

《生物实验设计》

第二章 资料整理与特征数计算

王超

广东药科大学

Email: wangchao@gdpu.edu.cn

2022-08-10

第二章 资料整理与特征数计算

第一节 资料的搜集与整理

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  
##      (45,50]      (50,55]      (55,60]  
## 0.033333333 0.013333333 0.003333333
```

```
addmargins(prop.table(fdt_wt))
```

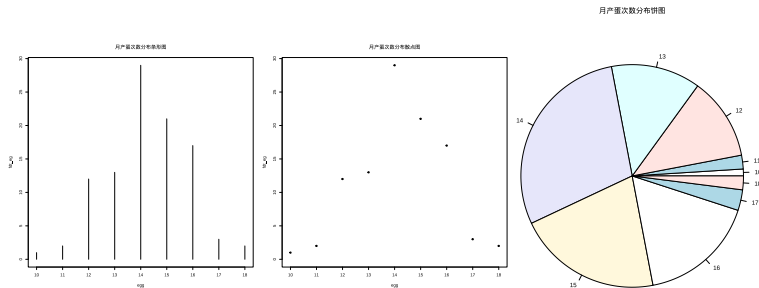
```
##
```

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

计量资料的整理

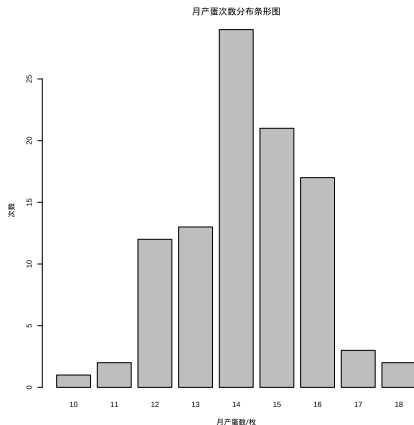
频数分布图

```
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")
```



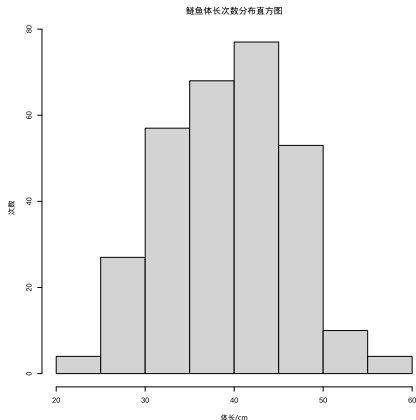
plot 之外的方法

```
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=" 体长/cm")
```



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:

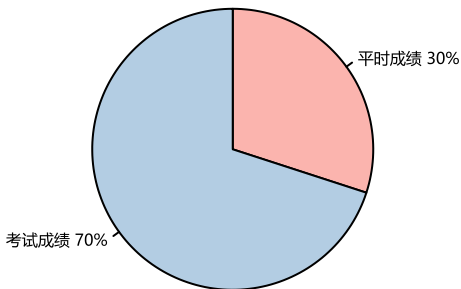
$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$$

算数平均数计算方法

1. 直接及算法
2. 加减常数法

课程考核

- 成绩评定
 - 平时成绩
 - 考试成绩
- 作业要求
 - 独立思考
 - 演算正确
 - 作图清楚
 - 书写整齐



学习重点

- 重点讲解统计方法在生物学中的应用；
- 了解公式的推导和证明；
- 及时完成作业，按时提交和反馈。

第一章 概论

第一节 生物统计学的概念

- 生物统计学是数理统计在生物学研究中的应用
- 用数理统计的原理和方法来分析和解释生物界各种现象和试验调查资料的科学
- 属于生物数学的范畴
 - 涉及到数列、排列、组合、矩阵、微积分等知识

为什么要学习统计学

第二节 统计学发展概况

- 统计实践随着计数活动开始（原始社会）
- 上升到理论成为系统的统计学（17 世纪英国）
 - 政治算数：Political Arithmetick, 1690, W. Petty.
 - 该书分为两部分：英法荷三国国力比较，英国国情国力和增长分析
- 发展经历三个阶段
 - 古典记录统计学
 - 近代描述统计学
 - 现代推断统计学

一、古典记录统计学

二、近代描述统计学

三、现代推断统计学

第三节 常用统计学术语

- 总体与样本
- 参数与统计数
- 变量与资料
- 因素与水平

一、总体与样本

二、参数与统计数

三、变量与资料

四、因素与水平

五、处理与重复

六、效应与互作

七、因素与水平

八、误差和错误

- 误差：也称为试验误差，是指观测值偏离真值的差异，分为随机误差和系统误差。
 - 随机误差：由于试验中许多无法控制的偶然因素所造成的试验结果与真实值之间的差异，是不可避免的。
 - 系统误差：由于试验处理以外的其他条件明显不一致所产生的带有倾向性的或定向性的偏差。
- 错误：在实验过程中，人为因素引起的差错。

第四节 生物统计学的内容与作用