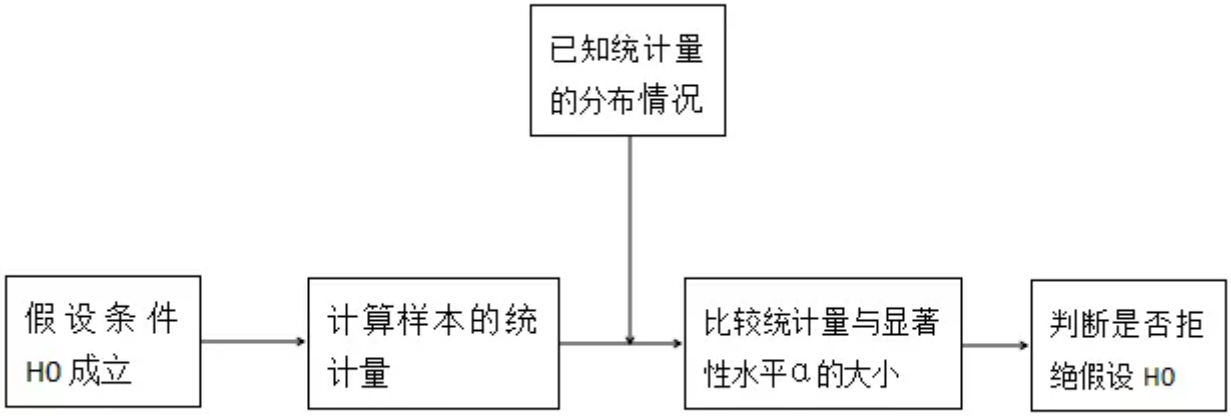
非参数正态性检验

原创 石头 机器学习算法那些事 2018-08-31

非参数正态性检验

前面两节介绍了采用Q-Q图和偏度与峰度来对采集样本进行正态性检验，本节介绍非参数性的正态性检验，非参数性的正态性检验算法思想大致相同，算法思想步骤为：首先假设条件H0成立，然后计算采集样本的统计量，最后在已知统计量分布的情况下比较统计量与显著性水平α的大小，根据比较结果判断是否拒绝检验假设H0（如下图）。



本文首先介绍了非参数正态性检验算法如 χ^2 拟合优度检验，K-S检验，S-W检验等，最后比较各非参数性正态检验的适用条件。

1、χ^2 拟合优度检验

χ^2 是在总体X的分布未知时，根据来自总体的样本，检验关于总体分布的假设的一种检验方法，比较样本的经验分布和所假设的理论分布之间的吻合程度来决定是否接受总体分布的原假设。比如，记录小明最近一年每天花在学习英语的时间，判断小明是否是英语爱好者。运用 χ^2 检验法来判断的步骤是：

（1）假设小明是英语爱好者。

（2）统计英语爱好者最近一年内每天学习英语的时间。

（3）计算英语爱好者每天学习英语的时间与小明每天学习英语的时间的差异，再计算这一年内学习英语时间的总差异，若总差异结果超过某一阈值，拒绝假设，即小明不是英语爱好者；反之，不拒绝假设，即小明是英语爱好者。

在用 χ^2 检验法检验假设H0时，需要用极大似然估计计算检验假设H0的参数，比如，若H0是正态分布，则需要用极大似然估计计算均值和方差；若H0是指数分布，则需要用极大似然估计计算均值；

χ^2**验统计量为**:

$$\chi^2=\sum_{i=1}^{k}\frac{n}{p_i}(\frac{f_i}{n}-p_i)^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}\sim\chi(k-1)$$

$$\chi^2=\sum_{i=1}^{k}\frac{n}{\hat{p}_i}(\frac{f_i}{n}-\hat{p}_i)^2=\sum_{i=1}^{k}\frac{(f_i-n\hat{p}_i)^2}{n\hat{p}_i}\sim\chi(k-r-1)$$

其中r是检验假设H0模型的被估参数，n为样本容量，离散化样本容量成k段，Pi为假设检验H0成立时第i个分段的频率，n*Pi，fi分别为第i段的理论频数和实际频数。

