

## R 语言-6

郑泽靖 zzjstat2023@163.com

北京师范大学统计学院

2025 年 12 月 5 日

# 从 Excel 文件中读取数据

在 R 中，推荐使用 `readxl` 包来读取 Excel 文件。

## 1. 安装和加载 `readxl` 包：

```
1 install.packages("readxl")    # 安装 readxl 包
2 library(readxl)                # 加载 readxl 包
```

## 2. 读取 Excel 文件：

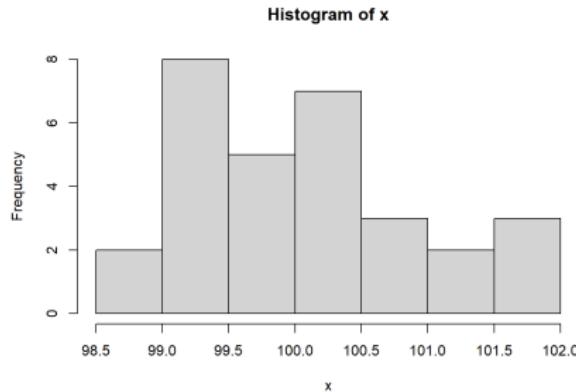
```
1 # 设置文件路径（注意使用正斜杠 / 或双反斜杠 \\）
2 file_path <- "path/to/your/file.xlsx"
3
4 # 读取 Excel 文件中的第一个工作表
5 data <- read_excel(file_path)
6
7 # 查看数据前几行
8 head(data)
```

## 直方图 (Histogram)

直方图用于展示连续变量的分布情况。使用 `hist()` 函数即可绘制。

例如，模拟生成 30 个服从正态分布的数据并绘制直方图：

```
1 x <- rnorm(30, mean=100, sd=1)
2 hist(x)
```



图：模拟正态分布数据的直方图

## hist 函数详解

常用命令格式：

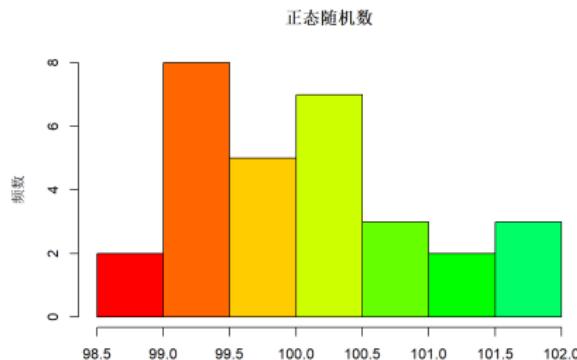
```
1 hist(x, breaks="Sturges", freq=FALSE)
```

参数	含义
x	数据向量（必须为数值型）
breaks	设置分组边界 (1) 默认为"Sturges"（自动计算最适组距） (2) 指定向量：自定义组边界点
freq	纵轴显示设置 TRUE (默认): 显示频数 (Frequency) FALSE: 显示频率/密度 (Density)

## 自定义直方图样式

可以使用 main、xlab、ylab 设置标签，用 col 设置颜色。

```
1 hist(x,
2     col = rainbow(15),      # 设置彩虹色
3     main = '正态随机数分布', # 主标题
4     xlab = '数值',          # x 轴标签
5     ylab = '频数')          # y 轴标签
```

