软件工程统计方法

非参数假设检验

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

内 容 提 纲

- ❑ 统计学导论
- ❑ 描述统计
- ❑ 概率计算基础
- ❑ 随机变量及其分布
- ❑ 统计量及其抽样分布
- ❑ 参数估计
- ❑ 参数假设检验
- ❑ 非参数假设检验
- ❑ 方差分析
- ❑ 回归分析

本 节 内 容

- 符号检验(Sign test)
- 符号秩检验(Sign rank test)
- 秩和检验(Rank sum test)
- 卡方拟合优度检验(Chi-square goodness-of-fit test)
- 偏度峰度检验(Skewness-Kurtosis test)
- 柯尔莫哥洛夫检验(Kolmogorov test)
- 斯米尔诺夫检验(Smirnov test)
- 独立检验(Independence test)

1 非参假设检验

- 在实际应用中, 我们对所研究的总体可能知之不多, 要给出或假设总体的分布十分困难, 或者总体的分布并不满足假定的前提, 或者不知道推断时需要的总体参数值, 或者没有足够多的样本。此时, 参数统计的方法不适用, 必须应用非参数统计的方法。
- 非参假设检验一般不涉及总体参数, 也不依赖于对总体分布作出假定, 往往仅依据数据的顺序量或等级资料等即可进行统计推断, 在实际中得到了极为广泛的应用。简而言之, 和数据本身的总体分布无关的检验称为非参数检验。
- 多根据数据观测值的相对大小建立检验统计量, 然后找到在零假设下这些统计量的分布, 看这些统计量的数据实现是否在零假设下属于小概率事件。
- 对总体假定较少。
- 可以处理所有类型数据。
- 易于理解，容易计算。

非 参 假 设 检 验

非参假设检验的优点:

- 在总体分布未知时, 如果还假定总体有诸如正态分布那样的已知分布, 在进行统计推断就可能产生错误甚至灾难。
- 非参数检验总是比传统检验安全(更不容易拒绝原假设)。
- 但是在总体分布形式已知时, 非参数检验不如传统方法精度高。

非参假设检验的适用场景:

- 如果需要对定性数据做假设检验, 则需要使用非参数方法。
- 如果需要对中位数做检验, 则需要使用非参数的方法。
- 如果需要对统计分布做检验, 例如检验数据是否来自正态总体, 检验两个总体的统计分布是否相同等, 则需要用非参数方法。
- 当参数检验需要的假设不成立时, 需要采用非参数检验方法。

例 子

我们首先通过一个简单的例子来比较参数假设检验和非参数假设检验之间的联系和区别。

Example 1 南京某区*16*座预售楼盘均价如下表（单位：元/平方米）

7800	*7400*	*7300*	*6700*	*7000*	*7800*	*8200*	*7400*
8300	*6800*	*7700*	*7700*	*7400*	*12900*	*7500*	*7700*

试推断南京市某区楼盘价格与媒体公布的*7900*元/平方米是否相符（显著性水平为$\alpha = 0.05$）？

例 子

我们首先按照参数假设检验中的小样本t检验方法进行推断。假设某区楼盘价格近似服从正态分布$\mathbb{N}(\mu, \sigma^2)$, 依照题意我们建立假设

$$H_0 : \mu = 7900, H_1 : \mu \neq 7900$$

我们根据样本数据计算可得

$$\overline{x} = 7850, s^2 = 2005333, s = 1416$$

我们计算检验统计量得

$$t = |\frac{\overline{x} - \mu_0}{s/\sqrt{n}}| = |\frac{7850 - 7900}{1461/\sqrt{16}}| = 0.1412 < t_{0.025}(15) = 2.1448$$

样本未落入拒绝域, 所以我们认为总体均值与7900在统计意义上差异不显著, 则在某种程度上接受了媒体公布的7900元/平方米。

例 子

我们仔细观察则可以发现：我们假设总体分布近似服从正态分布存在很大疑问。样本数据存在较大的偏性, 即数据偏小的多, 偏大的少。假如总体分布与正态分布偏离较大, 那我们使用t检验就是不够恰当的, 从而得出的结论也争议较大。

我们进一步观察数据发现：16个样本数据中，3个高于7900, 13个低于7900。由正态分布的对称性常识可以知道假设样本来自正态分布的风险较大。

符号检验与参数检验中相关样本显著性t检验相对应，当资料不满足参数检验条件时，可采用此法来检验两相关样本的差异显著性。

2 符号检验

我们首先建立假设

$$H_0: M_e = 7900, H_1: M_e \neq 7900 \tag{1}$$

其中M_e为总体的中位数。如果原假设H_0为真, 即7900为中位数, 则数据在中位数两篇的分布应该接近, 即各位8个左右。我们采用n_+记大于7900的样本数据个数, n_-记小于7900的样本数据个数。则$n_+ + n_- = 16$ (暂时不考虑等于7900情况)。

在原假设成立和独立抽样的前提下, n_+和n_-都是一个n重伯努利试验, 即近似服从二项分布$\mathbb{B}(n, 0.5)$。因此, n_+和n_-不能太大, 也不能太小, 否则就有理由拒绝原假设。由于n_+和n_-互补, 所以我们考虑其中之一即可, 实际应用中选择n_+进行检验。这就是符号检验的基本思想。

