## 生物统计学

#### 第三章 概率与概率分布

云南大学 生命科学学院



會澤百宗 至心天下

YN3010180007 1 / 79

- ❶ 概率基础
- 2 随机变量及其概率分布
- 3 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 2 / `

- ❶ 概率基础
- 2 随机变量及其概率分布
- ③ 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 3 / 79

### 3.1 概率基础 3.1.1 随机事件

#### 事件在概率论中可分为三类:

- 必然事件 (certain event), 在一定条件下必然出现的现象。常用  $\Omega$  表示。
- 不可能事件 (impossible event), 在一定条件下必不出现的现象。常用 ∅ 表示。
- 随机事件 (random event), 在一定条件下可能出现,也可能不出现的现象。常用大写英文字母  $A, B, C, \dots$  表示。

YN3010180007 4 / 79

3.1.1 随机事件

使随机现象得以实现和观察的全过程, 称为随机试验 (random experiment)。

#### 具有以下三个特征:

- 可重复性, 试验在相同条件下可以重复进行;
  - 可知性,每次试验的可能结果不止一个,但事先能明确所有可能的结果;
  - 随机性,在完成试验之前不能确定哪一个结果会出现,但必然出现结果中的一个。

YN3010180007 5 / 79

3.1.1 随机事件

概率论中,通常把单一的试验结果称为一个基本事件 (elementary event),

一些基本事件组合起来即构成复合事件 (compound event)。

YN3010180007 6 / 79

### 3.1 概率基础 3.1.1 随机事件

### 定义 (3.1)

随机试验的基本事件,也称样本点  $\omega_i$ ,构成样本空间  $\Omega$ 。样本空间中的任一子 集 A,称为随机事件,有  $A \subset \Omega$ 。

因为空集  $\emptyset$  和样本空间  $\Omega$  本身都是样本空间的子集,用集合论的语言来讲,不可能事件  $\emptyset$  和必然事件  $\Omega$  属于两类特殊的随机事件。

YN3010180007 7 / 79

3.1.2 频率

### 定义 (3.2)

设事件 A 在 n 次重复试验中发生了 m 次,有比值

$$W(A) = \frac{m}{n}, \quad 0 \le W(A) \le 1$$
 (3.1)

该比值称为事件 A 发生的频率 (frequency)。

YN3010180007 8 / 79

3.1.3 概率

### 定义 (3.3)

事件 A 在 n 次重复试验中,发生了 m 次,当试验次数 n 不断增大时,事件 A 发生的频率 W(A) 趋近某一确定值 p,则 p 为事件 A 发生的概率(probability),即

$$P(A) = p = \lim_{n \to \infty} \frac{m}{n} \tag{3.2}$$

用频率的极限形式来定义概率, 称为统计概率。

YN3010180007 9 / 79

3.1.3 概率

### 定义 (3.4)

在一个有 n 个等可能结果的随机试验中,事件 A 包含其中 m 个结果,则事件 A 的发生概率可定义为

$$P(A) = \frac{m}{n} \tag{3.3}$$

这就是概率的古典定义, 或称古典概率。

概率有以下性质:

- ① 非负性: 样本空间  $\Omega$  任意子集 A 的概率在 0 到 1 之间,即  $\forall A \subseteq \Omega, 0 \le P(A) \le 1$ 。
- ② 归一性: 样本空间  $\Omega$  的概率为 1, 即  $P(\Omega) = 1$ 。

YN3010180007 10 / 79

#### 3.1.4 事件的相互关系

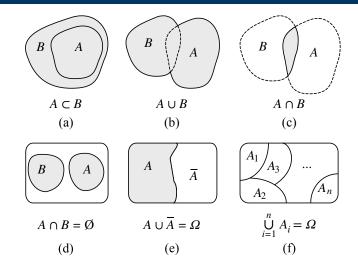


图 3.1 事件的相互关系

YN3010180007 11 / 79

3.1.5 概率的计算法则 ...... 加法法则

设有事件 A 和事件 B, 它们的和事件概率为

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
 (3.4)

若事件 A 和 B 互斥,则有  $P(A \cap B) = 0$ ,所以  $P(A \cup B) = P(A) + P(B)$ 。由此可得概率的加法定理 (additive law of probability)。

