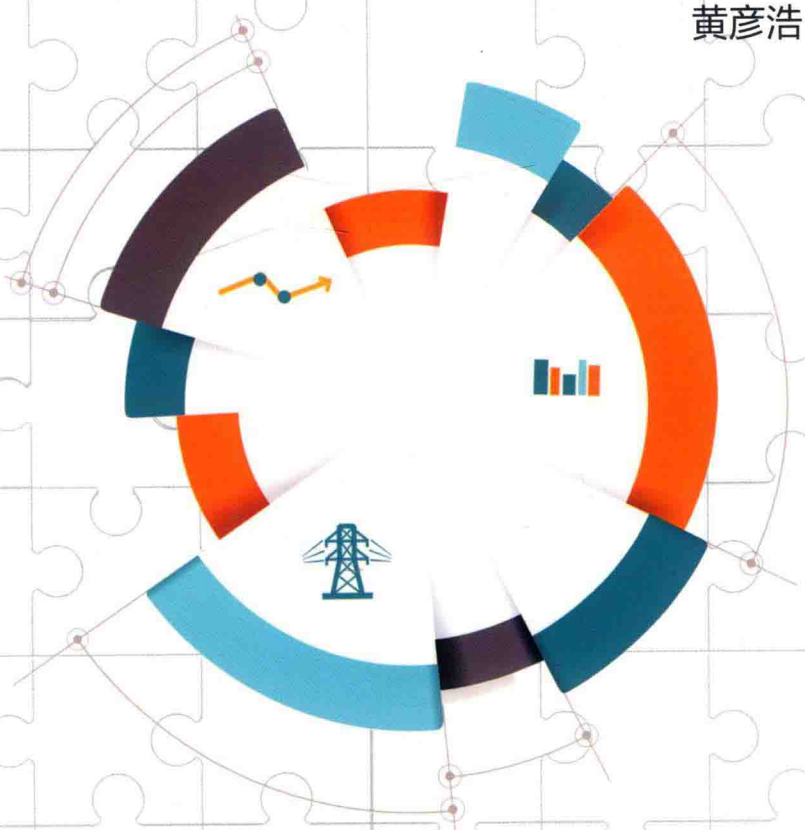
SHUJU FENXI

JIQI ZAI DIANLI XITONG ZHONG DE YINGYONG

数据分析及其在电力系统中的应用

黄彦浩　于之虹　高　波　编



中国电力出版社
CHINA ELECTRIC POWER PRESS

SHUJU FENXI

JIQI ZAI DIANLI XITONG ZHONG DE YINGYONG

数据分析及其在电力系统中的应用

黄彦浩　于之虹　高　波　编

中国电力出版社
CHINA ELECTRIC POWER PRESS

内 容 提 要

本书详细介绍了数据分析的基础、方法，及其在电力系统中的应用。全书共分十二章，包括数据分析的统计学基础、核方法、知识建模、关联分析、分类分析、聚类分析、数据分析的其他方法、数据可视化、电网仿真数据分析实践的思路和过程、电网仿真数据分析案例系统的搭建等。附录简介了 R 语言、Matlab、Python 语言以及其他数据分析工具，主要侧重于软件特性和基本使用方法的描述。

本书旨在强调对实际工作的帮助。内容在兼顾理论的同时，向应用侧重，力争以通俗易懂的方式介绍常见的理论要点和方法，不强调数学推导，可作为数据分析的快速上手向导和工作的参考指南。

本书可供从事电力系统大数据、人工智能研究和数据分析的工作人员，以及其他行业从事数据分析的相关技术人员参考使用。

图书在版编目（CIP）数据

数据分析及其在电力系统中的应用 / 黄彦浩，于之虹，高波编. —北京：中国电力出版社，2019.4
ISBN 978-7-5198-2531-7

Ⅰ. ①数… Ⅱ. ①黄… ②于… ③高… Ⅲ. ①数据处理–应用–电力系统–系统分析–研究 Ⅳ. ①TM711-39

中国版本图书馆 CIP 数据核字（2018）第 238787 号

出版发行：中国电力出版社
地　　址：北京市东城区北京站西街 19 号（邮政编码 100005）
网　　址：http://www.cepp.sgcc.com.cn
责任编辑：邓　春　刘　薇（010-63412787）
责任校对：黄　蓓　李　楠
装帧设计：赵姗姗
责任印制：石　雷

