统计学与 R 语言

第 12 讲　随机抽样, 概率计算

张敬信

2022 年 4 月 2 日

哈尔滨商业大学

R 语言就是因统计分析而生的编程语言，可以很方便地完成各种统计计算、统计模拟、统计建模等。

统计学是关于数据的科学，是一套有关数据收集整理（获取及预处理数据)、描述统计（汇总、图表描述)、分析推断（选择适当的统计方法研究数据，并从数据中提取有用信息进而得出结论）的方法。

```
library(tidyverse)
```

一. 若干概念

1. **随机变量**

当一件事情的结果无法预料时，就叫随机现象。表示随机现象一组结果的变量就是**随机变量**。

比如说，调查了 100 个人的身高，这 100 个身高的数据是随机变量身高的数据。并不是说这些身高值是不固定可变的，而是这 100 个身高值是一次调查的结果，再调查 100 个人就是另一组不同的 100 个身高值。

2. 概率分布

随机变量既然是这样随机的，还有必要研究它吗？有必要！因为把多个随机结果放在一起的时候，能发现一定的规律性。比如 100 人的身高可能对称地分布在 175cm 附近，离得越远人数越少，即表现出一种正态分布规律性。

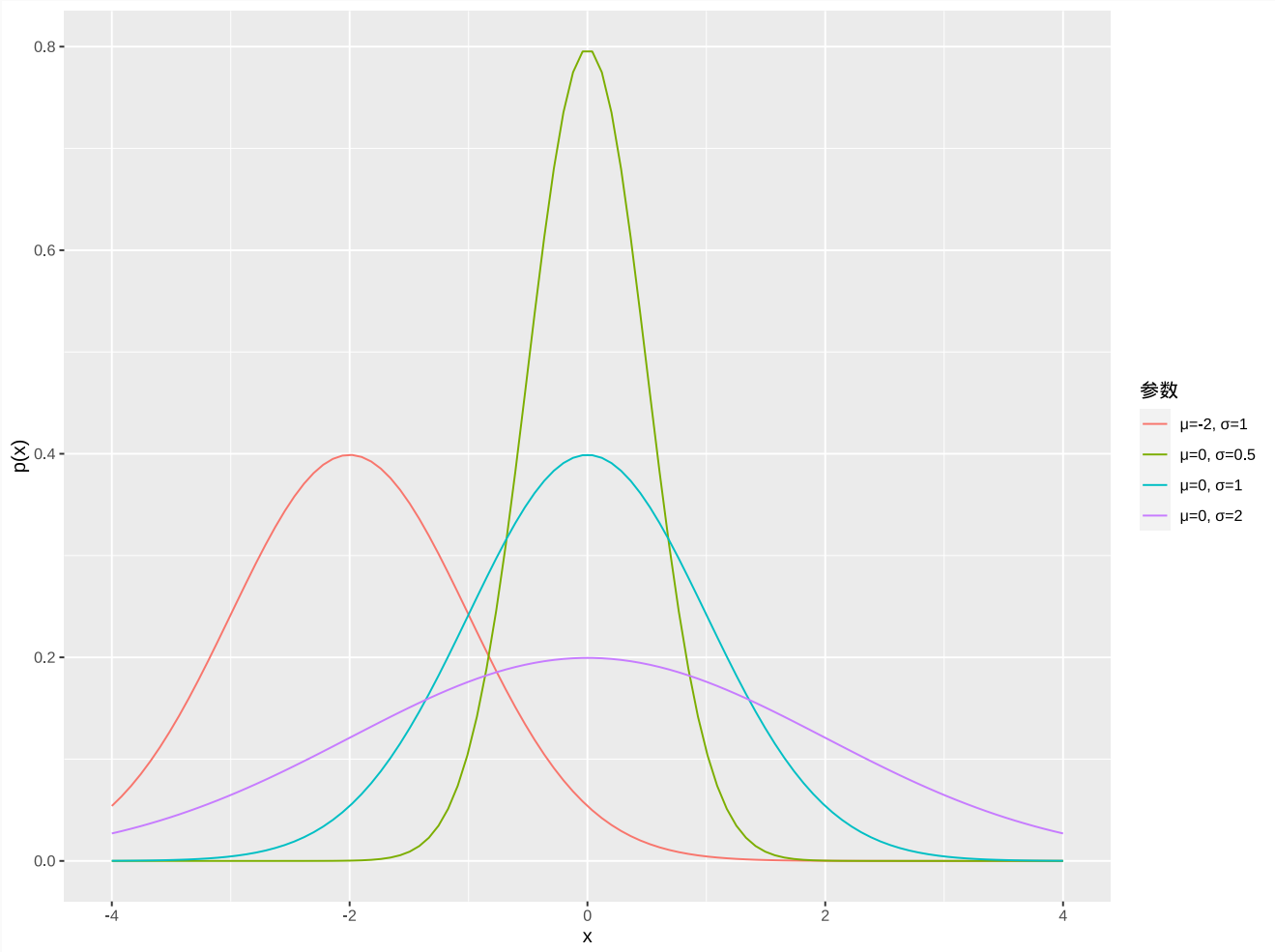
随机现象五花八门，但每一种随机现象表现出来的规律性是固定的，用数学语言表达出来就是**概率分布**。所以，不同概率分布就是不同随机现象规律性的数学描述。

