卡方分布与卡方检验

■ 房祥忠

你能想象出普鲁士军队中历年被马踢死的士兵数不是混乱的杂乱无章的，而是服从一个特定的分布规律吗？你能想象出历史上发生的战争次数以及二战德国发射到伦敦的导弹数也都有类似的规律吗？生活中看似偶然发生而略显杂乱无章的现象，背后总是被客观规律左右。而能够揭示这个规律的就是统计学。但所用到的统计工具却很简单，就是用于拟合优度检验的卡方检验方法。卡方检验用途很多，可以用于检验企业管理水平是否达到要求，可以判断对某种疾病影响最大的因素，还可以用于分析不同文化类型对水资源认知差异等等问题。

无论是研究统计学还是使用统计学的人，都不会不知道卡方分布和卡方检验。因为他们起到的作用实在太大，应用的范围实在太广，涉及的领域实在太多，有趣的实例也不胜枚举。我们今天就来谈谈这个卡方分布和卡方检验。

卡方分布是什么

卡方分布的定义非常简单，若干个相互独立且服从标准正态分布的随机变量平方之和所服从的分布就称为卡方分布。可以知道它的分布完全由自由度确定，求和项个数称为这个卡方分布的自由度，自由度本身等于卡方变量的均值，而自由度的两倍就是它的方差。虽然定义是从这种简单方式出发的，但多种形式的统计量其精确分布或者渐近分布却都服从卡方分布。正是由于卡方分布的广泛性，它的应用面很广，特别是在似然比检验、拟合优度检验和独立性检验问题中。

正态方差的区间估计和检验

我们知道，方差是一个随机变量的重要特征，它反映了随机变量取值参差不齐的情况。卡方分布最直接的用途是正态方差的区间估计和检验。这里之所以用到卡方分布是因为样本方差与总体方差的商就服从一个卡方分布，从而可以通过卡方分布的分位点来获得总体方差的置信区间和假设检验的方法。方差的大小可以用于评价产品质量的好坏，能够反映一个企业管理水平的高低和生产设备的精良程度，方差越小说明质量越稳定。对于电子产品来说，若其主要性能指标的方差小一倍，则产品的价格就可能高出几倍甚至十几倍。因此对于方差的区间估计或者检验方法常被用于监控产品质量。

指数均值的估计和检验

产品的使用寿命是贸易中买卖双方最感兴趣的量，因而对产品平均寿命的估计和检验历来都是企业界重视的问题。电子产品的寿命往往被认为服从指数分布，从而对于指数分布均值的区间估计和检验在实际当中应用很多，常常在经贸中被用于评价或者验收电子产品。令人想不到的是此时也需要用到卡方分布。这是因为若干个电子产品的实际寿命相加与平均寿命之商的两倍也服从卡方分布，其自由度为产品个数的两倍。这样的一个结论被用来获得平均寿命的区间估计或者构造检验方法。

似然比检验

似然比检验是构造假设检验问题的最一般方法，具有较大的优良性，可以用于广泛的分布类型和检验问题。所谓似然比就是在零假设下的似然函数最大值与备择假设下似然函数最大值之比。但似然比统计量的精确分布在一般情况下难以获得。理论上可以证明在极其一般条件下，似然比统计量的分布在样本量较大时可以通过卡方分布来近似。这项理论成果使得检验方法有了按部就班的操作步骤。但对于一些常用的分布来说，根据具体分布获得的方法更加有效。尽管如此，似然比卡方检验还是应用比较多的方法，在处理实际问题时显示出它的普适性。

拟合优度检验

拟合优度检验是检验样本是否来自于某一个或者某一族分布。关于正态分布和指数分布的拟合优度检验问题由于用到的比较多，已经被制定为国家标准。但对于大部分分布，很难得到精确的检验方法，一般采用近似方法。利用卡方分布进行检验是最常用的方法。对于取有限个值的离散型分布，通过比较每个单元的观测值和期望值构造一个检验统计量，这个统计量的渐近分布是卡方分布，自由度是取值单元的个数减一。这是一个基于大样本的方法，样本量越大越精确。对于取无限个值的离散型随机变量，要在合适的地方进行截断处理。对于连续型分布，可以通过离散化变成离散型分布近似处理。因此，这个方法几乎对所有分布都可以进行拟合优度检验。这方面有趣的例子很多，下面我们把开头提到的几个例子再稍微详细地说一说。