符号检验

考虑随机变量Y_i

$$Y_i = \begin{cases} 1 & X_i \geq M_e \\ 0 & X_i < M_e \end{cases}$$

$n_+ = \sum_{i=1}^{n} Y_i$, 其中$Y_i \sim \mathbb{B}(1,p)$, $n_+ \sim \mathbb{B}(n,p)$, $p = P\{X_i \geq M_e\}$。进而, 我们可以将假设检验转化为检验如下假设

$$H_0 : p = 0.5, H_1 : p \neq 0.5$$

在本例中, 根据二项分布公式, 我们可以计算在$n = 16$, $p = 0.5$条件下, 抽到样本$n_+ = 3$甚至更差的概率为

$$\mathbb{B}(3; 16, 0.5) = \sum_{i=0}^{3} C_{16}^{i}(0.5)^i(0.5)^{16-i} = 0.0213$$

上述概率即是假设检验的p值, 根据双边检验的规则

$$p\text{值} = 0.0213 < \frac{\alpha}{2} = 0.025$$

故拒绝原假设, 即认为总体中心位置与7900在统计意义上存在显著差异。

3 配对符号检验

检验两个总体的分布是否相同：符号检验法（正负号个数检验法）

检验两个总体的分布是否相同的符号法又称正负号个数检验法。它所要处理的问题是：假设两个总体的分布$F(x)$与$G(x)$相同，用两个总体的容量相同的配对样本$x_1, x_2, \cdots, x_n$与$y_1, y_2, \cdots, y_n$来检验它，即检验假设$H_0 : F(x) = G(x)$是否成立.

设两个总体的样本相互独立，当$H_0 : F(x) = G(x)$成立时，概率$P\{X_i < Y_i\}$应当与概率$P\{X_i > Y_i\}$相同，$i = 1, 2, \cdots, n$. 也就是说，对于样本观测值而言，$x_i - y_i > 0$的个数(记为$n+$)，应当与$x_i - y_i < 0$的个数(记为$n-$)基本相同(从样本观测值角度，不一定刚好相等). 如果两者相差很远，我们就有理由，拒绝假设$H_0 : F(x) = G(x)$.

配对符号检验

从分布函数为$F(x)$,$G(x)$的总体中分别取容量均为n的样本$X_1,\cdots,X_n$和$Y_1,\cdots,Y_n$,假设检验

$$H_0: F(x)=G(x), H_1: F(x)\neq G(x)$$

将数据配对排好, 列成表, 当$x_i > y_i$时, 取"+"号; 当$x_i < y_i$时, 取"-"号; 当$x_i = y_i$时,取"0", 并用n_+和n_-分别表示"+"号和"-"号的个数.

若H_0成立, 两总体分布相同, n_+与n_-应相差不大. 如若差异过大, 就认为不仅仅有实验误差, 而认为$F(x)$与$G(x)$有差异. n_+, n_-都服从二项分布$b(n,p)$, H_0成立时分布为$b(n,0.5)$. 因此n_+不宜太大或太小.

计算临界值c_1,c_2, 使得最大的c_1和最小的c_2满足:

$$P\{n_+\leq c_1\}=\sum_{i=0}^{c_1} C_n^i(0.5)^n\leq\frac{\alpha}{2}$$

$$P\{n_+\geq c_2\}=\sum_{i=c_2}^{n} C_n^i(0.5)^n\leq\frac{\alpha}{2}$$

拒绝域为$n_+\leq c_1$或者$n_+\geq c_2$.
备注: 可以证明$c_1+c_2=n$.

例子

Example 2 X, Y两个样本容量均为*40*, 计算$x_i - y_i$得$n_+ = 30$, $n_- = 10$. 试比较X, Y是否有显著差异*(*$\alpha = 0.05$*)?*

解: 需检验,

$$H_0 : X, Y\text{差异不显著}.$$

(方法一:) 因为$n_+ = 30, n = 40$. 查表(或计算)得$c_1 = 13$, $c_2 = 40 - 13 = 27$. $n_+ > 27$. 故拒绝H_0.

(方法二:) 因为$n_+ \sim b(n, p)$, 由中心极限定理可知, n较大时, $n_+ \sim N(np, np(1-p))$. 即当H_0成立, $p = 0.5$, 则

$$U = \frac{n_+ - np}{\sqrt{np(1-p)}} = \frac{2n_+ - n}{\sqrt{n}} \sim N(0, 1)$$

拒绝域为$|U| > z_{\alpha/2}$.

此例中, $U = \frac{|2*30-40|}{\sqrt{40}} = 3.16 > 1.96 = z_{0.025}$, 拒绝$H_0$.

