非参数检验包括:

(一)定性检验法:

正态概率纸检验

(二)数值检验法:

皮尔逊(Pearson) χ^2 拟合检验

柯尔莫哥洛夫检验

斯米尔诺夫检验

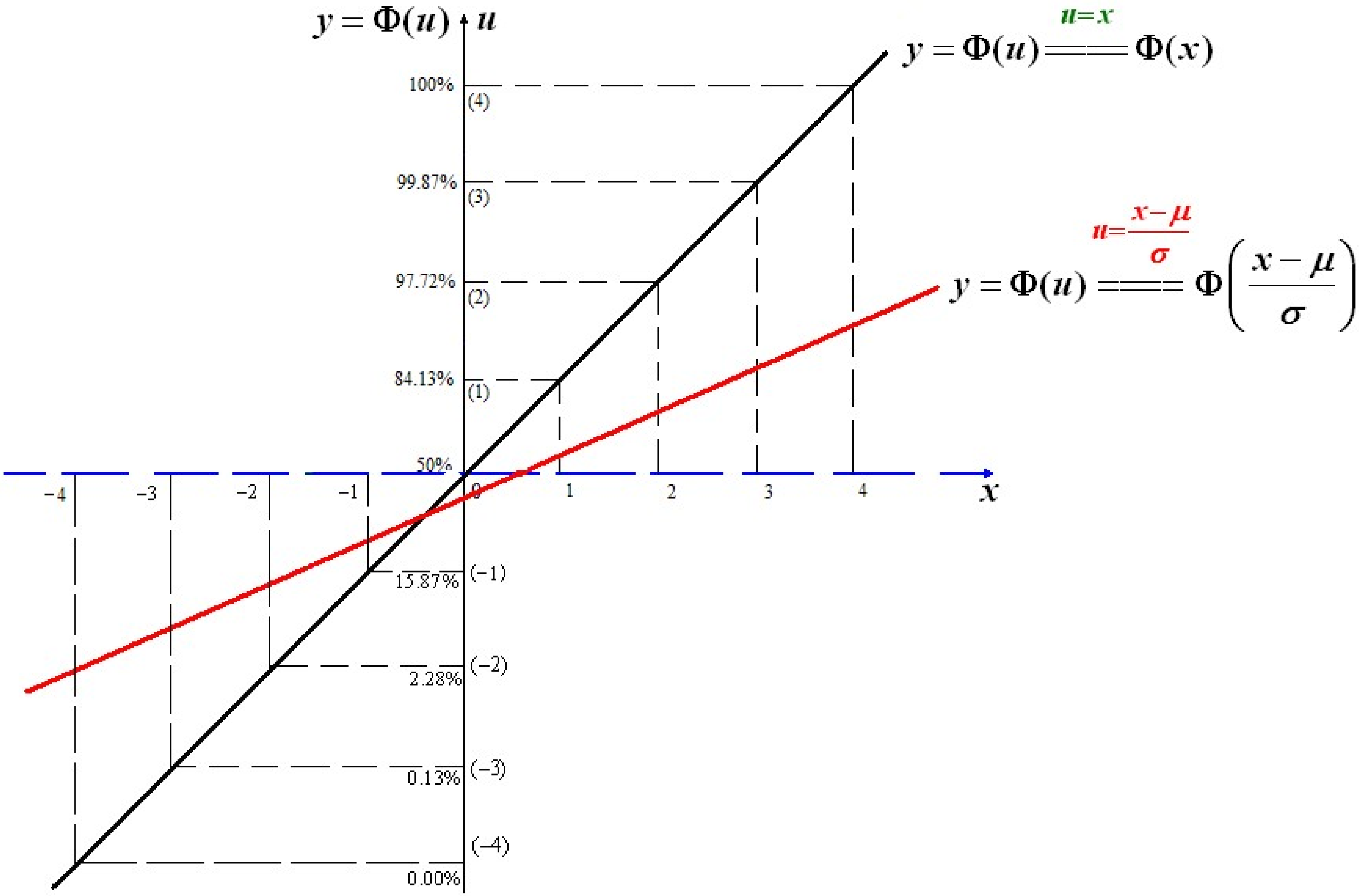
Shapiro-Wilk W 检验与**Agostino's D** 检验(不讲)

Wilcoxon秩和检验

(补) 偏度峰度检验.

一、正态概率纸检验

1. 正态概率纸的构造:



2. 检验方法:

检验假设: $\boxed{H_0: X \sim N(\mu,\sigma^2) \quad 或 \quad H_0: F(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)}$

(μ,σ^2是未知参数)

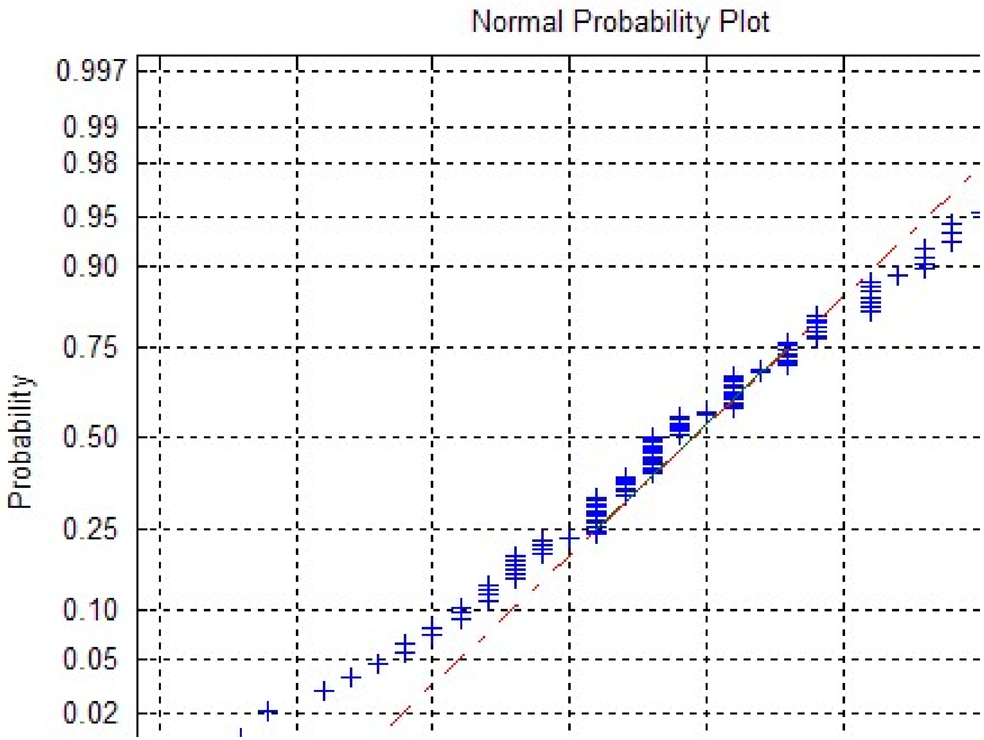
理论依据: 经验分布函数 $\xrightarrow[\text{近似}]{n\to\infty}$ 总体的分布函数.

检验方法: 当样本值算出的m个点$(x_i', F_n(x_i'))\ (i=1,2,\cdots,m)$近似在一直线上, 就接受$H_0$.

【例3.14(P_{99})】某工厂生产一种220伏25瓦的白炽灯泡, 其光通量用X表示. 现从该厂生产的灯泡中随机抽取120个灯泡测得其光通量的数据如下所示. 试问光通量X是否服从正态分布$N(\mu,\sigma^2)$.

216	203	197	208	206	209	206	208	202	203	206	213	218	207	208
202	194	203	213	211	193	213	208	208	204	206	204	206	208	209
213	203	206	207	196	201	208	207	213	208	210	208	211	211	214

220	211	203	216	224	211	209	218	214	219	211	208	221	211	218
218	190	219	211	208	199	214	207	207	214	206	217	214	201	212
213	211	212	216	206	210	216	204	221	208	209	214	214	199	204
211	201	216	211	209	208	209	202	211	207	202	205	206	216	206
213	206	207	200	198	200	222	203	208	216	206	222	213	209	219



注意1:中间的点离直线位置的偏差不能过大,两头的点的偏差可以允许大一些.否则就拒绝H_0.

二、皮尔逊 (Pearson) χ^2 拟合检验

(一) 单总体检验

【引例】我和一同学进行乒乓球比赛, 我宣称自己的胜率是0.8, 现打了10局, 我胜了3局, 可否认为我说了谎?

思路: 设 $X=\begin{cases}1, & \text{一局中我赢}\\ 0, & \text{一局中我输}\end{cases}$, 即检验 $H_0: X\sim b(1,0.8)$.

$X=i$	0	1
p_i	0.2	0.8
理论频数 np_i	2	8
实际频数 n_i	7	3

检验统计量及H_0的拒绝域为 $D=\sum_{i=0}^{1}\left|n_i-np_i\right|>F?_{1-\alpha}(x)$.

设F(x)—总体X的分布函数, 未知;

$F_0(x)$——一个完全已知或类型已知但参数未知的分布函数.

利用样本$X_1,\cdots,X_n$检验关于总体X的假设:

$$\boxed{H_0: F(x)=F_0(x)} \tag{3.14}$$

说明1: 此处X不论是连续或离散型, 也不论是一维或多维.

$F_0(x)$———X的理论分布.

检验原理: 通过样本的实际频数与理论频数之间差异的大小, 推断该样本是否来自给定分布函数的总体.

一) 总体X为离散型且取值有限

1. Pearson χ^2统计量:

(1) 设总体X是仅取有限个不同值$a_1,\cdots,a_k$的离散型随机变量，则原检验问题(3.14) $H_0:F(x)=F_0(x)$改为:

$$\boxed{H_0:P(X=a_i)=p_i,\quad i=1,\cdots,k} \tag{3.15}$$

(2) 构造检验统计量:

① 当F₀(x)中不含未知参数时,

$$\boxed{\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\ \underset{H_0\text{为真}}{\overset{n\to\infty}{\sim\!\sim\!\sim}}\ \chi^2(k-1)} \tag{3.16}$$

其中，n_i——a_i的观察频数 (即a_i在样本$x_1,\cdots,x_n$中出现的频数)

np_i——a_i的理论频数

χ^2——观察频数n_i与理论频数np_i相对差异的总和

② 当$F_0(x)$中未知参数的个数为 m 时，即理论分布为 $F_0(x;\theta_1,\cdots,\theta_m)$，

$$\boxed{\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}\ \underset{H_0\text{为真}}{\overset{n\to\infty}{\sim\!\sim\!\sim}}\ \chi^2(k-m-1)} \tag{3.17}$$

说明2: ★ 当 $\hat{\theta}_1,\cdots,\hat{\theta}_m$ 是按Fisher的条件求得时，(3.17)成立；

★ 实际应用时，$\hat{\theta}_1,\cdots,\hat{\theta}_m$ 一般用**MLE**代替，此时(3.17)不一定成立，但仍用(3.17*)作为H_0的否定域.

2. H_0的拒绝域:

$$\boxed{\chi^2>\chi^2_{1-\alpha}(k-m-1)} \tag{3.17*}$$

【★例3.15(P_{104})】某厂宣称自己产品的合格率为99%，检验人员从该厂的一批产品中抽查了100件，发现有两件次品. 在$\alpha=0.1$下，能否据此断定该厂谎报合格率.

解: 设 $X=\begin{cases}1, & \text{产品为正品} \\ 0, & \text{产品为次品}\end{cases}$, 则 $X\sim b(1,p)$.

◆(非参数检验法) 检验 $H_0: X\sim b(1,0.99)$.

$X=x_i$	0	1
p_i	0.01	0.99
理论频数 np_i	1	99
实际频数 n_i	2	98

检验统计量 $\chi^2=\sum_{i=1}^{2}\frac{(n_i-np_i)^2}{np_i}$,

H_0的拒绝域为 $\chi^2>\chi^2_{1-\alpha}(k-1)$,

代入样本检验: $\chi^2=\frac{(2-1)^2}{1}+\frac{(98-99)^2}{99}\approx 1.01<\chi^2_{0.9}(2-1)=2.706$, 故不拒绝$H_0$, 即认为该厂未谎报合格率.

◆(参数检验法)

检验 $H_0: p = p_0$, $p \neq p_0$ $(p_0 = 0.99)$.

利用中心极限定理构造检验统计量

$$u = \frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}} \underset{n\to\infty}{\overset{H_0\text{为真}}{\sim\sim\sim\sim\sim}} N(0,1),$$

H_0的拒绝域为 $|u| > u_{1-\alpha/2}$,

代入样本检验: 样本$(\underbrace{1,\cdots,1}_{98\text{个}},0,0)$, $|u| = \left|\dfrac{98-100\cdot 0.99}{\sqrt{100\cdot 0.99\cdot 0.11}}\right| \approx 1 < u_{0.95}$

$= 1.645$, 故不拒绝H_0, 即认为该厂未谎报合格率.

二)总体X为离散型但取值无限, 或X是连续型

不论X是一维还是多维, 如果X是离散型但取值无限、或X是连续型, 都可将其作离散化且取值有限处理.

【★例3.16(P_{106})】1991年某校工科研究生有60名以数理统计作为学位课, 考试成绩如下:

93	75	83	93	91	85	84	82	77	76	77	95	94	89	91
88	86	83	96	81	79	97	78	75	67	69	68	84	83	81
75	66	85	70	94	84	83	82	80	78	74	73	76	70	86
76	90	89	71	66	86	73	80	94	79	78	77	63	53	55

试问考试成绩是否服从正态分布 $(\alpha = 0.1)$.

注意2: χ^2拟合检验法要求: $n \geq 50$且$np_i \geq 5$或$n\hat{p}_i \geq 5$.

(二)列联表检验

设随机向量(X,Y), X的可能取值是1, 2, …, r, Y的可能取值是1, 2, …, s. 现在对(X,Y)进行了n次独立观察, 发现“$X=i,Y=j$”的次数为n_{ij}, 要据此检验假设:

$$\boxed{H_0: X与Y独立}$$

1. Pearson χ^2统计量:

(1) 绘制列联表 (Contingency Table):

X \ Y	1	2	…	s	$n_{i\cdot}$
1	n_{11}	n_{12}	…	n_{1s}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2s}	$n_{2\cdot}$
⋮	⋮	⋮	…	⋮	⋮
r	n_{r1}	n_{r2}	…	n_{rs}	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot s}$	n

此时检验问题"H_0: X与Y独立"变成

$$H_0: p_{ij}=p_{i\cdot}p_{\cdot j},\quad i=1,\cdots,r,\quad j=1,\cdots,s.$$ (其中 $p_{ij}, p_{i\cdot}, p_{\cdot j}$ 未知).

(2) 求得边缘分布的极大似然估计

$$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n},\ i=1,\cdots,r;\quad \hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n},\ j=1,\cdots,s.$$

(3) 构造检验统计量:

思路: 检验统计量及 H_0 的拒绝域为 $D=\sum_{i=1}^{r}\sum_{j=1}^{s}\left|\hat{p}_{ij}-\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right|> F?_{1-\alpha}(x)$.

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}=n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right)\overset{H_0:p_{ij}=p_{i\cdot}p_{\cdot j}\text{真}}{\underset{n\to\infty}{\sim\sim\sim}}\chi^2((r-1)(s-1))\tag{3.24}$$

2. H_0的拒绝域: $\chi^2>\chi^2_{1-\alpha}((r-1)(s-1))$.

说明4: 对X、Y是连续取值的情况, 可将其离散化处理.

【★例 3.17(P_{110})】某研究所推出一种感冒特效新药,为证明其疗效, 选择 200 名患者为志愿者. 将他们等分为两组, 分别不服药或服药, 观察三日后痊愈的情况, 得出下列数据:

是否痊愈 / 是否服药	痊愈者数	未痊愈者数	合计
未服药者数	48	52	100
服药者数	56	44	100
合计	104	96	200

问新药是否疗效明显? ($\alpha = 0.25$)

3. Pearson χ^2 检验的优缺点:

(1) 优点: 此法使用范围广:

不论总体是一维还是多维, 是离散型还是连续型;

也不论总体分布中参数是已知还是未知;

甚至不仅可以用于全样本, 也可用于截尾样本, 还可用于成群数据.

(2) 缺点: 由于分组处理样本的观察值, 有时虽然假设 $H_0: F(x) = F_0(x)$ 不成立, 但在某种划分之下, 并不影响统计量的观察值, 因而很容易犯第II类错误. 而柯尔莫哥洛夫检验可以克服这一缺点.