《生物实验设计》
第三章　概率和概率分布

王超

广东药科大学
Email: wangchao@gdpu.edu.cn

2022-09-04

第三章　概率与概率分布

目的和要求

为什么学习概率？

- 进行资料统计的目的不在于描述部分样本
- 而是通过样本统计数来推断数据总体的参数（统计推断）
- 统计推断的基础是：概率和概率分布

学习要求

- 掌握：事件、频率、概率的定义
- 熟悉：正态分布

（一）事件

在一定条件下，某种事物出现与否被称为是**事件**。

- 确定事件:
  - 必然事件 U: 在一定条件下必然出现的现象。
  - 不可能事件 V: 在一定条件下必然不出现的事件。
- 随机事件:
  - 有可能发生，也可能不发生。

（二）频率

- 在 n 次试验中，事件 A 出现的次数 m 称为事件 A 出现的频数，比值 $\frac{m}{n}$ 称为事件 A 出现的频率

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

为测定某批玉米种子的发芽率，分别取 10，20，50，100，200，500，1000 粒种子。在相同条件下进行发芽试验：

Table 1: 某批种子的发芽试验结果

种子总数	发芽种子总数	种子发芽率
10	9	0.900
20	19	0.950
50	47	0.940
100	91	0.910
200	186	0.930
500	459	0.918
1000	920	0.920

（三）概率

- 假设在相同的条件下，进行大量重复试验，若事件 A 的频率稳定地在某一确定值 p 的附近摆动，则称 p 为事件 A 出现的概率

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

不可能完全准确得到 p，在 n 充分大时，频率 $W(A)$ 作为 $P(A)$ 的近似值。

- 概率的基本性质：
  1. 任何事件的概率都在 0 和 1 之间 $0 \leqslant P(A) \leqslant 1$;
  2. 必然事件的概率等于 1 $P(U) = 1$
  3. 不可能事件的概率等于 0 $P(V) = 0$

（一）事件的相互关系

- 和事件
  - 事件 A 和事件 B 至少有一件发生而构成的新事件，$A + B$
- 积事件
  - 事件 A 和事件 B 同时发生而构成的新事件，$A \cdot B$
- 互斥事件
  - 事件 A 和事件 B 不能同时发生，$A \cdot B = V$
- 对立事件
  - 事件 A 和事件 B 必有一个事件发生，但二者不能同时发生，$A \cdot B = V, A + B = U, \bar{A} = B, \bar{B} = A$
  - 新生儿要么为男孩，要么为女孩

（一）事件的相互关系

- 独立事件
  - 事件 A 的发生与事件 B 的发生毫无关系
  - 独立事件群：多个事件 $A_1, A_2, A_3, \cdots, A_n$ 彼此独立
- 完全事件系
  - 多个事件 $A_1, A_2, A_3, \cdots, A_n$ 两两相斥，且每次试验结果必然发生其一

（二）概率计算法则

- **加法定理**

  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，称为**加法定理**。

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

  推理 1：如果 $A_1, A_2, A_3, \cdots, A_n$ 为 n 个互斥事件，则其和事件的概率为

  $P(A_1+A_2+A_3+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$

  推理 2：对立事件 $\bar{A}$ 的概率为 $P(\bar{A})=1-P(A)$

  推理 3：完全事件系和事件的概率等于 1

（二）概率计算法则

- **乘法定理**
  如果事件 A 和事件 B 为独立事件，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，称为**乘法定理**。

$$P(A \cdot B) = P(A) \cdot P(B)$$

推理：如果 $A_1, A_2, A_3, \cdots, A_n$ 彼此独立，则
$P(A_1 \cdot A_2 \cdot \ldots \cdot A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$

- 研究随机变量主要是研究变量的取值范围，也就是取值的概率
- 随机变量的概率分布：随机变量的取值与取这些值的概率之间的对应关系
- 随机变量的概率分布可以用分布函数表述
- 离散型变量的概率分布
  - 二项分布
  - 泊松分布
- 连续型变量的概率分布
  - 正态分布

