除以组内均方

```
Fstat = MSA / MSE
Fstat
#> [1] 3.41
```

该 F 统计量服从 $F(k-1, n-k)$ 分布，计算右侧检验的临界值和 p 值：

```
alpha = 0.05
qf(1-alpha, k-1, n-k)          # 临界值
#> [1] 3.13
1 - pf(Fstat, k-1, n-k)        # p 值
#> [1] 0.0388
```

上述结果通常整理在一个如下的方差分析表中：

误差来源	平方和 (SS)	自由度 (df)	均方 (MS)	F值	P值	F 临界值
组间 (因素影响)	*SSA*	*k*-1	*MSA*	$\frac{MSA}{MSE}$		
组内 (误差)	*SSE*	*n-k*	*MSE*			
总和	*SST*	*n*-1				

- 生成方差分析表：

```
anova(lm(投诉 ~ 行业, complaints))
#> Analysis of Variance Table
#>
#> Response: 投诉
#>           Df Sum Sq Mean Sq F value Pr(>F)
#> 行业       3   1457     486    3.41  0.039 *
#> Residuals 19   2708     143
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

2. 直接用 rstatix::anova_test()

```
complaints %>%
  anova_test(投诉 ~ 行业)
#> ANOVA Table (type II tests)
#>
#>   Effect DFn DFd    F     p p<.05  ges
#> 1   行业   3  19 3.41 0.039     * 0.35
```

注: 设置参数 type = 3, 结果同 SPSS.

3. 直接用 bruceR::MANOVA()

```
library(bruceR)
MANOVA(complaints, dv = " 投诉", between = " 行业",
       file = "temp.docx")
#>
#> ====== ANOVA (Between-Subjects Design) ======
#>
#> Descriptives:
#> ──────────────────────────────────
#>       " 行业"   Mean     S.D. n
#> ──────────────────────────────────
#>     航空公司 35.000 (10.416) 5
#>   家电制造业 59.000 (12.748) 5
#>       零售业 49.000 (10.801) 7
#>       旅游业 48.000 (13.594) 6
#> ──────────────────────────────────
```

ANOVA Table:

Dependent variable(s): 投诉

Between-subjects factor(s): 行业

Within-subjects factor(s): –

Covariate(s): –

	MS	*MSE*	df_1	df_2	*F*	*p*		η^2_p [90% CI]	η^2_G
行业	485.536	142.526	3	19	3.407	.039	*	.350[.014, .544]	.350

Note. MSE = Mean Square Error.

Descriptive Statistics:

"行业"	*M*	*SD*	*n*
航空公司	35.000	10.416	5
家电制造业	59.000	12.748	5
零售业	49.000	10.801	7
旅游业	48.000	13.594	6

Total sample size: $N = 23$

4. 两两比较

通过对总体均值之间的**两两比较**来进一步检验到底哪些均值之间存在差异。

若要做 Tukey'HSD 组间的两两比较（多重比较）：

```
tukey_hsd(complaints, 投诉 ~ 行业)
#> # A tibble: 6 x 9
#>   term  group1 group2 null.value estimate conf.low conf.hi
#> * <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>     <db
#> 1 行业  航空 ~  家电 ~         0     24.0     2.77
#> 2 行业  航空 ~  零售业         0     14.0    -5.66       3
#> 3 行业  航空 ~  旅游业         0     13.0    -7.33       3
#> # ... with 3 more rows
```

三. 两因素方差分析

1 个因变量，2 个影响因素，其模型可表示为：

$$\text{总差异 } Y_{ijk} = \text{平均差异 } \mu + \text{因素 1 差异 } \alpha_i + \text{因素 2 差异 } \beta_j + \text{因素 1,2 交互作用差异 } \gamma_{ij} + \text{随机差异 } \varepsilon_{ijk}$$

原理是类似的，只是构建方差分析表的过程分的更细化（略）。

案例 2: 豚鼠牙齿生长数据

```
df = as_tibble(ToothGrowth) %>%
  mutate(dose = factor(dose))
df
#> # A tibble: 60 x 3
#>     len supp  dose
#>   <dbl> <fct> <fct>
#> 1   4.2 VC    0.5
#> 2  11.5 VC    0.5
#> 3   7.3 VC    0.5
#> # ... with 57 more rows
```

牙齿长度（len）为因变量，关于喂食方法（supp）和剂量（dose）做两因素混合模型方差分析。

- 检验正态性

```
df %>%
  group_by(supp, dose) %>%
  shapiro_test(len)
#> # A tibble: 6 x 5
#>   supp  dose  variable statistic     p
#>   <fct> <fct> <chr>        <dbl> <dbl>
#> 1 OJ    0.5   len          0.893 0.182
#> 2 OJ    1     len          0.927 0.415
#> 3 OJ    2     len          0.963 0.815
#> # ... with 3 more rows
```

- 检验方差齐性