#### 定理 (3.1)

互斥事件 A 和 B 的和事件的概率,等于事件 A 和事件 B 的概率之和,即

$$P(A+B) = P(A) + P(B)$$
 (3.5)

YN3010180007 12 / 79

3.1.5 概率的计算法则 ...... 加法法则

加法定理也可推广至多个互斥事件的和事件的概率,即

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$$
 (3.6)

此即概率的第三条性质: 可列可加性。

YN3010180007 13 / 79

3.1.5 概率的计算法则 ...... 加法法则

因对立事件互斥, 所以将加法定理应用到一对对立事件上, 就有

$$P(\overline{A}) = 1 - P(A) \tag{3.7}$$

即事件 A 与其逆事件的概率之和为 1。

YN3010180007 14 / 79

3.1.5 概率的计算法则 ...... 乘法法则

若事件 A 的发生与事件 B 有关,则事件 A 在事件 B 发生的前提下发生的概率,被称为事件 A 发生的条件概率 (conditional probability),记作 P(A|B)。

根据概率的古典定义 (定义 3.4), 可知条件概率的计算方法:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.8}$$

所以两事件交的概率可以表示为  $P(A \cap B) = P(A|B)P(B)$ 。

YN3010180007 15 / 79

当事件 A 和 B 相互独立时,即 P(A|B) = P(A),可据此推出概率的乘法定理 (multiplicative law of probability)。

#### 定理 (3.2)

如果事件 A 和事件 B 相互独立,则事件 A 与事件 B 同时发生的概率等于事件 A 和事件 B 各自概率的乘积,即

$$P(A \cap B) = P(A) \times P(B) \tag{3.9}$$

其中 P(A|B) = P(A) 或 P(B|A) = P(B) 是事件独立关系的概率表达形式,可用于证明事件的独立关系。

YN3010180007 16 / 79

- 概率基础
- 2 随机变量及其概率分布

随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布 随机变量的数字特征

- 3 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 17 / 79

- 概率基础
- ② 随机变量及其概率分布 随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布
- ③ 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 18 / 79

3.2.1 随机变量及其类型

- 离散型随机变量 (discrete random variable)
   全部可能取值为有限个或可数无穷个,且取相应值的概率是确定的。
- 连续型随机变量 (continuous random variable) 全部可能取值为某范围内的任何值 (无穷且不可数), 且其中任一区间取值的概率是确定的。

YN3010180007 19 / 79

3.2.1 随机变量及其类型

概率分布是描述随机变量取值概率的函数, 主要涉及三种函数:

- 概率质量函数 (probability mass function, PMF)
   用于描述离散型随机变量取值及其概率的关系。
- 概率密度函数 (probability density function, PDF)
   用于描述连续型随机变量在特定值上的概率密度。
- 累积分布函数 (cumulative distribution function, CDF)
   有时简称为分布函数,用于描述随机变量取值小于等于特定值的概率。

YN3010180007 20 / 79

- 概率基础
- ②随机变量及其概率分布 随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布 随机变量的数字特征
- ③ 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 21 / 79

3.2.2 离散型随机变量的概率分布 ...... 概率质量函数

设 X 是某个离散型随机变量, 其概率质量函数可表示为

$$f(x) = P(X = x_i) = p_i, \quad i = 1, 2, \cdots$$
 (3.13)

其中x 是X 的某个可能的观测值, $p_i$  表示X 取到 $x_i$  的概率。

概率质量函数满足概率的三个性质:

- 非负性
- 2 归一性
- 3 可列可加性。

YN3010180007 22 / 79

3.2.2 离散型随机变量的概率分布 ...... 概率质量函数

表 3.1 离散型随机变量的概率分布表

变量 x	$x_1$	$x_2$	$x_3$	 $x_i$	 $x_n$
概率 p	$p_1$	$p_2$	$p_3$	 $p_{i}$	 $p_n$

3.2.2 离散型随机变量的概率分布 ...... 概率质量函数

#### 例 (3.2)

连续投掷 2 次 6 个面的均质骰子,所得点数之和也为一个随机变量,求该随机变量的概率质量函数,并用概率分布表表示。

YN3010180007 24 / 79

3.2.2 离散型随机变量的概率分布 ...... 概率质量函数

解

该离散型随机变量的概率质量函数为

$$f(x) = P(X_1 + X_2 = x) = \begin{cases} \frac{1}{36} \times (x - 1), & x \in \{2, 3, 4, 5, 6, 7\} \\ \frac{1}{36} \times (13 - x), & x \in \{8, 9, 10, 11, 12\} \end{cases}$$

