方差分析

单因素试验的方差分析

> 本小节内容摘录自知乎。

背景

首先来说说我们为什么要用单因素方差分析（one-way ANOVA)。在做一些实验时，我们通常会把样本分成不同的组，给予不同的对待。例如，我们想研究某种药物在不同剂量下对人们的作用。我们可能会将病人随机分为同等大小的三组，A 组每天吃一片，B 组每天吃两片，C 组每天吃三片。因为我们只研究这个药品计量对病人的影响，所以是单因素分析，如果想要加入别的因素，例如，年龄，就需要用到多因素分析了。在上述实验中，我们给了三种不同的计量，所以这个药物计量因素下有三个水平（level）。实验结束以后，你老板问你，这三组病人的表现有显著的区别吗？这个时候，你就可以使用 ANOVA 来回答你老板的问题啦。

虽然 ANOVA 叫做方差分析，但是他的目的是**检验每个组的平均数是否相同**（敲黑板！）。也就是说，ANOVA 的零假设（null hypothesis）是 $H_0: \mu_A = \mu_B = \mu_C$ 。现在，我们换一个角度考虑这个问题，如果这三组病人的表现并没有显著的区别，那他们其实是同一个总体的三次随机抽样。反过来说，我们想要分析，是不是有一组病人他们的表现非常与众不同，让这组病人不是来自同一个总体。

为什么是方差分析?

为什么不直接比较均值?

举个例子，A_1 组：29，30，31；A_2组：3，31，41。A_1 组均值为30，A_2 组均值 25，看起来 A_1 组大一些，但实际上 A_2 组有两个值都大于 A_2 组。

这是因为，不同组极端值可能会影响到均值，从而给判断造成误导。

为什么不用 t 检验?

我们有一个样本后进行一次 t 检验，在这里，每一组就相当于一个样本，那么比如有三组， C_3^1 就要做 3 次独立的 t 检验。但是 t 检验是每次给定一个显著性水平，比如我们给定 $\alpha = 0.05$，也就是每次犯错的概率为 0.05，那么每次不犯错的概率是 0.95，三次不犯错的概率为 $0.95^3 = 0.857375$，那么我们犯错的概率就高达 0.142625。

而方差分析是一次检验，犯错的概率就小很多，但方差分析也有局限性，它只能检验各组之间的均值是否有差异，并不能给出谁大谁小，所以，适当时候有必要方差分析后，再进行 t 检验。

前提假设

在具体说如何理解 ANOVA 之前，我们先来说 ANOVA 有哪些假设。如果你的实验不能满足 ANOVA 的假设，那你需要考虑别的分析方法或者改变实验设计。ANOVA 有主要有以下 3 个假设：

1. **方差的同质性（homogeneity of variance）**。可以理解为每组样本背后的总体（也叫族群）都有相同的方差；
2. **族群遵循正态分布**；
3. **每一次抽样都是独立的**。在我们的例子中，每一个病人只能提供一个数据。对于一些实验一个样本需要提供多个数据，有其他相应的 ANOVA 分析方法。