印　　刷：三河市万龙印装有限公司
版　　次：2019 年 4 月第一版
印　　次：2019 年 4 月北京第一次印刷
开　　本：787 毫米×1092 毫米　16 开本
印　　张：14.75
字　　数：330 千字
印　　数：0001—1500 册
定　　价：70.00 元

前言

近年来随着大数据和人工智能的兴起，电力行业也掀起了相关研究和应用的热潮，从多种电力大数据平台的提出和构建，到大量机器学习方法在电力各领域的研究和应用，都在预示着电力系统数据从分散到汇集、从人工分析向智能分析方向发展的趋势。

本书旨在为从事电力系统大数据、人工智能研究和数据分析的工作人员，以及其他行业从事数据分析的相关技术人员提供参考。由于介绍各类算法理论的书籍已经汗牛充栋，因此本书在编写时更强调对实际工作的帮助。全书内容覆盖数据分析的基础、方法、实践和相关工具，在兼顾理论的同时，向应用侧重，力争以通俗易懂的方式介绍常见的理论要点和方法，不强调数学推导。

本书所涉及的内容繁多，但我们并不打算将其写成面面俱到的长篇巨著，而是遵循“先选择方法再深入研究”的思路，在各章的主体部分为读者提供能够了解方法要点的简明解释，在小结和参考文献部分提供开展深入研究所需的相关指引。可以将本书看作是数据分析的快速上手向导和工作的参考指南。此外，尽管本书与应用有关的叙述和例子主要是围绕电力系统仿真数据分析，但其中所涉及的要点、关键环节等是通用的。事实上，不同领域的数据分析在很多时候技术是相通的，所遇到的问题也会有相似之处。

全书共分为十二章：

（1）在第一～三章中，概述了数据分析的基础知识，主要是统计学知识，并且加入了一些重要的、在传统教材中较少涉及的内容，如核方法。对于这些新知识，本书力争能够让读者较为容易地明确其概念、基本原理和作用，并为希望深入研究的读者提供指引。

（2）在第四～九章中，本书介绍了十多种常见的数据分析方法，落脚点依然是使读者能够较为容易地明确其概念、基本原理和作用，并为希望深入研究的读者提供指引，也为读者根据问题选择方法提供参考。

（3）在第十～十二章中，本书以电力系统仿真数据分析为例，描述了在实际中如何

处理原始数据、提取特征、进行模型训练，以及搭建一个能够自动运行的数据分析系统。希望通过该部分的内容使读者可以具备更强的动手能力，并且了解实际情况下可能存在的各种问题。

（4）在附录中，本书简介了 R 语言、Matlab、Python 语言、SAS、SPSS 等常用数据分析软件，主要侧重于对这些软件特性和基本使用方法的描述。附录内容只提供二维码，感兴趣的读者可以扫码获取相关内容。