其中  $X_1$  和  $X_2$  分别表示前后 2 次投掷所得点数, x 为点数之和。

表 3.2 两次掷骰子点数之和的概率分布表

	2	3	4	5	6	7	8	9	10	11	12
概率 p	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

YN3010180007 25 / 79

3.2.2 离散型随机变量的概率分布 ...... 累积分布函数

离散型随机变量 X 的累积分布函数为

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} P(X = x_i) = \sum_{x_i \le x} p_i$$
 (3.14)

该函数本质上描述的是一个 和事件 的概率。

随机变量每次只能取一个值,所以在  $X \leq x$  的范围内,X 取不同的值应为互斥事件。根据概率加法定理可知,互斥事件的和事件概率等于事件概率之和。

YN3010180007 26 / 79

3.2.2 离散型随机变量的概率分布 ...... 累积分布函数

### 例 (3.3)

结合例题 3.2, 试求点数之和小于或等于 8 的概率。

#### 解

据离散型随机变量的累积分布函数公式,以及 2 次掷骰子点数之和的概率质量函数,有

$$P(X \le 8) = \sum_{x=2}^{8} f(x)$$

$$= f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8)$$

$$= \frac{26}{36}$$

YN3010180007 27 / 79

- 概率基础
- ② 随机变量及其概率分布 随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布
  - 随机变量的数字特征
- ③ 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 28 / 79

3.2.3 连续型随机变量的概率分布 ...... 概率密度函数

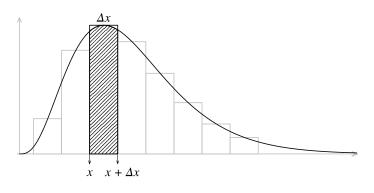


图 3.2 连续变量概率密度示意图

任意区间  $[x, x + \Delta x]$  内的取值概率可以表示为  $P(x \le X \le x + \Delta x)$ 。

YN3010180007 29 / 79

3.2.3 连续型随机变量的概率分布 ...... 概率密度函数

 $\frac{P(x \le X \le x + \Delta x)}{\Delta x}$  存在一个极限值,称为随机变量 X 在点 x 处的概率密度 (probability density),用 f(x) 表示,即

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \le X \le x + \Delta x)}{\Delta x}$$
 (3.15)

在随机变量 X 取值的全域内,所有概率密度构成了一条平滑的函数曲线,称为概率密度曲线,相应的函数即概率密度函数。

YN3010180007 30 / 79

3.2.3 连续型随机变量的概率分布 ...... 概率密度函数

由定积分的知识, 可知

$$P(x \le X \le x + \Delta x) = \int_{x}^{x + \Delta x} f(x) dx$$
 (3.16)

表示是在区间  $[x, x + \Delta x]$  内概率密度曲线与 x 轴所围的面积。

当  $\Delta x \rightarrow 0$  时,面积为 0,所以就有连续型随机变量 X 取特定值 x 的概率为 0的结果。因此有

$$P(x \le X \le x + \Delta x) = P(x < X < x + \Delta x) \tag{3.17}$$

YN3010180007 31 / 79

3.2.3 连续型随机变量的概率分布 ...... 累积分布函数

与离散型随机变量的累积分布函数一样,连续型随机变量的累积分布函数同样 表示随机变量 X 取小于某值 x 的概率,即

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x) dx$$
 (3.18)

YN3010180007 32 / 79

3.2.3 连续型随机变量的概率分布 ...... 累积分布函数

与离散型随机变量的累积分布函数一样,连续型随机变量的累积分布函数同样 表示随机变量 X 取小于某值 x 的概率,即

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x) dx$$
(3.18)

对概率密度函数积分得累积分布函数; 对累积分布函数求导得概率密度函数。

YN3010180007 32 / 79

3.2.3 连续型随机变量的概率分布 ...... 累积分布函数

#### 累积分布函数F(x) 有以下性质:

- 取值于 [0,1],  $ploseprepsize{0.5em} 0 \le F(x) \le 1$ .
- 单调不减函数, 即当  $x_1 < x_2$  时,  $F(x_1) \le F(x_2)$ 。
- 右连续函数, 即  $\lim_{n\to x_0^+} F(x) = F(x_0), \forall x_0 \in \mathbf{R}$ .