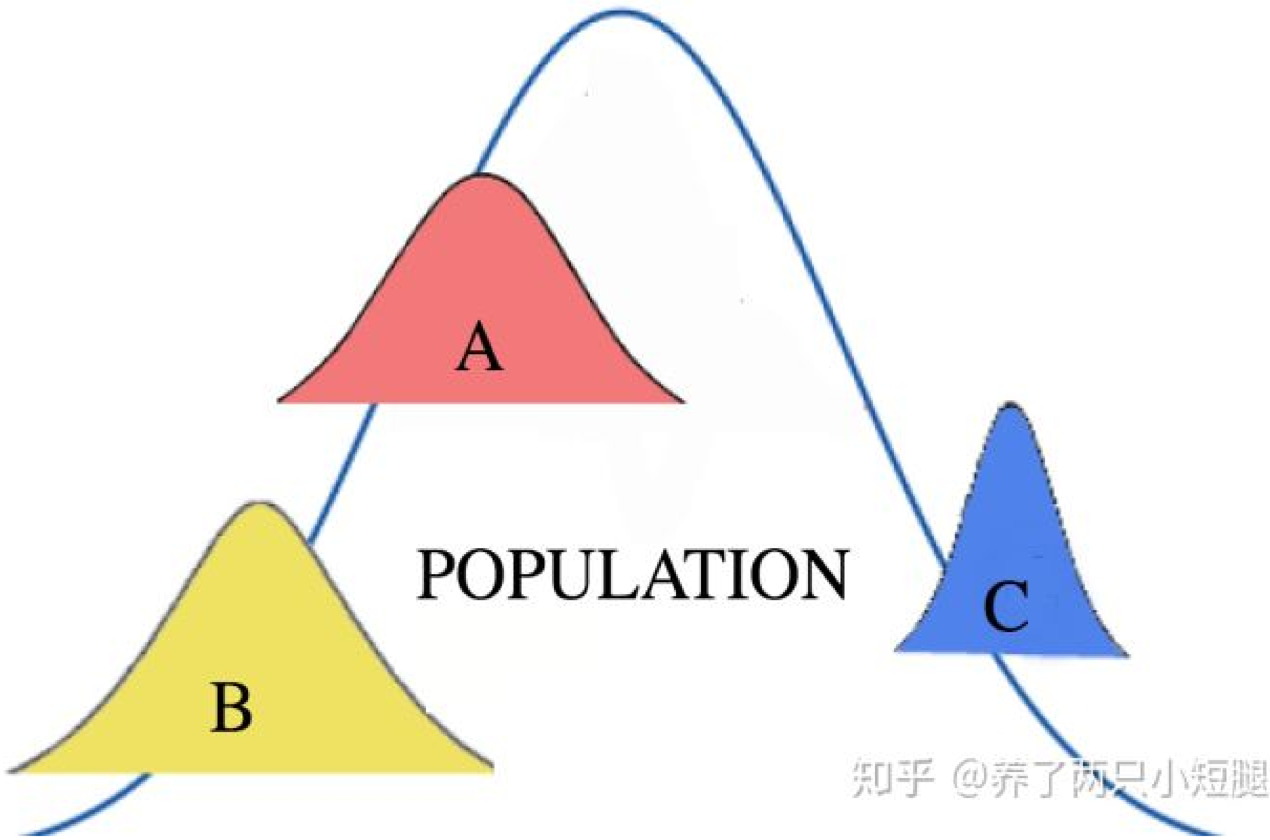
原假设

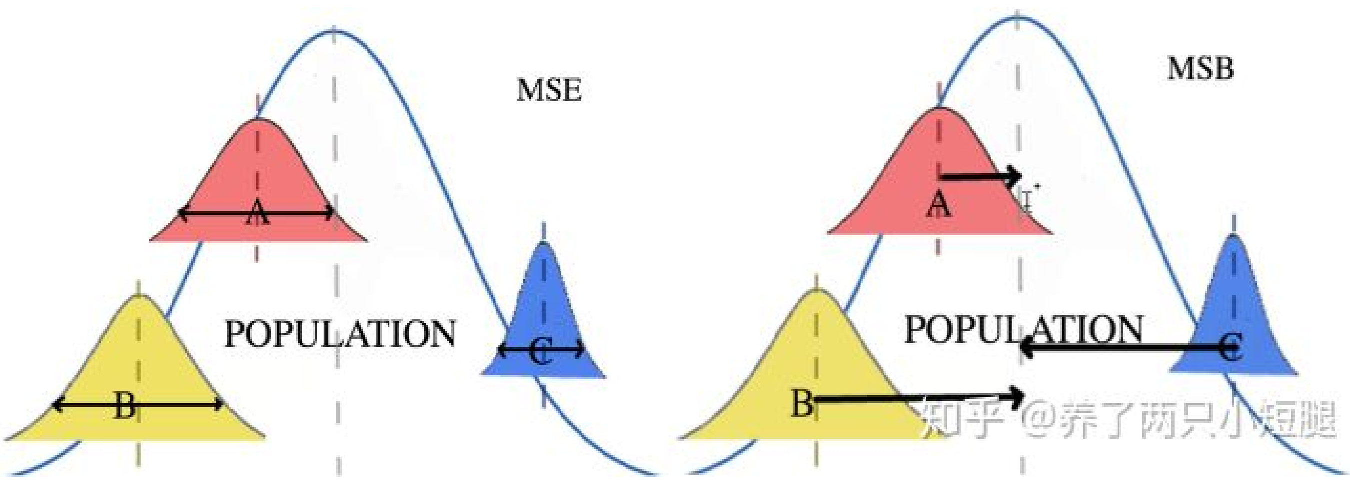
$H_0: \mu_1 = \mu_2 = \ldots = \mu_s$

$H_1: \mu_1, \mu_2, \ldots, \mu_s$ 不全相等

平方和的分解

假设我们得到的抽样结果是这样的：





现在，我们终于可以来看方差分析。首先我们来看**单因素试验方差分析表**：

方差来源	平方和	自由度	均方	F比
因素 A	S_A	s−1	$\bar{S}_A = \frac{S_A}{s-1}$	$F = \frac{\bar{S}_A}{\bar{S}_E}$
误差	S_E	n−s	$\bar{S}_E = \frac{S_E}{n-s}$	
总和	A_T	n−1(即总样本方差自由度)		

总偏差平方和：$S_T=\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}\right)^2$

我们可以将其分解：

$$\begin{aligned}S_T&=\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}\right)^2\\&=\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left[\left(X_{ij}-\bar{X}_{\bullet j}\right)+\left(\bar{X}_{\bullet j}-\bar{X}\right)\right]^2\\&=\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}_{\bullet j}\right)^2+\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(\bar{X}_{\bullet j}-\bar{X}\right)^2\\&=S_E+S_A\end{aligned}$$

S_E：误差平方和

S_A：效应平方和

自由度

S_E：**比较简单的理解方法是，每组（即每个 j ）是 n_j-1,s 个组一共** n−s；

S_A：**比较简单的理解方法是，将每组数据的均值看成一个数据**，共 s 个，求这 s 个数据的方差，方差自由度为 **s−1** （实际需要严谨的证明）；

