ANOVA 单因素方差分析

前置知识

定义

- **因子（factor）**：待研究的自变量
- **因子水平（factor level）**：因子的具体取值（记为 $A_1, A_2, ..., A_r$）

去解释这个定义，可看下面两个example：

eg1：调查问卷颜色与回收率

- 因变量 Y：回收率
- 因子 A：颜色
- 因子水平：
    - A_1：红色
    - A_2：绿色

eg2：商品定价与销售量

- 因变量 Y：销售量
- 因子 A：定价
- 因子水平：
  - A_1：50元
  - A_2：30元

Model

观测数据表示

- 因子 A 影响因变量 Y
- 因子水平 A_i（$i = 1, 2, ..., r$）下的观测值：$y_{i1}, y_{i2}, ..., y_{in_i}$
  - n_i：A_i 水平下的观测次数（各水平次数可不同）
  - 总观测次数：$n = \sum_{i=1}^{r} n_i$
  - y_{ij}：A_i 水平下的第 j 次观测值

假设

- 同一因子水平下的观测差异较小，均值近似相同；**不同因子水平下的均值存在差异**（即 μ_i 互不相同）
- 数据模型：$y_{ij} = \mu_i + \varepsilon_{ij}$
  - μ_i：A_i 水平下的总体均值
  - ε_{ij}：随机误差项
- ε_{ij} 之间**互相独立**
- ε_{ij} 服从**同分布**：$\varepsilon_{ij} \sim N(0, \sigma^2)$（正态分布，均值0，方差 σ^2）

参数与分析目标

- 待估参数：$\mu_1, ..., \mu_r$（各水平均值）、σ^2（误差方差）
- 分析目标：

1. 估计（estimate）：点估计（point estimate）、置信区间（CI）
2. 假设检验（hypothesis testing）

ANOVA 与线性回归模型的联系

- 通用回归形式：$Y = X\beta + \varepsilon$
  - X：连续型变量（自变量）
  - $\mu(x) \triangleq E[Y|X = x] = X\beta$：给定 $X = x$ 时 Y 的条件期望

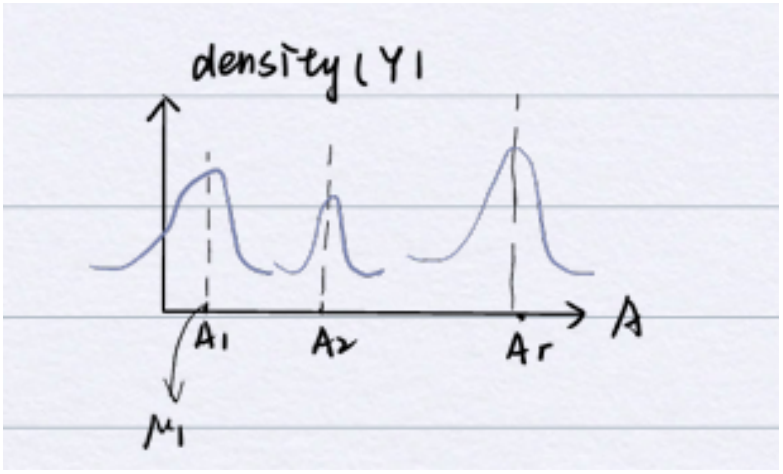
ANOVA可转化为**含指示变量的回归模型**（X 为非连续型）：

- **变量含义**：
  - Y（$n \times 1$）：观测值向量（如 $\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{rr} \end{bmatrix}$）
  - X：指示矩阵（列对应因子水平 $A_1, A_2, ..., A_r$，某行对应水平时该列取1，其余取0）
  - β（$r \times 1$）：各水平的总体均值 $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix}$
  - ε：随机误差项

$$
\underbrace{\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{r1} \\ \vdots \\ y_{rn_r} \end{bmatrix}}_{Y\ (n\times 1)} = \underbrace{\begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_r \\ \vdots \\ \mu_r \end{bmatrix}}_{\text{均值向量}} + \varepsilon = \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}}_{X\ \text{(非连续指示矩阵)}} \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix}}_{\text{水平均值向量}} + \varepsilon
$$

- X 的作用：通过“指示列”标记观测值所属的因子水平（如前 n_1 行对应 A_1，第1列全为1）
- 回归模型视角：$\mu(x)$ 是 X 的函数（连续型 X 对应平滑曲线）