你能够相信吗，普鲁士兵团中每年被马踢死的士兵的人数是服从泊松分布的。历史记录显示了1875—1894 历年普鲁士兵团中被马踢死的士兵的人数。近 20 年间，有 6 年没人被踢死，有 8 年每年被踢死 1 人，有 3 年每年被踢死 2 人，有 2 年每年被踢死 3 人，有 1 年被踢死 4 人，总共被马踢死了 24 人。可以计算出平均一年踢死的人数是 1.2 人。如果用一个泊松分布来描述，则这个泊松分布的均值参数就是 1.2。首先按照这个泊松分布计算取各个值的概率，再用总数 24 和概率相乘得到期望数，把记录数和期望数比较得到卡方统计量。可以经过统计的检验方法分析认为每年被踢死的士兵数确实服从参数为 1.2 的泊松分布。

在第二次世界大战时期德国总共向伦敦发射了 537 颗 V2 导弹。许多人认为中弹点有聚集倾向，怀疑是否有意为之。但经过统计分析后，得到的结论认为不存在聚集现象，而是服从空间散布的泊松过程。分析方法是将伦敦南部分为面积大约相等的 576 个区，则每个小区落下的导弹数有多有少，可以认为它是一个随机变量，人们怀疑它服从泊松分布。根据每个小区落下的导弹数的历史记录，对是否服从泊松分布进行拟合优度检验。首先计算平均

■ 房祥忠

每个小区落下的导弹数，537/576=0.9323，这可以做为泊松分布参数的估计值，然后利用泊松分布概率计算公式计算出取各个值的概率，再用总导弹数537乘以相应的概率得到期望数，比较实际数和期望数得到卡方统计量。经过检验认为小区落下的导弹数非常好地服从泊松分布。

自1500—1931年的432间，比较重要的战争在全世界共发生了299次，记录中给出了历年的战争次数，则同样可以使用卡方统计量检验证明每年发生的战争次数也是服从泊松分布的。

列联表的独立性检验

卡方分布可以用于列联表数据的统计分析。统计学的一个重要作用是研究不同变量之间的关系。利用相关系数可以刻画数值型变量之间的线性相关性。方差分析可以用来刻画数值型变量和离散型变量之间的相关关系；回归分析可以具体刻画变量之间的各类相关关系。列联表的独立性检验主要是对两个或者多个离散型随机变量检验他们之间的相关性。实际应用中是通过检验否定他们之间的独立性而得到科学发现的，即证明了某种相关性。在实施这个检验时，要构造检验统计量，这个统计量也是通过比较观测数和期望数得到的，并且这个统计量的渐近分布也服从卡方分布。

列联表分析简单易学，不用太多的统计知识就可以学会。但能够解决的实际问题却可能不简单，有时甚至可以有重要的科学发现。本质上就是要否定独立性而证明变量之间具有相关关系。目前比较流行的因果分析很多也是通过这个方法实现的。

抽烟和肺癌之间的关系研究是历史上一个最经典的例子。在实际生活中的例子更是数不胜数。我们在中文期刊数据库中搜列联表分析，马上发现很多运用列联表进行分析的有趣例子：土地细碎化对农户家庭投资农机作业社会化服务的影响；城市等级与本科毕业生就业性质以及就业地区流向的影响；通过列联表分析不同文化类型与水资源认知差异的相关性；影响柑橘黄龙病防治的显著性影响因素；代谢综合征与急性脑血管病进行列联表分析；运用列联表分析灾民受灾经历、个体属性与恢复期长短的关系；列联表分析我国制造业信息化技术来源地区和机构的差异……这些例子千差万别，但都来自于实践，十分生动具体。

结语

卡方分布是统计学中重要的分布，它的定义很简单，唯一的特征指标是它的自由度。卡方检验在不同的场景中形式不同，会出现在很多的统计工作中。卡方检验之所以被广泛运用，也是由于它的方法简单易学，只要稍微进行一下训练就可以掌握这些方法。这样的统计方法值得普及推广。

作者单位：北京大学