模型-评价主题-统计类评价-卡方检验 【gyj】
1. 模型名称
2. 适用范围
3. 形式
4. 求解方法
4.4.1 matlab:
4.4.2 spss:
5. 补充资料

模型-评价主题-统计类评价-卡方检验 【gyj】

1. 模型名称

卡方检验 (Chi-Squared Test) : χ^2

2. 适用范围

1. 适配度检定：可用于验证一组观察值的**次数分配**是否异于理论上的分配。
   - 换言之，即**统计样本的实际观测值与理论推断值之间的偏离程度**
   - 卡方值越大，二者偏离程度越大；卡方值为0，表明实际观测值完全符合理论值
2. 独立性检定：可用于检验**两个或两个以上因素**之间**是否相互影响**的问题

3. 形式

1. 适配度检定
   1. 样本容量 $n \geqslant 50$
   2. 每个事件里包含的样本个数 $np_i \geqslant 5$,否则相邻组要进行合并。
2. 独立性检定
   1. 从所感兴趣的总体中抽取的**简单随机样本**数据值
   2. 变量类型为分类型[1]或名义型[2]变量
      - ❌ 请勿对定义类别组合的连续型变量使用独立性检验
   3. 两个变量的每一个组合都要有**5个以上**的期望值。

4. 求解方法

1. 适配度检定

   4.1 概念

   如果样本总体X的分布为未知是，往往需要利用样本数据 $(x_1, x_2, \cdots, x_n)$ 来检验总体的分布是不是某一事先给定的函数 $F(x)$.即检验如下假设：

   $$H_0：总体X的分布函数为F(x),\ H_1：总体X的分布函数不是F(x)$$

   - 若总体X为**离散型**的随机变量，则上述假设相当于如下假设：

     $$H_0：总体X的分布列为P(X=x_i)=p_i, i=1,2,\cdots$$

   - 若总体X为**连续型**随机变量，则上述假设相当于如下假设：

     $$H_0：总体X的密度函数为f(x)$$

   - 在假设H_0为真时，总体X的分布函数 $F(x;\ \theta)$ 的形式已知，但若其中有位置参数 θ , 需要先用最大似然估计法[^3]估计出未知数 $(\hat{\theta})$,然后带入 $F(x;\theta)$ 中，此时分布函数 $F(x;\hat{\theta})$ 为已知函数

4.2 步骤

- **划分事件**：将样本空间Ω划分为k个互不相容的事件 A_1,A_2,...,A_k, 即

$$\Omega = A_1 \cup A_2 \cup \cdots \cup A_k$$

- **计算每个事件发生的概率**：在假设H_0为真时，计算概率 $p_i = P(A_i), i = 1, 2, \cdots, k$

- **根据题意写出n, f_i, k, r**：确定由实验数据确定 A_i 发生的频率 $\frac{f_i}{n}$

  一般来说，在H_0为真且试验次数很多时，理论频率(概率) $p_i, i = 1, 2, \cdots, k$ 与实际频率 $\frac{f_i}{n}$，$i = 1, 2, \cdots k$ 之间的大体差异不应该很大。

- **计算拒绝域**：查 χ^2 分布临界值表(卡方分布)

  在假设H_0成立的条件下，置信水平为α的拒绝域为

$$W = [\chi^2_\alpha(k - r - 1), +\infty)$$

- **计算 χ^2 的观察值**：采用如下统计量来衡量理论频率与实际频率事件的总体差异程度。

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

- 定理1：当假设H_0为真($X \sim F(x)$)及x充分大时，无论F(x)为什么分布函数，统计量 χ^2 总是近似服从自由度为 $k - r - 1$ 的 χ^2 的分布，即

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \overset{近似}{\sim} \chi^2(k - r - 1)$$

  r为F(x)中被估计参数的个数

❗【注意】由于定理1的结论为近似结果，故应用此方法时，要满足**3.形式**中规定的内容

4.3 实例

题目：生物学家孟德尔根据颜色与形状将豌豆分成4类：黄圆的，青圆的，黄有角，青有角，且运用遗传学的理论指出这4类豌豆之比为9：3：3：1。 他观察了556颗豌豆，发现各类的颗数分别为315，108，101，32. 试问可否认为孟德尔的分类推断是正确的？(α = 0.05)