> 皮尔逊证明了χ^2统计量的分布服从（k-r-1）个自由度的χ^2分布的前提是样本容量n足够大。使用χ^2拟合优度检验正态分布需要注意大样本容量和n*Pi不能太小（≥5）这两个条件，若某一段出现的频数太小，则与其他的分段合并，达到频数≥5的条件。

最后比较样本χ^2检验统计量数值与显著性水平α的大小，来判断假设是否成功。

【例】

例 7.5　下面是 150 名 10 岁儿童的 IQ 得分，请检验其是否服从正态分布

125.9	99.3	133.4	100.0	131.9	98.2	137.1	97.4	135.9	105.9
143.8	116.7	151.1	104.9	75.3	95.0	78.6	97.7	76.3	114.9
66.1	111.8	68.9	103.2	73.0	99.8	74.1	103.2	73.4	109.1
118.5	112.3	119.0	114.1	121.9	111.4	123.7	109.5	127.8	109.3
84.0	113.2	81.2	107.8	83.3	108.5	83.9	115.7	82.7	113.2
83.9	112.8	84.5	113.4	79.9	108.6	78.9	120.1	84.8	109.8
77.6	113.2	76.9	108.4	85.	105.9	89.6	115.7	92.6	105.5
90.9	110.8	87.6	113.7	88.6	109.3	93.6	108.2	93.8	106.8
86.6	118.6	93.6	113.9	89.1	113.2	87.6	113.1	89.9	119.7
85.5	122.5	88.2	112.5	93.6	113.1	90.1	114.1	93.4	95.9
92.6	92.3	86.6	121.7	94.6	115.9	87.3	99.8	89.2	107.7
93.7	95.8	87.6	123.6	93.3	124.7	89.6	101.4	94.6	109.2
102.0	104.1	88.6	108.0	86.9	109.6	103.2	104.1	95.2	98.9
98.0	57.5	99.5	103.9	98.6	99.1	95.8	99.0	101.8	103.0
99.4	104.1	104.2	95.0	104.3	101.4	96.8	102.3	97.0	103.5

解：H0：IQ得分服从正态分布，H1：不服从正态分布，α=0.05，X = 101.294，S =15.585

表 7.3 正态分布拟合优度χ^2检验的计算表

IQ 得分组限	实际观测频数 O_i	标准化组限 Z_i	累计概率	概率	理论频数 E_i		$\frac{\lvert O_i-E_i\rvert^2}{E_i}$	$\sum\frac{\lvert O_i-E_i\rvert^2}{E_i}$
(1)	(2)	(3)	(4)	(5)	(6)=150*(5)		(7)	(8)
55.0 ~	1	−2.97048~	0.00149~	0.00844	1.2660	6.6450	0.06261	0.06261
65.0 ~	5	−2.32882	0.00993~	0.03586	5.3790			
75.0 ~	15	−1.68717	0.04579~	0.10210	15.3150		0.00648	0.06909
85.0 ~	31	−1.04551	0.14789~	0.19527	29.2905		0.09977	0.16886
95.0 ~	39	−0.40386	0.34316~	0.25082	37.6230		0.05040	0.21926
105.0~	36	0.23780	0.59398~	0.21644	32.4660		0.38468	0.60394
115.0~	15	0.87945	0.81042~	0.12546	18.8190		0.77500	1.37894
125.0~	4	1.52111	0.93588~	0.04884	7.3260			
135.0~	3	2.16276	0.98472~	0.01276	1.9140	9.5760	0.25938	1.63832
145.0~155	1	2.80441~ 3.44607	0.99748~ 0.99972	0.00224	0.3360			

其中O_i为第i段的实际观测频数，E_i为第i段的理论频数。因为最后两组的观测频数过小，则合并最后三组成一组，该组频数为8。

自由度 v = 7-1-2 = 4，$\chi^2_{0.05,4}=9.49$，统计量χ^2=1.63832 < 9.49，所以不拒绝零假设H0，即IQ得分服从正态分布。