同一种概率分布，也不是都相同，这是由不同参数值决定和区分的。

统计学最常用到四大概率分布：正态分布、t 分布、卡方分布、F 分布。

比如正态分布 $N(\mu, \sigma^2)$, μ 和 σ 就是参数，它们只要取不同值，就是不同的分布形状：

```
tibble(
  x = seq(-4,4,length.out = 100),
  `μ=0, σ=0.5` = dnorm(x, 0, 0.5),
  `μ=0, σ=1` = dnorm(x, 0, 1),
  `μ=0, σ=2` = dnorm(x, 0, 2),
  `μ=-2, σ=1` = dnorm(x, -2, 1)
) %>%
  pivot_longer(-x, names_to = " 参数",
               values_to = "p(x)") %>%
  ggplot(aes(x, `p(x)`, color = 参数)) +
  geom_line()
```


0.8
0.6
0.4
0.2
0.0
p(x)
-4
-2
0
2
4
x
参数
μ=-2, σ=1
μ=0, σ=0.5
μ=0, σ=1
μ=0, σ=2

3. 概率论与数理统计

概率论就是研究随机现象规律性，即各种概率分布及性质的理论。数理统计所研究的数据是带有随机性的，所以就需要借助概率论中的概率分布理论加以描述和做出统计推断。所以说：