练 习

Example 3 *12*家超市对于牛奶促销前后的数据
促销前：*42, 57, 38, 49, 63, 36, 48, 58, 47, 71, 83, 27*
促销后：*40, 60, 38, 47, 65, 39, 49, 50, 47, 52, 72, 33*
试检验促销是否有效$(\alpha = 0.05)$。

$$U = \frac{2n_+ - n}{\sqrt{n}} \sim N(0,1)$$

当$n' = n_+ + n_-$较小时, 使用以下修正公式

$$U = \frac{2n_+ - n' + C}{\sqrt{n'}} \sim N(0,1)$$

其中$n_+ > n_-$时, $C = \frac{1}{2}$, 否则$C = -\frac{1}{2}$。

符号秩检验

在对称情况下，平均值=中位数。但是非对称情况呢？
让我们看以下数据(与均值的差)：

$$9, 13, -7, 10, -18, 4$$

- 考虑符号情况
- 考虑距离情况

我们首先将$|X_i|$进行排序，顺序统计量称为秩。

样本值	9	13	-7	10	-18	4
符号	$+$	$+$	$-$	$+$	$-$	$+$
秩	3	5	2	4	6	1

我们考虑+的秩和$R_+ = 3+5+4+1 = 13$, $-$的秩和$R_- = 2+6 = 8$。

符号秩检验

符号检验容易失败！尤其是存在一些离群点的情况下。

我考虑以下样本对$\mu = 12$进行检验。

$$9.3, 0.9, 9.0, 21.7, 11.5, 13.9$$

注意到数据中有两个离群值0.9, 21.7, 因此说明样本来自正态总体的可能性较小. 所以不能采用t检验法。考虑使用Wilcoxon符号秩检验法.

威尔科克森符号秩检验是由威尔科克森（F・Wilcoxon）于1945年提出的。该方法是在成对观测数据的符号检验基础上发展起来的，比传统的单独用正负号的检验更加有效。

Wilcoxon符号秩检验

Wilcoxon符号秩检验法步骤如下:

(1) 样本排序, 计算符号秩.

x	11.5	13.9	9.3	9.0	21.7	0.9
$x-12$	-0.5	1.9	-2.7	-3.0	9.7	-11.1
符号秩	-1	2	-3	-4	5	-6

(2) 计算正符号秩R_+, 作为统计量. 本例中$R_+=7$. 假设R_-为负符号秩, 则有$R_+ + |R_-| = \frac{n(n+1)}{2}$.

(3) 取$min\{R_+, R_-\}$作为检验统计量W。

(4) 判定结果

- 小样本：查Wilcoxon符号秩检验表进行判定。
- 大样本：

$$Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

练 习

试用Wilcoxon符号秩检验方法对以下问题进行检验。

Example 4 南京某区*16*座预售楼盘均价如下表（单位：元/平方米）

7800	*7400*	*7300*	*6700*	*7000*	*7800*	*8200*	*7400*
8300	*6800*	*7700*	*7700*	*7400*	*12900*	*7500*	*7700*

试推断南京市某区楼盘价格与媒体公布的*7900*元/平方米是否相符？

练 习

Example 5 *12*家超市对于牛奶促销前后的数据
促销前：*42, 57, 38, 49, 63, 36, 48, 58, 47, 71, 83, 27*
促销后：*40, 60, 38, 47, 65, 39, 49, 50, 47, 52, 72, 33*
试检验促销是否有效？

如何构建符号秩检验？

4 秩和检验

符号检验法在下面问题中常见到应用: 如,消费者对两种包装的评分, 或对两种产品品牌的评分; 学生对两门不同课程的成绩的反映(评分), 企业对两种政策的反映(评分)等等, 都存在两个总体的分布是否相同的检验问题.

是按照问题本身的属性，“天然”配对的。也就是说，不能各自独立地颠倒顺序。正负号检验的一个重要的前提是：样本x_i 或y_i 不能各自独立地颠倒顺序。

例：两套问卷的满分都是200分，两套问卷测得的结果如表：
卷A: 147, 150, 152, 148, 155, 146, 149, 148, 151, 150
卷B: 146, 151, 154, 147, 152, 147, 148, 146, 152, 150

(1) 用两套问卷测量10个管理人员的素质.

(2) 用两套问卷测量20个管理人员的素质.

以上两种场景分别用什么检验方法?

秩和检验法

检验两个总体的分布是否相同的另一种方法: 秩和检验法. 秩和检验方法最早是由Wilcoxon提出，后来Mann-Whitney将其应用到两样本容量不等（$n_1 \neq n_2$）的情况，因而又称为曼—惠特尼U检验。

设有两个总体的样本观测值$x_1, x_2, \cdots, x_n$ 与$y_1, y_2, \cdots, y_m$，可能$m \neq n$. 两组样本是可以各自独立颠倒顺序的.

不妨设$n \neq m$, 把两组样本放在一起, 按样本观测值的大小重新排序, 那么每个观测值就有一个序号, 称为秩. 把样本个数少的这组样本$x_1, x_2, \cdots, x_n$的秩（序号）加总起来, 记为W. 如果两个总体的分布相同, 那么样本$x_1, x_2, \cdots, x_n$ 与$y_1, y_2, \cdots, y_m$ 应当是均匀混合的, 也就是说, W不能太小, 也不能太大. W 太小, 说明样本$x_1, x_2, \cdots, x_n$较多地集中在左段. W太大, 说明样本$x_1, x_2, \cdots, x_n$ 较多地集中在右段.