YN3010180007 33 / 79

- 概率基础
- 2 随机变量及其概率分布

随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布 **随机变量的数字特征** 

- 3 常见的概率分布
- 4 大数定律与中心极限定理

YN3010180007 34 / 79

3.2.4 随机变量的数字特征 ......数学期望

数学期望的概念源于"分赌本问题"。

YN3010180007 35 / 79

3.2.4 随机变量的数字特征 ......数学期望

#### 数学期望的概念源于"分赌本问题"。

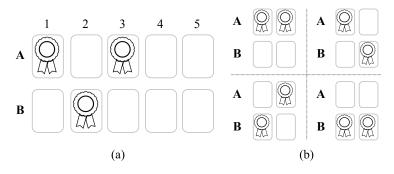


图 3.3 分赌本问题

YN3010180007 35 / 79

3.2.4 随机变量的数字特征 ......数学期望

引入一个随机变量 X,表示在当前的局面下 (A 两胜一负) 继续赌下去 A 的最终所得,X 有两个可能的取值 (100,0),取值概率分别为  $(\frac{3}{4},\frac{1}{4})$ 。

A 的期望所得等于 X 的可能值与其概率之积的累加(加权平均),即

$$100\times\frac{3}{4}+0\times\frac{1}{4}=75$$

YN3010180007 36 / 79

3.2.4 随机变量的数字特征 ......数学期望

## 定义 (3.5)

设**离散型**随机变量 X 可取有限个值  $(x_1,x_2,\cdots,x_n)$ ,取值概率分别为  $P(X=x_i)=p_i,\ i\in\{1,\cdots,n\}$ ,若级数  $\sum_{i=1}^n x_i p_i=x_1 p_1+x_2 p_2+\cdots+x_n p_n$  收敛,则称其为随机变量 X 的数学期望,记作 E(X),即

$$E(X) = \sum_{i=1}^{n} x_i p_i \tag{3.19}$$

#### 定义 (3.6)

设**连续型**随机变量 X 的概率密度函数为 f(x),若积分  $\int_{-\infty}^{+\infty} x f(x) dx$  收敛,则称其为随机变量 X 的数学期望,记作 E(X),或简记 EX,即

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$
 (3.20)

YN3010180007 37 / 79

3.2.4 随机变量的数字特征 ......数学期望

#### 数学期望具有以下性质:

- 常数 c 的期望等于常数本身, 即 E(c) = c。
- 随机变量数乘的期望等于随机变量期望的数乘, 即 E(aX) = aE(X)。
- 若干随机变量之和的期望等于各变量的期望之和,即

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

• 若干独立随机变量之积的期望等于各变量的期望之积,即

$$E(X_1X_2\cdots X_n)=E(X_1)E(X_2)\cdots E(X_n)$$

YN3010180007 38 / 79

3.2.4 随机变量的数字特征 ...... 方差

#### 定义 (3.7)

设随机变量 X, 若

$$Var(X) = E[(X - EX)^2]$$

(3.21)

存在,则称其为随机变量 X 的方差,记作  $\mathrm{Var}(X)$ 。

YN3010180007 39 / 79

3.2.4 随机变量的数字特征 ...... 方差

#### 方差具有以下性质:

- 常数 c 的方差为 0, 即 Var(c) = 0。
- 随机变量乘常数 c 的方差等于随机变量的方差乘以  $c^2$ , 即  ${\rm Var}(cX)=c^2{
  m Var}(X)$ 。
- 独立随机变量之和的方差等于各变量的方差之和, 即

$$\operatorname{Var}(X_1 + \dots + X_n) = \operatorname{Var}(X_1) + \dots + \operatorname{Var}(X_n)$$

• 结合性质 2 和 3, 有  $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$ 。

YN3010180007 40 / 79

- 概率基础
- 2 随机变量及其概率分布

#### 3 常见的概率分布

两点分布 二项分布 泊松分布 超几何分布 正太公布

4 大数定律与中心极限定理

YN3010180007 41 / 79

- 概率基础
- 2 随机变量及其概率分布
- ③ 常见的概率分布 两点分布

二项分布 泊松分布 超几何分布

4 大数定律与中心极限定理

YN3010180007 42 / 79

3.3.1 两点分布

#### 伯努利试验 (Bernoulli trial)

在同样条件下可重复、且相互独立的一种随机试验。

特点是试验只有两种可能的结果:发生或者不发生。

$$f(x) = P(X = x) = p^{x}(1-p)^{1-x}$$
 (x = 0,1) (3.24)