分布与期望

S_E

由于 $\frac{\sum_{i=1}^{n_1}\left(X_{ij}-\bar{X}_{\bullet j}\right)^2}{\sigma^2}\sim\chi^2\left(n_j-1\right)$，且各 X_{ij} 相互独立，由卡方分布可加性知 $\frac{S_E}{\sigma^2}\sim\chi^2\left(n-s\right)$ ，这也再次说明误差平方和的自由度为 **n−s** 。

故 $E\left(S_E\right)=\left(n-s\right)\sigma^2$

S_A

可以推出（过程略） $E\left(S_{A}\right)=(s-1)\,\sigma^{2}+\sum_{j=1}^{s} n_{j} \delta_{j}^{2}$ ，其中 ${ \delta }{ j }={ \mu }{ j }–\mu $。

进一步还有：

1. S_A 与 S_E 独立；
2. 当 H_0 为真时， $\dfrac{S_A}{\sigma^2} \sim \chi^2\left(s-1\right)$ 。

拒绝域

从上一小节分布和期望，我们可以总结以下几点：

1. S_A 与 S_E 独立；
2. $\dfrac{S_E}{\sigma^2} \sim \chi^2\left(n-s\right)$，因此无论 H_0 是否为真， $E\left(S_{E}\right)=(n-s)\,\sigma^{2}$；
3. 只有当 H_0 为真时， $\dfrac{S_A}{\sigma^2} \sim \chi^2\left(s-1\right)$ ， $E\left(S_{A}\right)=(s-1)\,\sigma^{2}$；而当 H_1 为真时， $E\left(S_{A}\right)=(s-1)\,\sigma^{2}+\sum_{j=1}^{s} n_{j} \delta_{j}^{2}>(s-1)\,\sigma^{2}$ ；

而两个独立方差一般用F检验，所以我们考虑**统计量**：

$$F=\frac{S_{A}/(s-1)}{S_{E}/(n-s)}=\frac{S_{A}/\sigma^{2}}{s-1}\Big/\frac{S_{E}/\sigma^{2}}{n-s}$$