黄彦浩撰写了本书的第二、三、九～十二章和附录 A、C、D，于之虹撰写了本书的第四～八章，高波撰写了本书的第一章和附录 B。

本书的撰写得到了华北电力大学研究生张春、王恬月和中国电力科学研究院有限公司研究生徐华廷等人的大力协助，在此深表感谢！

由于作者水平有限，书中不足之处难免，恳请读者批评指正。

编　者

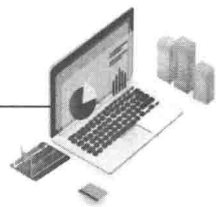
2018 年 8 月

目录

附录内容请扫码获取

第一章

数据分析的统计学基础



数据分析的本质在于通过对样本数据的归纳总结，来实现对总体未知数据的推测，因此参数估计和假设检验就是分析的主要工作之一。而方差分析和回归分析是数据分析中最具有广泛应用的内容，许多统计学习或机器学习算法都由此深化与拓展而来。本章主要为读者复习回顾一下统计学基础知识。需要说明的是，本部分内容强调的是概念的论述、结论的归纳，而不在于定义的精确阐述和公式的完整数学推导，这样可以让读者关注重点落于本书其他重要章节的了解与学习。

第一节 | 数据与统计的基本概念

一、随机事件

在个别试验中其结果呈现出不确定性，而在大量重复试验中其结果又具有统计规律性的现象称之为随机现象。在随机现象中发生的、满足一定或指定条件的任一个事件称之为随机事件。随机事件是这样一种事件，在发生条件相同的情况下，结果具有可重复性，但在事件未发生前，无法预测它的单次结果。

二、总体与样本、随机变量

在某一次随机试验中，所有试验的所有可能结果组成的集合称之为该试验的总体。统计学中，总体是指所有需要被研究的个体，在研究之前需要被严格而明确地定义。同时，在统计研究中，人们往往关心总体中每个个体的一项（或几项）数量指标和该数量指标在总体中的分布情况。这时，每个个体具有的数量指标的全体也可称之为总体。由于总体中的个体及其被关注的数量指标的出现带有随机性，因此通常定义 X，Y，Z，… 为随机变量。随机变量用以对随机试验结果或随机事件进行量化。随机变量的引入，使得我们可以用变量来描述各种随机现象和随机事件。因此随机变量首先是一个数学分析意义上的变量，在事先规定的有限或无限范围内取值，但不同的是其取值有一定的概率。

总体分为有限总体和无限总体。包含有限个或固定数目观察值的研究对象，称为有限总体；如果一个总体包含着无限多的单位或无数个观察值，则称为无限总体。但如果

总体数目较多且不便于全部考察，有时也把这样的有限总体视为无限总体。对有限总体的调查方式可以是全面调查，也可以调查其中一部分；而无限总体只能进行非全面调查，据以推断总体情况。本书以黑体字表示总体，即 $\boldsymbol{X}$，$\boldsymbol{Y}$，$\boldsymbol{Z}$，…。

以一定规则从总体中抽取若干个体进行观察实验，以获得总体的有关信息，这一过程称为抽样。抽取的个体集合或个体的数量指标的集合就是样本。通常记为 $(X_1, X_2, \cdots, X_n)$。样本中所包含的个体数量 n 称为样本容量。容量为 n 的样本可以看做 n 维随机变量。一旦取定一个样本，得到的是 n 个具体的数，即 $(x_1, x_2, \cdots, x_n)$，称此为样本的一次观测值，简称样本值。

样本来自于总体，应该被设计为可良好反映总体的情况，这就要求样本具有代表性和独立性。代表性指的是抽取的样本中的每一个随机变量 X_i 与总体 $\boldsymbol{X}$ 具有相同的分布；独立性指的是 $X_1, X_2, \cdots, X_n$ 之间相互独立，即每次观测结果互不影响。

概率的概念是，在一定条件下，重复做 n 次试验，n_A 为 n 次试验中事件 A 发生的次数，如果随着 n 逐渐增大，频率 n_A/n 逐渐稳定在某一数值 p 附近，则数值 p 称为事件 A 在该条件下发生的概率，记作 $P(A)=p$。随机变量的取值随试验的结果而定，而试验的各个结果出现存在一定的概率，因而随机变量的取值有一定的概率。

三、统计数据的分类

随机变量类型不同，其产生的统计数据也不同，并且采用的统计建模方法也不同。随机变量或者说统计数据通常有三种分类方法。第一种按计量层次可以分为分类的数据、顺序的数据和数值型的数据；第二种按收集方法可以分为观察的数据和实验的数据；第三种按时间状况可以分为截面的数据和时间序列的数据。

