《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

2022-11-25

第二章　资料整理与特征数计算

- 在试验及调查中能够获得大量的原始数据，这是在一定条件下对某种具体事物或现象观察的结果，称之为**资料**。
- 统计分析就是对资料的整理分析，列出统计表，绘出统计图，计算特征数。

对资料的分类整理，必须坚持**同质**的原则。

只有同质的试验数据，才能根据科学原理来分类，使资料能正确反映事物的本质和规律。

- 数量性状资料（定量）
- 质量性状资料（定性）

（一）数量性状资料

数量性状资料一般由计数和测量或度量得到的。

- 计数资料：由计数法得到的数据（非连续变量）。
- 计量资料：由测量或度量所得的数据（连续变量）
  - 依试验的要求和测量仪器或工具的精度而定。

（二）质量性状资料

对某种现象只能观察而不能测量的资料。

一般需要先把质量性状资料数量化，可以采用：

1. 统计次数法
   - 根据某一质量性状的类别统计其次数或频数，以次数和频数作为该质量性状的数据。
2. 评分法
   - 评分是用数字级别表示某现象在表现程度上的差别。
   - 根据评分将质量性状资料进行量化，然后参照计数资料的处理方法进行。

（一）调查

1. 全面调查/普查

- 对研究对象的每一个个体逐一进行调查
- 范围广、时间长、工作量大
- 极少数情况

2. 抽样调查

- 根据一定的原则对研究对象抽取一部分个体进行测量或度量
- 把得到的数据资料作为样本进行统计处理，然后利用样本特征数对总体参数进行推断
- 要想无偏差估计总体，重要的是采用科学的抽样方法

抽样调查

1. 随机抽样

在试验过程中对试验单位的抽样、分组等都必须遵守随机原则，避免人为主观因素的影响

2. 顺序抽样

- 按照既定顺序从总体中抽取一定数量的个体构成样本

3. 典型抽样

- 根据初步资料或经验判断，有意识、有目的地选择一个典型群体作为代表进行调查，以估计整个总体

可以混合地才采用以上集中抽样方法。

随机抽样

随机抽样必须满足两个条件：

- 总体中每个个体被抽中的机会是均等的
- 总体中任意一个个体是否被抽中是相互独立的（不受其他个体影响）

随机抽样的方法：

1. 简单随机抽样
2. 分层随机抽样
3. 整体抽样
4. 双重抽样

（二）试验

通过一定数量有代表性的试验单位，在一定的条件下进行的有探索性的研究工作。

试验处理原则：

1. 随机
2. 重复
3. 局部控制

试验设计方法在第九章介绍。

- 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。

R demo

```
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 47 46 43 43 48 48 38
##  [26] 34 45 42 36 38 41 46 29 38 34 48 48 41 41 39 38 46 32 30 40 34 42 38 31 43
##  [51] 52 47 41 49 42 44 40 48 37 43 31 31 38 41 34 40 32 48 56 43 39 49 43 40 49
##  [76] 39 47 50 50 46 33 53 48 32 25 44 49 50 59 40 45 42 43 37 60 36 27 45 42 43
```

```
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.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

```
set.seed(2022)
fish <- round(rnorm(150, mean = 55, sd = 9))
fish[1:100]
```

```
##   [1] 63 44 47 42 52 29 45 58 62 57 64 53 46 56 55 54 49 46 64 63 58 58 65 66 52
##  [26] 47 61 58 50 53 56 62 41 53 48 65 65 56 56 53 53 62 45 42 56 48 58 53 43 58
##  [51] 70 64 57 66 58 61 55 66 51 59 43 43 52 56 47 55 45 65 76 59 54 67 59 56 67
##  [76] 53 64 67 68 63 46 72 65 45 35 60 67 67 80 55 62 57 59 52 81 50 38 61 57 59
```

```
summary(fish)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   50.00   56.50   55.69   63.00   81.00
```

R demo

