[**卡方检验原理及应用**](https://segmentfault.com/a/1190000003719712)

* [数据分析](https://segmentfault.com/t/%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90/blogs)
* [数据挖掘](https://segmentfault.com/t/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98/blogs)

[**2shou**](https://segmentfault.com/u/2shou) 2015年09月08日发布

* **7.5k** 次浏览

卡方检验，或称x2检验，被誉为二十世纪科学技术所有分支中的20大发明之一，它的发明者卡尔·皮尔逊是一位历史上罕见的百科全书式的学者，研究领域涵盖了生物、历史、宗教、哲学、法律。之前做文本分类项目用过卡方值做特征选择（降维），后来听内部培训，另一个部门说他们有用卡方检验做异常用户的检测，于是就想把卡方检验再温习一次，同时把卡方检验和特征选择串起来理解。

**无关性假设**

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%

**理论值四格表**

第二步，根据无关性假设生成新的理论值四格表：

| **组别** | **属于娱乐** | **不属于娱乐** | **合计** |
| --- | --- | --- | --- |
| 不包含吴亦凡 | 43 \* 0.609 = 26.2 | 43 \* 0.391 = 16.8 | 43 |
| 包含吴亦凡 | 44 \* 0.609 = 26.8 | 44 \* 0.391 = 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。

**x2值的计算**

x2的计算公式为：

https://segmentfault.com/img/bVpLPk

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。

x2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

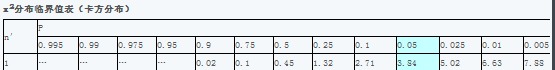
对上述场景可计算x2值为10.01。

**卡方分布的临界值**

既然已经得到了x2值，我们又怎么知道x2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：



显然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%。

**应用场景**

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，x2值描述了自变量与因变量之间的相关程度：x2值越大，相关程度也越大，所以很自然的可以利用x2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x2值，然后按x2值排序，取x2值最大的100个词。

**前言、什么是卡方检验**

卡方检验是一种用途很广的计数资料的假设检验方法。它属于非参数检验的范畴，主要是比较两个及两个以上样本率( 构成比）以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度或拟合优度问题。

它在分类资料统计推断中的应用，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。

**一、四格表资料的卡方检验**

例1.两组大白鼠在不同致癌剂作用下的发癌率如下表，问两组发癌率有无差别？

表1



（52 19 39 3） 这四个数据是整个表中的基本资料，其余数据均由此推算出来；这四格资料表就专称四格表（fourfold table），或称2行2列表（2×2 contingency table）。从该资料算出的两组发癌率分别为73.24%和92.86%，两者的差别可能是抽样误差所致，亦可能是两组发癌率（总体率）确有所不同。这里可通过卡方检验来区别其差异有无统计学意义，检验的基本公式为：

http://images.cnitblog.com/blog/502930/201310/14202346-b5742cacb64c4f4d8d07ad638179e471.jpg

式中A为实际数，以上四格表的四个数据就是实际数。T为理论数，是根据检验假设推断出来的；即假设这两组的发癌率本无不同，差别仅是由抽样误差所致。这里可将两组合计发癌率作为理论上的发癌率，即91/113=80.3%，以此为依据便可推算出四格表中相应的四格的理论数。以表1资料为例检验如下。

检验步骤：

1.建立检验假设：

H0：π1=π2；H1：π1≠π2；α=0.05

2.计算理论数（TRC），计算公式为：

TRC=nR.nC/n

式中TRC是表示第R行C列格子的理论数，nR为理论数同行的合计数，nC为与理论数同列的合计数，n为总例数。

第1行1列： 71×91/113=57.18

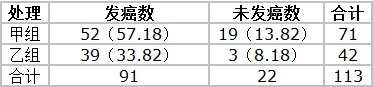
第1行2列： 71×22/113=13.82

第2行1列： 42×91/113=33.82

第2行2列： 42×22/113=8.18

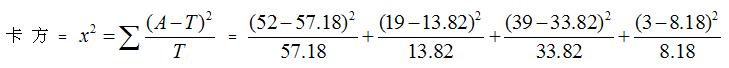
以推算结果，可与原四项实际数并列成表2：

表2



因为上表每行和每列合计数都是固定的，所以只要用TRC式求得其中一项理论数（例如T1.1=57.18），则其余三项理论数都可用同行或同列合计数相减，直接求出。

3.计算卡方值按公式代入

=0.47+1.94+0.79+3.28=6.48

4.查卡方值表求P值

在查表之前应知本题自由度。按卡方检验的自由度v=（行数-1）（列数-1），则该题的自由度v=（2-1）（2-1）=1，查卡方界值表，找到http://images.cnitblog.com/blog/502930/201310/14202723-653b727200d04843b43df4ebbf06bbe0.jpg，而本题卡方=6.48即卡方＞http://images.cnitblog.com/blog/502930/201310/14202803-1b55a8be5e44431f94a3ed2c961fbd34.jpg，P＜0.05，差异有显著统计学意义，按α=0.05水准，拒绝H0，可以认为两组发癌率有差别。