分类的数据指的是只能归于某一类别的非数字型数据，通常用符号或文字表述，比如人口的性别分为男、女两类，人口的籍贯等；顺序的数据指的是有序的非数字型数据，用于对事物类别顺序的测度，数据表现为类别，也用符号或文字表述，例如产品分为一等品、二等品、三等品，满意度分为非常满意、比较满意、不满意等；数值型的数据则是最常见的数据，比如身高、体重等可以用数字尺度测量的观测值，结果表现为具体的数值，可以实现对事物的精确测量。

观察的数据指的是调查或观测收集到的数据，它是在没有对事物人为控制的条件下而得到的数据，比如消费记录、调查问卷信息等，通常有关社会经济现象的统计数据几乎都是观察的数据；实验的数据指的是在实验中控制对象而收集到的数据，比如对一种新药疗效的实验、对一种新的农作物品种的实验，其一般通过参照组和实验组两个来对比分析，自然科学领域的数据大多数都为实验的数据。

截面的数据指的是在相同或近似相同的时间点上收集的数据，其用于描述现象在某一时刻的变化情况，比如某一年我国各地区的国内生产总值数据；时间序列的数据则是在不同时间上收集到的数据，用于描述现象随时间变化的情况，比如某个地区 2000～2015 年 GDP 的变化情况，股票随时间的涨跌幅度等。

四、离散与连续随机变量

离散随机变量是只取有限值或至多可列无限值的随机变量。如果离散随机变量的取值用 x_1，x_2，……表示，那么存在满足 $p(x_i)=P(X=x_i)$ 和 $\sum_i p(x_i)=1$ 的函数 p。我们称这个函数为随机变量 X 的概率密度函数，有时也利用随机变量的累积分布函数来表示：

$$F(x)=P(X\leqslant x),-\infty<x<+\infty \tag{1-1}$$

累积分布函数是非降的，并满足：

$$\lim_{x\to-\infty}F(x)=0\text{和}\lim_{x\to\infty}F(x)=1 \tag{1-2}$$

对于取值连续的随机变量，称之为连续随机变量。对应的概率密度函数 $f(x)\geqslant 0$，f 分段连续且 $\int_{-\infty}^{\infty}f(x)=1$。其在区间 (a,b) 上函数如下：

$$P(a<X<b)=\int_a^b f(x)\mathrm{d}x \tag{1-3}$$

其表征了随机变量 X 落入一个区间 $(a,\ b)$ 的概率。

需要注意的是这种定义的情况下，连续型随机变量 X 取特定值的概率为 0，即：

$$P(X=c)=\int_c^c f(x)\mathrm{d}x=0 \tag{1-4}$$

则有：

$$P(a<X<b)=P(a\leqslant X<b)=P(a<X\leqslant b) \tag{1-5}$$

其累积分布函数如下：

$$F(x)=P(X\leqslant x)=\int_{-\infty}^{x}f(u)\mathrm{d}x \tag{1-6}$$

五、随机变量的数字特征

随机变量当然用分布函数或概率密度可以全面详尽地进行表征，但有时在实际分析中并不需要全面了解随机变量的分布情况，或者有时根本不可能确定其分布状况，只需要使用一些具有可以部分或大概反映其特征的数字即可，比如均值或方差。如此虽然会在信息上有损失，但却可以直观地描述随机变量在某些方面的重要特征。这些特征在理论和实践上都具有重要意义。随机变量最常见的数字特征包括：期望、方差、相关系数和矩。

1. 数学期望

对于离散型随机变量 X，设其分布律 $P\{X=x_k\}=p_k$，$k=1,2,\cdots$，若级数 $\sum_{k=1}^{\infty}x_k p_k$ 绝对收敛，则称级数 $\sum_{k=1}^{\infty}x_k p_k$ 为随机变量 X 的期望。即：

$$E(X)=\sum_{k=1}^{\infty}x_k p_k \tag{1-7}$$

连续型随机变量 X 的概率密度为 $f(x)$，则 $\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$ 的值为随机变量 X 的数学期望，即：