由于$n \leq m$, W 应当比另一组样本的序号之和小一些. 也就是说, W应当在某两个数字之间: $W_1 < W < W_2$.
Wilcoxon 给出了W 的概率分布表, 对于给定的显著性水平α, 可以由秩和检验表, 依据n, m, 查出W_1, W_2 . 若$W \leq W_1$ 或$W \geq W2$, 则拒绝H_0 : $F(x) = G(x)$ (认为两个总体分布不同)。

例子

Example 6 两样本X, Y观测值如下:

X	2.36	3.14	7.52	3.48	2.76	5.43	6.54	7.41
Y	4.38	4.25	6.54	3.28	7.21	6.54		

是否两个样本来自同一总体$(\alpha = 0.05)$?

解: 首先数据统一编号计算相应的秩. Y容量较小, 计算统计量$T = 4 + 6 + 7 + 10 + 10 + 12 = 49$. $\alpha = 0.05, n_1 = 6, n_2 = 8$, 查表得$T_1 = 32$, $T_2 = 58$. $T_1 < 49 < T_2$, 接受H_0.

[大样本逼近]当样本较大时$(n_1, n_2 > 10)$, T近似服从$N(\mu, \sigma^2)$, 其中$\mu = \frac{n_1(n_1+n_2+1)}{2}$, $\sigma^2 = \frac{n_1 n_2(n_1+n_2+1)}{12}$. 则

$$U = \frac{T - \mu}{\sigma} \sim N(0, 1)$$

拒绝域为$|U| > z_{\alpha/2}$.

练 习

例：两套问卷的满分都是200分，两套问卷测得的结果如表：

卷A: 147, 150, 152, 148, 155, 146, 149, 148, 151, 150

卷B: 146, 151, 154, 147, 152, 147, 148, 146, 152, 150

- 符号检验
- 符号秩检验
- 秩和检验

5 分布函数拟合检验

在X的总体分布$F(x)$未知时, 根据样本$X_1, \cdots, X_n$来检验关于总体分布的假设:

$$H_0 : F(x) = F^*(x),$$

$$H_1 : F(x) \neq F^*(x).$$

- 卡方拟合优度检验
- 偏度峰度检验
- 柯尔莫哥洛夫检验

6 卡方拟合优度检验

卡方拟合优度检验思想与步骤:

1. 总体可以分为k类: $A_1, \cdots, A_k$;
2. 原假设$H_0 : P(A_i) = p_i, i = 1, \cdots, k$; 其中任意$p_i \geq 0$, 且$\sum_i p_i = 1$.
3. 对总体作n次观察, k个类出现的频数分别为$n_1, \cdots, n_k$, 且$\sum_i^k n_i = n$.
4. 构造统计量$(E_i = n_i, T_i = np_i)$

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(E_i - T_i)^2}{T_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n \tag{2}$$

若假设中的$F(X)$包含未知参数, 根据H_0求p_i的最大似然估计值$\hat{p}_i$, 然后构造统计量

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{n_i^2}{n\hat{p}_i} - n \tag{3}$$

卡方拟合优度检验

Theorem 1 若n充分大($n \geq 50$), 则当H_0为真时, 统计量(2)近似服从$\mathcal{X}^2(k-1)$分布; 而统计量(3)近似服从$\mathcal{X}^2(k-r-1)$分布, 其中r是被估计的参数的个数.

两种情况给出检验统计量和拒绝域:

- p_i均已知, 检验统计量$\mathcal{X}^2=\sum_{i=1}^{k}\frac{n_i^2}{np_i}-n$, 拒绝域为

$$\mathcal{X}^2 \geq \mathcal{X}_\alpha^2(k-1)$$

- p_i不完全已知, 检验统计量$\mathcal{X}^2=\sum_{i=1}^{k}\frac{n_i^2}{n\hat{p}_i}-n$, 拒绝域为

$$\mathcal{X}^2 \geq \mathcal{X}_\alpha^2(k-r-1)$$

  其中r为$p_1,\cdots,p_k$中估计参数的个数, $\hat{p}_i$为p_i的极大似然估计.

根据以上分布拟合检验的思想及步骤, 我们可对泊松分布, 指数分布, 正态分布等进行$\mathcal{X}^2$拟合检验.