```
df %>%
  group_by(supp) %>%
  levene_test(len ~ dose)
#> # A tibble: 2 x 5
#>   supp    df1   df2 statistic     p
#>   <fct> <int> <int>     <dbl> <dbl>
#> 1 OJ        2    27      1.84 0.178
#> 2 VC        2    27      2.17 0.134
```

- 两因素混合模型方差分析

```
anova_test(df, len ~ supp * dose)
#> ANOVA Table (type II tests)
#>
#>       Effect DFn DFd     F         p p<.05   ges
#> 1       supp   1  54 15.57 2.31e-04     * 0.224
#> 2       dose   2  54 92.00 4.05e-18     * 0.773
#> 3 supp:dose   2  54  4.11 2.20e-02     * 0.132
## bruceR 包实现
# MANOVA(df, dv = "len", between = c("supp", "dose"))
```

说明：`len ~ supp * dose` 是设定模型公式，遵从 R 的 formula 语法，~ 左边是因变量，右边是自变量公式，`supp * dose` 是 `supp + dose + supp:dose` 的简写，`supp:dose` 表示这两个变量的交互项。

方差分析结果的主效应 supp 和 dose 都非常显著（P 值都远小于 0.05），交互效应也显著（P 值 =0.022<0.05），表明 supp 和 dose 的协同变化下的各组均值显著不同[1]。

若要做 Tukey'HSD 组间的两两比较（多重比较）：

```
tukey_hsd(df, len ~ supp * dose)
#> # A tibble: 19 x 9
#>   term  group1 group2 null.value estimate conf.low conf.hi
#> * <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>     <db
#> 1 supp  OJ     VC              0    -3.70    -5.58     -1.
#> 2 dose  0.5    1               0     9.13     6.36     11.
#> 3 dose  0.5    2               0    15.5     12.7      18.
#> # ... with 16 more rows, and 1 more variable: p.adj.signif
```

[1]若交互作用不显著，可以只做去掉交互效应的方差分析.

四. 重复测量方差分析

方差分析要求观测之间相互独立，而重复测量数据是在分组因素之外，分别在组内不同的时间点上重复测量同一个体获得因变量的观测值，或者是通过重复测量同一个体的不同部位获得因变量的观测值。这就不再具有相互独立性，需要专门方法来处理，称为**重复测量方差分析**。

重复测量数据，常用来分析因变量在不同时间点上的变化性。分析前需要对重复测量数据之间是否存在相关性进行球形检验，若 P 值 <0.05 则说明存在相关性，应该做重复测量方差分析。

重复测量方差分析的模型公式一般形式为：

$$Y \sim B * W + \text{Error}(\text{Subject}/\text{W})$$

其中，B 为组间因素，W 为组内因素，Subject 为个体标记。

改造豚鼠数据：相当于 10 只豚鼠, 每只重复测量 6 次

```
df = df %>%
  mutate(ID = rep(1:10, 6))
```

```r
df %>%
  anova_test(len ~ supp * dose + Error(ID / (supp * dose)))
#> ANOVA Table (type III tests)
#>
#> $ANOVA
#>      Effect DFn DFd      F        p p<.05   ges
#> 1      supp   1   9  34.87 2.28e-04     * 0.224
#> 2      dose   2  18 106.47 1.06e-10     * 0.773
#> 3 supp:dose   2  18   2.53 1.07e-01       0.132
#>
#> $`Mauchly's Test for Sphericity`
#>      Effect     W     p p<.05
#> 1      dose 0.807 0.425
#> 2 supp:dose 0.934 0.761
#>
#> $`Sphericity Corrections`
#>      Effect   GGe      DF[GG]    p[GG] p[GG]<.05  HFe
#> 1      dose 0.838 1.68, 15.09 3.79e-09     * 1.01 3.02
```

球形检验结果表明，重复测量数据存在相关性，两个主效应都很显著，交互效应不显著。

注：重复测量方差分析也要求满足方差齐性，若不满足可以考虑用`lme4::lmer()`拟合混合效应模型。

若方差分析的前提：正态性、方差齐性不满足，则可以用相应的非参数检验：

- (独立) `kruskal_test()`: Kruskal-Wallis 秩和检验
- (相关) `friedman_test()`: Friedman 检验

本篇主要参阅 (张敬信, 2022), (贾俊平, 2018), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

贾俊平 (2018). 统计学. 中国人民大学出版社, 北京, 7 edition.

黄湘云 (2021). *Github: R-Markdown-Template*.