模型-评价主题-统计类评价-KS检验 【gyj】

1. 模型名称

KS检验 (Kolmogorov-Smirnov Test)

2. 适用范围

- 单样本KS检验：用于比较**样本**与**理论分布** (预先给定的分布) 是否一致
- 双样本KS检验：用于比较**两个样本**概率分布是否相同

换言之，就是**观测得到的样本，声称其服从某一分布是否可信**

3. 形式

单一样本或者两个样本

4. 求解方法

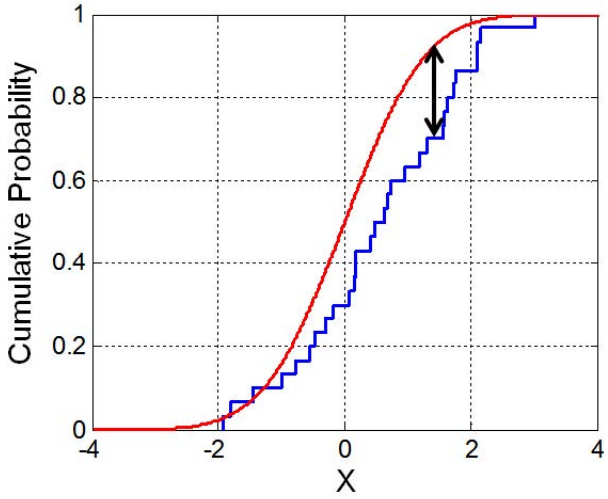
4.1 步骤

① 单样本KS检验

1. 设置零假设 H_0: 样本所来自的总体X分布服从理论分布$F(x)$
2. 画出$F(x)$与样本累计概率函数$F_n(x)$, 并求$F(x)$与$F_n(x)$的差的最大值D

$$D = max||F(x) - F_n(x)||$$

这代表着样本所属总体的分布$F_n(x)$与给定分布F(x)之间的**距离**



3. D(n, α)公式与常用近似表

【D(n, α) 表示在显著性水平为α，样本容量为n时，D的**拒绝域临界值**】

$$D(n,\alpha)=\frac{1}{\sqrt{n}}K_\alpha$$

期中K_α为Kolmogorov分布的置信度为1-α 的置信上限，下图为常用*D(n, α)*近似表

n\α	0.4	0.2	0.1	0.05	0.04	0.01
5	0.369	0.447	0.509	0.562	0.580	0.667
10	0.268	0.322	0.368	0.409	0.422	0.487
20	0.192	0.232	0.264	0.294	0.304	0.352
30	0.158	0.190	0.217	0.242	0.250	0.290
50	0.123	0.149	0.169	0.189	0.194	0.225
>50	$\frac{0.87}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.37}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

$$1-\alpha=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2K_\alpha^2}=\frac{\sqrt{2\pi}}{K_\alpha}\sum_{i=1}^{\infty}e^{\frac{-(2i-1)^2\pi^2}{8K_\alpha^2}}$$