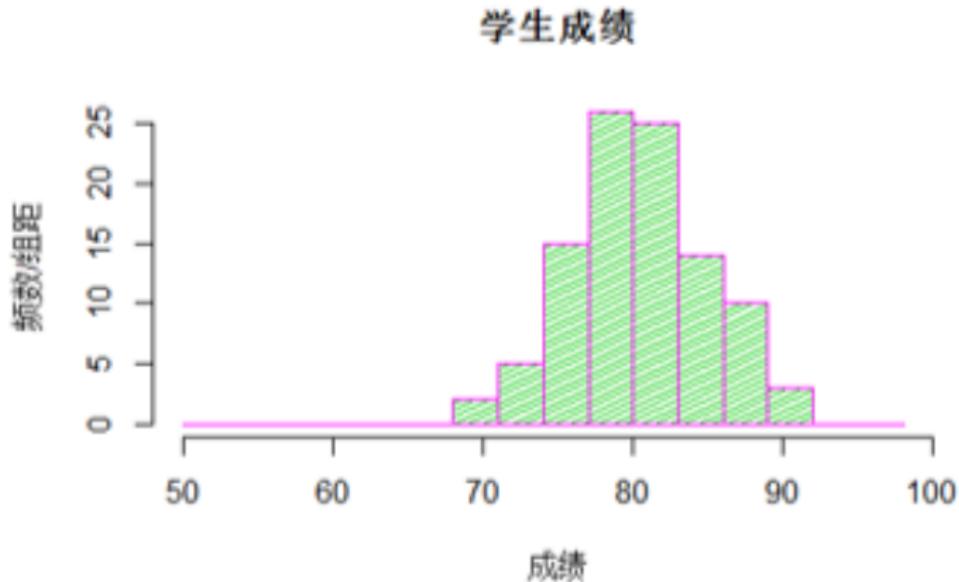


图：自定义样式的直方图

# 直方图示例：频率直方图

## 例一：绘制频率直方图 (Density)

```
1 set.seed(123)
2 x <- rnorm(100, 80, 5)          # 生成 100 个均值为 80 的成绩数据
3 x <- x[x > 50 & x <= 100]    # 截取 50-100 分之间的数据
4 g <- seq(50, 100, 3)           # 定义组边界，组距为 3
5
6 hist(x, breaks = g,            # 设置自定义组距
7       freq = FALSE,             # FALSE 表示绘制频率(密度)直方
8       main = "学生成绩分布",
9       xlim = c(50, 100),         # 设置 x 轴范围
10      xlab = "成绩",
11      ylab = "频率 / 组距")
```

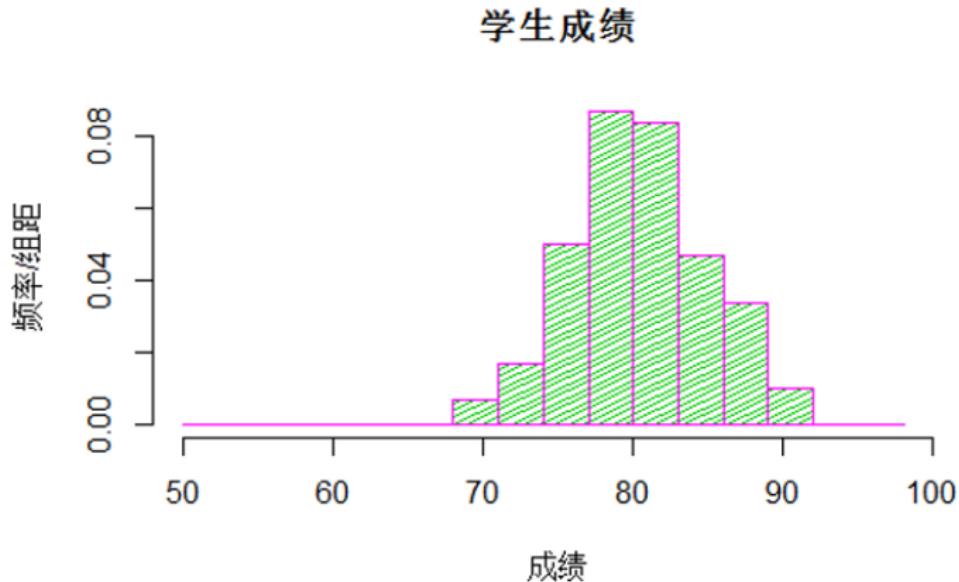


图：学生成绩的频率直方图

# 直方图示例：频数直方图

## 例二：绘制频数直方图 (Frequency)

```
1 set.seed(123)
2 x <- rnorm(100, 80, 5)
3 x <- x[x > 50 & x <= 100]
4 g <- seq(50, 100, 3)
5
6 hist(x, breaks = g,
7       freq = TRUE,           # TRUE 表示绘制频数直方图
8       main = "学生成绩分布",
9       xlim = c(50, 100),
10      xlab = "成绩",
11      ylab = "频数")
```



图：学生成绩的频数直方图

## 分类变量的频数统计

对于分类变量（因子），可以使用 `table()` 函数统计各类别的频数。

```
1 # 创建示例数据
2 x <- c("女", "男", "女", "女", "女", "男", "女", "男",
3       "女", "女", "男", "男", "女", "男", "女", "女",
4       "女", "女", "男", "男", "男", "男")
5 y <- as.factor(x)
6
7 u <- table(y)      # 计算频数
8 print(u)
9 # y
10 # 男   女
11 # 12  17
```

使用 `addmargins()` 添加合计行：

```
1 addmargins(u)
2 # y   女   男   Sum
3 #     17  12  29
```

## 分类变量的频率统计

使用 `prop.table()` 计算比例（频率）：

```
1 prop.table(u)
2 # 男          女
3 # 0.4137931 0.5862069
```

