

# 假设检验

## 概念定义

统计量 ( statistic )

样本的一个已知函数

零假设 ( null hypothesis )

又称原假设，指进行统计检验时预先建立的假设，通常记为  $H_0$

备择假设 ( alternate hypothesis )

与原假设对立的假设，通常指包含关于总体分布的一切使原假设不成立的命题（不一定是原假设的否命题），通常记为  $H_1$

第一类错误

原假设正确，但却当成错误的加以拒绝，即第一类错误

第二类错误

原假设错误，但却当成正确的接受，即第二类错误

显著性水平 ( significance level  $\alpha$  )

犯第一类错误的概率就是显著性水平  $\alpha$ ，通常取值为0.05或0.01

拒绝域 ( rejection region / critical region )

拒绝域  $\mathcal{R}$  为满足  $P(X \in \mathcal{R} | H_0) \leq \alpha$  的区域 ( $\mathcal{R}$  人为选定，例如单边检验、双边检验)

p值 ( p-value )

零假设的分布为  $F$ ，p值则为  $F(X)$ ，其中X为样本，可以理解为在零假设成立时样本对零假设的支持度

置信水平 ( confidence level )

置信区间内包含真实参数的概率就是置信水平，在假设检验下，置信水平就是不犯第二类错误的概率，即  $1 - \alpha$

置信区间 CI ( confidence interval )

在参数估计中，置信区间  $[L(X), R(X)]$  满足

$$P_{\theta}\{L(X) \leq \theta \leq R(X)\} \geq 1 - \alpha$$

其中  $X$  为样本， $L, R$  为样本的函数

## 假设检验框架

1. 选定零假设  $H_0$  与备择假设  $H_1$ ，2, 3, 4, 5 全是在假设  $H_0$  成立的前提下得到
2. 考虑对样本进行测试的统计学假设
3. 基于 (2) 的假设选择相关的合适的统计量和检验，常见的如 t-test，F-test， $\chi^2$ -test
4. 选定显著性水平  $\alpha$ ，给出拒绝域  $\mathcal{R}$ ， $1 - \alpha$  置信区间
5. 通过样本计算统计量的值  $S_{obs}$
6. 若  $S_{obs} \in \mathcal{R}$  则拒绝零假设（接受备择假设），否则不拒绝零假设

## 检验方法

### Z检验 Z-test

利用大数定律，用正态分布进行检验，需要方差已知或者样本量足够大能获得方差的良好估计，以下假设方差未知，样本量足够大，方差已知的情况直接将真实方差代替样本方差即可。Z检验是用标准正态分布的理论来推断差异发生的概率，从而比较两个平均数的差异是否显著。

单总体样本  $X : \{x_1, x_2, \dots, x_n\}$

统计学假设：样本服从正态分布  $N(\mu, \sigma^2)$ ， $\sigma$  未知

零假设  $H_0 : \mu = \mu_0$

备择假设  $H_1 : \mu \neq \mu_0$ ，或单边  $H_1 : \mu \leq, \geq \mu_0$

$$\text{统计量: } z = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim N(0, 1), \text{ 其中 } \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$

拒绝域： $\mathcal{R} = (-\infty, Z_{\alpha/2}) \cup (Z_{1-\alpha/2}, \infty)$ ，单边则为  $\mathcal{R} = (-\infty, Z_{\alpha})$  或  $\mathcal{R} = (Z_{1-\alpha}, \infty)$

检验：若  $|z| > Z_{1-\alpha/2}$  则拒绝零假设，单边则条件改为  $u \leq Z_{\alpha}$  或  $u \geq Z_{1-\alpha}$

双总体  $X_1 : \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$ ， $X_2 : \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$

统计学假设： $X_1$  服从  $N(\mu_1, \sigma_1^2)$ ， $X_2$  服从  $N(\mu_2, \sigma_2^2)$

零假设  $H_0 : \mu_1 = \mu_2$

备择假设  $H_1 : \mu_1 \neq \mu_2$  , 做单边测试也可以是  $\leq, \geq$  , 此后不再赘述

统计量:  $z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim N(0, 1)$  , 其中  $\bar{X}_k = \frac{1}{n} \sum_{i=1}^n x_{k,i}$  ,

$$S_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{k,i} - \bar{X}_k)^2} , k = 1, 2$$