概率论是数理统计的理论基础，数理统计是概率论的一种应用

4. 区分数据类型

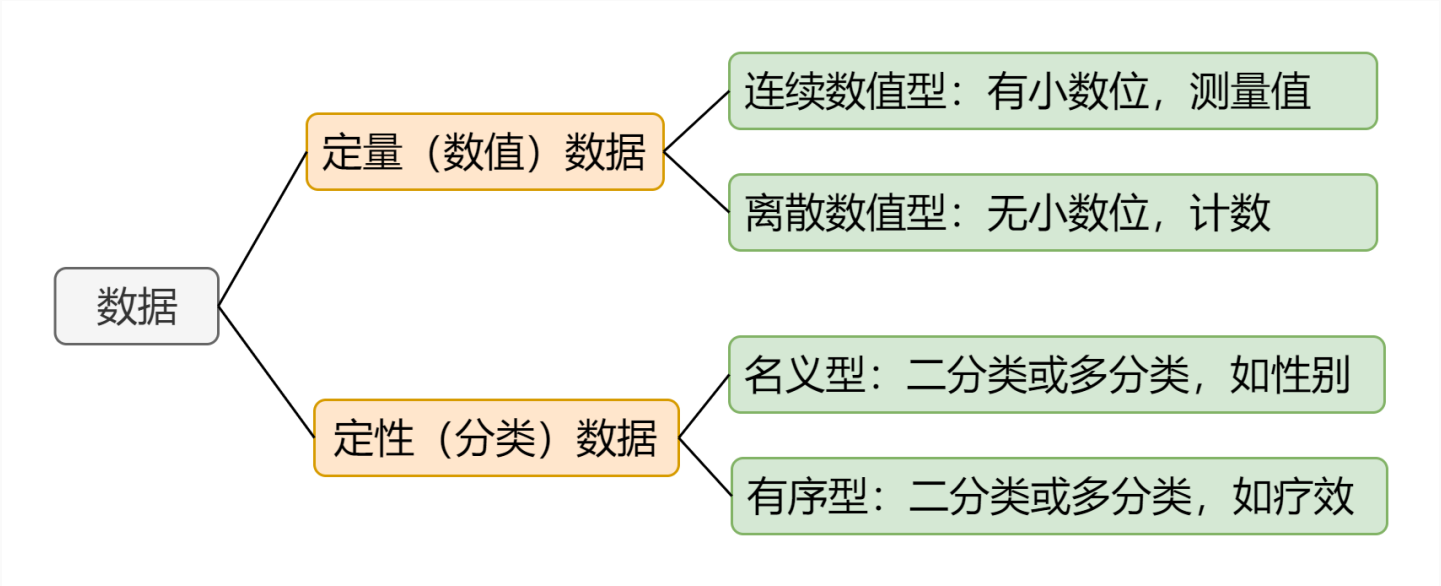


图 1: 常见的数据类型

区分数据类型非常有必要，因为不同数据类型适用的统计分析方法是不同的！

5. 总体和样本

- **总体** (population)：是包含所研究的全部个体（数据）的集合。
- **样本** (sample)：从总体中抽取的一部分个体的集合，样本包含个体的数目称为样本量。

抽样的目的是根据样本数据提供的信息推断总体的特征，或者说，用样本统计量推断总体参数。

比如，要研究哈尔滨市成年男性的身高，则所有哈尔滨市成年男性的身高数据就是总体，但实际上不可能把所有这些身高都测量一遍，只能是随机抽取一部分，比如 100 人，测得身高数据，这就是样本，样本量是 100。

抽样调查结果的可靠性不在于样本数量大不大（当然也不能太少），更主要的是科学抽样，使样本足够代表总体。

身高数据大致服从正态分布，所有哈尔滨市成年男性身高的均值 μ 和标准差 σ，就是总体参数。用样本的 100 人的平均身高作为 μ 的估计，就是用样本统计量推断总体参数。

6. 参数与统计量

- **参数** (parameter)：用来描述总体特征的概括性值，是研究者想要了解的总体的某种特征值，如总体均值 (μ)、总体方差 (σ^2)、总体比例 (π) 等。
- **统计量** (statistic)：是用来描述样本特征的概括性数字度量，是根据样本数据计算出来的量，由于抽样是随机的，因此统计量是样本的函数。与上面总体参数对应的统计量是样本均值 ($\bar{x}$)、样本标准差 (s^2)、样本比例 (p) 等。

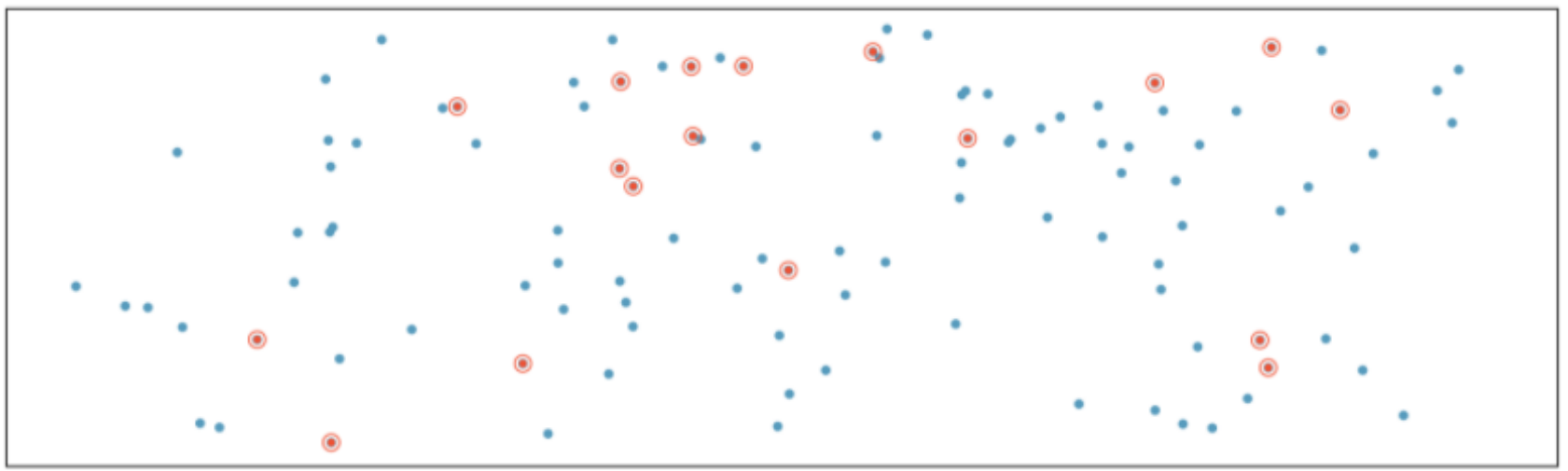
由于总体数据通常是不知道的，故参数是未知常数。所以才进行抽样，根据样本计算出相应统计量值去估计总体参数值。