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x \tag{1-8}$$

无论是离散型还是连续型随机变量的数学期望都有如下性质：

（1）常数 C 的期望 $E(C)=C$。

（2）随机变量与常数乘积的期望为随机变量期望与常数的乘积，即：

$$E(CX)=CE(X) \tag{1-9}$$

（3）随机变量和的期望等于各个随机变量期望的和，即：

$$E(\sum X_i)=\sum E(X_i),i=1,2,\cdots,n \tag{1-10}$$

（4）有限个相互独立的随机变量之积的期望等于各个随机变量期望之积，即：

$$E(\prod X_i)=\prod E(X_i),i=1,2,\cdots,n \tag{1-11}$$

（5）若随机变量 X 的函数 $Y=r(X)$ 是连续函数，那么 Y 的期望可以直接用 X 计算或表达。对于离散变量 X，有 $E(Y)=E\left[r(X)\right]=\sum_{k=1}^{\infty}r(x_k)p_k$；对于连续变量 X，有 $E(Y)=E\left[r(X)\right]=\int_{-\infty}^{\infty}r(x)f(x)\mathrm{d}x$。

2. 方差

为了刻画随机变量取值的分散程度，采用变量值与其均值的偏离程度来计算，即：

$$E\left\{\left[X-E(X)\right]^2\right\} \tag{1-12}$$

通常记为 $D(X)$ 或 $\mathrm{Var}(X)$，经常使用的还有标准差或均方差 $\sigma(X)$，其为方差的开根号值。方差或标准差可理解为是一个距离单位，总体或样本数据以此衡量自己在数据集中的位置。

方差的计算除了按照定义还有两种办法：

$$D(X)=E\left(X^2\right)-E^2(X) \tag{1-13}$$

或者，对于离散变量 X，其分布律为 $P\{X=x_k\}=p_k$，$k=1,2,\cdots$，则：

$$D(X)=\sum_{k=1}^{\infty}\left[x_k-E(X)\right]^2 p_k \tag{1-14}$$

对于连续变量 X，有：

$$D(X)=\int_{-\infty}^{\infty}\left[x-E(X)\right]^2 f(x)\mathrm{d}x \tag{1-15}$$

无论随机变量是离散型还是连续型，其都有如下计算性质：

（1）常数 C 的方差为零，即 $D(C)=0$。

（2）常数 C 与随机变量 X 乘积的方差等于 X 的方差乘以常数 C 的平方，即：

$$D(CX)=C^2D(X) \tag{1-16}$$

（3）两个随机变量 X,Y 和的方差为：

$$D(X+Y)=D(X)+D(Y)+2E\left\{\left[X-E(X)\right]\left[Y-E(Y)\right]\right\} \tag{1-17}$$

（4）有限多个相互独立的随机变量之和的方差等于各个随机变量方差的和，即：

$$D(\sum X_i)=\sum D(X_i),\ i=1,2,\cdots,n \tag{1-18}$$

3. 相关系数

两个不相互独立的变量必然存在某种关系。相关系数可以用来衡量两者之间是否存在线性关系或存在线性关系的强度。相关系数定义如下：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E\left\{\left[X-E(X)\right]\left[Y-E(Y)\right]\right\}}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{1-19}$$

$\mathrm{Cov}(X,Y)$ 通常称为两者的协方差。$|\rho_{XY}|\leqslant 1$，当 $|\rho_{XY}|$ 较大时，表明两个随机变量存在较强的线性关系，特别地，当 $|\rho_{XY}|=1$，表明这两个随机变量以 100%概率存在着线性关系；$|\rho_{XY}|=0$ 时，称 X 和 Y 不相关。