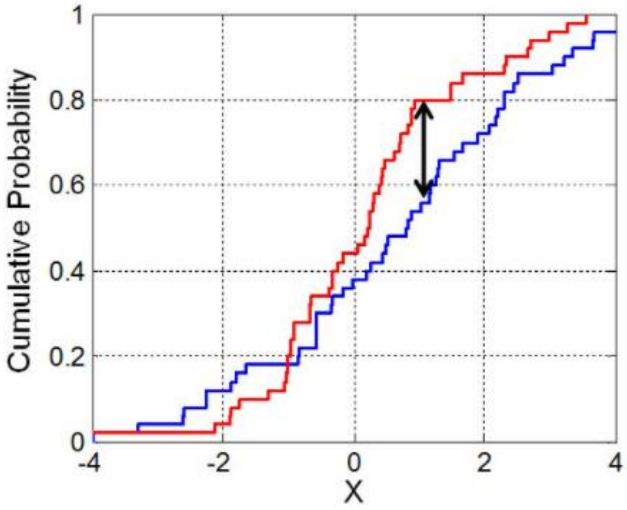
【❗关于置信度等概念的整理，详情参考第五部分补充资料】

4. 当实际观测*D<D(n, α)*，则接受零假设H_0， 反之则拒绝零假设H_0

②双样本KS检验

1. 设置零假设H_0：两样本概率分布相等，即$F(x_1) = F(x_2)$
2. 画出$F_n(x_1)$与$F_n(x_2)$，并求$F_n(x_1)$与$F_n(x_2)$的差的最大值D

$$D = max||F_n(x_1)| - F_n(x_2)|$$



3. 置信度表

Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided)

Table gives critical D-values for $\alpha = 0.05$ (upper value) and $\alpha = 0.01$ (lower value) for various sample sizes. * means you cannot reject H_0 regardless of observed D.

$n_2 \backslash n_1$	3	4	5	6	7	8	9	10	11	12
1	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	16/16	18/18	20/20	22/22	24/24
	*	*	*	*	*	*	*	*	*	*
3	*	*	15/15	18/18	21/21	21/24	24/27	27/30	30/33	30/36
	*	*	*	*	*	24/24	27/27	30/30	33/33	36/36
4		16/16	20/20	20/24	24/28	28/32	28/36	30/40	33/44	36/48
		*	*	24/24	28/28	32/32	32/36	36/40	40/44	44/48
5			*	24/30	30/35	30/40	35/45	40/50	39/55	43/60
			*	30/30	35/35	35/40	40/45	45/50	45/55	50/60
6				30/36	30/42	34/48	39/54	40/60	43/66	48/72
				36/36	36/42	40/48	45/54	48/60	54/66	60/72
7					42/49	40/56	42/63	46/70	48/77	53/84
					42/49	48/56	49/63	53/70	59/77	60/84
8						48/64	46/72	48/80	53/88	60/96
						56/64	55/72	60/80	64/88	68/96
9							54/81	53/90	59/99	63/108
							63/81	70/90	70/99	75/108
10								70/100	60/110	66/120
								80/100	77/110	80/120
11									77/121	72/132
									88/121	86/132
12										96/144
										84/144

For larger sample sizes, the approximate critical value D_α is given by the equation

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where the coefficient is given by the table below.

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

Examples: (1) At $\alpha = 0.05$ and samples sizes 5 and 8, $D_\alpha = 30/40 = 0.75$.

(2) At $\alpha = 0.01$ and samples sizes 15 and 28, $D_\alpha = 1.63\sqrt{\frac{15+28}{15 \cdot 28}} = 0.522$.

4. 当实际观测*D<D(n, α)*，则接受零假设H_0， 反之则拒绝零假设H_0

4.3 实例

检验样本X是否服从正态分布

$$X = [87, 77, 92, 68, 80, 78, 84, 77, 81, 80, 80, 77, 92, 86, 76, 80, 81,$$
$$75, 77, 72, 81, 72, 84, 86, 80, 68, 77, 87, 76, 77, 78, 92, 75, 80, 78]$$

1. 计算样本X的均值与方差

$$\overline{X} = \frac{87 + 77 + 92 + \ldots + 80 + 78}{35} = 79.7429$$
$$S^2 = \frac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \ldots + (x_n - \overline{X})^2}{n} = 35.2555$$

2. 由此，得到理论分布函数：N (79.7429 , 35.2555)
3. 计算D

- 计算样本累计频率 $F_n(x)$

序号	数字	次数	累计次数	样本累计频率 Fn(x)
1	68	2	2	2/35=0.0571
2	72	2	4	4/35=0.1143
3	75	2	6	6/35=0.1714
4	76	2	8	8/35=0.2286
5	77	6	14	14/35=0.4000
6	78	3	17	17/35=0.4857
7	80	6	23	23/35=0.6571
8	81	3	26	26/35=0.7429
9	84	2	28	28/35=0.8000
10	86	2	30	30/35=0.8571
11	87	2	32	32/35=0.9143
12	92	3	35	35/35=1.0000

- 计算理论累积频率 F(x)
  - 方法一：用积分计算样本所拟合的正态分布的理论累计频率（存在一定误差）

    a) x处正态分布概率为f(x)，其中σ 是标准差，μ是数学期望

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

b) 用积分算出从无穷小到x值的累积频率F(x)，F(x)和Fn(x)的差的绝对值得到D

- 方法二：用自带函数拟合正态分布并得到理论累积频率F(x)，F(x)和Fn(x)的差的绝对值得到D

累积频率	理论累积频率	D
0.0571	0.0224	0.0348
0.1143	0.0928	0.0215
0.1714	0.2087	0.0373
0.2286	0.2611	0.0325
0.4000	0.3195	0.0805
0.4857	0.3829	0.1029
0.6571	0.5171	0.1400
0.7429	0.5847	0.1581
0.8000	0.7664	0.0336
0.8571	0.8575	0.0003
0.9143	0.8925	0.0218
1.0000	0.9819	0.0181

结论：

$$D_{max}=0.1581$$

4. 计算D(n , α)

- 设α = 0.05
- 求拒绝域：

$$W=[D(35,0.05),+\infty]\subseteq[D(50,0.05),+\infty]=[0.189,+\infty]$$

5. 比较D与D(n , α)的大小并得出结论

$$D\notin W$$

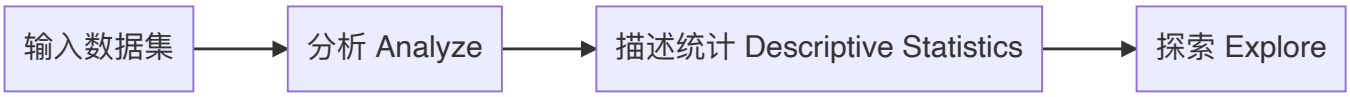
故认为H_0为真，即X服从正态分布

4.4 代码实现

4.4.1 SPSS 🌟

步骤：

SPSS检验正态分布



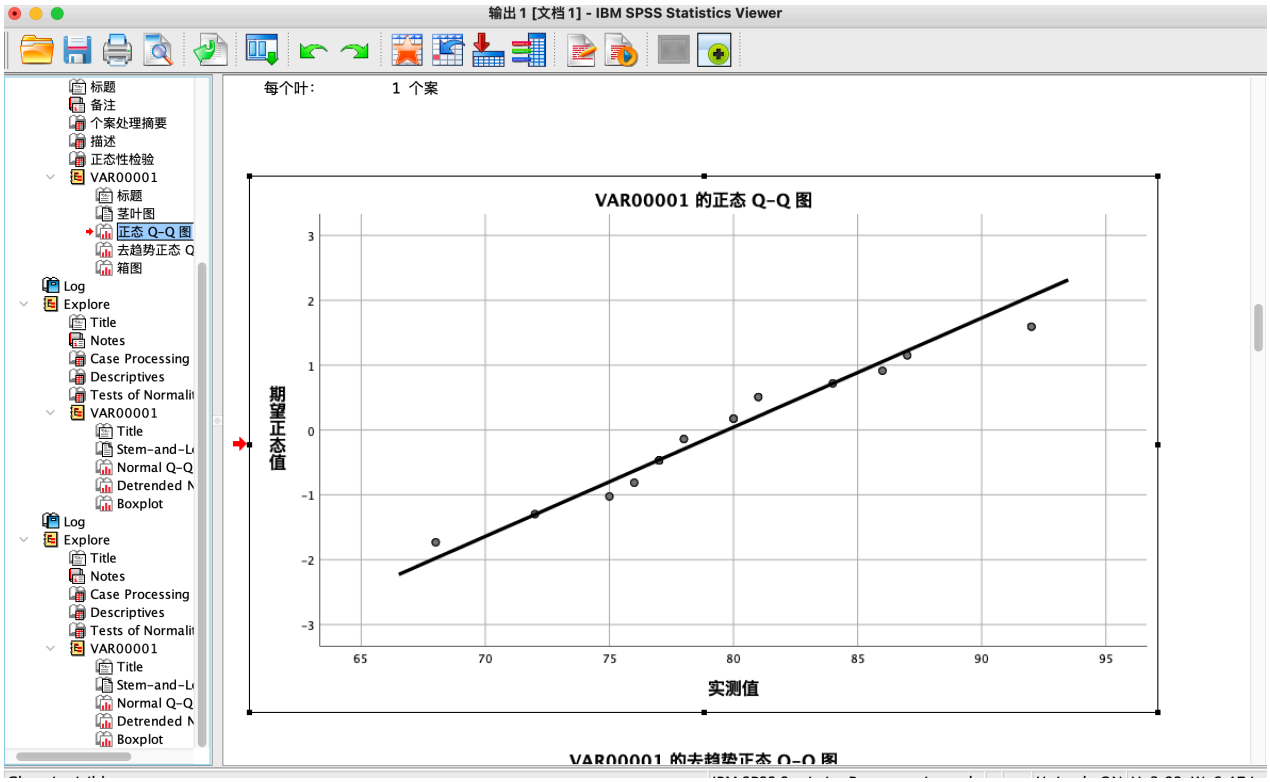
结果：

$$D = 0.159018$$

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
VAR00001	0.159018	35	.025	.949	35	.104

a. Lilliefors Significance Correction



4.4.3 Matlab利用内置函数

代码：

KS2.m

```
%KS检验
%注意事项
%需要输入样本，即为X，检查X是否符合正态分布，会输出D
%查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同。反之，则分布不同

%X为待检验的样本，检验其是否服从正态分布
X=[87 77 92 68 80 78 84 77 81 80 80 77 92 86 76 80 81 75 77 72 81 72 84 86 80 68 77 87
76 77 78 92 75 80 78];
N=length(X);
tmp1=unique(X);%将X中重复的样本除去
tmp2=histc(X,tmp1);%确定每一种情况的数目
%下面为计算样本不同情况的累积频率
F=0; %初始化累计次数
Fn=zeros(1,length(tmp1));%初始化数组，Fn为样本每种情况的累计频率
for i=1:length(tmp1)
    F=F+tmp2(i);
    Fn(i)=F/N;
end

%下面为计算样本所拟合的正态分布的理论累积频率
[mu,sigma]=normfit(X);%拟合正态分布，mu为期望，sigma为标准差
A=normcdf(tmp1,mu,sigma);%计算理论上分布的累积频率

D = max(abs(Fn-A)) %得到样本与理论分布函数差的最大值
%查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同
%反之，则分布不同
```

