

基于 R 语言的基础统计分析方法

t 检验、ANOVA 与线性回归

郭书谏

同济大学

guoshujian@tongji.edu.cn

2025年12月17日

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

- ① 独立样本 t 检验
- ② 配对样本 t 检验
- ③ ANOVA 方差分析
- ④ 简单线性回归
- ⑤ 多元线性回归

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

1. 独立样本 t 检验 (Independent Samples t-test)

原理

用于检验两个独立组别的均值是否存在显著差异。

案例：汽车油耗分析

数据集： R 内置 mtcars 数据集。

目标： 比较自动挡 (am=0) 和手动挡 (am=1) 汽车的油耗 (mpg) 是否有显著差异。

假设：

- H_0 : 两种变速箱类型的平均油耗相等。
- H_1 : 两种变速箱类型的平均油耗不相等。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

独立样本 t 检验：R 语言实现

</> R Code: Independent t-test

```
1 # 加载数据系统内置 ()
2 data(mtcars)
3
4 # 假设方差齐性 (var.equal = TRUE) 进行 t 检验
5 # 公式：响应变量 ~ 分组变量
6 result_t <- t.test(mpg ~ am, data = mtcars, var.equal = TRUE)
7
8 # 查看结果
9 print(result_t)
```

- mpg: Miles/(US) gallon (油耗)
- am: Transmission (0 = automatic, 1 = manual)
- 如果 p-value < 0.05, 则拒绝原假设, 认为油耗有显著差异。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

2. 配对样本 t 检验 (Paired Samples t-test)

原理

用于检验同一受试对象在处理前后，或配对组之间的均值差异。

案例：药物睡眠效果

数据集： R 内置 sleep 数据集。

目标： 比较两种药物 (group 1 vs 2) 对同一组病人 (ID) 增加睡眠时间 (extra) 的影响。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

配对样本 t 检验：R 语言实现

</> R Code: Paired t-test

```
1 data(sleep)
2
3 # 关键参数: paired = TRUE
4 # 此时 R 会自动根据 ID 匹配观测值
5 result_paired <- t.test(extra ~ group,
6                           data = sleep,
7                           paired = TRUE)
8
9 print(result_paired)
```

- extra: 睡眠时间的增加量
- group: 药物组别 (1 或 2)
- paired = TRUE: 告诉 R 这是一个配对设计。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

3. 单因素方差分析 (One-way ANOVA)

原理

用于比较三个或更多组别的均值是否存在显著差异。

案例：植物生长实验

数据集： R 内置 PlantGrowth 数据集。

目标： 比较对照组 (ctrl) 和两个处理组 (trt1, trt2) 的植物干重 (weight) 是否有差异。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

</> R Code: One-way ANOVA

```
1 data(PlantGrowth)
2
3 # 使用 aov() 函数
4 # 公式: 连续变量 ~ 分类因子
5 fit_aov <- aov(weight ~ group, data = PlantGrowth)
6
7 # 使用 summary() 查看方差分析表
8 summary(fit_aov)
9
10 # 如果显著, 可以进行事后检验例如 ( TukeyHSD)
11 # TukeyHSD(fit_aov)
```

- 查看 Pr(>F) 列。如果值 < 0.05 , 说明至少有一组与其他组存在显著差异。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

4. 简单线性回归 (Simple Linear Regression)

原理

研究一个连续型自变量和一个连续型因变量之间的线性关系。

$$Y = \beta_0 + \beta_1 X + \epsilon$$

案例：汽车刹车距离

数据集：R 内置 cars 数据集 (1920年代的数据)。

目标：研究汽车速度 (speed) 对刹车距离 (dist) 的影响。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

简单线性回归：R 语言实现

</> R Code: Linear Regression

```
1 data(cars)
2
3 # 使用 lm() 函数 (Linear Model)
4 # 公式: 因变量 ~ 自变量
5 fit_lm <- lm(dist ~ speed, data = cars)
6
7 # 查看回归系数、R-squared 和 p-value
8 summary(fit_lm)
9
10 # 简单的绘图代码仅供参考 ()
11 # plot(cars$speed, cars$dist)
12 # abline(fit_lm, col="red")
```

- 重点关注 Coefficients 表中的 Estimate (斜率) 和 Pr(>|t|) (显著性)。
- Multiple R-squared: 模型解释了多少数据的变异。

(c) <http://shuangguo.com> | guoshujian@tongji.edu.cn

5. 多元线性回归 (Multiple Linear Regression)

原理

研究一个因变量与多个自变量（包含连续型和分类型）之间的线性关系。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

案例：综合因素对油耗的影响

数据集：R 内置 mtcars 数据集。

目标：研究车重 (wt)、马力 (hp) 和 变速箱类型 (am, 分类变量) 共同对油耗 (mpg) 的影响。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

多元线性回归：R 语言实现

</> R Code: Multiple Linear Regression

```
1 data(mtcars)
2
3 # 将 am 转换为因子分类变量 ()
4 # 这是处理分类变量的关键步骤
5 mtcars$am <- factor(mtcars$am, labels = c("Auto", "Manual"))
6
7 # 建立多元回归模型
8 # 公式: mpg ~ wt + hp + am
9 fit_multi <- lm(mpg ~ wt + hp + am, data = mtcars)
10
11 # 查看结果
12 summary(fit_multi)
```

- R 会自动将分类变量处理为虚拟变量 (Dummy Variables)。
- Estimate 显示在其他变量保持不变时，某变量每增加一个单位对 mpg 的平均影响。
- Adjusted R-squared: 模型拟合优度。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

谢谢观看！

欢迎提问

guoshujian@tongji.edu.cn

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn