卡方拟合检验

设计与制作：刘琼荪

引例1

有差异！

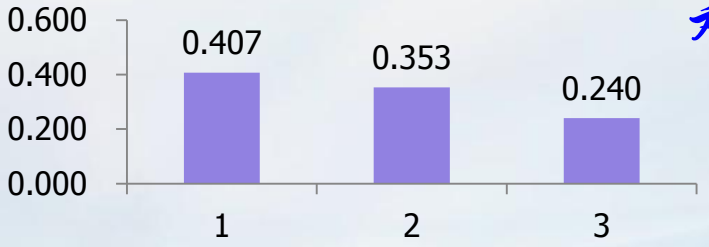
是否显著？

统计表

品牌	1	2	3
购买人数	61	53	36

150

频率直方图

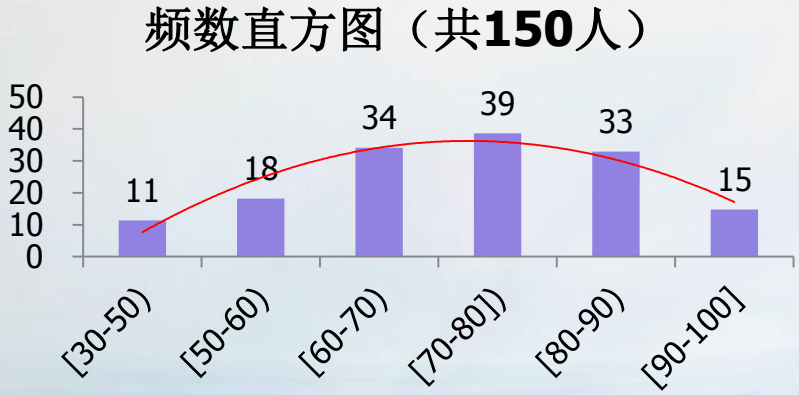


H_0：无显著差异，　H_1：有显著差异

or

$$H_0: X \sim \begin{pmatrix} 1 & 2 & 3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}, \quad H_1: X \sim \begin{pmatrix} 1 & 2 & 3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

引例2　重庆大学某次概率论与数理统计课程考试后，学生成绩分类统计如下



问能否认为学生成绩呈正态分布?

$H_0 : X \sim N(\mu, \sigma^2), \quad H_1 : X \nsim N(\mu, \sigma^2)$

共同特征

经验分布

$H_0: X \sim F_0(x), \quad H_1: X$不服从分布$F_0(x)$

采用卡方检验法

1. 离散(分类)变量的卡方检验

设X表示离散型随机变量，取值记为$1,2,\cdots,r$

提出问题：

$$H_0: X \sim \begin{pmatrix} 1 & 2 & \cdots & r \\ p_1 & p_2 & \cdots & p_r \end{pmatrix}, \quad H_1: X \not\sim \begin{pmatrix} 1 & 2 & \cdots & r \\ p_1 & p_2 & \cdots & p_r \end{pmatrix}$$

第i个类别	实际频数	v_i
	理论频数	np_i

（在H0下）

距离平方和：$\sum_{i=1}^{r}(v_i - np_i)^2 \Rightarrow \min$

卡方检验统计量：

$$\chi^2=\sum_{i=1}^{r}\frac{(v_i-np_i)^2}{np_i}$$

在H_0下，$\chi^2=\sum_{i=1}^{r}\frac{(v_i-np_i)^2}{np_i}\sim\chi^2(r-1)$

H0的拒绝域为：

$$K_0=\{\chi^2>\chi^2_{1-\alpha}(r-1)\}$$

英国数学家
Pearson

2. 连续型变量的卡方拟合分布检验

$$H_0: F(x) = F_0(x), \quad H_1: F(x) \neq F_0(x)$$

假设连续型随机变量X，抽取观测值 $x_1, x_2, \ldots, x_n$

并且 $(x_{(1)}, x_{(n)}) \subset [a, b]$ 划分区间$[a, b]$

第i个区间 $[a_i, a_{i+1})$	实际频数	v_i
	理论频数	np_i

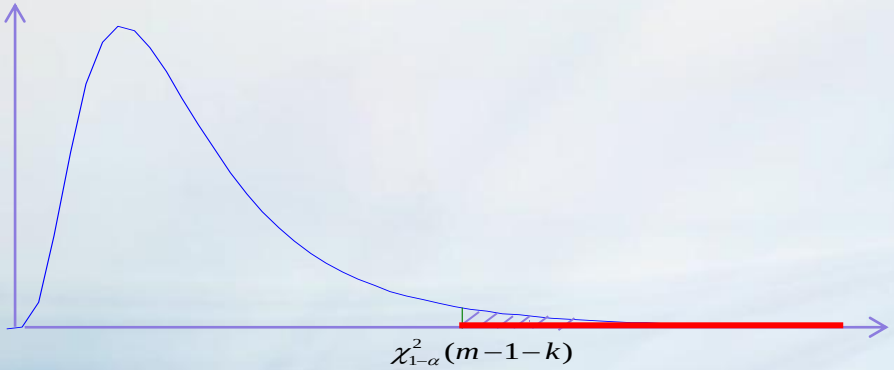
$i = 1, 2, \ldots, m$

$$p_i = P\{a_i \le X < a_{i+1}\} \xrightarrow{\text{在}H_0\text{下}} F_0(a_{i+1}) - F_0(a_i)$$

在H_0下， 构造检验统计量 $\chi^2 = \sum_{i=1}^{m} \frac{(v_i - np_i)^2}{np_i} \sim \chi^2(m-1-k)$

其中，k为分布$F_0(x)$中未知参数个数。

H_0的拒绝域　$K_0=\{\chi^2>\chi^2_{1-\alpha}(m-1-k)\}$



卡方检验步骤：

1. 提出问题　$H_0: F(x) = F_0(x);\quad H_1: F(x) \neq F_0(x)$

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

3. H_0的拒绝域　$K_0 = \{\chi^2 > \chi^2_{1-\alpha}(m-1-k)\}$

4. 计算χ^2，另一方面查表$\chi^2_{1-\alpha}(m-1-k)$

5. 如果$\chi^2 > \chi^2_{1-\alpha}(m-1-k)$,则拒绝$H_0$，否则，接受$H_0$.

引例1　问顾客对这三种矿泉水的喜好是否存在差异?

$$H_0: p_1 = p_2 = p_3 = \frac{1}{3}, \quad H_1: \text{至少一个} p_i \neq \frac{1}{3}$$

品牌	1	2	3
实际频数	61	53	36
理论频数	50	50	50
$(v_i - np_i)^2$	121	9	196

$$\chi^2 = \frac{121}{50} + \frac{9}{50} + \frac{196}{50} = 2.42 + 0.18 + 3.92 = 6.52$$

另一方面，$\chi^2_{0.95}(2) = 5.991$

结论：拒绝H0，顾客对三种矿泉水的喜爱存在显著差异。

引例2　学生成绩是否服从正态分布?

$$H_0: X \sim N(\mu,\sigma^2), \quad H_1: X \sim N(\mu,\sigma^2)$$

$$\mu=?\quad \sigma^2=?$$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{9} v_i x_i, \qquad x_i \text{ 表示组中值,}$$

$$s^2=\frac{1}{n-1}\sum_{i=1}^{9} v_i (x_i-\bar{x})^2 \qquad (*)$$

组中值 x_i	40	55	65	75	85	95	求和	平均
组数*组中值	454	999	2219	2898	2801	1404	10775	72
$(x_i-\bar{x})^2$	1013	283	47	10	173	537		
$v_i(x_i-\bar{x})^2$	11489	5146	1593	388	5716	7932	32263	217

$$\hat{\mu}=72, \quad \hat{\sigma}\approx 14.715$$

引例2

分数段	[30-50)	[50-60)	[60-70)	[70-80])	[80-90)	[90-100]	合计
频数	11	18	34	39	33	15	150

pi	0.065	0.140	0.239	0.261	0.183	0.082	
理论频数	9.794	20.992	35.782	39.108	27.408	12.313	

$$\chi^2=\sum_{i=1}^{6}\frac{v_i^2}{np_i}-150 \approx 6.954$$

另一方面，$\chi^2_{0.95}(6-1-2)=7.815$

结论：接受H0，成绩分布服从正态分布。

小结

卡方分布检验法