例 子

Example 7 在一实验中, 每隔一段时间观察一次由某铀所放射的α例子个数X, 共观察了100次, 得如下结果:

i	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
n_i	1	5	16	17	26	11	9	9	2	1	2	1	0
A_i	A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	$\geq A_{12}$

其中n_i是观察到有i个粒子的次数, 从理论上X应服从泊松分布

$$P\{X=i\}=\frac{\lambda^i e^{-\lambda}}{i!}, i=0,1,2,3$$

现问在显著性水平0.05下检验假设

$$H_0: \text{总体服从泊松分布} P\{X=i\}=\frac{\lambda^i e^{-\lambda}}{i!}$$

例 子

先估计λ. 由最大似然估计法得$\hat{\lambda} = \overline{x} = 4.2$. 在$H_0$假设下, 将$X$分成$A_0, \cdots, A_{12}$, 则$P\{X = i\}$有估计

$$\hat{p}_i = \frac{4.2^i e^{-4.2}}{i!}$$

其中, 我们将$n\hat{p}_i < 5$的组合并, 得下表

A_i	$A_0 - A_1$	A_2	A_3	A_4	A_5	A_6	A_7	$A_8 - A_{12}$
n_i	6	16	17	26	11	9	9	6
$\hat{p}_i$	0.078	0.132	0.185	0.194	0.163	0.114	0.069	0.065
$\frac{n_i^2}{n\hat{p}_i}$	4.615	19.394	15.622	34.845	7.423	7.105	11.739	5.538

$$\sum_{i=1}^{n} \frac{n_i^2}{n\hat{p}_i} = 106.281$$

合并以后$k = 8, r = 1, 106.281 - 100 < \mathcal{X}_{0.05}^2(6) = 12.592.$
所以在0.05水平下不能拒绝H_0.

例 子

Example 8 把一颗骰子重复抛掷*300* 次, 结果如下:

点数	1	2	3	4	5	6
次数	40	70	48	60	52	30

试检验这颗骰子的六个面是否匀称? $\alpha = 0.05$

例 子

Example 9 *84个Etruscan人的男子头颅的最大宽度(毫米)如下:*

141 148 132 138 154 142 150 146 155 158 150 140 147 148 144 150 149 145 149 158 140
143 141 144 144 126 140 144 142 141 140 145 135 147 146 141 136 140 146 142 137 137
148 154 137 139 143 140 131 143 141 149 148 135 148 152 143 144 141 143 147 146 152
150 132 142 142 143 153 149 146 149 138 142 149 142 137 134 144 146 147 140 142 145

现问在显著性水平$\alpha = 0.1$下检验假设

$$H_0: \text{ 总体服从正态分布} f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

我们可以先粗略估计

组	n_i	n_i/n	累计频率
124.5-129.5	01	0.0119	0.0119
129.5-134.5	04	0.0476	0.0595
134.5-139.5	10	0.1191	0.1786
139.5-144.5	33	0.3929	0.5715
144.5-149.5	24	0.2857	0.8572
149.5-154.5	09	0.1071	0.9524
154.5-159.5	03	0.0357	1

将上述表格画成直方图, 可见大致成正态分布.

例 子

先估计μ, σ. 由最大似然估计法得$\hat{\mu} = 143.8, \hat{\sigma}^2 = 6^2$. 在$H_0$假设下, 将$X$分成$A_1, \cdots, A_7$, 计算理论概率。例如:

$$\hat{p}_2 = \hat{P}\{129.5 < X \leq 134.5\}$$

$$= \Phi(\frac{134.5 - 143.8}{6}) - \Phi(\frac{129.5 - 143.8}{6}) = \Phi(-1.55) - \Phi(-2.38) = 0.0519$$

分别查表可得

A_i	n_i	$\hat{p}_i$	$n\hat{p}_i$	$n_i^2/(n\hat{p}_i)$
$(-\infty, 129.5]$	01	0.0087	0.73	4.91
$[129.5, 134.5]$	04	0.0519	4.36	-
$[134.5, 139.5]$	10	0.1752	14.72	6.79
$[139.5, 144.5]$	33	0.3120	26.21	41.55
$[144.5, 149.5]$	24	0.2811	23.61	24.40
$[149.5, 154.5]$	09	0.1336	11.22	10.02
$[154.5, \infty)$	03	0.0375	3.15	$\sum = 87.67$

合并以后$k = 5, r = 2, 87.67 - 84 < \mathcal{X}_{0.1}^2(2) = 4.605$.
所以在0.1水平下不能拒绝H_0.

非 参 假 设 检 验
符号检验
配对符号检验
秩和检验
分布函数拟合检验
卡方拟合优度检验
偏 度 峰 度 检 验
*K-S*检验
*K-S*检验
独立性检验
小结
第 31 页 共 100 页
返 回
全 屏
关 闭

例 子

Example 10 *84*个*Etruscan*人的男子头颅的最大宽度*(*毫米*)*如下*(*排序后*):*
126 131 132 132 134 135 135 136 137 137 137 137 138 138 139
140 140 140 140 140 140 140 141 141 141 141 141 141 142 142
142 142 142 142 142 142 143 143 143 143 143 143 144 144 144
144 144 144 145 145 145 146 146 146 146 146 146 147 147 147
147 148 148 148 148 148 149 149 149 149 149 149 150 150 150
150 152 152 153 154 154 155 158 158
现问在显著性水平$\alpha = 0.1$下检验假设

$$H_0: \text{ 总体服从正态分布} f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

试对121-125,126-130,$\cdots$,156-160分组后进行检验。$(\hat{\mu} = 144, \hat{\sigma^2} = 6^2)$