- ANOVA视角：不同因子水平 A_i 对应独立的分布（各分布均值为 μ_i）



参数点估计 μ_i~$\hat{\mu}_i$

some notation

- y_{ij}：第 i 个因子水平下的第 j 次观测值（$i=1,2,...,r$；$j=1,2,...,n_i$）
- $y_{i\cdot}=\sum_{j=1}^{n_i} y_{ij}$：第 i 水平下的观测值总和
- $\bar{y}_{i\cdot}=\frac{1}{n_i}y_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$：第 i 水平下的样本均值
- $y_{\cdot\cdot}=\sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij}=\sum_{i=1}^{r} y_{i\cdot}$：所有观测值的总和
- $\bar{y}_{\cdot\cdot}=\frac{1}{n}y_{\cdot\cdot}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij}$（$n=\sum_{i=1}^{r} n_i$）：总样本均值

1. 最小二乘估计（LSE）

- **目标**：最小化残差平方和 $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2$
  （在ANOVA模型中的估计 $\hat{y}_{ij}=\hat{\mu}_i$，即第 i 水平的拟合值为其均值）
- 转化为：$\hat{\mu}_i=\arg\min_{\mu_i}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$
- **求导求解**：对 μ_i 求偏导并令其为0：

$$\frac{\partial}{\partial\mu_i}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2=-2\sum_{j=1}^{n_i}(y_{ij}-\mu_i)=0$$

解得：$\hat{\mu}_i^{\text{LSE}}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}=\bar{y}_{i\cdot}$

2. 最优线性无偏估计（BLUE）

- **无偏性验证**：设线性估计为 $\hat{\mu}_i=\sum_{j=1}^{n_i}d_jy_{ij}$（$d_j$ 为常数），则：

$$E[\hat{\mu}_i]=\sum_{j=1}^{n_i}d_jE[y_{ij}]=\mu_i\sum_{j=1}^{n_i}d_j$$

要求无偏，需 $\sum_{j=1}^{n_i}d_j=1$。

- **方差最小化**：在约束 $\sum_{j=1}^{n_i}d_j=1$ 下，最小化 $\text{Var}(\hat{\mu}_i)=\sigma^2\sum_{j=1}^{n_i}d_j^2$。
- **通过拉格朗日对偶问题求解**：最优解为 $d_j=\frac{1}{n_i}$，即 $\hat{\mu}_i^{\text{BLUE}}=\bar{y}_{i\cdot}$（与LSE一致）。

3. 极大似然估计（MLE）

- **假设**：$y_{i1},y_{i2},\ldots,y_{in_i}$ 独立同分布于 $N(\mu_i,\sigma^2)$。
- **联合概率密度（pdf）**：

$$f(y_{i1},\ldots,y_{in_i})=\frac{1}{(\sqrt{2\pi}\sigma)^{n_i}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2\right\}$$

- **极大似然目标**：$\hat{\mu}_i^{\text{MLE}}=\arg\max_{\mu_i}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$
- 解得：$\hat{\mu}_i^{\text{MLE}}=\bar{y}_{i\cdot}$（与LSE、BLUE一致）

$\hat{\mu}_i$ 性质

- 无偏性

$$E[\hat{\mu}_i] = \mu_i$$

- 方差

$$\mathrm{Var}(\bar{y}_{i\cdot}) = \mathrm{Var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\right) = \frac{1}{n_i^2}\sum_{j=1}^{n_i}\sigma^2 = \frac{\sigma^2}{n_i}$$

- 分布（正态性）

$$\hat{\mu}_i \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$$

（注：因 $\hat{\mu}_i = \bar{y}_{i\cdot}$ 是正态变量的线性组合，故服从正态分布）

参数点估计σ^2~$\hat{\sigma}^2$

MLE

因为假设如下

- 各水平下的观测独立且服从正态分布：$(y_{i1}, y_{i2}, ..., y_{in_i}) \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$ $(i = 1, 2, ..., r)$
- 误差项 ε_{ij} 互相独立

所以我们可以得到联合概率分布

总样本量 $n = \sum_{i=1}^{r} n_i$，样本的**联合概率密度**为：

$$L(\sigma^2, \mu_1, ..., \mu_r) = \prod_{i=1}^{r}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\mu_i)^2\right\}$$