拒绝域:  $\mathcal{R} = (-\infty, Z_{\alpha/2}) \cup (Z_{1-\alpha/2}, \infty)$

检验: 若  $|z| > Z_{1-\alpha/2}$  则拒绝零假设

## t检验 Student's t-test

t检验 (Student's t test) , 主要用于样本含量较小 (例如  $n < 30$ ) , 总体标准差  $\sigma$  未知的正态分布。t检验是用t分布理论来推论差异发生的概率, 从而比较两个平均数的差异是否显著。

单总体样本  $X : \{x_1, x_2, \dots, x_n\}$

统计学假设: 样本服从正态分布  $N(\mu, \sigma^2)$  ,  $\sigma$  未知

零假设  $H_0 : \mu = \mu_0$  , 备择假设  $H_1 : \mu \neq \mu_0$

统计量:  $t = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim t_{n-1}$  , 其中  $\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$  ,

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2} , t_{n-1} \text{ 是自由度为 } n-1 \text{ 的 t 分布}$$

拒绝域:  $\mathcal{R} = (-\infty, t_{n-1, \alpha/2}) \cup (t_{n-1, 1-\alpha/2}, \infty)$

检验: 若  $|t| > t_{n-1, 1-\alpha/2}$  则拒绝零假设

同方差双总体  $X_1 : \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$  ,  $X_2 : \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$

统计学假设:  $X_1$  服从  $N(\mu_1, \sigma^2)$  ,  $X_2$  服从  $N(\mu_2, \sigma^2)$

零假设  $H_0 : \mu_1 = \mu_2$  , 备择假设  $H_1 : \mu \neq \mu_0$

统计量:  $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}} \sim t_{df}$

其中  $\bar{X}_k = \frac{1}{n} \sum_{i=1}^n x_{k,i}$  ,  $S_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{k,i} - \bar{X}_k)^2}$  ,  $k = 1, 2$  ,  $df = n_1 + n_2 - 2$

拒绝域： $\mathcal{R} = (-\infty, t_{\text{df}, \alpha/2}) \cup (t_{\text{df}, 1-\alpha/2}, \infty)$

检验：若  $|t| > t_{\text{df}, 1-\alpha/2}$  则拒绝零假设

异方差双总体  $X_1 : \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$  ,  $X_2 : \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$

统计学假设： $X_1$  服从  $N(\mu_1, \sigma_1^2)$  ,  $X_2$  服从  $N(\mu_2, \sigma_2^2)$

零假设  $H_0 : \mu_1 = \mu_2$  , 备择假设  $H_1 : \mu \neq \mu_0$

统计量： $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t_{\text{df}}$

其中  $\bar{X}_k = \frac{1}{n} \sum_{i=1}^n x_{k,i}$  ,  $S_k = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{k,i} - \bar{X}_k)^2}$  ,  $k = 1, 2$  ,

$\text{df} \simeq \frac{\sqrt{S_1^2/n_1 + S_2^2/n_2}}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$

拒绝域： $\mathcal{R} = (-\infty, t_{\text{df}, \alpha/2}) \cup (t_{\text{df}, 1-\alpha/2}, \infty)$

检验：若  $|t| > t_{\text{df}, 1-\alpha/2}$  则拒绝零假设

## 卡方检验 $\chi^2$ test

在十九世纪，统计分析方法主要被用于生物数据分析,当时主流意见认为正态分布普遍适用于此类数据。直到十九世纪末期，皮尔森指出了部分数据具有明显的偏态，正态分布并不是普遍适用。为了更好地对这些观察数据进行建模，皮尔森在1893年至1916年发表的系列文章中提出了一个包含正态分布以及众多偏态分布的连续概率分布族——皮尔森分布族，皮尔森卡方检验是最有名的卡方检验，有两种用途，分别是“适配度检验”和“独立性检验”。“适配度检验”验证一组观察值的次数分配是否异于理论上的分配，“独立性检验”验证从两个变量抽出的配对观察值组是否互相独立。

皮尔森卡方检验评估的是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合；卡方值越小，偏差越小，越趋于符合，若两个值完全相等时，卡方值就为0，表明理论值完全符合。

