概率论与数理统计基础理论与应用研究

崔伯熹
广东广雅中学　广东　广州　510160

[摘　要]概率论与数理统计作为一种研究随机现象的重要数学工具，在数值天气预报、金融保险等工农业生产生活中得到了广泛的应用。本文以高中生视角从生活中的随机事件引入，梳理了概率论与数理统计基本理论及其发展历程，对统计学中描述样本数据特征的基本数学指标（期望、方差）进行了介绍。论文最后以保险金计算、基于最小二乘的人工智能为例对概率论与数理统计相关理论在现代科技与工程中所取得的辉煌成就进行了归纳和总结。

[关键词]概率论；统计学；发展历程；正态分布
【DOI】10.12252/j.issn.2096-6288.2021.12.331

1 引言

1.1 研究背景与意义

如今，人们的生活已离不开概率，正如概率诞生于对赌博的预测一样，人类对未知世界的探索总倾向于知道事物的发生可能性为多大，因而使得概率论成为一门非常实用的科学。由于涉及到数据，因此数理统计和概率是密不可分的，在对随机现象的大量数理统计中总结的规律性便是概率，而基于概率理论又可以对统计结果进行分析预测、对统计数据的特征进行描述。本文聚焦于数理统计与概率理论，在阐述其本质原理的基础上归纳了概率论与数理统计的应用。

1.2文献研究现状

概率论与数理统计诞生已有百年历史，对人类生产和生活进步起到了重要的指导作用，对概率论和数理统计方面的研究已见有多篇文献报道。蒙国往从施工风险评估出发，将数理统计与地铁工程相结合，提出应用于施工风险评估的模糊数理方法，实现工程质量、人员安全全程管控与实时评价[1]。邓阳斌等将概率论中的蒙特卡洛方法创新应用于碳化硅材料外壳失效评价方面[2]，在实验碳化硅外壳失效计算模型的基础上，利用概率论中的蒙特卡洛方法计算碳化硅外壳失效概率。潘小峰等以数学期望为入手点讨论了概率论中数学期望的基本概念[3]，文章对概率论发展历程进行了简要归纳，并通过数据分析实例揭示了生活中存在的随机事件现象。

2 概率论与数理统计的发展简史

公认的第一部有关概率论的著作是伯努利的《推测术》，书中系统性的介绍了排列与组合的基本原理，并利用排列组合讨论了有关骰子的赌博游戏，伯努利在这本书中提出了著名的伯努利大数定律，翻开了概率论发展的宏伟篇章。1901年，在众多数学家的努力下终于严格的证明了中心极限定理，而后数学家正利用中心极限定理第一次科学地解释实际生活中随机变量呈近似正态分布背后的原理。1906年，俄国数学家马尔科夫提出了所谓“马尔科夫链”的数学模型。1934年，前苏联数学家辛钦又提出一种在时间中均匀进行着的平稳过程理论。20世纪初完成的勒贝格测度与积分理论及随后发展的抽象测度和积分理论，为概率公理体系的建立奠定了基础。

3 概率论与数理统计的基本原理

3.1 概率的定义及排列组合基本原理

概率描述了随机事件发生的可能性。在计算上，概率可以通过随机事件某种结果在总结果中的占比来进行计算。式（1）给出了概率计算公式：

$$P(A)=N_A/N \quad (1)$$

其中，P（A）表示随机事件A发生的概率，N、N_A分别为随机事件的总结果数以及事件A发生的结果数。

当随机事件结果较多时，通过列举的方法计算N_A和N已几乎不可能，此时需要使用排列组合原理计算N_A和N。所谓排列是指从n个元素中取出m个元素排成一排的总结果数，可记作A_n^m，其数值应为：

$$A_n^m=n(n-1)(n-2)\dots(n-m+1)=\frac{n!}{(n-m)!} \quad (2)$$

与排列相对应的概念是组合，所谓组合是指从n个元素中取出m个元素的总结果数，可记作C_n^m。组合C_n^m的数值应为：

$$C_n^m=\frac{A_n^m}{m!}=\frac{n!}{m!(n-m)!} \quad (3)$$

3.2 随机变量的期望与方差

3.2.1随机变量的期望

通过排列与组合，可以方便计算出式（1）中的N与N_A，从而计算某随机事件发生的概率。若将随机变量与它们的概率相乘，可得到式（4）：

$$E(X)=\sum X_i P_i \quad (4)$$

它描述了随机变量的平均特征，称为数学期望。可见，数学期望是随机变量对于其概率的加权平均，数学期望趋于概率较大的随机变量值。由式（4）可知，掷骰子其正面朝上的数学期望值为：