取**对数似然**：

$$\ell = \log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$$

对 σ^2 求导并令导数为0：

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = 0$$

整理得 σ^2 的MLE：

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$$

代入 μ_i 的MLE（$\hat{\mu}_i = \bar{y}_{i\cdot}$），最终：

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$$

- 通过一个基础的重要统计结论可得

$$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2 \sim (n_i-1)\sigma^2\chi^2(n_i-1)$$

- 证明思路

  设 $X=(x_1,...,x_n)^T \overset{\text{iid}}{\sim} N(\mu,\sigma^2)$，令 $Y = X - \mu\mathbf{1}_n$（$\mathbf{1}_n$ 为全1向量），则 $\bar{Y} = \bar{X} - \mu$。

  样本离均差平方和可表示为：

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \|Y-\bar{Y}\mathbf{1}_n\|_2^2 = Y^T\left(I-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)Y$$

  其中矩阵 $I-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ 是**对称幂等矩阵**，秩为 $n-1$，因此：

$$\|Y-\bar{Y}\mathbf{1}_n\|_2^2 \sim \sigma^2\chi^2(n-1)$$

推广到第 i 水平的观测，即得结论。

MLE的有偏性

计算 $\hat{\sigma}^2_{\text{MLE}}$ 的期望：

$$E\left[\hat{\sigma}^2_{\text{MLE}}\right] = \frac{1}{n}\sum_{i=1}^{r} E\left[\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2\right] = \frac{1}{n}\sum_{i=1}^{r}(n_i-1)\sigma^2 = \frac{n-r}{n}\sigma^2$$

可见 $\hat{\sigma}^2_{\text{MLE}}$ 是**有偏估计**（Biased）。

修正后的无偏估计

对MLE修正（除以自由度 $n-r$），得到无偏估计：

$$\hat{\sigma}^2 = \frac{n}{n-r}\cdot\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$$

也可表示为各水平样本方差的加权平均：

$$\hat{\sigma}^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2 \quad \left(S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2\right)$$

分布性质

因各水平的 $(n_i-1)S_i^2 \sim \sigma^2\chi^2(n_i-1)$ 且互相独立，根据 χ^2 分布的可加性：

$$\hat{\sigma}^2 = \frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2 \sim \frac{\sigma^2}{n-r}\chi^2(n-r)$$

置信区间 μ_i

根据μ_i的性质：

- 点估计：$\hat{\mu}_i = \bar{y}_{i\cdot}$（第 i 水平的样本均值）
- 分布性质：$\bar{y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$

由正态分布的标准化性质：

$$\frac{\bar{y}_{i\cdot}-\mu_i}{\sigma/\sqrt{n_i}}\sim N(0,1)$$

但 σ 是未知参数，需用其**无偏估计** s（$s=\sqrt{\hat{\sigma}^2}$，其中 $\hat{\sigma}^2=\frac{1}{n-r}\sum_{i=1}^{r}(n_i-1)S_i^2$）替代，得到：

$$\frac{\bar{y}_{i\cdot}-\mu_i}{\sigma/\sqrt{n_i}}\cdot\frac{\sigma}{s}=\frac{\bar{y}_{i\cdot}-\mu_i}{s/\sqrt{n_i}}$$

该统计量服从 **t分布**，但需先证明 $\bar{y}_{i\cdot}$ 与 s 独立。

$\bar{y}_{i\cdot}$ 与 s 独立的证明

要证 $\bar{y}_{i\cdot}\perp s$（对所有 i），需分两步：

1. $s^2=\frac{1}{n-r}\sum_{j=1}^{r}(n_j-1)S_j^2$，其中 S_j^2 是第 j 水平的样本方差；
2. 对 $\bar{y}_{i\cdot}$ 而言，$S_j^2\perp\bar{y}_{i\cdot}$（$j\neq i$）显然成立，因此只需证 $S_i^2\perp\bar{y}_{i\cdot}$。

下面也是一个经典的结论及证明

设 $Y_i=(y_{i1},y_{i2},...,y_{in_i})^T$（第 i 水平的观测向量），则：

$$y_{ij}-\bar{y}_{i\cdot}=e_j^TY_i-\frac{1}{n_i}\mathbf{1}_{n_i}^TY_i=\left(e_j^T-\frac{1}{n_i}\mathbf{1}_{n_i}^T\right)Y_i$$