（一）离散型随机变量的概率分布

【复习】离散型变量/非连续变量：在变量数列中仅能取得固定数值，并且通常是整数

- 离散型随机变量 x 所有可能的取值为 $x = x_i (i = 1, 2, \cdots, n)$
- 对于任意一个 x_i，都有一个相应的概率为 $p_i (i = 1, 2, \cdots, n)$

可以用下式表示为，

$$P(x = x_i) = p_i \quad (i = 1, 2, \cdots, n)$$

- x_i 与 p_i 为数值，表示事件“变量 x 取值为 x_i 时”的概率等于 p_i

并且，

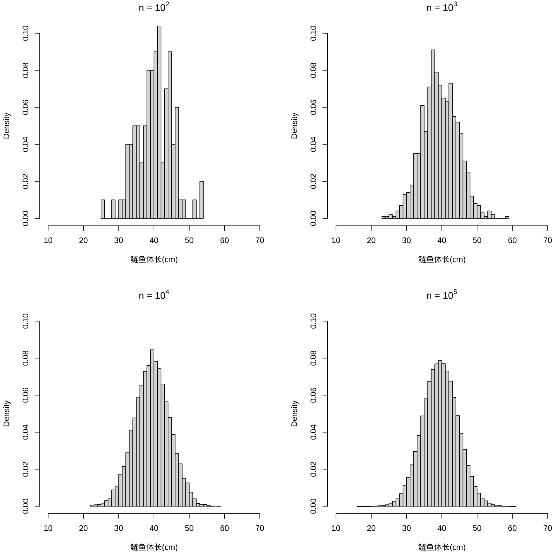
$$\sum_{i=1}^{n} p_i = 1$$

（二）连续型随机变量的概率分布

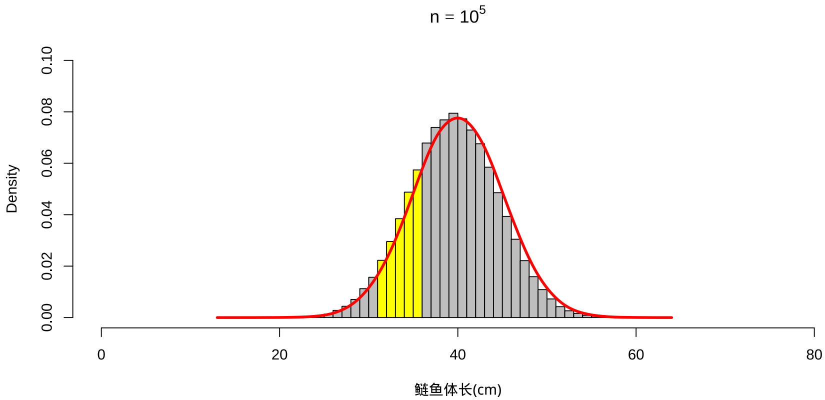
【复习】连续变量：在变量范围内可抽出某一范围内的所有值，变量之间是连续的、无限的

- 对于连续型随机变量，可以通过分组整理成次数分布表
- 如果从总体中抽取样本的容量 n 相当大，则频率分布就趋于稳定，近似地看成总体的概率分布
- 对连续型随机变量的次数分布表作直方图，直方图中同一间距内的频率密度是相等的
- 当 n 无限大，频率转化为概率，频率密度转化为概率密度，直方图逼近光滑连续曲线
    - 概率密度曲线（曲线下的总面积为 1）
    - 概率密度函数 $f(x)$

（二）连续型随机变量的概率分布



（二）连续型随机变量的概率分布



对于一个连续型变量 x，取值于区间内的概率即黄色阴影部分的面积，也就是概率密度函数 $f(x)$ 的积分，即

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

- 事件 A 发生的频率 $W(A)$ 和概率 $P(A)$ 之间的关系，实际上就是样本统计数和总体参数的关系。
- 当 n 足够大的时候，为什么可以用样本中的 $W(A)$ 代替?
- 大数定律是阐述大量随机现象平均结果稳定性的一系列定律的总称。