```
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
```

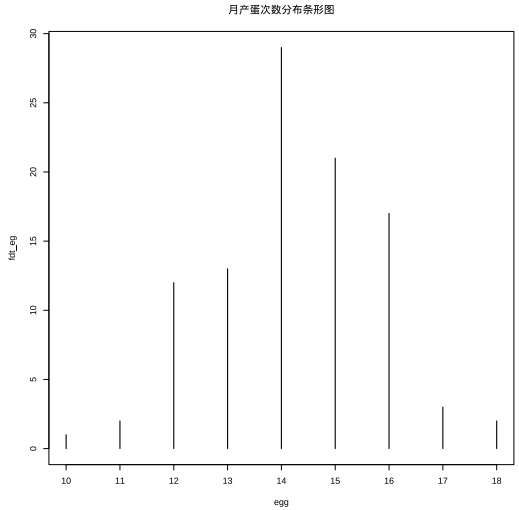
```
##
##     [15,20]     (20,25]     (25,30]     (30,35]     (35,40]
## 0.003333333 0.010000000 0.090000000 0.190000000 0.226666667 0
##     (45,50]     (50,55]     (55,60]
## 0.176666667 0.033333333 0.013333333
```

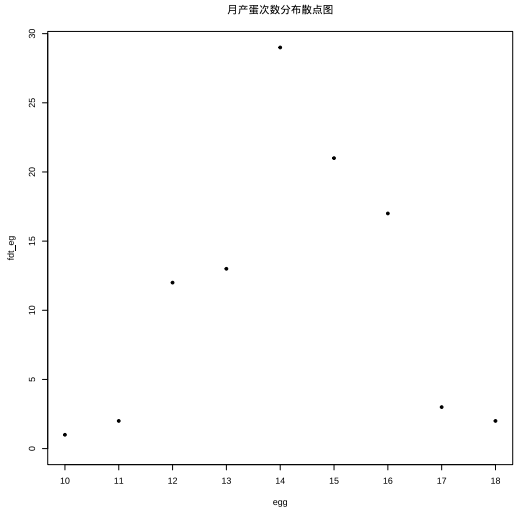
```
addmargins(prop.table(fdt_wt))
```

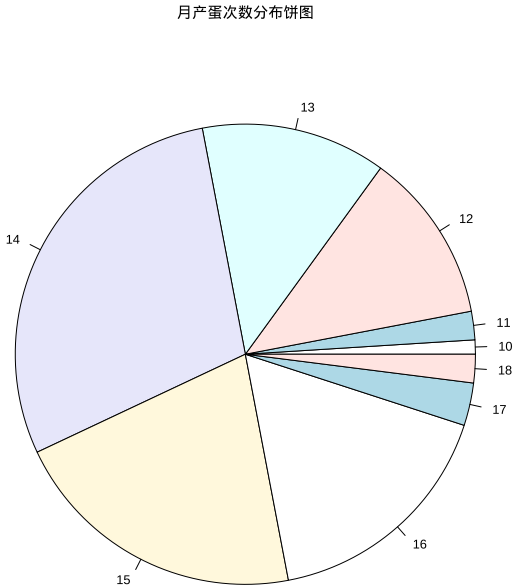
```
##
##     [15,20]     (20,25]     (25,30]     (30,35]     (35,40]
## 0.003333333 0.010000000 0.090000000 0.190000000 0.226666667 0
##     (45,50]     (50,55]     (55,60]         Sum
## 0.176666667 0.033333333 0.013333333 1.000000000
```

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")
```