二. 随机抽样

几乎所有的统计方法都是基于隐含随机性。如果数据不是在一个随机的框架内从总体中收集的，这些统计方法–估计值和与估计值相关的误差–就不可靠。

通常有四种随机抽样技术：**简单、分层、整群、多阶段**。

(1) 简单抽样（Simple sampling)

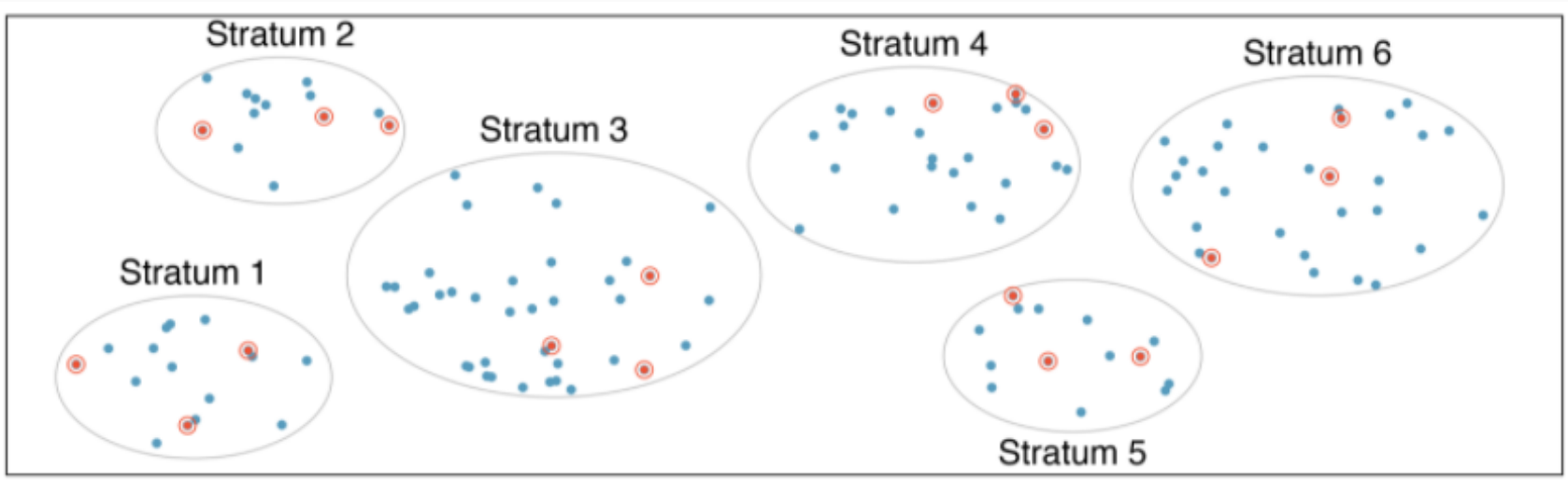


总体中的每个体都有相同的机会被抽到，并且个体被抽到是相互独立的。

```
(X = str_c(1:6, " 点"))
#> [1] "1 点" "2 点" "3 点" "4 点" "5 点" "6 点"
sample(X, 1)
#> [1] "1 点"
sample(X, 10, replace = TRUE)
#>  [1] "4 点" "2 点" "3 点" "5 点" "2 点" "6 点" "4 点" "5 点
sample(c(" 正"," 反"), 10, replace = TRUE)
#>  [1] " 反" " 正" " 正" " 正" " 正" " 正" " 反" " 反" " 反"
sample(c(" 正"," 反"), 10, replace = TRUE,
       prob = c(0.9, 0.1))
#>  [1] " 正" " 正" " 正" " 正" " 反" " 正" " 正" " 正" " 反"
```

```r
df = as_tibble(iris)
df %>%
  slice_sample(n = 5)
#> # A tibble: 5 x 5
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
#>          <dbl>       <dbl>        <dbl>       <dbl> <fct>
#> 1          5.6         2.5          3.9         1.1 versic
#> 2          6.2         3.4          5.4         2.3 virgin
#> 3          5.4         3.9          1.7         0.4 setosa
#> 4          7.2         3            5.8         1.6 virgin
#> 5          5           3.3          1.4         0.2 setosa
```