- **划分事件**：此时的样本空间是这556颗豌豆，由题意可以划分成4个互不相容的事件：A_1 表示黄圆，A_2表示青圆，A_3表示黄有角，A_4表示青有角

  由题意，我们需要检验

$$H_0 : P(A_1) = \frac{9}{16}, P(A_2) = \frac{3}{16}, P(A_3) = \frac{3}{16}, p(A_4) = \frac{1}{16}.$$

- **根据题意写出相关变量**：本题中

  $n = 556, f_1 = 315, f_2 = 108, f_3 = 101, f_4 = 32, k = 4, r = 0$

  【k=4是因为k表示样本可以被划分成多少个互不相容的事件，r为理论函数(给定函数)被估计参数的个数，这里并没有估计故为0】

- **计算拒绝域**：(本题给定置信水平α = 0.05)

  查表得：$W = [\chi^2_{0.05}(4 - 0 - 1), +\infty) = [\chi^2_{0.05}(3) + \infty) = [7.81, +\infty)$

  【 χ^2 后面的(3)表示临界值表里的纵列应该读哪个数】

- **计算 χ^2 的观察值**：

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

$$= \frac{(315 - 556 \times \frac{9}{16})^2}{556 \times \frac{9}{16}} + \frac{(108 - 556 \times \frac{3}{16})^2}{556 \times \frac{3}{16}} + \frac{(101 - 556 \times \frac{3}{16})^2}{556 \times \frac{3}{16}} + \cdots$$

$$+ \frac{(32 - 556 \times \frac{1}{16})^2}{556 \times \frac{1}{16}} = 0.47 \notin W$$

不在拒绝域内，接受假设H_0, 孟德尔的论断是正确的。

4.4 代码实现

4.4.1 Matlab

```
%使用适合性检验，需要输入情况的种类数(对应题目中的k)，每种情况的实际频数(fi)和需检验的分布的频率(pi)
%适应性检验就是检验一个样本的分布是否符合某个分布的一种假设检验方法。比如说检验数据是否正态分布，是否成二项分布或者平均分布等等。
%设共有i种情况，A为每种情况的实际数目，n为总频数，p为待检验的分布的理论频率。
A=[315 108 101 32];          %根据实际情况输入每种情况的实际频数
p=[9/16 3/16 3/16 1/16];     %根据实际情况输入每种情况待检验分布的理论频率，
                             %可以是一特殊的分布，如正态分布等
i=length(A);                 %共有多少种情况
n=sum(A);                    %计算总频数
k_2=sum((A-p*n).^2./(p*n))   %计算卡方分布的观测值
df=i-1                       %理论上应服从的卡方分布的参数，对应k-r-1
%查卡方分布表，若k_2<参数为df的卡方分布水平为0.05(即表示概率大于95%)的值，则认为A服从待检验分布
```

4.4.2 Python

```
#使用适合性检验，需要输入情况的种类数，每种情况的实际频数和需检验的分布的频率
def compatibility():
    #就是检验一个样本的分布是否符合某个分布的一种假设检验方法。
    #比如说检验数据是否正态分布，是否成二项分布或者平均分布等等。
    #设共有i种情况，A为每种情况的实际数目，n为总频数，p为待检验的分布的理论频率。
    A=[315,108,101,32]          #根据实际情况输入每种情况的实际频数
    p=[9/16,3/16,3/16,1/16]     #根据实际情况输入每种情况待检验分布的理论频率
                                #可以是一特殊的分布，如正态分布等
    i=len(A)                    #共有多少种情况
    n=sum(A)                    #计算总频数
    #以下为计算卡方分布的观测值
    k_2=0
    for j in range(i):
        k_2+=pow((A[j]-p[j]*n),2)/(n*p[j])
    #求理论上应服从的卡方分布的参数
    df=i-1;
    print("k_2=",k_2)
    print("df=",df)
    print("查卡方分布表，若k_2<参数为df的卡方分布水平为0.05(即表示概率大于95%)的值，则认为A服从待检验分布")
    #查卡方分布表，若k_2<参数为df的卡方分布水平为0.05(即表示概率大于95%)的值，则认为A服从待检验分布

if __name__ == '__main__':
    compatibility()
```