二维正态分布的随机变量 X,Y，X 和 Y 不相关与 X 和 Y 相互独立是等价的。

4. 矩

设有随机变量 X 和 Y，若 $E\left(X^k\right)$，$E\left(Y^k\right)$，$k=1,2,\cdots$ 存在，则称两者分别为 X 和 Y 的 k 阶矩；

若 $E\left\{\left[X-E(X)\right]^k\right\}$，$E\left\{\left[Y-E(Y)\right]^k\right\}$，$k=1,2,3,\cdots$ 存在，则称两者分别为 X 和 Y 的 k 阶中心矩；

若 $E\left(X^kY^l\right),k=1,2,\cdots$ 存在，称之为 X 和 Y 的 $k+l$ 阶混合矩；

若 $E\left\{\left[X-E(X)\right]^k\left[Y-E(Y)\right]^l\right\}, k=1,2,\cdots$ 存在，称之为 X 和 Y 的 $k+l$ 阶混合中心矩。

对于一般情况下，若 n 维随机变量 $(X_1,X_2,\cdots,X_n)$ 的二阶混合中心矩

$$c_{ij}=\mathrm{Cov}(X_i,X_j)=E\left\{\left[X_i-E(X_i)\right]^k\left[X_j-E(X_j)\right]^l\right\}, i,j=1,2,\cdots$$

存在，则称矩阵 $\boldsymbol{C}=\begin{bmatrix} c_{11} & c_{12}\cdots & c_{1n} \\ c_{21} & c_{22}\cdots & c_{2n} \\ \vdots & \vdots\ \cdots & \vdots \\ c_{n1} & c_{n2}\cdots & c_{nn} \end{bmatrix}$ 为 n 维随机变量 $(X_1,X_2,\cdots,X_n)$ 的协方差矩阵。

5. 重要定理与不等式

（1）马尔科夫不等式。如果 X 是只取非负值的随机变量，那么对于任意 $a>0$，有：

$$P(X\geqslant a)\leqslant\frac{E(X)}{a} \tag{1-20}$$

（2）切比雪夫不等式。如果 X 是具有均值 μ 和方差 σ^2 的随机变量，那么对于任意 $k>0$，有：

$$P(|X-\mu|\geqslant k)\leqslant\frac{\sigma^2}{k^2} \tag{1-21}$$

马尔科夫不等式和切比雪夫不等式的重要性在于，当概率分布的均值或者均值和方差已知时，就能得到概率的上界。

（3）辛钦定理。设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，服从同一分布，且都具有数学期望 $E(X_k)=\mu(k=1,2,\cdots)$，则对于任意正数 ε，有：

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}=1 \tag{1-22}$$

（4）独立同分布中心极限定理。若 $X_1,X_2,\cdots,X_n,\cdots$ 是一系列独立同分布的随机变量，每个具有均值 μ 和方差 σ^2，那么当 $n\to\infty$ 时，$\dfrac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ 的分布趋于标准正态分布，即当 $n\to\infty$ 时：

$$P\left(\frac{\sum_{k=1}^{n}X_k-n\mu}{\sigma\sqrt{n}}\leqslant a\right)\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}t \tag{1-23}$$

该定理的另一种形式可以描述为 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 近似服从标准正态分布。

（5）独立非同分布中心极限定理。设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，它们具有不同的数据期望和方差

$$E(X_k)=\mu_k,\ D(X_k)=\sigma_k^2,\ k=1,2,\cdots$$

若存在正数 $\delta>2$，使得当 $n\to\infty$ 时 $\dfrac{1}{\sum\limits_{k=1}^{n}\sigma_k^{\delta}}\sum\limits_{k=1}^{n}E\left\{\left|X_k-\mu_k\right|^{\delta}\right\}\to 0$，则随机变量之和 $\sum\limits_{k=1}^{n}X_k$ 的标准化变量为：$Z_n=\dfrac{\sum\limits_{k=1}^{n}X_k-E\left(\sum\limits_{k=1}^{n}X_k\right)}{\sqrt{D\left(\sum\limits_{k=1}^{n}X_k\right)}}=\dfrac{\sum\limits_{k=1}^{n}X_k-\sum\limits_{k=1}^{n}\mu_k}{\sqrt{\sum\limits_{k=1}^{n}\sigma_k^2}}$。当 n 很大时，近似服从标准正态分布 $N(0,1)$。