2、K-S正态性检验

S检验是通过比较样本经验分布函数与给定分布函数来推断该样本是否来自给定分布函数的总体。比较容量n的经验分布函数Fn(x)与给定分布函数$F_0(x)$的间隔，构造统计量D为两个分布函数的间隔最大值，如下图。

$$Fn(x)=\frac{\sum_i (i\le x)_{\text{频数}}}{n}$$

$$D=\max|F_n(x)-F_0(x)|$$

步骤：

（1）提出假设：H0：Fn(x)= F0(x)，H1：Fn(x)≠$F_0(x)$。

（2）计算统计量D。

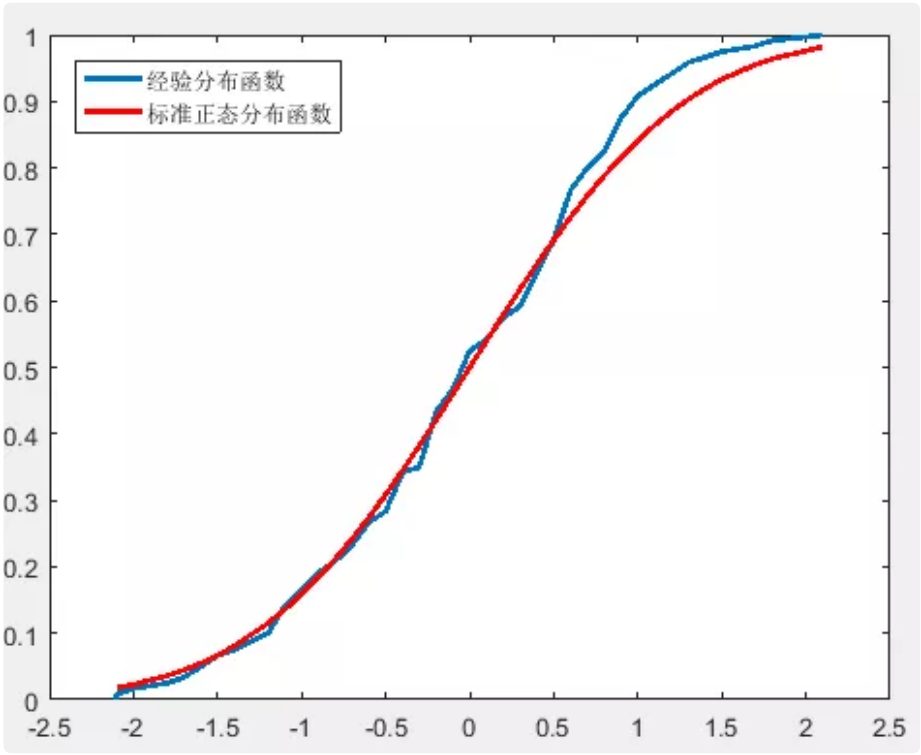
（3）根据给定的显著性水平α和样本数据个数n，确定单样本K-S检验的临界值Dα(n)。