（e_j 是第 j 个分量为1的单位向量，$\mathbf{1}_{n_i}$ 是全1向量）

计算协方差：

$$\operatorname{cov}\left(y_{ij}-\bar{y}_{i\cdot},\bar{y}_{i\cdot}\right)=\left(e_j^T-\frac{1}{n_i}\mathbf{1}_{n_i}^T\right)\operatorname{cov}(Y_i)\cdot\frac{1}{n_i}\mathbf{1}_{n_i}^T$$

因 $\operatorname{cov}(Y_i)=\sigma^2I_{n_i}$，代入得：

$$\operatorname{cov}\left(y_{ij}-\bar{y}_{i\cdot},\bar{y}_{i\cdot}\right)=\sigma^2\cdot\frac{1}{n_i}\left(1-\frac{n_i}{n_i}\right)=0$$

由于 $y_{ij}-\bar{y}_{i\cdot}$ 与 $\bar{y}_{i\cdot}$ 均为正态变量，**协方差为0等价于独立**，因此 $S_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$ 与 $\bar{y}_{i\cdot}$ 独立。

综上，$\bar{y}_{i\cdot}\perp s$，故：

$$\frac{\bar{y}_{i\cdot}-\mu_i}{s/\sqrt{n_i}}\sim t(n-r)$$

所以最终可得到置信区间估计

对置信水平 $1-\alpha$，由t分布的分位数性质：

$$P\left(\left|\frac{\bar{y}_{i\cdot}-\mu_i}{s/\sqrt{n_i}}\right|\leq t_{n-r}\left(\frac{\alpha}{2}\right)\right)=1-\alpha$$

整理得 μ_i 的置信区间：

$$\left[\bar{y}_{i\cdot}-\frac{s\cdot t_{n-r}\left(\frac{\alpha}{2}\right)}{\sqrt{n_i}},\bar{y}_{i\cdot}+\frac{s\cdot t_{n-r}\left(\frac{\alpha}{2}\right)}{\sqrt{n_i}}\right]$$

置信区间$\mu_i-\mu_j$（$i\neq j$）

记两水平的均值差为 $D_{ij}\triangleq\mu_i-\mu_j$，其**点估计**为样本均值差：

$$\hat{D}_{ij}=\bar{y}_{i\cdot}-\bar{y}_{j\cdot}$$

由 $\bar{y}_{i\cdot}\sim N\left(\mu_i,\frac{\sigma^2}{n_i}\right)$、$\bar{y}_{j\cdot}\sim N\left(\mu_j,\frac{\sigma^2}{n_j}\right)$，且不同水平的观测独立，得：

$$\hat{D}_{ij}\sim N\left(\mu_i-\mu_j,\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n_j}\right)$$

对 $\hat{D}_{ij}$ 标准化（利用正态分布性质）：

$$\frac{\hat{D}_{ij}-D_{ij}}{\sigma\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}}\sim N(0,1)$$

因 σ 未知，用其无偏估计 $s=\sqrt{\hat{\sigma}^2}$（$\hat{\sigma}^2=\frac{1}{n-r}\sum_{k=1}^{r}(n_k-1)S_k^2$）替代，需先验证 $s\perp\hat{D}_{ij}$：

- 由之前的结论，$s\perp\bar{y}_{i\cdot}$ 且 $s\perp\bar{y}_{j\cdot}$，故 $s\perp\bar{y}_{i\cdot}-\bar{y}_{j\cdot}=\hat{D}_{ij}$。

因此，标准化统计量服从 **t分布**：

$$\frac{\hat{D}_{ij} - D_{ij}}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t(n-r)$$

对置信水平 $1-\alpha$，由t分布的分位数性质：

$$P\left(\left|\frac{\hat{D}_{ij} - D_{ij}}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}\right| \leq t_{n-r}\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

整理得 $D_{ij} = \mu_i - \mu_j$ 的置信区间：

$$\hat{D}_{ij} \pm s \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \cdot t_{n-r}\left(\frac{\alpha}{2}\right)$$