六、常见概率分布

1. 伯努利分布

伯努利随机变量只可取两个值，即 0 和 1，各自的取值概率分别为 $1-p$ 和 p。它的密度函数表达式为：

$$\begin{cases}p(1)=p\\ p(0)=1-p\\ p(0)=0,\ \text{若}x\neq 0\text{且}x\neq 1\end{cases}\tag{1-24}$$

又可以表示为：

$$p(x)=\begin{cases}p^x(1-p)^{1-x},\ \text{若}x=0\text{或}x=1\\ 0,\ \text{其他}\end{cases}\tag{1-25}$$

2. 二项分布

若进行 n 次独立试验，n 固定，试验结果包括且仅包括 2 种可能：A 或 B。若每次试验结果为 A 的概率是 p，结果为 B 的概率是 $1-p$。试验所有可能发生的结果情况 X 将是一个参数为 n 和 p 的二项随机变量，其发生 k 次结果为 A 的概率可以表达如下：

$$p(k)=C_n^k p^k(1-p)^{n-k}\tag{1-26}$$

服从二项分布的随机变量可以利用相互独立的伯努利随机变量表示，即可以令 $X_1, X_2, \cdots, X_n$ 是相互独立、$p(X_i=1)=p$ 的伯努利随机变量，那么 $Y=X_1+X_2+\cdots+X_n$ 是一个二项随机变量。

3. 几何分布

无穷次伯努利试验可以构造几何分布。若每次试验发生 A 的概率是 p，X 表示直到第一次发生 A 时所做的试验次数，那么 X 的概率密度分布函数是：

$$p(k)=P(X=k)=(1-p)^{k-1}p,\ k=1,2,3\cdots\tag{1-27}$$

并且有：

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p=1 \tag{1-28}$$

如果考虑连续独立地试验直到 A 发生 r 次时，试验次数 X 的概率分布，可以写为如下：

$$p(k)=P(X=k)=C_{k-1}^{r-1}p^{r}(1-p)^{k-r} \tag{1-29}$$

有时称其为负二项分布。

4. 超几何分布

产品检验中经常遇到一类实际问题，假定在 N 件产品中有 M 件不合格产品，从中随机、不重复抽取 n 件产品，能有 $k(k\leqslant n)$ 件不合格的概率密度分布函数：

$$p(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n},\ k=0,1,\cdots,n \tag{1-30}$$

需要注意的是超几何分布是不重复抽样。

5. 泊松分布

二项分布当满足试验次数 n 趋于无穷，试验结果为 A 的概率 p 极小而趋于 0，且满足 $np=\lambda$ 时，二项分布可以由泊松分布良好地近似表示，其密度函数为：

$$P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\ k=0,1,2,\cdots \tag{1-31}$$

注意到 $\mathrm{e}^{\lambda}=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$，因此 $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=1$。

泊松分布常用来分析电话系统或保险公司的模型，比如单位时间内到达交换机的呼叫次数可以用泊松分布来建模，或者给定时间区间内大群体中的小概率事件发生模型等。

6. 均匀分布

区间 $[a,b]$ 内任何实数都是一个可能的试验结果，这样的模型为均匀随机分布，其密度函数表示为：

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases} \tag{1-32}$$

其累积分布函数为：

$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\leqslant x\leqslant b\\ 1, & x\geqslant b\end{cases} \tag{1-33}$$

7. 威布尔分布

非负连续型随机变量 T 具有如下定义的概率密度函数时，称此随机变量服从威布尔