4.2.3 C++

```
#include <iostream>
#include <cmath>
using namespace std;
//若使用适合性检验，需要输入情况的种类数，每种情况的实际频数和需检验的分布的频率

void compatibility()//适合性检验
{
    //就是检验一个样本的分布是否符合某个分布的一种假设检验方法。
    //比如说检验数据是否正态分布，是否成二项分布或者平均分布等等。
    //设共有i种情况，A为每种情况的实际数目，n为总频数，p为待检验的分布的理论频率。
    const int i=4;                                  //根据实际情况输入共有多少种情况
    double A[i]={315,108,101,32};                   //根据实际情况输入每种情况的实际频数
    double p[i]={9./16,3./16,3./16,1./16};  //根据实际情况输入每种情况待检验分布的
理论频率
                                                    //可以是一特殊的分布，如正态分布等
    //注意输入分数应注意加.
    //计算总频数
    double n=0;
    for(int j=0;j<i;j++)
        n +=A[j];
    //以下为计算卡方分布的观测值
    double k_2=0;
    for(int j=0;j<i;j++)
        k_2+=pow((A[j]-p[j]*n),2)/(n*p[j]);
    //求理论上应服从的卡方分布的参数
    int df=i-1;
    cout<<"k_2="<<k_2<<endl;
    cout<<"df="<<df<<endl;
    cout<<"查卡方分布表，若k_2<参数为df的卡方分布水平为0.05(即表示概率大于95%)的值，则
认为A服从待检验分布"<<endl;
    //查卡方分布表，若k_2<参数为df的卡方分布水平为0.05(即表示概率大于95%)的值，则认为A
服从待检验分布
}

int main()
{
    compatibility();
    return 0;
}
```

4.4.4 SPSS

步骤：[spss卡方适合性检验](#)

2. 独立性检定

4.1 概念

要验证两变量A，B是否独立，可利用独立样本四格表法。

作如下假设 H_0 ： 总体A与总体B无关； H_1 ： 总体A与总体B相关。

4.2 步骤

- **划分事件**：划分样本空间 ω 并计算概率 p_{ij}

$$p_{ij}, i = 1, 2, \cdots, s; j = 1, 2, \cdots, k,$$

s, k 分别为A、B变量的分区数，即可以把A和B各分成多少个互不包含的事件

- 自由度： $(s-1)(k-1)$

- **确定事件发生的频率，写出观察值**：确定事件 A_iB_j 发生的频率 $\frac{f_{ij}}{n}$,作下标：观察值就是每个单元格中的值（ o_i ）

	B_1	B_2	$\cdots$	B_k	**总计**
A_1	$X_{11}(p_{B_1}S_{A_1})$	$X_{12}(p_{B_2}S_{A_1})$	$\cdots$	$X_{1k}(p_{B_k}S_{A_1})$	$S_{A1}=\sum_{j=1}^{k}X_{1j}$
A_2	$X_{21}(p_{B_1}S_{A_2})$	$X_{22}(p_{B_2}S_{A_2})$	$\cdots$	$X_{2k}(p_{B_k}S_{A_2})$	$S_{A2}=\sum_{j=1}^{k}X_{2j}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_s	$X_{s1}(p_{B_1}S_{A_s})$	$X_{s2}(p_{B_2}S_{A_s})$	$\cdots$	$X_{sk}(p_{B_k}S_{A_s})$	$S_{As}=\sum_{j=1}^{k}X_{sj}$
总计	$\sum_{i=1}^{s}X_{i1}$	$\sum_{i=1}^{s}X_{i2}$	$\cdots$	$\sum_{i=1}^{s}X_{ik}$	$\sum_{i=1}^{s}\sum_{j=1}^{k}X_{ij}$

其中 X_{ij} 表示同时满足 A_iB_j 的数量； p_{B_j} 表示 H_0 为真时 B_j 的概率；括号中为 H_0 为真时 np_{ij} 数。

- **计算期望值**：对每个单元格 (e_i)

$$\text{期望计数} = \frac{\text{总行数} \times \text{总列数}}{\text{总数}}$$

- **计算拒绝域**：查 χ^2 分布临界值表(卡方分布)

  在假设 H_0 成立的条件下，置信水平为 α 的拒绝域为：

$$W = [\chi^2_\alpha((s-1)(k-1)), +\infty)$$

- **计算 χ^2 的观察值**：

$$\chi^2 = \sum_{i=1}^{k}\frac{(o_i - e_i)^2}{e_i}$$

【注】k为一共有多少个 A_iB_j 组合

4.3.1 实例(1)

