Probability

Haoqi ZHAO

November 2023

1 Introduction

2 概率分布关系总览

本文档旨在梳理和解释不同概率分布之间的关系，以及它们如何从贝努利实验衍生而来。

2.1 贝努利实验

贝努利实验是最基础的概率实验，只有两种可能的结果。例如，抛硬币的实验，正面出现的概率为 p，反面出现的概率为 $1-p$。
期望：p
方差：$p(1-p)$

2.2 二项分布

二项分布描述了连续进行 n 次贝努利实验时，观察到特定结果（如正面）出现 k 次的概率，公式为：

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

期望：np
方差：$np(1-p)$

2.3 泊松分布

当二项分布中的试验次数 n 很大且成功概率 p 很小时（特别是 np 固定），二项分布逼近泊松分布。泊松分布的公式为：

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

其中 $\lambda=np$。
期望：λ
方差：λ

2.4 几何分布

几何分布是负二项分布的特殊情况，描述的是在重复贝努利实验中，首次成功（如正面出现）所需的试验次数 Y，其概率质量函数为：

$$P(Y = k) = (1-p)^{k-1}p$$

期望：$\frac{1}{p}$
方差：$\frac{1-p}{p^2}$

2.5 负二项分布

在贝努利实验中，要让特定结果（如正面）出现确切的 r 次所需的试验次数 X，服从负二项分布，其概率质量函数为：

$$P(X = k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}$$

期望：$\frac{r}{p}$
方差：$\frac{r(1-p)}{p^2}$

2.6 贝塔分布

贝塔分布是定义在区间 $[0, 1]$ 上的连续概率分布，具有两个参数 α 和 β。其概率密度函数为：

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

期望：$\frac{\alpha}{\alpha+\beta}$
方差：$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

2.7 指数分布与伽马分布

伽马分布可以看作是多个独立同分布的指数分布随机变量的和，即等待事件发生 r 次的总时间。伽马分布的概率密度函数为：

$$f(x;\alpha,\beta)=\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}$$

期望：$\frac{\alpha}{\beta}$
方差：$\frac{\alpha}{\beta^2}$

2.8 指数分布

指数分布是几何分布在连续时间上的对应物，描述的是等待第一个事件发生所需的时间 T，其概率密度函数为：

$$f(t;\lambda)=\lambda e^{-\lambda t}$$

期望：$\frac{1}{\lambda}$
方差：$\frac{1}{\lambda^2}$

2.9 几何分布与负二项分布

负二项分布可以看作是多个几何分布的总和，即多次重复几何分布实验，直到事件发生 r 次。

3 Helpful example

我们有两个离散随机变量 X 和 Y，联合概率分布如下表所示：

$X\backslash Y$	a	b
1	0.1	0.3
2	0.2	0.4

计算 Y 的边缘概率分布 $p_Y(y)$：

$$p_Y(a)=p_{X,Y}(1,a)+p_{X,Y}(2,a)=0.1+0.2=0.3$$
$$p_Y(b)=p_{X,Y}(1,b)+p_{X,Y}(2,b)=0.3+0.4=0.7$$

接着计算 X 的条件期望 $E[X|Y=y]$：

$$E[X|Y=a]=\sum_x x\cdot p_{X|Y}(x|a)=1\cdot\frac{0.1}{0.3}+2\cdot\frac{0.2}{0.3}=\frac{1}{3}+\frac{4}{3}=\frac{5}{3}$$
$$E[X|Y=b]=\sum_x x\cdot p_{X|Y}(x|b)=1\cdot\frac{0.3}{0.7}+2\cdot\frac{0.4}{0.7}=\frac{3}{7}+\frac{8}{7}=\frac{11}{7}$$

最后，根据定理计算 X 的总期望 $E(X)$：

$$\begin{aligned}
E(X)&=\sum_y p_Y(y)\cdot E[X|Y=y]\\
&=p_Y(a)\cdot E[X|Y=a]+p_Y(b)\cdot E[X|Y=b]\\
&=0.3\cdot\frac{5}{3}+0.7\cdot\frac{11}{7}\\
&=\frac{5}{3}\cdot\frac{3}{10}+\frac{11}{7}\cdot\frac{7}{10}\\
&=\frac{5}{10}+\frac{11}{10}=\frac{16}{10}=1.6
\end{aligned}$$

因此，随机变量 X 的期望值 $E(X)$ 是 1.6。

4 中心极限定理（Central Limit Theorem, CLT）

中心极限定理是统计学中一个非常重要的概念，它表明，对于足够大的样本量，独立且同分布的随机变量之和（或平均值）的分布将近似为正态分布，无论原始随机变量的分布如何。

4.1 定义

如果 $X_1, X_2, \ldots, X_n$ 是一系列独立同分布的随机变量，且具有均值 μ 和标准差 σ，那么当样本量 n 足够大时，样本均值

$$\bar{X}=\frac{1}{n}(X_1+X_2+\ldots+X_n)$$

的分布接近正态分布 $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$。

4.2 应用场景

- 样本均值的分布估计

- 置信区间的计算
- 假设检验
- 品质控制

4.3 相关公式

- 标准化变量：$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$，其中 $Z \sim N(0,1)$。
- 样本总和：$S_n = X_1 + X_2 + \ldots + X_n$，其中 $E(S_n) = n\mu$，标准差为 $\sigma\sqrt{n}$。

5 额外说明

中心极限定理的应用非常广泛，特别是在样本量较大时。在实际应用中，通常认为当样本量大于或等于 30 时，样本均值的分布可以较好地近似为正态分布。该定理为许多统计方法提供了理论基础，尤其是在原始数据分布未知或非正态分布的情况下。