检验独立性 二元随机变量  $(X, Y) : \{(x_1, y_1), \dots, (x_n, y_n)\}$

$x$  取值有  $n_x$  种 ,  $y$  取值有  $n_y$  种 ,  $O_{i,j} = \sum_{k=1}^N \mathbf{1}_{\{x_k=i \text{ and } y_k=j\}}$  ,  $O_{i,j}$  即观测到  $x_k = i$  且  $y_k = j$  的出现频数。

记  $O_{i,\cdot} = \sum_{k=1}^{n_y} O_{i,k}$  , 即  $x_k = i$  的总数 ,  $O_{\cdot,j} = \sum_{k=1}^{n_x} O_{k,j}$  ,  $y_k = j$  的总数

如果  $X, Y$  独立 , 那么理论上应该有期望出现次数  $E_{i,j} = np_i p_j = n \frac{O_{i,\cdot}}{n} \frac{O_{\cdot,j}}{n} = \frac{O_{i,\cdot} O_{\cdot,j}}{n}$

$O_{i,j}$  列表称为列联表, 例如X为惯用手, Y为性别, 可以得到列联表:

X\Y	男	女	总计
右	$O_{1,1} = 43(E_{1,1} = 45.24)$	$O_{1,2} = 44(E_{1,2} = 41.76)$	$O_{1,\cdot} = 87$
左	$O_{2,1} = 9(E_{2,1} = 6.76)$	$O_{2,2} = 4(E_{2,2} = 6.24)$	$O_{2,\cdot} = 13$
总计	$O_{\cdot,1} = 52$	$O_{\cdot,2} = 48$	$n = 100$

统计学假设: 假设实验中从总体中随机取样得到的  $n$  个观察值被划分为  $n_r \cdot n_c$  个互斥的分类, 这样每个分类都有一个对应的实际观察次数。各个观察值落入第  $(i, j)$  个分类的概率为  $p_{i,j}$  , 从而获得了对应所有第  $(i, j)$  分类的理论期望次数  $E_{i,j} = np_{i,j}$  , 从而构造的特殊统计量服从卡方分布。

零假设  $H_0 : X, Y$  独立

备择假设  $H_1 : X, Y$  不独立

统计量:  $X^2 = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2_{(n_x-1)(n_y-1)}$  服从自由度为  $(n_x - 1)(n_y - 1)$  的卡方分布

拒绝域:  $\mathcal{R} = (\chi^2_{(n_x-1)(n_y-1), 1-\alpha}, \infty)$

检验:  $X^2 > \chi^2_{(n_x-1)(n_y-1), 1-\alpha}$  则拒绝零假设

检验拟合优度  $X : \{x_1, ..., x_n\}$

$x$  取值有  $k$  种 ,  $O_i = \sum_{k=1}^n \mathbf{1}_{\{x_k=i\}}$  , 根据拟合模型得到理论值记为  $E_i$

统计学假设: 假设实验中从总体中随机取样得到的  $n$  个观察值被划分为  $k$  个互斥的分类, 这样每个分类都有一个对应的实际观察次数。构造统计量服从卡方分布, 从而检验拟合优度。

零假设  $H_0 : O - E = 0$

备择假设  $H_1 : O - E \neq 0$

统计量:  $X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$  服从自由度为  $k - 1$  的卡方分布

拒绝域:  $\mathcal{R} = (\chi^2_{k-1, 1-\alpha}, \infty)$

检验： $X^2 > \chi_{k-1, 1-\alpha}^2$  则拒绝零假设

## 卡方检验注意事项

当用皮尔森卡方检定做独立性检定时，若任何一个字段的期望次数小于5，或者自由度为 1 且若期望次数 <10，则近似于卡方分配的假设不可信，统计值会系统性地偏高，导致过度地拒绝虚无假设。此时可以将每个观察值的离差减去 0.5 之后再做平方，即[叶氏连续性修正](#)。