(2) **分层抽样**（Stratified sampling)

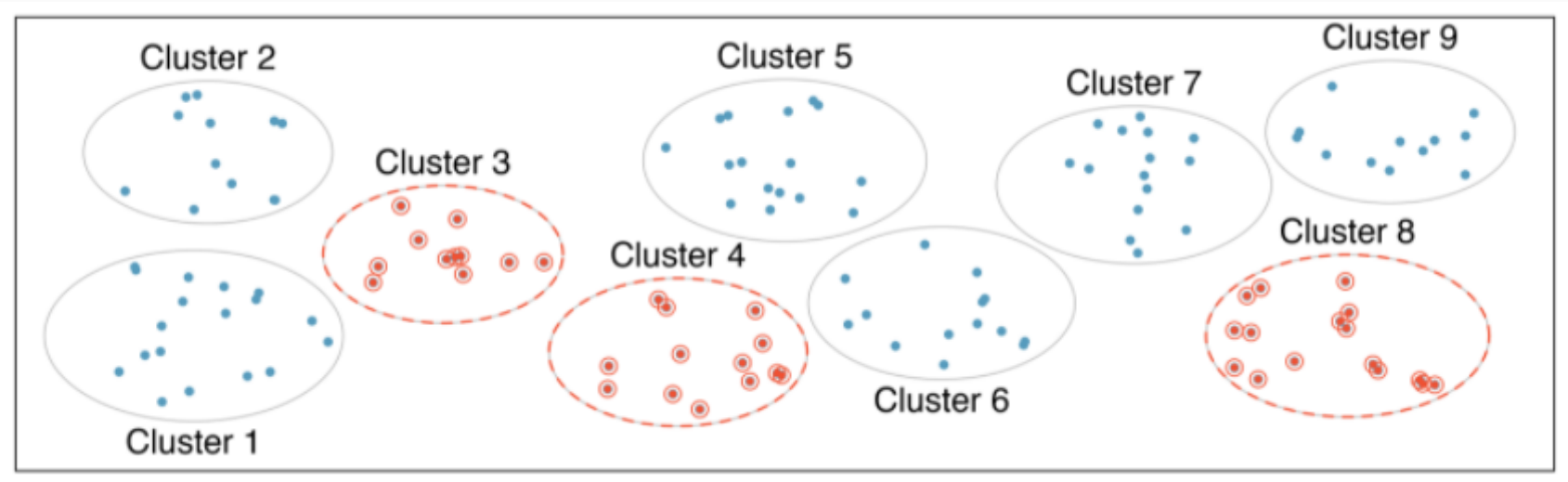


分层抽样是一种分而治之的抽样策略。将总体划分为若干组，以将类似的个体分组，然后在每个分组中采用简单随机抽样。

```
df %>%
  group_by(Species) %>%
  slice_sample(n = 5)
#> # A tibble: 15 x 5
#> # Groups:   Species [3]
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
#>          <dbl>       <dbl>        <dbl>       <dbl> <fct> 
#> 1          4.7         3.2          1.3         0.2 setosa
#> 2          4.4         3.2          1.3         0.2 setosa
#> 3          5.2         3.5          1.5         0.2 setosa
#> 4          5           3.3          1.4         0.2 setosa
#> 5          5.1         3.3          1.7         0.5 setosa
#> # ... with 10 more rows
```

```
df %>%
  group_by(Species) %>%
  slice_sample(prop = 0.1)
#> # A tibble: 15 x 5
#> # Groups:   Species [3]
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
#>          <dbl>       <dbl>        <dbl>       <dbl> <fct> 
#> 1          5           3.6          1.4         0.2 setosa
#> 2          4.3         3            1.1         0.1 setosa
#> 3          4.5         2.3          1.3         0.3 setosa
#> 4          5.4         3.4          1.5         0.4 setosa
#> 5          5.1         3.7          1.5         0.4 setosa
#> # ... with 10 more rows
```

(3) **整群抽样**（Cluster sampling)