即：

$$\left[\bar{y}_{i\cdot} - \bar{y}_{j\cdot} - s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \cdot t_{n-r}\left(\frac{\alpha}{2}\right),\ \bar{y}_{i\cdot} - \bar{y}_{j\cdot} + s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \cdot t_{n-r}\left(\frac{\alpha}{2}\right)\right]$$

置信区间：线性组合 $L = \sum_{i=1}^{r} c_i\mu_i$

设 $L = \sum_{i=1}^{r} c_i\mu_i$（$c_i$ 为已知常数），其**点估计**为样本均值的线性组合：

$$\hat{L} = \sum_{i=1}^{r} c_i\bar{y}_{i\cdot}$$

已知 $\bar{y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$，且不同水平的 $\bar{y}_{i\cdot}$ 相互独立，因此 $\hat{L}$ 作为正态变量的线性组合，服从正态分布：

$$\hat{L} \sim N\left(\sum_{i=1}^{r} c_i\mu_i,\ \sum_{i=1}^{r} c_i^2 \cdot \frac{\sigma^2}{n_i}\right)$$

用向量形式表示更直观：

$$\hat{L} = (c_1,\ c_2,\ ...,\ c_r) \begin{pmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{r\cdot} \end{pmatrix} \sim N\left(\sum_{i=1}^{r} c_i \mu_i,\ \sum_{i=1}^{r} \frac{c_i^2 \sigma^2}{n_i} \right)$$

对 $\hat{L}$ 标准化（利用正态分布性质）：

$$\frac{\hat{L} - L}{\sqrt{\sum_{i=1}^{r} \frac{c_i^2 \sigma^2}{n_i}}} \sim N(0,1)$$

因 σ 未知，用其无偏估计 $s = \sqrt{\hat{\sigma}^2}$（$\hat{\sigma}^2 = \frac{1}{n-r} \sum_{i=1}^{r} (n_i - 1) S_i^2$）替代，且由之前的结论：

- $s^2 \perp \bar{y}_{i\cdot}$（对所有 i），故 $s^2 \perp \hat{L}$（$\hat{L}$ 是 $\bar{y}_{i\cdot}$ 的线性组合）。

因此，标准化统计量服从 **t分布**：

$$\frac{\hat{L} - L}{\sqrt{\sum_{i=1}^{r} \frac{c_i^2 \sigma^2}{n_i}}} \cdot \frac{\sigma}{s} = \frac{\hat{L} - L}{s \cdot \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}}} \sim t(n-r)$$

对置信水平 $1-\alpha$，由t分布的分位数性质：

$$P\left(\left| \frac{\hat{L} - L}{s \cdot \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}}} \right| \leq t_{n-r}\left(\frac{\alpha}{2}\right) \right) = 1 - \alpha$$

整理得 L 的置信区间：

$$\hat{L} \pm s \cdot t_{n-r}\left(\frac{\alpha}{2}\right) \cdot \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

即：

$$\left[\sum_{i=1}^{r} c_i \bar{y}_{i\cdot} - s \cdot t_{n-r}\left(\frac{\alpha}{2}\right) \cdot \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}},\ \sum_{i=1}^{r} c_i \bar{y}_{i\cdot} + s \cdot t_{n-r}\left(\frac{\alpha}{2}\right) \cdot \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}} \right]$$

单因素ANOVA的假设检验（$H_0：\mu_1 = \mu_2 = ... = \mu_r$）

总平方和的分解（SST）

总平方和（反映所有观测的总变异）：

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$$

将 $y_{ij} - \bar{y}_{\cdot\cdot}$ 拆分为 $(y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$，代入得：

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{n_i}[(y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})]^2$$

展开后交叉项为：

$$2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = 2\sum_{i=1}^{r}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) \cdot \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})$$

由于 $\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot}) = 0$，交叉项为0，因此总平方和分解为：

$$SST = \underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}_{S_e\text{（组内平方和）}} + \underbrace{\sum_{i=1}^{r} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{S_A\text{（组间平方和）}}$$