伯努利大数定律：m 是 n 次独立试验中事件 A 出现的次数，p 是事件 A 在每次试验中出现的概率，对于任意小的正数 ϵ，有：

$$\lim_{n\to\infty} P\left\{\left|\frac{m}{n}-p\right|<\epsilon\right\}=1$$

以上，P 为实现 $\left|\frac{m}{n}-p\right|<\epsilon$ 这一事件的概率，$P=1$ 是必然事件。

- 设一个随机变量 x_i，是由一个总体平均数 μ 和随机误差 ϵ_i 构成
  $x_i = \mu + \epsilon_i$
- 从总体中抽取 n 个随机变量构成一组样本，样本的平均数是

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(\mu + \epsilon_i) = \mu + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i$$

- 当样本容量 n 越来越大，$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i$ 就越小，使得 $\bar{x}$ 逼近 μ
- 样本容量越大，样本统计数与总体参数之差越小

随机变量的概率分布可以用分布函数来表述。

- 离散型变量的概率分布
  - 二项分布
  - 泊松分布
- 连续型变量的概率分布
  - 正态分布

- 二项分布是一种离散型随机变量的分布。
  - 每次试验只有两个对立结果，A 和 $\bar{A}$，出现的概率分别记为 p 和 q $(q=1-p)$。
  - 试验具有重复性和独立性。
    - 重复性：每次试验条件不变，在每次试验中事件 A 出现的概率都是 p
    - 独立性：任何一次试验中事件 A 的出现与其余各次试验中出现的任何结果无关。

$$P(x)=\mathrm{C}_n^x p^x q^{(n-x)}$$

二、泊松分布

三、正态分布

第三节　统计数的分布

二、样本平均数的分布

三、样本平均数差数的分布

四、t 分布

五、χ^2 分布

六、F 分布

$$F = \frac{s_1^2}{s_2^2}$$

三、概率的分布

一、资料的类型

- 数量性状资料
- 质量性状资料

二、资料的搜集

- 调查
- 试验

- 原始资料的检查与核对
- 频数分布表
- 频数分布图

频数分布表

100 只鸡每月产蛋数（用 rnorm 随机生成这样一组数据）

```
set.seed(2022)
egg <- round(rnorm(100, mean = 14, sd = 1.5))
egg
```

```
##   [1] 15 12 13 12 14 10 12 14 15 14 16 14 13 14 14 14 13 13 1
##  [26] 13 15 14 13 14 14 15 12 14 13 16 16 14 14 14 14 15 12 1
##  [51] 17 15 14 16 14 15 14 16 13 15 12 12 14 14 13 14 12 16 1
##  [76] 14 16 16 16 15 12 17 16 12 11 15 16 16 18 14 15 14 15 1
```

```
summary(egg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   13.00   14.00   14.25   15.00   18.00
```

利用 summary 可以大致了解数据的分布情况。

```
fdt_eg <- table(egg) # 次数统计
addmargins(fdt_eg)
```

```
## egg
##  10  11  12  13  14  15  16  17  18 Sum
##   1   2  12  13  29  21  17   3   2 100
```

```
prop.table(fdt_eg) # 频率统计
```

```
## egg
##   10   11   12   13   14   15   16   17   18
## 0.01 0.02 0.12 0.13 0.29 0.21 0.17 0.03 0.02
```

```
addmargins(prop.table(fdt_eg))
```

```
## egg
##   10   11   12   13   14   15   16   17   18  Sum
## 0.01 0.02 0.12 0.13 0.29 0.21 0.17 0.03 0.02 1.00
```

分组统计

300 个麦穗的每穗穗粒数

```
set.seed(2022)
wheat <- round(rnorm(300, mean = 40, sd = 7))
wheat[1:100]
```

```
##   [1] 46 32 34 30 38 20 33 42 45 42 47 39 33 41 40 39 35 33 4
##  [26] 34 45 42 36 38 41 46 29 38 34 48 48 41 41 39 38 46 32 3
##  [51] 52 47 41 49 42 44 40 48 37 43 31 31 38 41 34 40 32 48 5
##  [76] 39 47 50 50 46 33 53 48 32 25 44 49 50 59 40 45 42 43 3
```

```
summary(wheat)
```