## F检验 F-test

F检验（F-test），最常用的别名叫做联合假设检验，此外也称方差比率检验、方差齐性检验。它是一种在零假设之下，统计值服从F-分布的检验。其通常是用来分析用了超过一个参数的统计模型，以判断该模型中的全部或部分参数是否适合用来估计母体。F检验还可以用于三组或者多组之间的均值比较，但是如果被检验的数据无法满足均是正态分布的条件时，该数据的稳健型会大打折扣，特别是当显著性水平比较低时。但是，如果数据符合正态分布，而且alpha值至少为0.05，该检验的稳健型还是相当可靠的。

双总体  $X_1 : \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$  ,  $X_2 : \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$

统计学假设：假设一系列服从正态分布的总体，都有相同的标准差这是最典型的F检验，即  $X_i \sim N(\mu_i, \sigma_i)$  ,  $\sigma_1 = \sigma_2 = \dots = \sigma_k$

零假设  $H_0 : \sigma_1 = \sigma_2$  , 备择假设  $H_1 : \sigma_1 \neq \sigma_2$

统计量： $F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$

其中  $S_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{k,i} - \bar{x}_k)^2$  ,  $\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{k,i}$  ,  $k = 1, 2$

拒绝域： $\mathcal{R} = (F_{n_1-1, n_2-1, 1-\alpha}, \infty)$

检验： $F > F_{n_1-1, n_2-1, 1-\alpha}$  则拒绝零假设

## 曼-惠特尼U检验 Mann-Whitney U-test (又称秩和检验 Wilcoxon rank-sum test)

它假设两个样本分别来自除了总体均值以外完全相同的两个总体，目的是检验这两个总体的均值是否有显著的差别，是一种非参数统计方法。秩和检验的优点是（1）不受总体分布限制，适用面广；（2）适用于等级资料及两端无确定值的资料；（3）易于理解，易于计算。缺点是符合参数检验的资料，用秩和检验，则不能充分利用信息，检验功效低。

非参数统计：即不考虑总体分布类型是否已知，不比较总体参数，只比较总体分布的位置是否相同的统计方法。

双总体  $X_1 : \{x_{1,1}, x_{1,2}, \dots, x_{1,n_1}\}$  ,  $X_2 : \{x_{2,1}, x_{2,2}, \dots, x_{2,n_2}\}$

将两组样本数据混合，并按照数据大小的升序编排等级。最小的数据等级为1，第二小的数据等级为2，以此类推，那么得到每个样本的秩(Rank)，记  $x_{i,j}$  的秩为  $r_{i,j}$  , 记分组和为  $R_1 = \sum_j r_{1,j}$  ,  $R_2 = \sum_j r_{2,j}$  , 调整为  $U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$  ,  $U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$  ,

零假设  $H_0$  : 两总体来自同一分布 , 备择假设  $H_1$  : 两总体来自不同分布

统计量:  $U = \min(U_1, U_2)$  , 对于充分大的样本,  $U \sim N(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12})$

拒绝域:  $\mathcal{R} = (-\infty, U_{\alpha/2}) \cup (U_{1-\alpha/2}, \infty)$  , 小样本查表, 大样本U可以用正态分布的分位数估计

检验:  $|U| > U_{1-\alpha/2}$  则拒绝零假设

## K-S检验 Kolmogorov-Smirnov test

Kolmogorov-Smirnov是比较一个经验分布  $F_n(x)$  与理论分布  $F(x)$  或者两个观测值分布的检验方法。

样本  $X : \{x_1, x_2, \dots, x_n\}$

经验分布:  $F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i)$

Kolmogorov distribution

$$P(K < x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2} = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8x^2)}$$

零假设  $H_0$  : 两个分布一致  $F_n(x) = g(x)$ , 备择假设  $H_1$  : 两个分布不一致

统计量:  $D_n = \sup_x |F_n(x) - F(x)|$  ,  $\sqrt{n} D_n \xrightarrow{n \rightarrow \infty} \sup_t |B(F(t))|$

其中 B(t) 为 [Brownian bridge](#),  $K = \sup_{t \in [0,1]} |B(t)| \sim \text{Kolmogorov distribution}$  , 由此可以估计  $\mathcal{D}_{n,\alpha}$

拒绝域:  $\mathcal{R} = (\mathcal{D}_{n,\alpha}, \infty)$

检验:  $D_n > \mathcal{D}_{n,\alpha}$  则拒绝零假设