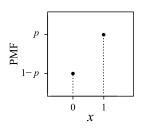


图 3.4 两点分布

YN3010180007 43 / 79

- ① 概率基础
- 2 随机变量及其概率分布

#### 3 常见的概率分布

两点分布

#### 二项分布

泊松分布 超几何分布

正态分布

4 大数定律与中心极限定理

YN3010180007 44 / 79

3.3.2 二项分布

假设现有某品种小麦的种子 12 粒,对其进行发芽试验。为每粒种子准备一样的营养土和独立的小花盆,种植一段时间后、假设得到以下试验结果。

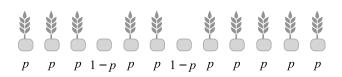


图 3.6 某品种小麦发芽试验

YN3010180007 45 / 79

3.3.2 二项分布

n 重伯努利试验中事件发生的概率分布, 称为二项分布 (binomial distribution), 记作 B(n,p)。有两个参数: 试验次数 n和事件发生的概率 p。

$$f(x; n, p) = P(X = x) = C_n^x p^x (1 - p)^{n - x}, \quad x \in \{0, 1, 2, \dots, n\}$$
 (3.25)

服从二项分布的随机变量 X 有数学期望 np,方差 np(1-p)。

YN3010180007 46 / 79

#### 3.3.2 二项分布

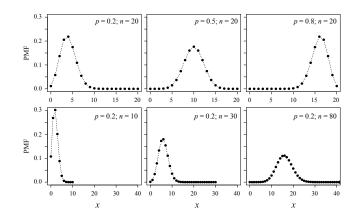


图 3.5 不同 n 和 p 值的二项分布

YN3010180007 47 / 79

3.3.2 二项分布

## 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

YN3010180007 48 / 79

3.3.2 二项分布

## 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

#### 解 (问题 1)

$$P(x \ge 2) = 1 - C_{100}^{0} \times 0.0056^{0} (1 - 0.0056)^{100} - C_{100}^{1} \times 0.0056^{1} (1 - 0.0056)^{99}$$
$$= 1 - 0.5703108 - 0.3211726$$
$$\approx 0.109$$

YN3010180007 48 / 79

3.3.2 二项分布

## 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

YN3010180007 49 / 79

3.3.2 二项分布

## 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

### 解 (问题 2)

$$1 - P(0) = 0.95$$

$$C_n^0 \times 0.0056^0 (1 - 0.0056)^n = 0.05$$

$$0.9944^n = 0.05$$

YN3010180007 49 / 79

3.3.2 二项分布

## 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

YN3010180007 50 / 79

3.3.2 二项分布

#### 例 (3.4)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

#### 解 (问题 2)

取对数后得

$$n \lg 0.9944 = \lg 0.05$$
 
$$n = \frac{\lg 0.05}{\lg 0.9944} = 533.4529$$

YN3010180007 50 / 79

- 概率基础
- 2 随机变量及其概率分布

#### 3 常见的概率分布

两点分布 二项分布 **泊松分布** 超几何分布

4 大数定律与中心极限定理

YN3010180007 51 / 79

3.3.3 泊松分布

假设经过某路口的汽车数量为 n。根据历史数据,该路口发生事故的年平均次数为  $\lambda$ ,那么单辆汽车发生事故的概率可表示为  $p=\frac{\lambda}{n}$ 。

代入二项分布的概率公式 3.25 可得

$$P(X=x) = C_n^x (\frac{\lambda}{n})^x (1 - \frac{\lambda}{n})^{n-x}, \quad x = 0, 1, 2, \dots, n$$
 (3.27)

YN3010180007 52 / 79

3.3.3 泊松分布

假设经过某路口的汽车数量为 n。根据历史数据,该路口发生事故的年平均次数为  $\lambda$ ,那么单辆汽车发生事故的概率可表示为  $p=\frac{\lambda}{n}$ 。