平方和的含义

- **组内平方和** S_e：反映同一水平内观测的随机误差（组内变异）。
- **组间平方和** S_A：反映不同水平均值的差异（组间变异，包含因子效应）。

因子效应的定义

设 $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{r} n_i\mu_i$（总均值），定义**主效应**：

$$a_i = \mu_i - \bar{\mu}$$

此时观测值可表示为：

$$y_{ij} = \mu_i + \varepsilon_{ij} = \bar{\mu} + a_i + \varepsilon_{ij}$$

组内离均差可表示为误差的离均差：

$$y_{ij} - \bar{y}_{i\cdot} = (\mu_i + \varepsilon_{ij}) - (\mu_i + \bar{\varepsilon}_{i\cdot}) = \varepsilon_{ij} - \bar{\varepsilon}_{i\cdot} \triangleq e_{ij}$$

其中 $\bar{\varepsilon}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}$，且 $\sum_{j=1}^{n_i} e_{ij} = 0$（对任意 i）。

组内平方和 S_e 的分布

由之前的结论，$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 \sim (n_i - 1)\sigma^2\chi^2(n_i - 1)$，且不同水平的 S_e 独立，因此：

$$S_e \sim \sigma^2\chi^2(n - r)$$

自由度 $df_e = n - r$（线性无约束）。

独立性

S_e 与 S_A 相互独立（组内变异与组间变异独立）。

均方定义

- 组内均方（误差均方）：$MSE = \frac{S_e}{n-r} = s^2$，是 σ^2 的**无偏估计**。
- 组间均方（因子均方）：$MSA = \frac{S_A}{r-1}$。

组间平方和 S_A 的分布与期望分析

组间平方和反映不同因子水平均值的差异，形式为：

$$S_A = \sum_{i=1}^{r} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

其中：

- $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$（第 i 水平的样本均值）；

- $\bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{y}_{i\cdot}$（总样本均值）；
- $n = \sum_{i=1}^{r} n_i$（总观测数）。

$E[S_A]$ 的推导

在**固定效应模型**（因子水平 A_i 是固定的）中，定义**主效应** $a_i = \mu_i - \bar{\mu}$（$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{r} n_i \mu_i$ 为总均值），则观测值可表示为：

$$y_{ij} = \mu_i + \varepsilon_{ij} = \bar{\mu} + a_i + \varepsilon_{ij}$$

其中 $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$（随机误差），且满足 $\sum_{i=1}^{r} n_i a_i = 0$（主效应的约束）。

将样本均值拆分为“主效应 + 误差均值”：

$$\bar{y}_{i\cdot} = \mu_i + \bar{\varepsilon}_{i\cdot} = \bar{\mu} + a_i + \bar{\varepsilon}_{i\cdot}$$

$$\bar{y}_{\cdot\cdot} = \bar{\mu} + \bar{\varepsilon}_{\cdot\cdot} \quad \left(\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{\varepsilon}_{i\cdot}\right)$$

因此：

$$\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} = a_i + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})$$

将 $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$ 代入 S_A，展开平方项：

$$S_A = \sum_{i=1}^{r} n_i \left[a_i + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})\right]^2 = \sum_{i=1}^{r} n_i a_i^2 + 2\sum_{i=1}^{r} n_i a_i (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}) + \sum_{i=1}^{r} n_i (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2$$

对各项取期望：

1. **第一项**：$E\left[\sum_{i=1}^{r} n_i a_i^2\right] = \sum_{i=1}^{r} n_i a_i^2$（$a_i$ 是固定常数）。
2. **第二项**：利用 误差期望为0，交叉项期望为 0。
3. **第三项**：需计算 $E\left[\sum_{i=1}^{r} n_i (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2\right]$。

利用误差的分布性质（$\bar{\varepsilon}_{i\cdot} \sim N\left(0, \frac{\sigma^2}{n_i}\right)$、$\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{\varepsilon}_{i\cdot}$），推导得：

$$E\left[\left(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}\right)^2\right]=\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}$$

因此误差项的总期望为：

$$E\left[\sum_{i=1}^{r} n_i\left(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}\right)^2\right]=\sum_{i=1}^{r} n_i\left(\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}\right)=(r-1)\sigma^2$$