也就是说，当 H_0 不真 H_1 为真时，分子的取值有偏大的趋势，于是是**拒绝域形式**：

$$F=\frac{S_{A}/(s-1)}{S_{E}/(n-s)} \geq k$$

而由 S_A 与 S_E 独立，当 H_0 为真时，统计量所满足的分布:

$$F \sim F\left(s-1, n-s\right)$$

于是，我们加上弃真概率 α，可以得到**拒绝域**：

$$F=\frac{S_{A}/(s-1)}{S_{E}/(n-s)} \geq F\left(s-1, n-s\right)$$

其实，这里的分子和分母就是其他常见解释中的 MSB 和 MSE：

平方和	表达式	均方	方差	简称	表达式	缩写
S_A	$\sum_{i=1}^{s}\sum_{j=1}^{n_j}\left(X_{ij}-\bar{X}_{\bullet j}\right)^2$	$\frac{S_A}{s-1}$	MSB	组间方差	$\frac{\sum_{i=1}^{s}\sum_{j=1}^{n_j}\left(X_{ij}-\bar{X}_{\bullet j}\right)^2}{n-s}$	Mean Square Between
S_E	$\sum_{i=1}^{s}\sum_{j=1}^{n_j}\left(\bar{X}_{\bullet j}-\bar{X}\right)^2$	$\frac{S_E}{n-s}$	MSE	组内方差	$\frac{\sum_{i=1}^{s}\sum_{j=1}^{n_j}\left(\bar{X}_{\bullet j}-\bar{X}\right)^2}{s-1}$	Mean Square Error

双因素试验的方差分析

双因素等重复试验的方差分析—有相互作用

前提

1. A, B 两因素作用与试验的指标；
2. A 有 r 个水平；
3. B 有 s 个水平；
4. 对 A, B 的水平的每对组合都做 $t(t \geq 2)$次试验（称为等重复试验）。
5. A, B 之间可能有相互作用。

参数与记号

$$X_{ijk} \sim N\left(\mu_{ij}, \sigma^2\right), i=1,2,\ldots,r; j=1,2,\ldots,s; k=1,2,\ldots t$$

μ_{ij}，σ^2 均为未知参数

总平均 $\mu=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ij}$

$$\mu_{i\bullet}=\frac{1}{s}\sum_{j=1}^{s}\mu_{ij}, i=1,2,\ldots r$$

$$\mu_{\bullet j}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij}, j=1,2,\ldots s$$

水平 A_i 的效应 $\alpha_i=\mu_{i\bullet}-\mu, i=1,2,\ldots r$

水平 B_j 的效应 $\beta_j=\mu_{\bullet j}-\mu, j=1,2,\ldots s$

$$\mu_{ij}=\mu+\alpha_i+\beta_j+\left(\mu_{ij}-\mu_{i\bullet}-\mu_{\bullet j}+\mu\right)=\mu+\alpha_i+\beta_j+\gamma_{ij}$$

$$\sum_{i=1}^{r}a_i=0$$

$$\sum_{j=1}^{s}\beta_j=0$$

$$\sum_{i=1}^{r}\gamma_{ij}=0, j=1,2,\ldots,s$$

$$\sum_{j=1}^{s}\gamma_{ij}=0, i=1,2,\ldots,r$$

总结:

$$X_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}$$

$\varepsilon_{ijk}\sim N\left(0,\sigma^2\right)$，各 ε_{ij} 独立

$$i=1,2,\ldots,r; j=1,2,\ldots,s; k=1,2,\ldots t$$

$$\sum_{i=1}^{r}a_i=0,\sum_{j=1}^{s}\beta_j=0,\sum_{i=1}^{r}\gamma_{ij}=0,\sum_{j=1}^{s}\gamma_{ij}=0$$

假设

$$H_{01}: \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$$
$$H_{11}: \alpha_1, \alpha_2, \ldots \alpha_r \text{不全为} 0$$

$$H_{02}: \beta_1 = \beta_2 = \ldots = \beta_s = 0$$
$$H_{12}: \beta_1, \beta_2, \ldots \beta_s \text{不全为} 0$$

$$H_{03}: \gamma_{11} = \gamma_{12} = \ldots = \gamma_{rs} = 0$$
$$H_{13}: \gamma_{11}, \gamma_{12}, \ldots \gamma_{rs} \text{不全为} 0$$

双因素试验的方差分析表

与单因素情况类似，对这些问题的检验方法也是建立在平方和的分解上的，思路是一样的，但由于较复杂，我们直接给出方差分析表：

方差来源	平方和	自由度	均方	F比
因素A	S_A	r−1	$\bar{S}_A = \frac{S_A}{r-1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$
因素B	S_B	s−1	$\bar{S}_B = \frac{S_B}{s-1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$
交互作用	$S_{A\times B}$	(r−1)(s−1)	$\bar{S}_{A\times B} = \frac{S_{A\times B}}{(r-1)(s-1)}$	$F_{A\times B} = \frac{\bar{S}_{A\times B}}{\bar{S}_E}$
误差	S_E	rs(t−1)	$\bar{S}_E = \frac{S_E}{rs(t-1)}$	
总和	S_T	rst−1		