（4）若$D<D(\alpha,n)$，则不拒绝假设H0；反之，拒绝假设H0。

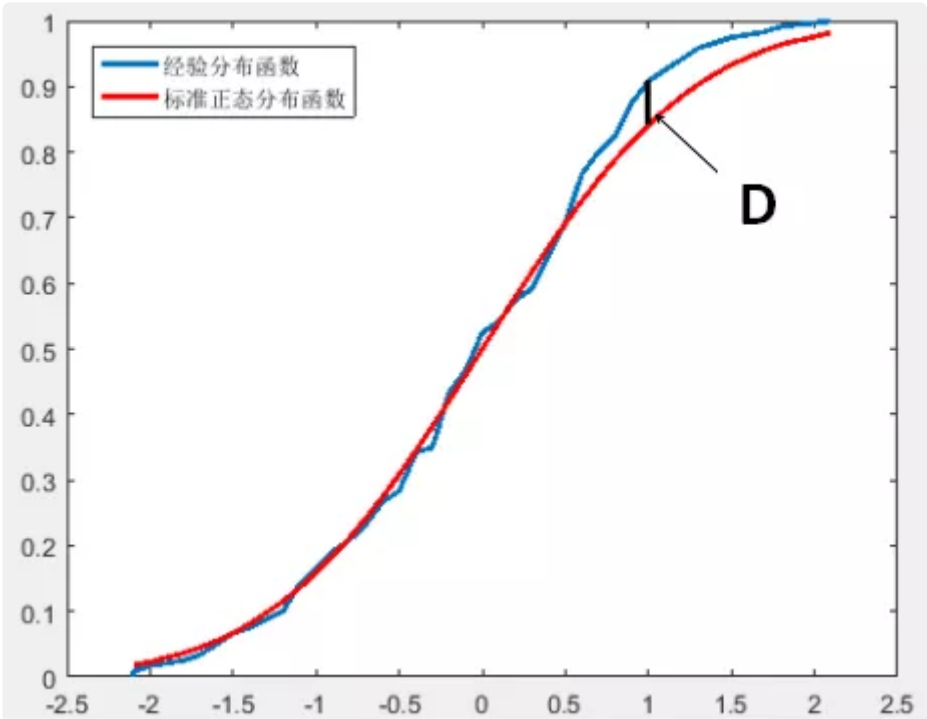
这个检验需要给定$F_0(x)$，因此非参数检验的K-S正态性检验只能做标准正态检验。

【例】验证一组39例抽样数据是否符合标准正态分布

（1）假设抽样数据符合标准正态分布；
（2）画出经验分布函数和标准正态分布函数的曲线图；



（3）确定统计量D；



（4）显著性水平α=0.05，样本容量n = 39，确定统计量的拒绝域最小值$D(\alpha,n)$，$D(\alpha,n)$可通过查表可得。

（5）比较统计量D与$D(\alpha,n)$的大小，若大于，则拒绝假设，反之，则不拒绝；

3、Lilliefor正态性检验

Lilliefor正态性检验是对K-S检验的修正，非参数K-S检验只能作标准正态分布检验，Lilliefor提出用样本均值和标准差代替总体的期望和标准差，然后再用K-S正态性检验法，步骤相同，不同点在于单样本K-S检验只能检测标准正态分布，Lilliefor检验能检测一般性的正态分布。

4、S-W正态性检验

S-W检验正态分布的思想与K-S检验一致，关键点在于如何求样本的统计量以及确定统计量的分布情况。

S-W检验称为W检验，统计量W定义为：

$$W=(\sum a_i y_i)^2/\sum(y_i-\bar{y})^2$$

其中$\bar{y}$是样本均值，$a=(a1,a2,...an)^T$，σ是样本来自正态分布的标准差，a的确切值是：

$$a=(m^T V^{-1}V^{-1}m)^{-\frac{1}{2}}m^T V^{-1}$$

其中V矩阵是n个标准正态分布的随机变量的顺序统计量的协方差矩阵。

给定显著性水平α和样本容量n，可以知道拒绝域的临界值$W_\alpha(n)$，比较统计量结果W与$W_\alpha(n)$的大小，判断是否拒绝原假设。

【例】 用函数rnorm获得一个标准正态分布的随机样本，然后用W检验它的正态性。

```
> y = rnorm(1:200)
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.99469, p-value = 0.7035
```

结果显示p-value值大于显著性水平0.05，因此不能拒绝零假设，即样本来自正态分布。

5、非参数检验算法的比较

（1）Lilliefor检验是对K-S检验的改进，可用于一般的正态性检验，而非参数检验的K-S检验只能做标准正态检验。

（2）χ2拟合优度检验的检验结果依赖于分组，而其他方法的检验结果与区间划分无关。

（3）拟合优度检验和K-S检验都采用实际频数和期望频数进行检验，前者既可用于连续总体，又可用于离散总体，而Kolmogorov-Smirnov检验只适用于连续和定量数据。

（4）SPSS规定：当样本含量3≤n≤5000时，结果以S—W(W 检验)为准，当样本含量n>5000结果以K-S检验(D检验)为准。

参考

https://blog.csdn.net/suncherrydream/article/details/51073001
http://www.docin.com/p-2006164716.html
http://www.dxy.cn/bbs/topic/26366190