最终 $E[S_A]$

综合以上结果，组间平方和的期望为：

$$E[S_A]=\sum_{i=1}^{r} n_i a_i^2+(r-1)\sigma^2$$

- 若原假设 H_0：$\mu_1=\mu_2=\ldots=\mu_r$ 成立（即 $a_i=0$），则：

$$E[S_A]=(r-1)\sigma^2$$

- 若 H_0 不成立（即存在 $a_i\neq 0$），则：

$$E[S_A]>(r-1)\sigma^2$$

H_0 成立时组间平方和 S_A 的分布

当原假设 H_0：$\mu_1=\mu_2=\cdots=\mu_r$ 成立时，$\bar{y}_{i\cdot}\sim N\left(\mu_0,\frac{\sigma^2}{n_i}\right)$，误差均值 $\bar{\varepsilon}_{i\cdot}=\bar{y}_{i\cdot}-\mu_0\sim N\left(0,\frac{\sigma^2}{n_i}\right)$。

组间平方和 S_A 的原始形式为：

$$S_A=\sum_{i=1}^{r} n_i\left(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot}\right)^2$$

定义新变量 $\mathcal{S}_i = \sqrt{n_i}\bar{\varepsilon}_{i\cdot}$，则 $\mathcal{S}_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$（独立同分布正态变量），且总误差均值 $\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r} n_i\bar{\varepsilon}_{i\cdot} = \frac{1}{n}\sum_{i=1}^{r}\sqrt{n_i}\mathcal{S}_i$。

将 $\bar{\varepsilon}_{i\cdot} = \frac{\mathcal{S}_i}{\sqrt{n_i}}$、$\bar{\varepsilon}_{\cdot\cdot} = \frac{1}{n}\sum_{j=1}^{r}\sqrt{n_j}\mathcal{S}_j$ 代入 S_A，变形得：

$$
\begin{aligned}
S_A &= \sum_{i=1}^{r} n_i\left(\frac{\mathcal{S}_i}{\sqrt{n_i}} - \frac{1}{n}\sum_{j=1}^{r}\sqrt{n_j}\mathcal{S}_j\right)^2 \\
&= \sum_{i=1}^{r}\left(\mathcal{S}_i - \sum_{j=1}^{r}\frac{\sqrt{n_i n_j}}{n}\mathcal{S}_j\right)^2
\end{aligned}
$$

将 S_A 表示为**正态向量的二次型**：
令 $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_r)^T \sim N(\mathbf{0}, \sigma^2 I_r)$（$I_r$ 为 r 阶单位矩阵），构造矩阵 A，其元素为：

$$
A_{ij} = \frac{\sqrt{n_i n_j}}{n}
$$

则 S_A 可写为：

$$
S_A = \mathcal{S}^T(I_r - A)\mathcal{S}
$$

经典投影矩阵结论：$I_r - A$ 的关键性质

1. **对称幂等性**：
   - 对称性：$(I_r - A)^T = I_r - A^T = I_r - A$（因 $A^T = A$）；
   - 幂等性：$(I_r - A)^2 = I_r - 2A + A^2 = I_r - A$（因 $A^2 = A$，可通过 A_{ij} 的定义验证）。
2. **秩为 $r-1$**：
   矩阵的迹 $\mathrm{tr}(I_r - A) = r - \mathrm{tr}(A)$，而 $\mathrm{tr}(A) = \sum_{i=1}^{r} A_{ii} = \sum_{i=1}^{r}\frac{n_i}{n} = 1$，故 $\mathrm{tr}(I_r - A) = r - 1$。
   幂等矩阵的秩等于其迹，因此 $\mathrm{rank}(I_r - A) = r - 1$。
3. **特征值仅为0或1**：
   幂等矩阵的特征值只能是0或1，结合秩为 $r-1$，可知 $I_r - A$ 有 $r-1$ 个特征值为1，1个特征值为0。

由于 $\mathcal{S} \sim N(\mathbf{0}, \sigma^2 I_r)$，且 $I_r - A$ 是秩为 $r-1$ 的对称幂等矩阵，根据**正态变量二次型的分布定理**：