拒绝域

同单因素方差分析类似，这里只做总结：

1. 当 H_{01} 为真时，可以证明 $F_A = \frac{S_A/(r-1)}{S_E/(rs(t-1))} \sim F(r-1, rs(t-1))$

   取显著性水平为 α，得到假设 H_{01} 的拒绝域为： $F_A = \frac{S_A/(r-1)}{S_E/(rs(t-1))} \geq F(r-1, rs(t-1))$

2. 当 H_{02} 为真时，可以证明 $F_B=\dfrac{S_B/(s-1)}{S_E/(rs(t-1))}\sim F(s-1,rs(t-1))$

   取显著性水平为 α，得到假设 H_{02}的 拒绝域为： $F_B=\dfrac{S_B/(s-1)}{S_E/(rs(t-1))}\geq F(s-1,rs(t-1))$

3. 当 H_{03} 为真时，可以证明 $F_{A\times B}=\dfrac{S_{A\times B}/((r-1)(s-1))}{S_E/(rs(t-1))}\sim F((r-1)(s-1),rs(t-1))$

   取显著性水平为 α，得到假设 H_{03} 的拒绝域为：

   $F_{A\times B}=\dfrac{S_{A\times B}/((r-1)(s-1))}{S_E/(rs(t-1))}\geq F((r-1)(s-1),rs(t-1))$

双因素无重复试验的方差分析—无相互作用

前提

1. A，B两因素作用与试验的指标；
2. A有r个水平；
3. B有s个水平；
4. 对A，B的水平的每对组合都做 1 次试验。
5. A，B之间不存在相互作用或很小可以忽略。

参数与记号

$X_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij}$

$\varepsilon_{ijk}\sim N\left(0,\sigma^2\right)$，各 ε_{ijk} 独立

$i=1,2,\ldots,r;j=1,2,\ldots,s$

$\sum_{i=1}^{r}a_i=0,\sum_{j=1}^{s}\beta_j=0$

假设

$$
\begin{gathered}
H_{01}: \alpha_1=\alpha_2=\ldots=\alpha_r=0 \\
H_{11}: \alpha_1, \alpha_2, \ldots \alpha_r \text{不全为0}
\end{gathered}
$$

$$
\begin{gathered}
H_{02}: \beta_1=\beta_2=\ldots=\beta_s=0 \\
H_{12}: \beta_1, \beta_2, \ldots \beta_s \text{不全为0}
\end{gathered}
$$

双因素试验的方差分析表

与单因素情况类似，对这些问题的检验方法也是建立在平方和的分解上的，思路是一样的，但由于较复杂，我们直接给出方差分析表：

方差来源	平方和	自由度	均方	F比
因素A	S_A	r–1	$\bar{S}_A=\frac{S_A}{r-1}$	$F_A=\frac{\bar{S}_A}{\bar{S}_E}$
因素B	S_B	s–1	$\bar{S}_B=\frac{S_B}{s-1}$	$F_B=\frac{\bar{S}_B}{\bar{S}_E}$
误差	S_E	(r–1)(s–1)	$\bar{S}_E=\frac{S_E}{(r-1)(s-1)}$	
总和	S_T	rs–1		

拒绝域

同单因素方差分析类似，这里只做总结：

1. 当 H_{01} 为真时，可以证明 $F_A=\frac{S_A/(r-1)}{S_E/((r-1)(s-1))} \sim F(r-1,(r-1)(s-1))$
   取显著性水平为 α，得到假设 H_{01} 的拒绝域为： $F_A=\frac{S_A/(r-1)}{S_E/((r-1)(s-1))} \geq F(r-1,(r-1)(s-1))$
2. 当 H_{02} 为真时，可以证明 $F_B=\frac{S_B/(s-1)}{S_E/((r-1)(s-1))} \sim F(s-1,(r-1)(s-1))$
   取显著性水平为 α，得到假设 H_{02} 的拒绝域为： $F_B=\frac{S_B/(s-1)}{S_E/((r-1)(s-1))} \geq F(s-1,(r-1)(s-1))$