通过实例计算，读者对卡方的基本公式有如下理解：若各理论数与相应实际数相差越小，卡方值越小；如两者相同，则卡方值必为零，而卡方永远为正值。又因为每一对理论数和实际数都加入卡方值中，分组越多，即格子数越多，卡方值也会越大，因而每考虑卡方值大小的意义时同时要考虑到格子数。因此自由度大时，卡方的界值也相应增大。

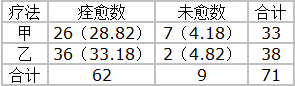
**二、四格表卡方值的校正**

卡方值表是数理统计根据正态分布中http://images.cnitblog.com/blog/502930/201310/14202914-b81bc02058ef4f58a3f9cda1f6c44828.jpg的定义计算出来的。是一种近似。在自由度大于1、理论数皆大于5时，这种近似很好；当自由度为1时，尤其当1＜T＜5，而n＞40时，应用以下校正公式：

http://images.cnitblog.com/blog/502930/201310/14203036-b5ae2d65925445b3a5ad50436af9f0d0.jpg

例2.某医师用甲、乙两疗法治疗小儿单纯性消化不良，结果如表3.试比较两种疗法效果有无差异？

表3



从表3可见，T1.2和T2.2数值都＜5，且总例数大于40，故宜用校正公式检验。步骤如下：

1.检验假设：

H0：π1=π2；H1：π1≠π2；α=0.05

2.计算理论数：（已完成列入四格表括弧中）

3.计算卡方值：应用校正公式运算如下：

http://images.cnitblog.com/blog/502930/201310/14203212-87ba39fa4e60423bb0e602f5bfa139c9.jpg

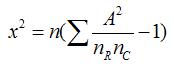
查卡方界值表，http://images.cnitblog.com/blog/502930/201310/14204036-b6782b300a534864ae8dd4ca240f61ef.jpg，故卡方＜http://images.cnitblog.com/blog/502930/201310/14204047-6264d75f9b2045d2bc628bdac57e00ed.jpg，P＞0.05。按α=0.05水准，接受H0，两种疗效差异无统计学意义。

如果不采用校正公式，而用原基本公式，算得的结果卡方=4.068，则结论就不同了。

如果观察资料的T＜1或n＜40时，四格表资料用上述校正法也不行，可参考预防医学专业用的医学统计学教材中的Fisher精确检验法直接计算概率以作判断。

**三、行×列表的卡方检验**

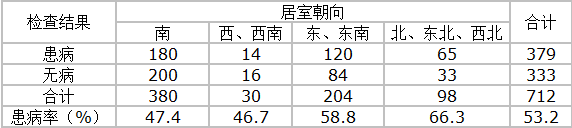
适用于两个组以上的率或百分比差别的显著性检验。其检验步骤与上述相同，简单计算公式如下：



式中n为总例数；A为各观察值；nR和nC为与各A值相应的行和列合计的总数。

例3.北方冬季日照短而南移，居宅设计如何适应以获得最大日照量，增强居民体质，减少小儿佝偻病，实属重要。胡氏等1986年在北京进行住宅建筑日照卫生标准的研究，对214幢楼房居民的婴幼儿712人体检，检出轻度佝偻病333例，比较了居室朝向与患病的关系。现将该资料归纳如表4作行×列检验。

表4



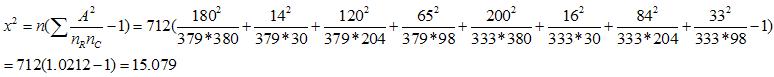
该表资料由2行4列组成，称2×4表，可用行×列卡方公式检验。

（一）检验步骤

1.检验假设

H0：四类朝向居民婴幼儿佝偻病患病率相同；H1：四类朝向居民婴幼儿佝偻病患率不同；α=0.05

2.计算卡方值



3.确定P值和分析

本题v=（2-1）（4-3）=3，据此查卡方界值表：

http://images.cnitblog.com/blog/502930/201310/14204556-7f925ecfef0243e99538bcc6953fe839.jpg，本题卡方=15.08，卡方＞http://images.cnitblog.com/blog/502930/201310/14204655-8681296fd91f4b2a99514a1e922e57ba.jpg ，P＜0.05，拒绝H0，可以认为居室朝向不同的居民，婴幼儿佝偻病患病率有差异。

（二）行×列表卡方检验注意事项

1.一般认为行×列表中不宜有1/5以上格子的理论数小于5，或有小于1的理论数。当理论数太小可采取下列方法处理：①增加样本含量以增大理论数；②删去上述理论数太小的行和列；③将太小理论数所在行或列与性质相近的邻行邻列中的实际数合并，使重新计算的理论数增大。由于后两法可能会损失信息，损害样本的随机性，不同的合并方式有可能影响推断结论，故不宜作常规方法。另外，不能把不同性质的实际数合并，如研究血型时，不能把不同的血型资料合并。

2.如检验结果拒绝检验假设，只能认为各总体率或总体构成比之间总的来说有差别，但不能说明它们彼此之间都有差别，或某两者间有差别。