代入二项分布的概率公式 3.25 可得

$$P(X=x) = C_n^x (\frac{\lambda}{n})^x (1 - \frac{\lambda}{n})^{n-x}, \quad x = 0, 1, 2, \dots, n$$
 (3.27)

 $\downarrow \downarrow$ 

$$f(x;\lambda) = P(X=x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots, n$$
 (3.36)

YN3010180007 52 / 79

3.3.3 泊松分布

泊松分布 (Poisson distribution),记作  $P(\lambda)$ ,有一个参数  $\lambda$ 。 服从泊松分布的随机变量 X 的数学期望和方差同为  $\lambda$ 。

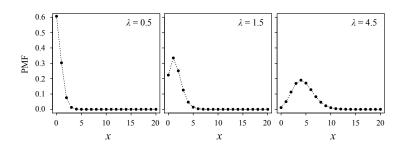


图 3.7 不同  $\lambda$  值的泊松分布

YN3010180007 53 / 79

3.3.3 泊松分布



图 3.8 长度为 1000 bp 的 DNA 序列 (箭头表示突变)

YN3010180007 54 / 79

3.3.3 泊松分布

#### 例 (3.5)

某水稻品种的田间自然变异概率为 0.0056, 试计算:

- (1) 调查 100 株, 获得 2 株或 2 株以上变异植株的概率是多少?
- (2) 期望有 0.95 的概率获得 1 株或 1 株以上的变异植株,至少应调查多少株?

### 解 (问题 1)

$$P(x \ge 2) = 1 - e^{-0.56} \frac{0.56^{0}}{0!} - e^{-0.56} \frac{0.56^{1}}{1!}$$
$$= 1 - 0.5712091 - 0.3198771$$
$$\approx 0.109$$

YN3010180007 55 / 79

3.3.3 泊松分布

#### 定理 (3.3)

在伯努利试验中,如事件 A 的概率 p 与试验总次数 n 有关,且当  $n\to\infty$  时有  $np\to\lambda$ ,也就是 n 很大而 p 很小时,泊松分布可近似代替二项分布,即:

$$C_n^k p^k (1-p)^{n-k} \approx e^{-\lambda} \frac{\lambda^k}{k!}$$
 (1)

由法国数学家 Simeon-Denis Poisson 于 1837 年提出。

YN3010180007 56 / 79

- 概率基础
- 2 随机变量及其概率分布

#### 3 常见的概率分布

两点分布 二项分布 泊松分布 超几何分布

正态分布

4 大数定律与中心极限定理

YN3010180007 57 / 79

3.3.4 超几何分布

一个个体数量为 N 的总体,假如其中的个体可以根据某种性状分为两类(比如发病动物和不发病动物),其中一类性状(如发病动物)的个体数量为 M。现在从该总体中随机抽取 n 个作为样本,试问抽中某类性状个体(如发病动物)的个数为 x 的概率是多少?

YN3010180007 58 / 79

3.3.4 超几何分布

根据概率的古典定义, 抽中发病动物个数为x的概率为

$$f(x; n, M, N) = P(X = x) = \frac{C_M^x C_{N-M}^{n-x}}{C_N^n}, \quad \max\{0, M+n-N\} < x < \min\{M, n\}$$
(3.39)

其中, max 和 min 分别表示求最大和最小值。

超几何分布 (hypergeometric distribution),记作 H(M,N,n),有三个参数 M,N,n。

服从超几何分布的随机变量 X 的数学期望为  $E(X)=\frac{nM}{N}$ , 方差  $\mathrm{Var}(X)=\frac{nM}{N}(1-\frac{M}{N})\frac{N-n}{N-1}\, \circ$ 

YN3010180007 59 / 79

#### 3.3.4 超几何分布

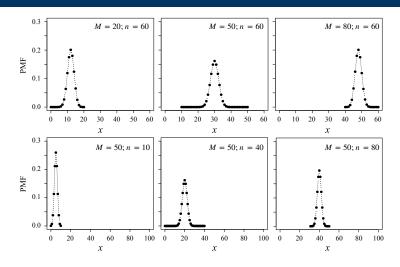


图 3.10 不同参数的超几何分布 (N = 100)