$$
\frac{1}{\sigma^2}\mathcal{S}^T(I_r - A)\mathcal{S} \sim \chi^2(r-1)
$$

因此 S_A 的分布为：

$$S_A \sim \sigma^2 \chi^2(r-1)$$

H_0下 F检验的构造

结合组内平方和 $S_e \sim \sigma^2 \chi^2(n-r)$（且 $S_A \perp S_e$），构造F统计量：

$$F = \frac{MSA}{MSE} = \frac{S_A/(r-1)}{S_e/(n-r)}$$

在 H_0 下，$F \sim F(r-1, n-r)$（F分布，分子自由度 $r-1$，分母自由度 $n-r$）。

检验决策：若 $F \geq F_\alpha(r-1, n-r)$（$F_\alpha$ 为F分布的上 α 分位数），则**拒绝** H_0。

ANOVA表（单因素方差分析）

ANOVA表是单因素方差分析的核心结果汇总，包含**平方和、自由度、均方、F统计量**等关键信息，格式及含义如下：

变异来源	平方和（SS）	自由度（df）	均方（MS）	F统计量
总变异	SST	$df_T = n-1$	-	-
组间（因子）	S_A	$df_A = r-1$	$MSA = \frac{S_A}{df_A}$	$F = \frac{MSA}{MSE}$
组内（误差）	S_e	$df_e = n-r$	$MSE = \frac{S_e}{df_e}$	-

单因素ANOVA的“全模型-简化模型”框架

全模型（Full Model）的矩阵表示

单因素ANOVA的**全模型**描述了观测值与因子水平均值的关系，矩阵形式为：

$$\underbrace{\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{rn_r} \end{bmatrix}}_{Y_{n\times 1}} = \underbrace{\begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_r} \end{bmatrix}}_{X_{n\times r}} \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix}}_{\boldsymbol{\mu}_{r\times 1}} + \underbrace{\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{rn_r} \end{bmatrix}}_{\boldsymbol{\varepsilon}_{n\times 1}}$$

其中：

- $\mathbf{1}_{n_i}$ 是 n_i 维全1向量（表示第 i 水平的指示变量）；
- $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$（误差项独立同分布正态）。

全模型的残差平方和 SSE_F

全模型的**残差平方和**（反映模型未解释的变异）为：

$$SSE_F = Y^T (I_n - H) Y$$

其中 $H = X(X^TX)^{-1}X^T$ 是投影矩阵（$X^TX = \mathrm{diag}(n_1, n_2, \dots, n_r)$，故 $(X^TX)^{-1} = \mathrm{diag}(1/n_1, 1/n_2, \dots, 1/n_r)$）。

展开 SSE_F：

$$\begin{aligned} SSE_F &= \sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij}^2 - Y^T H Y \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{r} n_i \bar{y}_{i\cdot}^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 = S_e \end{aligned}$$

即 SSE_F 等价于**组内平方和** S_e。

简化模型（Reduced Model）：H_0 成立时的模型

原假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_r = \mu_0$ 成立时，模型简化为“所有水平均值相同”，即：

$$Y = \mathbf{1}_n \mu_0 + \boldsymbol{\varepsilon}$$

其中 $\mathbf{1}_n$ 是 n 维全1向量（对应简化模型的设计矩阵 $X_R = \mathbf{1}_n$）。

简化模型的残差平方和 SSE_R

简化模型的残差平方和为：

$$SSE_R = Y^T (I_n - H_R) Y$$

其中 $H_R = X_R (X_R^T X_R)^{-1} X_R^T = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ 是简化模型的投影矩阵。

展开 SSE_R：

$$\begin{aligned} SSE_R &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij}^2 - Y^T H_R Y \\ &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij}^2 - n \bar{y}_{\cdot\cdot}^2 \\ &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = SST \end{aligned}$$

即 SSE_R 等价于**总平方和** SST。

模型比较与F检验的本质

全模型与简化模型的残差平方和之差，对应**组间平方和** S_A：

$$SSE_R - SSE_F = SST - S_e = S_A$$

F检验的统计量本质是“模型拟合度的提升幅度”与“残差变异”的比值：

$$F = \frac{(SSE_R - SSE_F)/(r-1)}{SSE_F/(n-r)} = \frac{S_A/(r-1)}{S_e/(n-r)} = \frac{MSA}{MSE}$$

其中：

- 分子自由度：简化模型对全模型的约束数 $m = r - 1$（即 H_0 中“均值相等”的约束数）；
- 分母自由度：全模型的残差自由度 $n - r$。

若 $F \geq F_\alpha(r-1, n-r)$，说明全模型比简化模型的拟合度提升更显著，因此**拒绝** H_0。

本质其实仍是上面证明的F检验