7 偏度峰度检验

根据中心极限定理可知,

所谓随机变量X的偏度和峰度是指X的标准化变量$\frac{X-E(X)}{\sqrt{D(X)}}$的三阶中心矩和四阶中心矩, 其中$E(X), D(X)$分别是随机变量$X$的均值和方差. X偏度nu_1和峰度ν_2的计算公式定义如下:

$$\nu_1 = E[(\frac{X-E(X)}{\sqrt{D(X)}})^3], \nu_2 = E[(\frac{X-E(X)}{\sqrt{D(X)}})^4]$$

偏度描述了随机变量分布相对其均值的不对称程度. 峰度反映了与正态分布相比, 随机变量分布的尖锐度或平坦度. 当随机变量X服从正态分布时, 其偏度$v_1 = 0$、峰度$v_2 = 3$. 检验时，我们采用样本的偏度峰度进行分析。

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$$

样 本 偏 度 和 峰 度

设$X_1, \cdots, X_n$是来自总体X的一个样本, 是这一样本的观察值, B_k是样本k阶中心矩, 则样本偏度G_1和样本峰度G_2:

$$G_1 = \frac{B_3}{B_2^{3/2}}, G_2 = \frac{B_4}{B_2^2}$$

若总体X为正态分布随机变量, $X_1, \cdots, X_n$是来自总体X的样本, 则可以证明当n充分大时, 样本偏度G_1、样本峰度G_2分别依概率收敛于总体偏度v_1和总体峰度v_2. 即, 当总体X为正态变量且n充分大时, G_1与的v_1偏离不应太大, 而G_2与v_2的偏离也不应太大. 记

$$\sigma_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}, \sigma_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}, \mu_2 = 3 - \frac{6}{n+1}$$

$$U_1 = G_1/\sigma_1, U_2 = (G_2 - \mu_2)/\sigma_2$$

当H_0为真, n充分大时, 近似有$U_1 \sim N(0,1), U_2 \sim N(0,1)$. 构造拒绝域为

$$|u_1| \geq z_{\alpha/4} \text{ 或} |u_2| \geq z_{\alpha/4}$$

例 子

Example 11 将84个Etruscan人的男子头颅的最大宽度(mm)如教材(P247).现问在显著性水平0.1下, 采用偏度、峰度检验法检验数据是否来自正态总体.

$$H_0 : X\text{来自正态总体.}$$

解: $n = 84$, 则计算$\sigma_1 = 0.2579$, $\sigma_2 = 0.4892$, $\mu_2 = 2.9294$. $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, 计算可得$A_1 = 143.773$, $A_2 = 20706.13$, $A_3 = 2987099$, $A_4 = 4.316426 * 10^8$. 根据

$$B_2 = A_2 - A_1^2$$

$$B_3 = A_3 - 3A_2A_1 + 2A_1^3$$

$$B_4 = A_4 - 4A_3A_1 + 6A_2A_1^2 - 3A_1^4$$

计算得$B_2 = 35.2246$, $B_3 = -28.5$, $B_4 = 3840$. 则$g_1 = -0.1363$, $g_2 = 3.0948$. 查表得$z_{0.025} = 1.96$.

计算u_1, u_2

$$|u_1| = 0.5285 < 1.96, |u_2| = 0.3381 < 1.96$$

接受H_0.

练 习

试用偏度峰度检验以下数据是否正态分布？$\alpha = 0.1$

$$147, 150, 152, 148, 155, 146, 149, 148, 151, 150$$

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, G_1 = \frac{B_3}{B_2^{3/2}}, G_2 = \frac{B_4}{B_2^2}$$

8 K-S检验

卡方拟合优度检验往往依赖于区间的划分, 不同的区间划分法可能得到差异较大的结果.柯尔莫哥洛夫检验(D_n检验)采取更加精细的方法克服这一缺点. 但其要求被检验的分布连续.
柯尔莫哥洛夫检验的基本思想：利用经验分布函数$F_n(x)$与理论分布函数的偏差的最大值D_n来判断总体分布$(0 < D_n < 1)$. 即利用$F_n(x)$与$F(x)$之间的偏差的上确界构造一个统计量:

$$D_n = sup_{-\infty<x<+\infty}|F_n(x) - F^*(x)|$$

当$H_0 : F(x) = F^*(x)$为真时,D_n的值一般应较小(n充分大),若D_n的值较大就应该拒绝H_0.
Theorem 2 设总体的分布函数$F(x)$连续,$X_1, \cdots, X_n$为X的一个样本, 样本容量为n, 当$H_0 : F(x) = F^*(x)$为真时, 有

$$\lim_{n\to\infty} P\{(D_n < \frac{\lambda}{\sqrt{n}}\} = k(\lambda)$$

其中$k(\lambda) = \sum_{k=-\infty}^{\infty}(-1)^k e^{(-2k^2\lambda^2)}, \lambda > 0; k(\lambda) = 0, \lambda \leq 0.$

查D_n分位表可得$D_{n,\alpha}$, 使得$P(D_n > D_{n,\alpha}) = \alpha$.