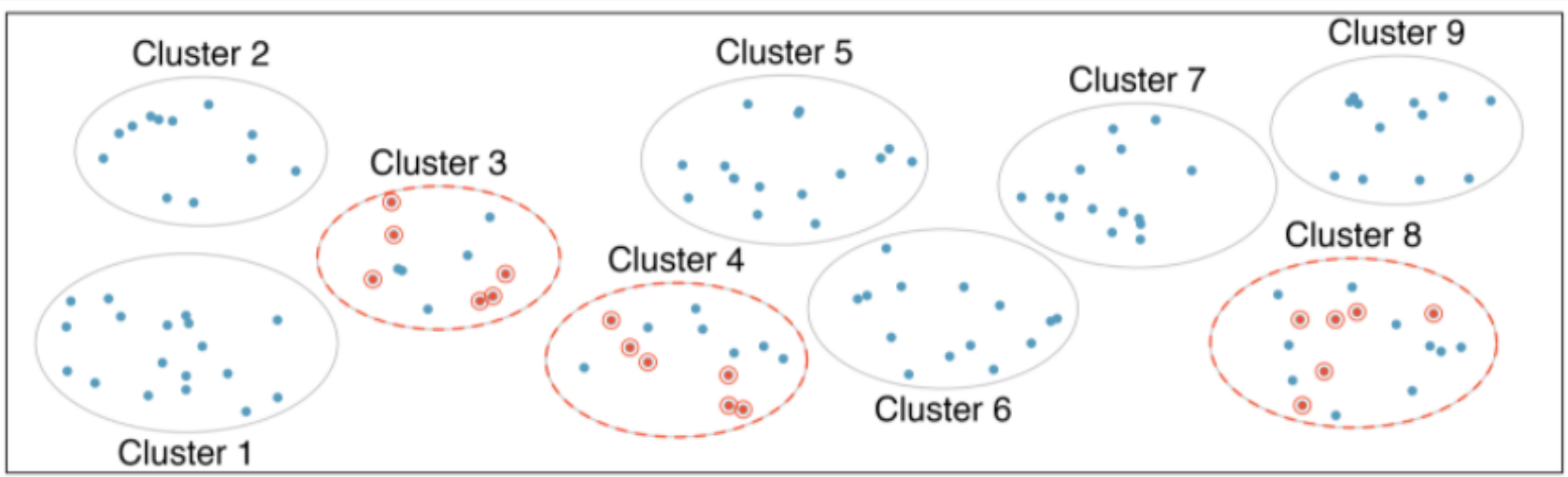


将总体分成许多群组，然后随机抽取一定数量的群组，并将每个群组的所有个体都纳入样本中。

```
df %>%
  filter(Species %in% sample(levels(Species), 2))
#> # A tibble: 100 x 5
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
#>          <dbl>       <dbl>        <dbl>       <dbl> <fct> 
#> 1            7         3.2          4.7         1.4 versic
#> 2          6.4         3.2          4.5         1.5 versic
#> 3          6.9         3.1          4.9         1.5 versic
#> 4          5.5         2.3          4           1.3 versic
#> 5          6.5         2.8          4.6         1.5 versic
#> # ... with 95 more rows
```

(4) 多阶段抽样（Multistage sampling）

多阶段抽样与整群抽样一样，但不是保留每个群组中的所有个体，而是在每个选定的群组中随机抽取若干个体。



```r
df %>%
  filter(Species %in% sample(levels(Species), 2)) %>%
  group_by(Species) %>%
  slice_sample(n = 3)
#> # A tibble: 6 x 5
#> # Groups:   Species [2]
#>   Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
#>          <dbl>       <dbl>        <dbl>       <dbl> <fct>
#> 1          4.5         2.3          1.3         0.3 setosa
#> 2          4.9         3.1          1.5         0.2 setosa
#> 3          5.4         3.9          1.7         0.4 setosa
#> 4          6.3         2.5          5           1.9 virgin
#> 5          6.3         3.3          6           2.5 virgin
#> # ... with 1 more row
```

三. 随机试验

研究者将治疗方法分配给病例的研究称为**试验**。当这种分配包括随机性时，例如，用抛硬币的方式来决定病人接受哪种治疗，就称为**随机试验**。

要研究两个变量之间的因果关系时，就需要随机试验。

(1) **控制（Controlling)**

研究者将治疗方法分配给病例，并尽力控制各组中的任何其他差异。例如，当病人吃药时，医生指示每个病人都同时喝 50 毫升的水。

(2) **随机化（Randomization)**

研究者将病人随机分为治疗组，以考虑到无法控制的变量。例如，由于饮食习惯的原因，一些病人可能比其他病人更容易感染某种疾病。本例中，饮食习惯是一个**混杂变量**，它被定义为与解释变量和响应变量都有关联的变量。将病人随机分为治疗组或对照组有助于平衡这种差异。