题目：考察性别与化妆与否是否有关，数据如下表：

	男	**女**	**合计**
化妆	15	95	110
不化妆	85	5	90
合计	100	100	200

根据题意，需检验假设：H_0：性别与化妆与否无关

- **划分事件**：性别可以分成两个事件，故s = 2；化妆与否也可以分成两个胡不包含的事件，故k = 2. 所以，自由度为1.
- **写出观察值**：观察值就是表格中每个单元格的数值

  $o_{11}=15, o_{12}=95, o_{21}=85, o_{22}=5$
- **计算期望值**： $e_{11}=55,\ e_{12}=55, e_{21}=45, e_{22}=45$

- **计算拒绝域**：令α为0.01，通过查 χ^2 的分布表得到

$$W=[\chi^2_{0.01}((2-1)\centerdot(2-1)),+\infty)=[\chi^2_{0.01}(1),+\infty)=[6.63,+\infty)$$

- **计算卡方统计量** χ^2

$$\chi^2=\frac{(15-55)^2}{55}+\frac{(95-55)^2}{55}+\frac{(85-45)^2}{45}+\frac{(5-45)^2}{45}=129.3\in W$$

故拒绝H_0假设，即性别与化妆有关。

-

4.3.2 实例(2)

题目：考察电影类型与观众是否购买零食有关与否。

- 检验该样本是否可以用独立性卡方检验
  - 样本为观影600人的简单随机样本
  - 变量是电影类型和是否购买零食，均为分类型变量
  - 确定每个组合要有5个以上的期望值(寻找期望计数)
    - 得到汇总的数据：

电影类型	有零食	无零食
操作	50	75
喜剧	125	175
家庭片	90	30
恐怖片	45	10

$$o_{11}=50,o_{12}=75,o_{21}=125,o_{22}=175,o_{31}=90,o_{32}=30,o_{41}=45,o_{42}=10$$

- 计算行总计数和列总计数（寻找期望计数）

电影类型	有零食	无零食	行总计
操作	50	75	125
喜剧	125	175	300
家庭片	90	30	120
恐怖片	45	10	55
列总计	310	290	**总体总计=600**

  - 利用公式计算每个单元格的期望计数

    对于"操作-有零食"这一组合，它的期望计数为：

$$\delta_{11}=\frac{125\times310}{600}=65$$

    【注】我们将答案四舍五入到最接近的**整数**。该数据表示，如果电影类型与购买零食之间没有关系，我们预计会有65个人在观看操作电影时购买零食

$$e_{11}=65,e_{12}=60,e_{21}=155,e_{22}=145,e_{31}=62,e_{32}=58,e_{41}=28,e_{42}=27$$

- **计算拒绝域**：（α=0.01）

$$W=[\chi^2_{0.01}((2-1)\cdot(4-1)),+\infty)=[\chi^2_{0.01}(3),+\infty)=[11.34,+\infty)$$

- **计算卡方统计量 χ^2**

$$\chi^2=\sum_{i=1}^{8}\frac{(o_i-e_i)^2}{e_i}=72.6\in W$$

故看电影的种类与观众是否会购买零食有关。

4.4 代码实现

4.4.1 matlab:

```
%独立性检验
%考虑X和Y是否相关
%下列pinshu中两列分别代表X1和X2，两行分别代表Y1和Y2，可以有更多列或更多行
%pinshu_ij为符合Xi和Yj的频数
pinshu=[15 95
        85 5];
rate=mean(pinshu,2);                                    %假设独立，应该服从均匀分布
fitness=sum(sum((pinshu-rate).^2./rate),2)              %计算拟合度
free_degree=(size(pinshu,1)-1)*(size(pinshu,2)-1)   %计算卡方分布的自由度
%查卡方分布表，若fitness>参数为free_degree的卡方分布水平为0.05(即表示概率大于95%)的值，
%则认为X和Y不独立
```

4.4.2 spss:

步骤：卡方独立性检验(R×C列联表)

5. 补充资料

1. 数模官网 - 卡方检验
2. JMP - 独立性卡方检验
3. 卡方独立性检验原理 - 知乎
4. 上文中用到的其他知识点

---

1. 分类型变量：与"数量型变量"相对，不可进行加减法等数学操作的变量. 例：电影的种类、喂动物食品的种类 ↩
2. 名义型变量：在现有的前提或条件下确定的数值。与"实际变量"相对，实际变量表示在现有的前提或条件发生改变后的数值 ↩