

《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

2022-08-25

第三章 概率与概率分布

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

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

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
10	9
20	19
50	47
100	91
200	186
500	459
1000	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

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

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 $0 \leq P(A) \leq 1$;
- 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
- 积事件
事件 A 和事件 B 同时发生而构成的新事件
- 互斥事件
事件 A 和事件 B 不能同时发生
- 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
- 独立事件
事件 A 的发生与事件 B 的发生毫无关系
- 完全事件系
多个事件 $A_1, A_2, A_3, \dots, A_n$ 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**。

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

- 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**。

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

(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 $x = x_i (i = 1, 2, \dots, n)$
- 对于任意一个 x_i ，都有一个相应的概率为 $p_i (i = 1, 2, \dots, n)$

可以用下式表示为，

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

- 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

第一节 概率基础知识

三、概率分布

- 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：

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

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。

第二节 几种常见的理论分布

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

第二节 几种常见的理论分布 一、二项分布

- 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。

二、泊松分布

三、正态分布

第三节 统计数的分布

一、抽样试验与无偏估计

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

三、资料的整理

- 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
```

R demo

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

```
##
```

```
## [15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55]
```

##	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	0.176666667	0.033333333	0.013333333
----	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------

```
## (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	0.176666667	0.033333333	0.013333333	1.000000000
----	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------

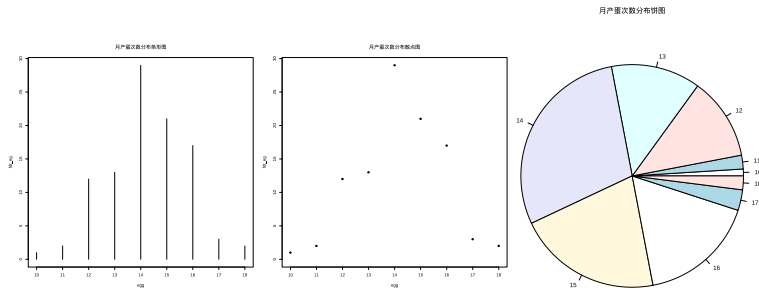
```
## (45,50] (50,55] (55,60] Sum
```

##	0.176666667	0.033333333	0.013333333	1.000000000					
----	-------------	-------------	-------------	-------------	--	--	--	--	--

计量资料的整理

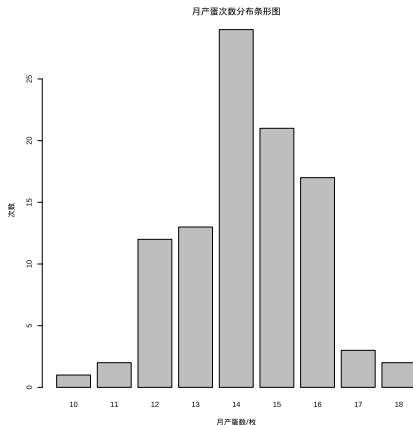
频数分布图

```
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")
```



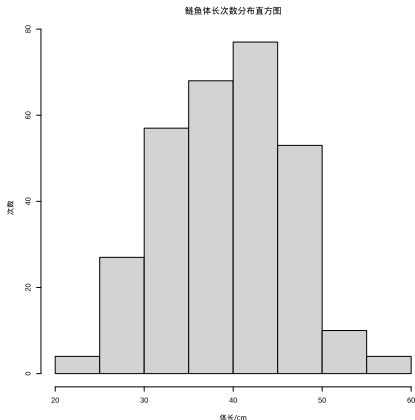
plot 之外的方法

```
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
```



第二节 资料特征数的计算

一、平均数

- 算数平均数:

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

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

- 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:

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

算数平均数计算方法

1. 直接及算法
2. 加减常数法