2.9 试述卡方检验过程。

- **答案**:

> 参考来源卡方检验原理及应用(白话)

> 卡方检验是研究数据比率上的差异性，用于比较定类与定类数据的关系情况。

通过实例简述四格表χ^2检验过程

1.建立无关性假设，通过数据构建四格表

举个例子，假设我们有一堆新闻标题，需要判断标题中包含某个词（比如吴亦凡）是否与该条新闻的类别归属（比如娱乐）是否有关，我们只需要简单统计就可以获得这样的一个四格表：

组别	属于娱乐	不属于娱乐	合计
不包含吴亦凡	19	24	43
包含吴亦凡	34	10	44
合计	53	34	87

> 通过这个四格表我们得到的第一个信息是：标题是否包含吴亦凡确实对新闻是否属于娱乐有统计上的差别，包含吴亦凡的新闻属于娱乐的比例更高，但我们还无法排除这个差别是否由于抽样误差导致。

那么首先假设标题是否包含吴亦凡与新闻是否属于娱乐是独立无关的，随机抽取一条新闻标题，属于娱乐类别的概率是：(19 + 34) / (19 + 34 + 24 +10) = 60.9%

2.根据假设生存新的理论四格表

在第一步中，我们计算得出了新闻属于娱乐类别的概率是60.9%，通过此概率可以计算得到新的理论值四格表

组别	属于娱乐	不属于娱乐	合计
不包含吴亦凡	$43 \times 60.9\% = 26.2$	$43 \times 39.1\% = 16.8$	43
包含吴亦凡	$44 \times 60.9\% = 26.8$	$44 \times 39.1\% = 17.2$	44

显然，如果两个变量是独立无关的，那么四格表中的理论值与实际值的差异会非常小。

3.计算χ^2的值

χ^2的计算公式为：

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

> 其中A为实际值，也就是第一个四格表里的4个数据，T为理论值，也就是理论值四格表里的4个数据。

> χ^2用于衡量实际值与理论值的差异程度（也就是卡方检验的核心思想），包含了以下两个信息：
>
> - 实际值与理论值偏差的绝对大小（由于平方的存在，差异是被放大的）
> - 差异程度与理论值的相对大小

对上述场景可计算χ^2值为10.01。

4.根据χ^2值查询卡方分布的临界值表

> 既然已经得到了χ^2值，我们又怎么知道χ^2值是否合理？也就是说，怎么知道无关性假设是否可靠？
> 答案是，通过查询卡方分布的临界值表。

这里需要用到一个自由度的概念，自由度等于V = (行数 - 1) * (列数 - 1)，对四格表，自由度V = 1。

对V = 1，卡方分布的临界概率是：

n′	P												
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	…	…	…	…	0.02	0.1	0.45	1.32	2.71	3.84	5.02	6.63	7.88

自由度=1时卡方分布的临界概率表

可以看到10.01>7.88,也就是标题是否包含吴亦凡与新闻是否属于娱乐无关的可能性小于0.5%，反过来，就是两者相关的概率大于99.5%。

卡方检验到此结束

> 卡方检验的一个典型应用场景是衡量特定条件下的分布是否与理论分布一致，比如：特定用户某项指标的分布与大盘的分布是否差异很大，这时通过临界概率可以合理又科学的筛选异常用户。

> 另外，χ^2值描述了自变量与因变量之间的相关程度：χ^2值越大，相关程度也越大，所以很自然的可以利用χ^2值来做降维，保留相关程度大的变量。

> 再回到刚才新闻分类的场景，如果我们希望获取和娱乐类别相关性最强的100个词，以后就按照标题是否包含这100个词来确定新闻是否归属于娱乐类，怎么做？很简单，对娱乐类新闻标题所包含的每个词按上述步骤计算χ^2值，然后按χ^2值排序，取χ^2值最大的100个词。

2.10 试述在Friedman检验中使用下两式的区别。

式1：$$\tau_{\chi^2}=\frac{k-1}{k}\cdot\frac{12N}{k^2-1}\sum_{i=1}^{k}(r_i-\frac{k+1}{2})^2$$

$$=\frac{12N}{k(k+1)}(\sum_{i=1}^{k}r_i^2-\frac{k(k+2)^2}{4})$$

式2：$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}$$

答案：式1是标准的弗里德曼检验计算公式，但是在样本量较小的情况下，式1计算结果明显偏离卡方分布。此时需利用式2计算，然后通过专门弗里德曼检验临界值表进行检验。
对于涉及6个以上总体的小样本量Friedman检验，如果不能从有关书籍中查到临界值，便只能采用卡方检验了 [1] 。

> [1] 陶澍．应用数理统计方法：中国环境科学出版社，1994年08月第1版