(3) **复制**（Replication）

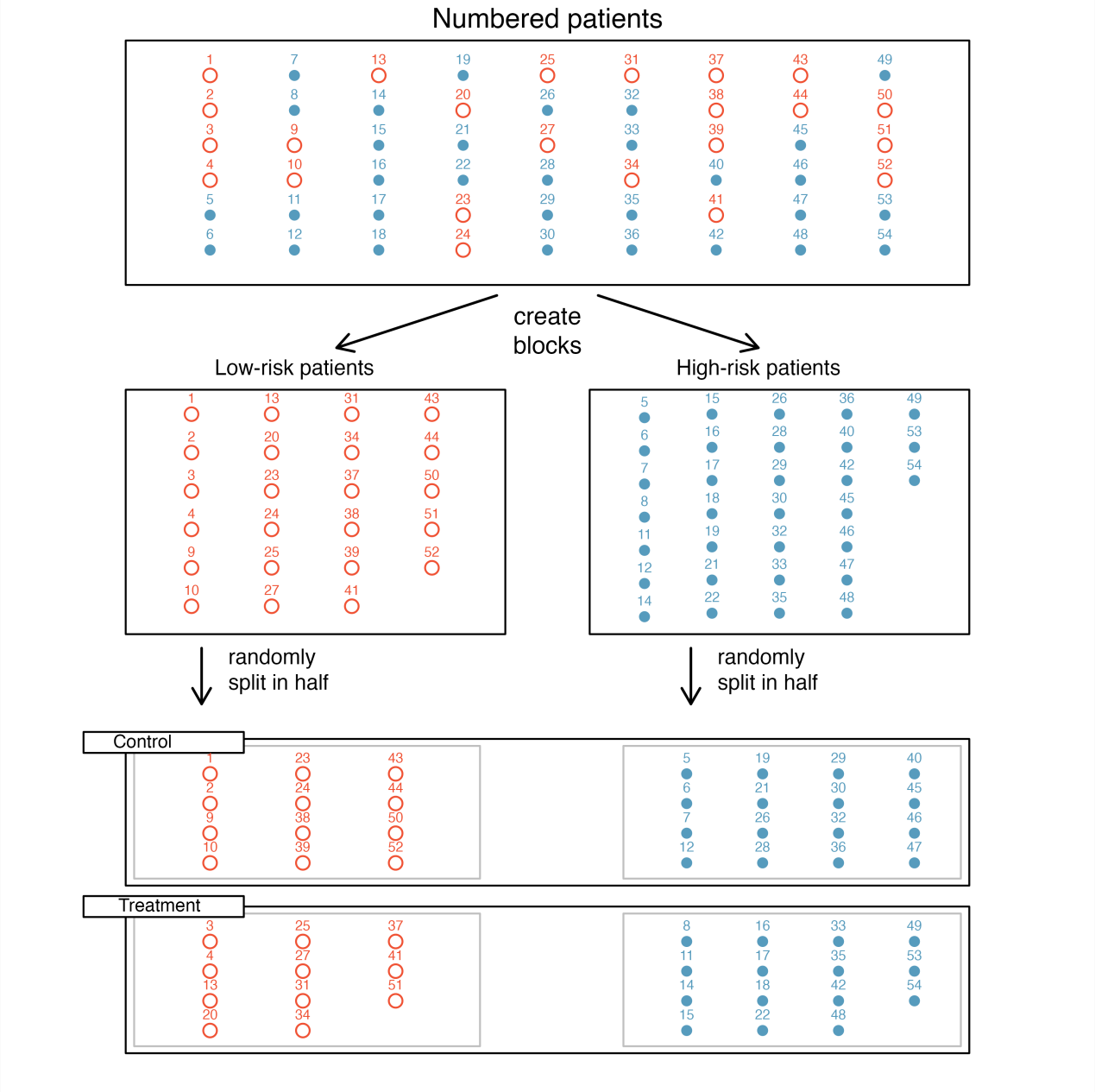
研究者观察的病例越多，就能更准确地估计解释变量对响应变量的影响。在一项研究中，通过收集足够大的样本进行复制。足够大的样本因试验而异，但至少希望每个治疗组有多个受试者。

实现复制的另一种方式是复制整个研究以验证早期的发现。当不同处理下的单个观测结果严重依赖彼此时，就会出现假性复制。例如，假设某个实验中有 50 名受试者，在整个研究过程中的 10 个时间点进行血压测量，则有 $50 \times 10 = 500$ 个测量值。报告有 500 个观测值将被认为是假性复制，因为一个特定个体的血压测量值并不是相互独立的。当错误的实体被复制时，假性复制经常发生，而且报告的样本量被夸大了。