YN3010180007 60 / 79

- 概率基础
- 2 随机变量及其概率分布

#### 3 常见的概率分布

两点分布 二项分布 泊松分布 超几何分布 正态分布

4 大数定律与中心极限定理

YN3010180007 61 / 79

3.3.7 正态分布

设随机变量 X 服从数学期望为  $\mu$ , 方差为  $\sigma^2$  的正态分布,记作  $N(\mu,\sigma^2)$ 。

概率密度函数:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
(3.52)

累积分布函数:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
 (3.53)

其中, e 为自然常数, π 为圆周率。

YN3010180007 62 / 79

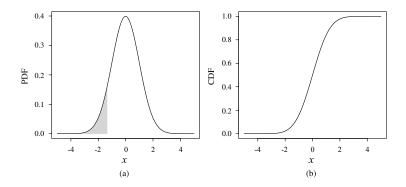


图 3.12 正态分布的概率密度函数 (a) 和累积分布函数 (b)

YN3010180007 63 / 79

3.3.7 正态分布

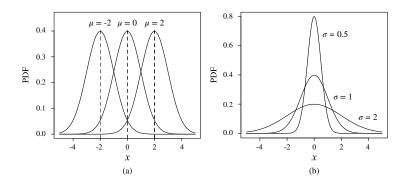


图 3.13 不同数学期望 (a) 和方差 (b) 的正态分布

YN3010180007 64 / 79

#### 正态分布具有以下几个重要性质:

- ① 密度函数 f(x) 为非负函数,以 x 轴为渐近线,分布从  $-\infty$  至  $+\infty$ ;
- ② 密度曲线是关于平均数 μ 对称的钟形曲线;
- ③ 密度函数在  $x = \mu$  处达到极大值  $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$ ;
- ④ 密度曲线在  $x = \mu \pm \sigma$  处各有一个拐点 (knee point), 通过拐点时曲线改变方向;
- $\mathbf{5}$  概率的归一性决定了密度曲线与x 轴所夹的面积为1;
- 6 密度曲线的位置由平均数  $\mu$  决定, 峰度由  $\sigma^2$  决定。

YN3010180007 65 / 79

3.3.7 正态分布 ...... 标准正态分布

平均数  $\mu$  和方差  $\sigma^2$  分别取值 0 和 1 时,该特定的正态分布称为标准正态分布 (standard normal distribution),有概率密度函数:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2}$$

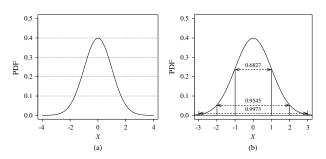


图 3.14 标准正态分布 (a) 及数据分布范围 (b)

YN3010180007 66 / 79

# 3.3 常见的概率分布

3.3.7 正态分布 ...... 标准正态分布

根据正态分布的特征,任意一个正态分布的密度曲线都可以通过位置平移和横 向缩放转换成标准正态分布。

几何意义上的转换对应到代数上可表示为:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

这种变换称为正态分布的标准化 (standardization)。

随机变量 z 称为标准正态离差 (standard normal deviate),表示变量取值离开平均数  $\mu$  有几个标准差  $\sigma$ 。

YN3010180007 67 / 79

# 3.3 常见的概率分布

3.3.7 正态分布 ...... 标准正态分布

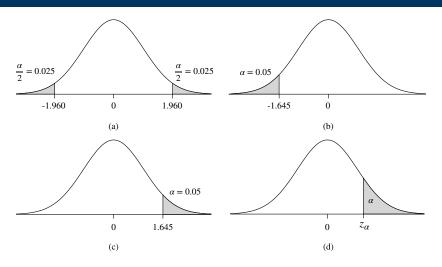


图 3.15 标准正态分布的上侧与下侧分位数

YN3010180007 68 / 79

#### 3.3 常见的概率分布

#### 

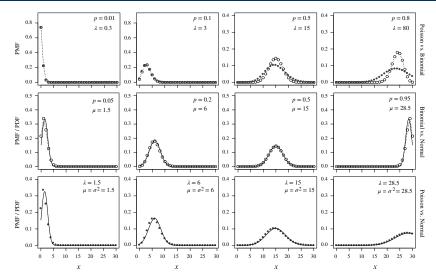


图 3.16 泊松分布 (实心点)、二项分布 (空心圆) 与正态分布 (实线) 的关系

YN3010180007 69 / 79

- 概率基础
- 2 随机变量及其概率分布
- ③ 常见的概率分布
- 4 大数定律与中心极限定理

大数定律

中心极限定理

YN3010180007 70 / 79

- 概率基础
- 2 随机变量及其概率分布
- ③ 常见的概率分布
- ◆ 大数定律与中心极限定理 大数定律

中心极限定理

YN3010180007 71 / 79

### 定理 (3.4)

设  $m \in n$  次独立试验中事件 A 出现的次数,而 p 是事件 A 在每次试验中出现的概率,则对于任意小的正数  $\epsilon$ ,有如下关系:

$$\lim_{n \to \infty} P(|\frac{m}{n} - p| < \epsilon) = 1 \tag{3.58}$$

YN3010180007 72 / 79

模拟一组投掷 10 次硬币的试验:

```
> rbinom(n = 1, size = 10, prob = 0.5)
[1] 4
```

YN3010180007 73 / 79

#### 模拟一组投掷 10 次硬币的试验:

```
> rbinom(n = 1, size = 10, prob = 0.5)
[1] 4
```

#### 模拟一组投掷 10000 次硬币的试验:

```
> rbinom(n = 1, size = 10000, prob = 0.5)
[1] 5071
```

YN3010180007 73 / 79

3.4.1 大数定律 ...... 辛钦大数定理

# 定理 (3.5)

设  $x_1, x_2, \dots, x_n$  是来自同一个平均数为  $\mu$  的总体,且相互独立的随机变量,对于任意小的正数  $\epsilon$ ,有如下关系:

$$\lim_{n \to \infty} P(|\frac{\sum_{i=1}^{n} x_i}{n} - \mu| < \epsilon) = 1$$

其含义是: 当 n 足够大时,随机变量的算术平均数无限接近于总体平均数的概率趋近于1,即样本平均数趋于总体平均数。

YN3010180007 74 / 79

3.4.1 大数定律 ...... 辛钦大数定理

模拟从正态分布中生成 10 个随机数作为一组样本,然后计算样本平均数。

```
> sample <- rnorm(n = 10, mean = 0, sd = 1)
> mean(sample)
[1] -0.01645363
```

YN3010180007 75 / 79

3.4.1 大数定律 ...... 辛钦大数定理

模拟从正态分布中生成 10 个随机数作为一组样本, 然后计算样本平均数。

```
> sample <- rnorm(n = 10, mean = 0, sd = 1)
> mean(sample)
[1] -0.01645363
```

将样本的观测值数增加到 10000, 重复模拟试验。

```
> sample <- rnorm(n = 10000, mean = 0, sd = 1)
> mean(sample)
[1] 0.003749129
```

YN3010180007 75 / 79

- 概率基础
- 2 随机变量及其概率分布
- ③ 常见的概率分布
- ◆ 大数定律与中心极限定理 大数定律

中心极限定理

YN3010180007 76 / 79

# 定理 (3.6)

在 n 重伯努利试验中,设事件 A 在每次试验中发生的概率为 p, X 为 n 次试验中事件 A 出现的次数,则  $\forall x \in \mathbf{R}$ ,有

$$\lim_{n \to \infty} P\left(\frac{X - np}{\sqrt{np(1 - p)}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{4}$$

其含义是: 当 n 足够大时,服从二项分布 B(n,p) 的随机变量 X 经标准化后近似服从标准正态分布 N(0,1),即  $X \sim N(np,np(1-p))$ 。

YN3010180007 77 / 79

# 定理 (3.7)

设  $\{X_i,\ i=1,\cdots,n\}$  是独立同分布随机变量序列,存在  $E(X_i)=\mu$ ,  ${\rm Var}(X_i)=\sigma^2$ ,且  $0<\sigma^2<\infty$ ,则有

$$\lim_{n \to \infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$
 (5)

其含义是: 当 n 足够大,独立同分布随机变量之和经标准化后近似服从标准正态分布 N(0,1),即  $\sum X_i \sim N(n\mu,\sigma^2n)$ 。

YN3010180007 78 / 79

# 本章小结

- 1 概率基础
- ② 随机变量及其概率分布

随机变量及其类型 离散型随机变量的概率分布 连续型随机变量的概率分布 随机变量的数字特征

3 常见的概率分布

两点分布 二项分布 泊松分布 超几何分布 正态分布

4 大数定律与中心极限定理

大数定律 中心极限定理