检验步骤

柯尔莫哥洛夫检验法: 对于样本观察值$(x_1, x_2, \cdots, x_n)$, 计算统计量D_n, 如果$D_n > D_{n,\alpha}$, 则拒绝假设$H_0 : F(x) = F^*(x)$, 否则接受H_0.
假如$F_0(x)$参数未知, 可用一个大容量样本值来估计参数, 然后进行D_n检验. 具体步骤如下:

1. 从总体进行抽样($n \geq 50$), 并得到排序样本值$x_1 \leq \cdots \leq x_n$.
2. 计算经验分布函数$F_n(x)$:
$$F_n(x) = \begin{cases} 0 & x \leq x_1 \\ \frac{i}{n} & x_i \leq x < x_{i+1}, i = 1, \cdots, n-1 \\ 1 & x \geq x_n. \end{cases}$$
3. 计算统计量
$$D_n = sup_{-\infty<x<+\infty}|F_n(x) - F^*(x)|$$
D_n可以在每个x_i处找经验分布和理论分布偏差最大的一个. 即计算
$$\delta_i = max\{F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F^*(x_i)\}$$
K-S统计量D_n就是所有δ_i中最大的一个.
4. 查表$D_{n,\alpha}$. 拒绝域为$D_n > D_{n,\alpha}$.

例子

Example 12 某设备进行寿命检验, 试验得*10*次无故障工作时间排序后分别为*: 420, 500, 920, 1380, 1510, 1650, 1760, 2100, 2300, 2350* 小时. 请问该设备的无故障工作时间是否服从$\frac{1}{\theta} = 1500$的指数分布$(\alpha = 0.10)$?

解: 需检验

$$H_0 : F^*(x) = 1 - e^{-\frac{x}{1500}}(x > 0)$$

计算统计量D_n, 得以下分析表:

	1	2	3	4	5	6	7	8	9	10
x_i	420	500	920	1380	1510	1650	1760	2100	2300	2350
$F^*(x_i)$	0.24	0.28	0.46	0.57	0.63	0.67	0.69	0.75	0.79	0.80
$\frac{i-1}{n}$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\frac{i}{n}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
δ_i	0.24	0.18	0.26	0.27	0.23	0.17	0.09	0.05	0.11	0.20

$D_n = max\{\delta_i\}$=0.27$< D_{10,0.10} = 0.37$. 接受H_0.

例 子

Example 13 *84个Etruscan人的男子头颅的最大宽度(毫米)如下(排序后):*
126 131 132 132 134 135 135 136 137 137 137 137 138 138 139
140 140 140 140 140 140 140 141 141 141 141 141 141 142 142
142 142 142 142 142 142 143 143 143 143 143 143 144 144 144
144 144 144 145 145 145 146 146 146 146 146 146 147 147 147
147 148 148 148 148 148 149 149 149 149 149 149 150 150 150
150 152 152 153 154 154 155 158 158
现问在显著性水平$\alpha = 0.1$下检验假设

$$H_0: \text{总体服从正态分布} f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

- n充分大时, $D_{n,\alpha}$采用$1.36/\sqrt{n}$近似计算。

9 K-S检验

设$(X_1,\cdots,X_{n_1})$, $(Y_1,\cdots,Y_{n_2})$使来自具有连续分布函数$F(x)$和$G(x)$的总体的样本, 且假定两个样本相互独立, 要检验假设

$$H_0: F(x)=G(x); H_1: F(x)\neq G(x)$$

设$F_{n_1}(x)$和$G_{n_2}(x)$分别是这两个样本所对应的经验分布函数, 作统计量

$$D_{n_1,n_2}=\sup_{-\infty<x<\infty}|F_{n_1}(x)-G_{n_2}(x)|$$

$$n=\frac{n_1n_2}{n_1+n_2}$$

斯米尔诺夫在分部函数$F(x)$和$G(x)$连续的条件下, 当H_0成立时, 导出了统计量D_{n_1,n_2}的极限分布:

$$\lim_{n_1\to\infty,n_2\to\infty}P\{\sqrt{n}D_{n_1,n_2}<\lambda\}=k(\lambda)$$

其中$k(\lambda)=\sum_{k=-\infty}^{\infty}(-1)^k e^{(-2k^2\lambda^2)},\lambda>0;k(\lambda)=0,\lambda\leq 0.$

对于给定显著水平α,令$n=\frac{n_1n_2}{n_1+n_2}$, 查表得$\hat{D}_{n,\alpha}$. 拒绝域为$\hat{D}_{n_1,n_2}\geq D_{n,\alpha}$.

例子

Example 14 两台机器生产同一型号产品, 甲抽*150*个样品, 乙抽*100*个样品, 样品与标准的偏差数据如下:

偏差X区间	甲样本频数	乙样本频数
$[-15,-10)$	*10*	*0*
$[-10,-5)$	*27*	*7*
$[-5,0)$	*43*	*17*
$[0,5)$	*38*	*30*
$[5,10)$	*23*	*29*
$[10,15)$	*8*	*15*
$[15,20)$	*1*	*1*
$[20,25)$	*0*	*1*
$\sum$	n_1=*150*	n_2=*100*

问甲乙是否来自同一总体($\alpha=0.05$)?