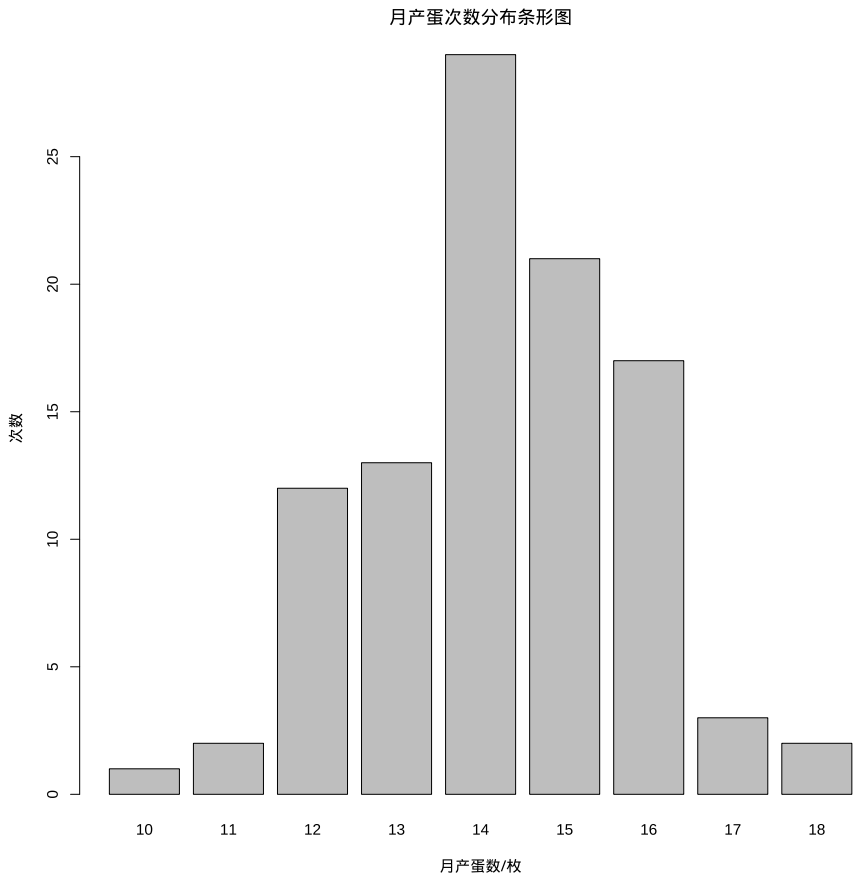






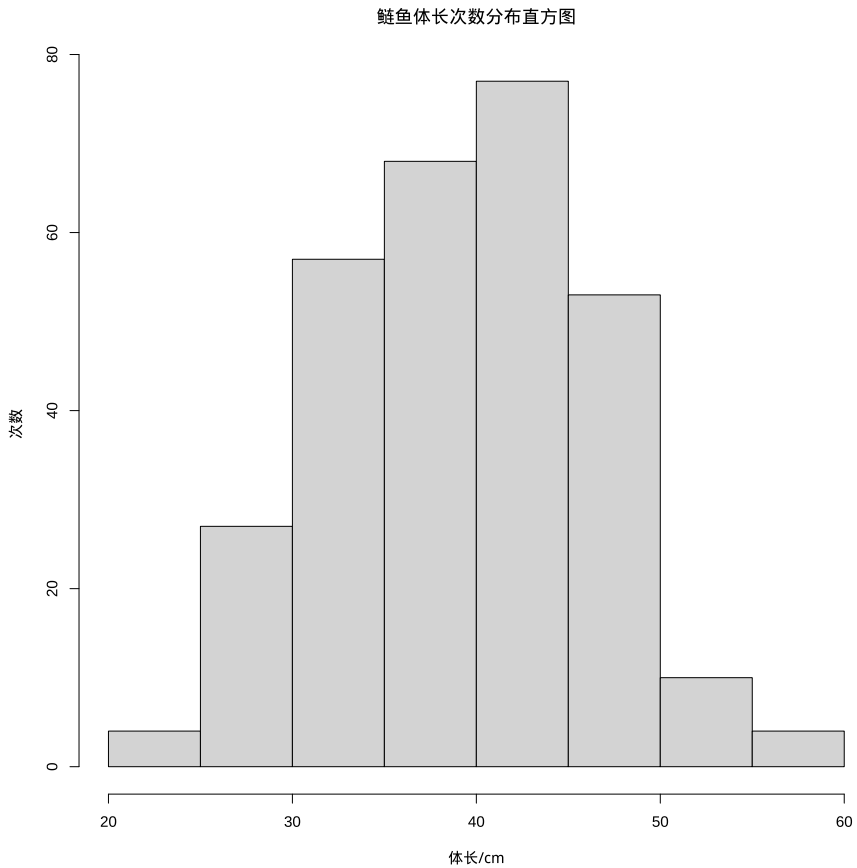
plot 之外的方法

```
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：

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

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

- 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

（二）方差

- 对 n 个观测值 $x_1, x_2, x_3, \cdots, x_4$ 的样本，可以用各观测值的离均差来表示
  - 离均差 $\sum(x-\bar{x})$
  - $\sum(x-\bar{x})=0$
- 先平方在求和
  - 离均差平方和 $\sum(x-\bar{x})^2$
  - 跟样本容量大小有关
- 用样本容量 n 来除以离均差平方和 $\sum(x-\bar{x})^2$
  - 方差/均方：$s^2=\frac{\sum(x-\bar{x})^2}{n-1}$
  - $n-1$ 为自由度

（三）标准差

- 由于离均差取平方，与原始数据的数值和单位不适应，需要开方还原
- 方差的平方根值就是标准差 $s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
- 标准差的计算：

因为 $\bar{x}=\frac{\sum x}{n}$

$$\begin{aligned}\sum(x-\bar{x})^2 &= \sum(x^2-2x\bar{x}+\bar{x}^2)\\ &= \sum x^2-2\bar{x}\sum x+n\bar{x}^2\\ &= \sum x^2-\frac{(\sum x)^2}{n}\end{aligned}$$

所以

$$s=\sqrt{\frac{\sum x^2-\frac{(\sum x)^2}{n}}{n-1}}$$

（三）标准差

- 标准差的特性
  - 标准差的大小受到多个观测值的影响
  - 如果将各个观测值加上或者减去一个常数 α，标准差不变；如果将各个观测值乘以或者除以一个常数 α，标准差扩大或缩小 α 倍；
  - 在正态分布下，样本变量的分布可以做出正态估计
- 标准差的作用
  - 表示变量分布的离散程度
  - 利用标准差可以估计各类观测值在总体中所占的比例
  - 估计平均数的标准误
  - 进行平均数的区间估计和变异系数计算

标准差（SD，Standard Deviation）*vs.* 标准误（SEM，Standard Error of the Mean）

- 标准差代表样本的单个观测值与均值之间的变化度或者离散度
- 标准误衡量样本的平均值可能与真实的总体平均值有多少的差距，标准误总是比标准差小
  - 标准误的含义包括基于抽样分布的统计推断
  - 标准误是样本平均值的抽样分布的标准差
- 计算方法
  - 标准差 $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
  - 标准误 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

“Mean ± SEM” or “Mean (SD)”?

Jaykaran

Use of descriptive statistics is very common in articles published in various medical journals. For the ratio and interval data following the normal distribution, the most common descriptive statistics is mean and standard deviation (SD) and for data not following the normal distribution, it is median and range. It is, however, observed in various medical journals that mean and standard error of mean (SEM) are used to describe the variability within the sample.[1] We, therefore, need to understand the difference between SEM and SD.

The SEM is a measure of precision for an estimated population mean. SD is a measure of data variability around mean of a sample of population. Unlike SD, SEM is not a descriptive statistics and should not be used as such. However, many authors incorrectly use the SEM as a descriptive statistics to summarize the variability in their data because it is less than the SD, implying incorrectly that their measurements are more precise. The SEM is correctly used only to indicate the precision of estimated mean of population. Even then however, a 95% confidence interval should be preferred.[1,2] Further, while reporting mean and SD, instead of writing “mean ± SD” the better way of representation would be “mean (SD)” as it will decrease the chance of confusion with confidence interval.[2]

Indian J Pharmacol. 2010 Oct;42(5):329. doi: 10.4103/0253-7613.70402.