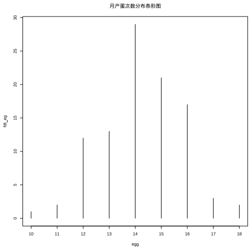
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   34.00   40.00   39.69   45.00   60.00
```

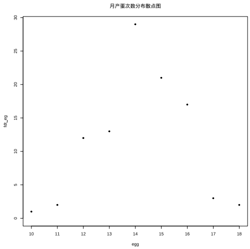
```
fdt_wt <- table(cut(wheat, breaks = seq(15, 60, 5), include.lowe
addmargins(fdt_wt)

##
## [15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,5
##       1       3      27      57      68      77      53

prop.table(fdt_wt)  # 频率统计

##
##     [15,20]     (20,25]     (25,30]     (30,35]     (35,40]
## 0.003333333 0.010000000 0.090000000 0.190000000 0.226666667 0
##     (45,50]     (50,55]     (55,60]
## 0.176666667 0.033333333 0.013333333

addmargins(prop.table(fdt_wt))

##
##     [15,20]     (20,25]     (25,30]     (30,35]     (35,40]
## 0.003333333 0.010000000 0.090000000 0.190000000 0.226666667 0
##     (45,50]     (50,55]     (55,60]         Sum
## 0.176666667 0.033333333 0.013333333 1.000000000
```

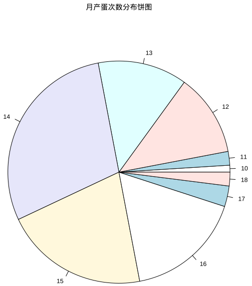
计量资料的整理

频数分布图

```
plot(fdt_eg, type = "h", main = " 月产蛋次数分布条形图")
plot(fdt_eg, type = "p", main = " 月产蛋次数分布散点图", pch = 20)
pie(prop.table(fdt_eg), main = " 月产蛋次数分布饼图")
```

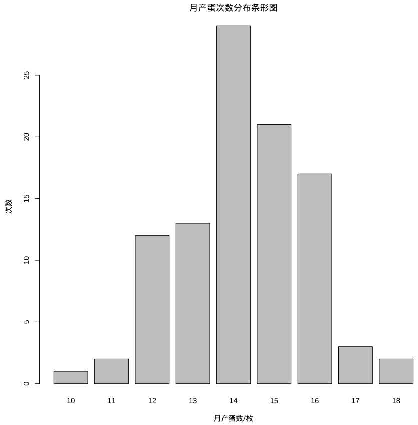






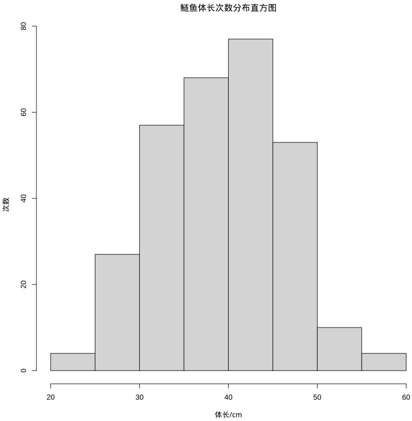
plot 之外的方法

```
barplot(fdt_eg, main = " 月产蛋次数分布条形图",
        ylab = " 次数", xlab = " 月产蛋数/枚")
```



直方图

```
hist(wheat, main=" 鲢鱼体长次数分布直方图", ylab=" 次数", xlab=" 体
```



一、平均数

- 算数平均数:

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 中位数:

$$M_d = x_{\frac{n+1}{2}}$$

或者

$$M_d = (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})/2$$

- 众数: M_o
- 几何平均数:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

算数平均数计算方法

1. 直接及算法 2. 加减常数法