例子

解: 需检验假设

$$H_0 : F(x) = G(x)$$

计算统计量D_{n_1,n_2}, 得以下分析表:

X	甲累积频数	乙累积频数	$F_{n_1}(x)$	$G_{n_2}(x)$	$\lvert F_{n_1}(x) - G_{n_2}(x)\rvert$
-10	10	0	0.067	0.000	0.067
-5	37	7	0.247	0.070	0.177
0	80	24	0.533	0.240	0.293
5	118	54	0.787	0.540	0.247
10	141	83	0.940	0.830	0.110
15	149	98	0.993	0.980	0.013
20	150	99	1.000	0.990	0.010
25	150	100	1.000	1.000	0.000

$D_{n_1,n_2} = 0.293$, $n = \frac{150\times100}{150+100} = 60$, $D_{60,0.05} = 0.17231$.

$D_{n_1,n_2} > D_{60,0.05}$, 拒绝原假设$H_0$.

10 独立性检验

设有两个总体X,Y,给定显著性水平α,检验非参数假设:

$$H_0: X,Y\text{相互独立}, H_1: X,Y\text{不独立}$$

将X的所有可能取值分为r个不同组$A_1,\cdots,A_r$;将Y的所有可能值分为s个不同的组$B_1,\cdots,B_s$,对(X,Y)进行n次独立观测,分别记录事件$(X\in A_i, Y\in B_j)$出现的频数$n_{ij}(i=1,\cdots,r;j=1,\cdots,s)$,将所得结果列成$r\times s$的表格.

$X\|Y$	y_1	$\cdots$	y_s	$n_{i\cdot}$
x_1	n_{11}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rs}	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot s}$	n

$$n_{i\cdot}=\sum_{j=1}^{s}n_{ij}, n_{\cdot j}=\sum_{i=1}^{r}n_{ij}$$

独立性检验

基本思想: 独立$\Leftrightarrow$联合概率等于边缘概率之积.

记E_{ij}和T_{ij}分别为频数的实验值和理论值. 检验统计量为

$$\chi^2 = \sum \frac{(E_{ij} - T_{ij})^2}{T_{ij}}$$

其中$E_{ij} = n_{ij}$,

$$T_{i,j} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

所以统计量为

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n})^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

则n很大时,

$$\chi^2 \sim \chi^2((r-1)(s-1))$$

拒绝域为:

$$\chi^2 \geq \chi^2_{\alpha}((r-1)(s-1))$$

例子

Example 15 (文化水平与居民收入关系分析) 我 们 将 收 入 水 平X分 为: *(a)2000*以下, *(b)2000-3000*, *(c)3000-4000*, *(d)4000-5000*, *(e)5000*以上; 文化水平Y分为: *(1)*研究生, *(2)*本科生, *(3)*本科以下. 具体如下表. 试检验两者是否独立$(\alpha = 0.01)$.

	(1)	(2)	(3)	合计
(a)	186	38	35	259
(b)	227	54	45	326
(c)	219	78	78	375
(d)	355	112	140	607
(e)	653	285	259	1197
合计	1640	567	557	2764

解: 建立假设$H_0 : X, Y$互相独立. 我们计算频数理论值(括号中)如下:

	(1)	(2)	(3)	合计
(a)	186(153.68)	38(53.13)	35(52.19)	259
(b)	227(193.27)	54(66.87)	45(65.70)	326
(c)	219(222.50)	78(76.93)	78(75.57)	375
(d)	355(360.16)	112(124.52)	140(122.32)	607
(e)	653(710.23)	285(245.55)	259(241.22)	1197
合计	1640	567	557	2764

计算可得$\chi^2 = 47.9$, 查表得$\chi^2_{0.01}(8) = 20.090$.故拒绝原假设$H_0$.

练 习

Example 16 (成绩统计) 南京大学软件学院*2004*年*2*班《软件工程统计方法》课程的期末考试成绩按男生女生的分组统计如下表。

性别—等级	不及格	及格	中	良	优	合计
男生	*4*	*13*	*8*	*4*	*3*	*32*
女生	*1*	*5*	*3*	*5*	*4*	*18*
合计	*5*	*18*	*11*	*9*	*7*	*50*

试检验这次考试成绩跟性别是否独立？

11 小结

非参数检验与传统的参数检验相比, 有以下优点:

1. 非参数检验方法要求的假定条件比较少, 因而它的适用范围比较广泛.
2. 多数非参数检验方法要求的运算比较简单, 可以迅速完成计算取得结果, 因而比较节约时间.
3. 大多数非参数检验方法在直观上比较容易理解, 不需要太多的数学基础知识和统计学知识.
4. 大多数非参数检验方法可用来分析如象由等级构成的数据资料, 而对计量水准较低的数据资料, 参数检验方法却不适用.

但非参数统计方法也有以下缺点:

- 由于方法简单, 用的计量水准较低, 因此, 如果能与参数检验方法同时使用时, 就不如参数统计方法敏感. 若为追求简单而使用非参数检验方法, 其检验功效就要差些. 这就是说, 在给定的显著性水平下进行检验时, 非参数检验方法与参数检验方法相比, 第II类错误的概率β要大些.