结果：

```
>> KS2

D =

    0.1590
```

4.4.3 Matlab用积分近似

代码：

KS3.m

```
%KS检验
%注意事项
%需要输入样本，即为X，检查X是否符合正态分布，会输出maxD
%查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同。反之，则分布不同

%X为待检验的样本，检验其是否服从正态分布
X=[87 77 92 68 80 78 84 77 81 80 80 77 92 86 76 80 81 75 77 72 81 72 84 86 80 68 77 87
76 77 78 92 75 80 78];
N=length(X);
```

```
tmp1=unique(X);%将X中重复的样本除去
tmp2=histc(X,tmp1);%确定每一种情况的数目
%下面为计算样本不同情况的累积频率
F=0; %初始化累计次数
Fn=zeros(1,length(tmp1));%初始化数组，Fn为样本每种情况的累计频率
for i=1:length(tmp1)
    F=F+tmp2(i);
    Fn(i)=F/N;
end

%A=zeros(1,length(tmp1));%初始化数组，A为理论每种情况的累计频数
for i=1:length(tmp1)
    l=min(min(tmp1),mu-10*sigma);
    x=l:sigma/1000:tmp1(i);
    f=1/(sigma*sqrt(2*pi))*exp(-(x-mu).^2/(2*sigma^2));
    A(i)=trapz(x,f);
end


    u=mean(X);
    d=std(X,1);
    A=zeros(1,length(tmp1));
    for i=1:length(tmp1)
        l=min(min(tmp1),u-10*d);
        x=l:d/1000:tmp1(i);
        f=1/(d*sqrt(2*pi))*exp(-(x-u).^2/(2*d^2));
        A(i)=trapz(x,f);
    end


maxD=max(abs(Fn-A))
%end
```

结果：

```
>> KS3

maxD =

    0.1581
```

4.4.4 Python用积分近似

代码：

KS4.py

```
#KS检验
from math import *
#由于积分是用步长计算的，存在一定误差，推荐使用matlab版本或使用SPSS软件
```

```
#需要输入样本，即为X，检查X是否符合正态分布，会输出maxD，查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同。反之，则分布不同
#使用python3.6

if __name__ == '__main__':
    #X为待检验的样本，检验其是否服从正态分布
    X=
[87,77,92,68,80,78,84,77,81,80,80,77,92,86,76,80,81,75,77,72,81,72,84,86,80,68,77,87,76,77,78,92,75,80,78]
    N=len(X)
    #下面为计算样本不同情况的累积频率
    X.sort()#对X排序
    #下面为将X中重复的数字去掉并确定每一种数字的数目
    tmp1=[]#tmp1为将X中重复数字去掉并排序的序列
    tmp2=[]#tmp2为tmp1中对应数字的频数
    num=1
    for i in range(N-1):
        if(X[i]!=X[i+1]):
            tmp1.append(X[i])
            tmp2.append(num)
            num=1
        else:
            num=num+1
    tmp1.append(X[N-1])
    tmp2.append(num)
    #下面为计算样本不同情况的累积频率
    Fn=[]
    F=0#初始化累计次数
    for i in range(len(tmp1)):
        F +=tmp2[i]
        Fn.append(F/N)
    #下面为求解样本拟合的正态分布
    #mu为期望，sigma为标准差
    mu=sum(X)/N
    sigma=0
    for i in range(N):
        sigma +=(X[i]-mu)*(X[i]-mu)
    sigma=sqrt(sigma/(N-1))
    #用积分计算样本所拟合的正态分布的理论累积频率
    #由于积分是用步长计算的，存在一定误差
    A=[]
    p=0#初始化累计频率
    l=min(min(tmp1),mu-10*sigma)#计算积分从该值来替代负无穷大
    for i in range(len(tmp1)):
        while l<tmp1[i]:
            p +=1/(sigma*sqrt(2*pi))*exp(-(l-mu)*(l-mu)/(2*sigma*sigma))*sigma/1000
            l +=sigma/1000
        A.append(p)
    #得到样本与理论分布函数差
```

```
    D=[]
    for i in range(len(tmp1)):
        D.append(abs(Fn[i]-A[i]))
    maxD=max(D)
    print("样本与理论分布函数差的最大值为:")
    print(maxD)
    #查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同
    #反之，则分布不同
    print("查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同.反之，则分布不同")
```

结果：

```
样本与理论分布函数差的最大值为:
0.15910570552390602
查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同.反之，则分布不同
```