$$E(X)=\frac{1+2+3+4+5+6}{6}=3.5 \quad (5)$$

可见，期望描述了随机变量的平均特征在多次投掷骰子之后，其正面朝上的均值约为3.5左右。

3.2.2随机变量的方差

期望描述了随机变量的平均特征，而方差则描述了随机变量偏离平均值的离散程度。式（6）给出了方差的定义：

$$D(X) = \sum [X_i - E(X)]^2 \quad (6)$$

可见，方差为每个随机变量与期望值之差的平方和，它反映了随机变量偏离均值的程度，方差越小则表明随机变量越接近数学期望，统计数据之间分布的离散程度也就越小。

3.3 随机变量的分布特性

当对一个随机变量进行独立重复实验时，会得到随机变量与概率之间的函数关系，称为随机变量的分布特性。本文介绍生活中常见的两类概率分布，即二项分布与正态分布。

3.3.1随机变量的二项分布

二项分布又称为伯努利分布，若在n次重复的独立实验中某事件A发生的概率为P，则将n次重复独立实验中事件A发生的次数作为随机变量，所得到的分布列称为二项分布。若在n次独立重复实验中，A事件发生的次数为k，则说明随机事件A发生了k次未发生（n-k）次，因而可得n次独立重复实验中A事件发生的次数为k的概率为：

$$P(X = k) = C_n^k P^k (1 - P)^{(n-k)} \quad (7)$$

图1给出投掷100次硬币，正反面概率均为0.5的二项分布示意图，可见投掷100次硬币最有可能的是出现50次正面朝上和50次反面朝上。

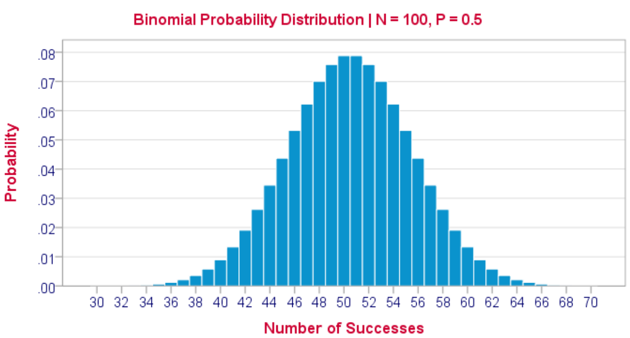

图1 投掷硬币正面朝上次数概率分布图

3.3.2随机变量的正态分布

正态分布又称为高斯分布，由数学家高斯得出。在自然界中的众多随机现象都服从正态分布规律，例如一个班级的身高统计数据就服从正态分布曲线，一个省份的高考成绩也服从正态分布规律。式（8）给出了正态分布概率密度公式：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8)$$

式中μ和σ分别为随机变量x的期望和标准差，随机变量x的取值不同，其对应的概率密度由式（8）计算。

在正态分布中，x的值介于[μ-3σ， μ+3σ]之间曲线下的面积经计算为0.9974，即随机变量x的值介于[μ-3σ， μ+3σ]的概率为99.74%，而在该区间之外的数据发生的概率仅有0.26%，认为该区间内数据是“异常”的，这便是著名的3σ原则。

4 概率论与数理统计的应用实例

保险作为金融产品的一种，为参保人提供了一种有效的投资手段。在保险中有一种为参保人身体健康承保的险种，称为寿险。参保人以一定金额购买寿险产品后，若在一定有效期范围内发生保险合同内约定的疾病甚至意外身故等都将获得一定量的赔偿。假设某保险约定了6种疾病，发生这6种疾病中的任意一种都将获得一定的理赔金，如表1所示。

表1 某人寿保险概率分布列

赔付金额	疾病1	疾病2	疾病3	疾病4	疾病5	疾病6
概率P	P_1	P_2	P_3	P_4	P_5	P_6

若已知人群中上述各类疾病发生的概率P_i，则可计算出该寿险产品的理赔期望A，如公式（9）所示：

$$A = E(X) = \sum X_i P_i \quad (9)$$

也就是说，若某参保人发生疾病，则保险公司平均理赔金额为A元。若在投保人群中发生理赔的概率为P，共有N个人参保，则N个人中有k个人发生理赔的概率服从3.3.1节中的二项分布，其分布率由式（7）确定。因而，根据二项分布的性质，保险公司的理赔金额期望应为AP，保费至少应设置在AP/N以上才能确保保险公司盈利。

5 结论

本文聚焦于数学中的概率论与数理统计，作为一种研究随机现象的重要数学工具，概率论与数理统计在数值天气预报、金融保险等工农业生产生活中得到了广泛的应用。文章以高中生视角从生活中的随机事件引入，梳理了概率论与数理统计基本理论及其发展历程，对统计学中描述样本数据特征的基本数学指标进行了介绍，在方差概念基础上对基于统计理论的最小二乘法进行了理论探究。论文最后以保险金计算对概率论与数理统计相关理论在现代科技与工程中所取得的辉煌成就进行了归纳和总结。

参考文献

[1]蒙国往，黄劲松，吴波，农忠建，许杰，韦汉.基于数理统计的地铁车站深基坑施工风险评估[J].城市轨道交通研究，2021，24(12)：55-60.

[2]邓阳斌，殷园，巫英伟，田文喜，秋穗正，苏光辉.基于蒙特卡罗方法的碳化硅包壳失效概率论评价[J].核科学与工程，2021，41(06)：1215-1222.

[3]潘小峰，胡坤.从概率论的发展史谈数学期望[J].中学数学月刊，2021(08)：65-66.