Kolmogorov-Smirnov 检验法

问题的提出

- 在进行累计概率统计的时候，如何区分组之间是否有显著差异？
- Kolmogorov-Smirnov检验（K-S检验）基于累积分布函数，用以检验一个经验分布是否符合某种理论分布或比较两个经验分布是否有显著性差异。
- 两样本K-S检验由于对两样本的经验分布函数的位置和形状参数的差异都敏感而成为比较两样本的最有用且常规的非参数方法之一。

单样本K-S检验

- 单样本的K-S检验是用来检验一个数据的观测经验分布是否是已知的理论分布。当两者间的差距很小时，推断该样本取自已知的理论分布。
- 作为零假设的理论分布一般是一维连续分布 F（如正态分布、均匀分布、指数分布等），有时也用于离散分布（如Poisson分布）。即H_0：总体 X 服从某种一维连续分布 F。
- 检验统计量为

$$Z = \sqrt{n}\max_i\left(|F_n(x_{i-1}) - F(x_i)|, |F_n(x_i) - F(x_i)|\right)$$

- H_0真，Z依分布收敛于Kolmogonov 分布。

即，当样本取自一维连续分布F时，

$$Z \xrightarrow{d} K = \sup_x |B(F(x))|$$

- 注：当F是连续分布时，随机变量K 的分布不依赖于F。

Kolmogonov 分布

- 维纳过程W(t)：W(0)=0；具有平稳独立增量；且

$$W(t) \sim N(0, \sigma^2 t), t \geq 0$$

- 布朗桥：$B(t) = (W(t) | W(1) = 0), t \in [0,1]$

- 考虑随机变量 $K = \sup_{t \in [0,1]} | B(t) |$，其分布函数为

$$P(K \leq x) | = 1 - 2\sum_{i=1}^{+\infty} (-1)^{i-1} e^{-2i^2 x^2}$$

称之为Kolmogonov 分布。

例1. 对一台设备进行寿命检验，记录10次无故障工作时间（数据如下）。检验其是否服从1/1500的指数分布？

```
X=c(420, 500, 920, 1380, 1510, 1650, 1760, 2100,
    2300, 2350)
ks.test(X,"pexp", 1/1500)
```

Output:

```
D = 0.3015, p-value = 0.2654
alternative hypothesis: two-sided
```

结论： p值大于0.05，不拒绝原假设，认为此设备无故障工作时间服从1/1500的指数分布。

两样本K-S检验

- 假定有分别来自两个独立总体的两个样本。要想检验它们背后的总体分布相同的零假设，可以进行两独立样本的K-S检验。原理完全和单样本情况一样。只不过把检验统计量中零假设的分布换成另一个样本的经验分布即可。假定两个样本的样本量分别为n_1和n_2，用$F_1(X)$和$F_2(X)$分别表示两个样本的累积经验分布函数。再记 $D_j=F_1(X_j)-F_2(X_j)$。检验统计量近似正态分布，表达式为

$$Z=\max_j |D_j| \sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

例2. 有分别从两个总体抽出的25个和20个观察值的随机样本（数据如下）。检验其是否可以认为来自同一分布?

```
X=scan()
0.61 0.29 0.06 0.59 -1.73 -0.74 0.51 -0.56 0.39 1.64
    0.05 -0.06 0.64 -0.82 0.37 1.77 1.09 -1.28 2.36 1.31
    1.05 -0.32 -0.40 1.06 -2.47

Y=scan()
2.20 1.66 1.38 0.20 0.36 0.00 0.96 1.56 0.44 1.50 -0.30
    0.66 2.31 3.29 -0.27 -0.37 0.38 0.70 0.52 -0.71

ks.test(X,Y)
```

Output:

Two-sample Kolmogorov-Smirnov test

data: X and Y

D = 0.23, p-value = 0.5286

alternative hypothesis: two-sided

结论：p值大于0.05，不拒绝原假设，可以认为两个样本来自同一分布。

K-S检验与卡方检验的比较

- 相同点：都是采用实际频数和期望频数之差进行检验。
- 不同点：卡方检验主要用于类别数据，而K-S检验主要用于有计量单位的连续和定量数据。

- 卡方检验也可以用于定量数据，但必须先将数据分组才能获得实际的观测频数，而K-S检验法可以直接对原始数据的n个观测值进行检验，所以它对数据的利用较完整。

K-S检验的优势和劣势

- 作为一种非参数方法，具有稳健性；
- 不依赖均值的位置；
- 对尺度化不敏感；
- 适用范围广（不像 t 检验仅局限于正态分布，当数据偏离正态分布太多时t 检验会失效；
- 比卡方更有效；
- 如果数据确实服从正态分布，没有 t 检验敏感（或有效）。