4.4.5 c++用积分近似

代码：

KS5.cpp

```
#include <iostream>
#include <math.h>
#include<algorithm>
#define pi 3.1415926
using namespace std;
//由于积分是用步长计算的，存在一定误差，推荐使用matlab版本或使用SPSS软件
//需要输入样本，即为X，检查X是否符合正态分布，会输出maxD，
//查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同。反之，则分布不同
int main()
{
    //X为待检验的样本，检验其是否服从正态分布
    //N为样本的数目，请先输入
    const int N=35;
    int n=0;//n为样本中不重复的数字的数目，后面将求解
    double X[N]=
{87,77,92,68,80,78,84,77,81,80,80,77,92,86,76,80,81,75,77,72,81,72,84,86,80,68,77,87,76
,77,78,92,75,80,78};
    //下面为计算样本不同情况的累积频率
    sort(X,X+N);//对X排序
    //下面为将X中重复的数字去掉并确定每一种数字的数目
    double tmp1[N],tmp2[N];
    int num=1;
    for(int i=0;i<N-1;i++)
        if(X[i]!=X[i+1])
        {
            tmp1[n]=X[i];
            tmp2[n]=num;
            num=1;
            n++;
```

```
    }
    else
        num++;
tmp1[n]=X[N-1];
tmp2[n]=num;
n++;
//下面为计算样本不同情况的累积频率
double *Fn;
Fn=new double[n];
double F=0;//初始化累计次数
for(int i=0;i<n;i++)
{
    F +=tmp2[i];
    Fn[i]=F/N;
}
//下面为求解样本拟合的正态分布
//mu为期望，sigma为标准差
double mu=0,sigma=0;
for(int i=0;i<n;i++)
    mu +=tmp1[i]*tmp2[i];
mu /=N;
for(int i=0;i<N;i++)
    sigma +=pow((X[i]-mu),2);
sigma=sqrt(sigma/(N-1));
//用积分计算样本所拟合的正态分布的理论累积频率
//由于积分是用步长计算的，存在一定误差
double *A;
A=new double[n];
double p=0,l;//初始化累计频率
l=tmp1[0]<(mu-10*sigma)?tmp1[0]:(mu-10*sigma);
for(int i=0;i<n;i++)
{
    while(l<tmp1[i])
    {
        p +=1/(sigma*sqrt(2*pi))*exp(-(l-mu)*(l-mu)/(2*sigma*sigma))*sigma/1000;
        l +=sigma/1000;
    }
    A[i]=p;
}
//得到样本与理论分布函数差
double *D;
D=new double[n];
double maxD=0;
for(int i=0;i<n;i++)
{
    D[i]=(Fn[i]-A[i])>0?(Fn[i]-A[i]):(-Fn[i]+A[i]);
    if(D[i]>maxD)
        maxD=D[i];
}
```

```
    cout<<"样本与理论分布函数差的最大值为:"<<maxD<<endl;
    cout<<"查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同.反之，则分布不同";
    //查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同.反之，则分布不同
    return 0;
}
```

结果：

```
样本与理论分布函数差的最大值为:0.159106
查D的参数表，当D<D(N,α)时（N为样本容量，α为显著性水平），分布相同.反之，则分布不同%
```

5. 补充资料

1. 数模官网 - KS检验
2. 标准正态分布表
3. 上文中用到的其他知识点

- **假设检验**：假设检验是围绕对原假设内容的审定而展开的。如果**原假设正确**且我们**接受**了（这意味着我们也同时拒绝了备择假设），或**原假设错误**我们**拒绝**了，这表明我们作出了正确的决定。但由于假设检验是根据样本提供的信息进行推断的，也就有犯错误的可能。
- **显著性水平**：假设检验这种方式本身会产生一种情况，就是原假设正确，但我们却把它当作错误加以拒绝。犯这种错误的概率用α表示，统计上把α成为假设检验中的显著性水平，也是决策中面临的风险。换句话说，α表示“原假设为真时，拒绝原假设的概率”。
    - α通常取0.05或0.01，这表明，当做出接受原假设的决定时，其正确的可能性（概率）为95%或99%。
- **置信度（置信水平）**：1-α，表明了区间估计的可靠性。