(4) **分块（**Blocking**)**

研究者有时知道或怀疑除治疗外的其他变量会影响响应变量。在这种情况下，他们可能首先根据该变量将个体分成几个组，然后将每个组内的病例随机分配到治疗组。这种策略通常被称为分块。

例如，研究某种药物对心脏病发作的影响，可能首先将研究中的病人分成低风险区和高风险区，然后将每个区的一半病人随机分配到对照组，另一半分配到治疗组。这种策略确保每个治疗组有相同数量的低危病人和高危病人。


Numbered patients
create blocks
Low-risk patients
High-risk patients
randomly split in half
randomly split in half
Control
Treatment

1. 排列组合

```
prod(3:7)                       # 连乘
#> [1] 2520
factorial(5)                    # 计算阶乘
#> [1] 120
choose(10, 6)                   # 计算组合数
#> [1] 210
combn(LETTERS[1:4], 2)          # 取出所有可能的组合
#>      [,1] [,2] [,3] [,4] [,5] [,6]
#> [1,] "A"  "A"  "A"  "B"  "B"  "C"
#> [2,] "B"  "C"  "D"  "C"  "D"  "D"
```

2. prob 包实现各种概率论中的计算

```
library(prob)
permsn(LETTERS[1:3], 2)    # 取出所有可能的排列
#>      [,1] [,2] [,3] [,4] [,5] [,6]
#> [1,] "A"  "B"  "A"  "C"  "B"  "C"
#> [2,] "B"  "A"  "C"  "A"  "C"  "B"
permutations = function(n, m) factorial(n) / factorial(n-m)
permutations(5, 2)          # 排列数
#> [1] 20
```

(1) 抽球问题

	ordered = TRUE	ordered = FALSE
replace = TRUE	n^k	$\frac{(n-1+k)!}{(n-1)!k!}$
replace = FALSE	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

- 计数

```
# 共 3 个球，抽 2 次，可重复，有顺序
nsamp(n=3, k=2, replace = TRUE, ordered = TRUE)
#> [1] 9
```

- 所有可能结果

```
urnsamples(1:3, size = 2, replace = TRUE, ordered = TRUE)
#>    X1 X2
#> 1   1  1
#> 2   2  1
#> 3   3  1
#> 4   1  2
#> 5   2  2
#> 6   3  2
#> 7   1  3
#> 8   2  3
#> 9   3  3
```

(2) 样本空间与概率空间

```
tosscoin(3)                    # 抛 3 次硬币的样本空间
#>   toss1 toss2 toss3
#> 1     H     H     H
#> 2     T     H     H
#> 3     H     T     H
#> 4     T     T     H
#> 5     H     H     T
#> 6     T     H     T
#> 7     H     T     T
#> 8     T     T     T
# rolldie(2, nsides = 6)    # 掷 2 次骰子的样本空间
```

- 独立重复试验

```
iidspace(c("H","T"), ntrials = 3, probs = c(0.7, 0.3))
#>   X1 X2 X3 probs
#> 1  H  H  H 0.343
#> 2  T  H  H 0.147
#> 3  H  T  H 0.147
#> 4  T  T  H 0.063
#> 5  H  H  T 0.147
#> 6  T  H  T 0.063
#> 7  H  T  T 0.063
#> 8  T  T  T 0.027
probspace(tosscoin(1), probs = c(0.70, 0.30))
#>   toss1 probs
#> 1     H   0.7
#> 2     T   0.3
```

(3) 概率与条件概率

```
S = cards(makespace = TRUE)
A = subset(S, suit == "Heart")
B = subset(S, rank %in% 7:9)
Prob(A)
#> [1] 0.25
Prob(A, given = B)
#> [1] 0.25
```

本篇主要参阅 (张敬信, 2022), (冯国双, 2018), (Mine Çetinkaya Rundel, 2022), (贾俊平, 2018), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

Mine Çetinkaya Rundel, J. H. (2022). *Introduction to Modern Statistics*. CRC, 1 edition.

冯国双 (2018). 白话统计. 电子工业出版社, 北京, 1 edition.

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

贾俊平 (2018). 统计学. 中国人民大学出版社, 北京, 7 edition.

黄湘云 (2021). *Github: R-Markdown-Template*.