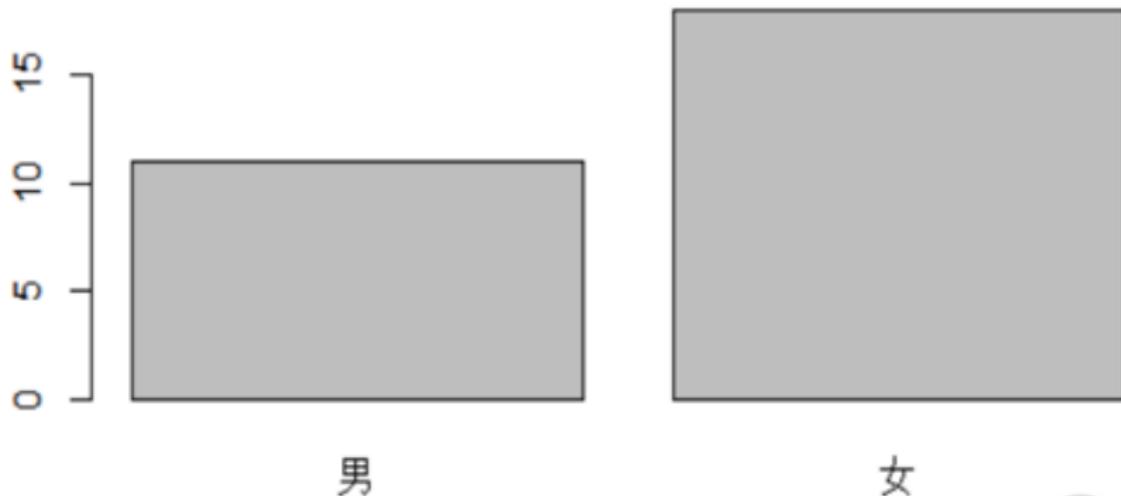
结合 `addmargins()` 显示合计频率：

```
1 v <- prop.table(u)
2 addmargins(v)
3 # 男          女          Sum
4 # 0.4137931 0.5862069 1.0000000
```

## 分类数据的条形图 (Bar Plot)

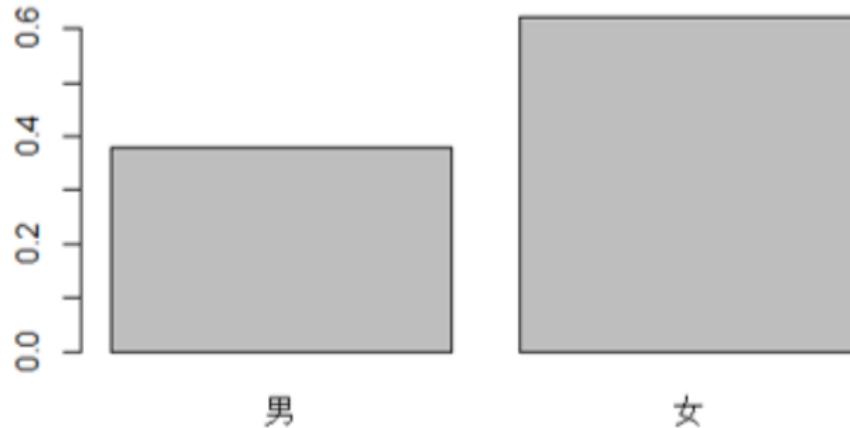
条形图和饼图常用于展示分类数据。在 R 中使用 `barplot()` 绘制条形图。

```
1 # 绘制频数条形图  
2 barplot(u, main = "性别频数分布", ylab = "人数")
```



## 分类数据的条形图

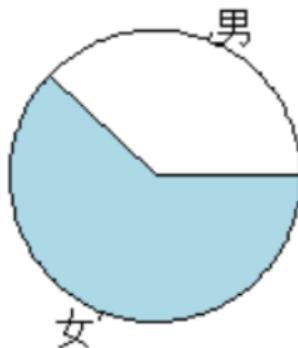
```
1 # 绘制频率条形图  
2 barplot(v, main = "性别频率分布", ylab = "占比")
```



## 饼图 (Pie Chart)

饼图通过扇形面积比例展示各分类变量的频率。使用 `pie()` 函数绘制。

```
1 print(u)
2 # y
3 # 男 女
4 # 12 17
5
6 pie(u, main = "性别分布饼图")
```



## 案例 4.5: 学生性别与专业分类

---

背景: 某大学入学一年后分专业。调查了 69 名学生的志愿选择情况。

数据统计:

- 文科: 29 人 (男生 12, 女生 17)
- 理科: 40 人 (男生 19, 女生 21)

此处涉及两个分类变量:

- 变量  $X$ : 专业 (文科、理科)
- 变量  $Y$ : 性别 (男、女)

## 多变量条形图

展示两个分类变量的关系时，常用的条形图类型：

- ① **堆积条形图 (Stacked Bar Plot)**：分为等高（展示比例）和非等高（展示总量）。
- ② **并列条形图 (Dodged Bar Plot)**：便于直接比较各组数值。

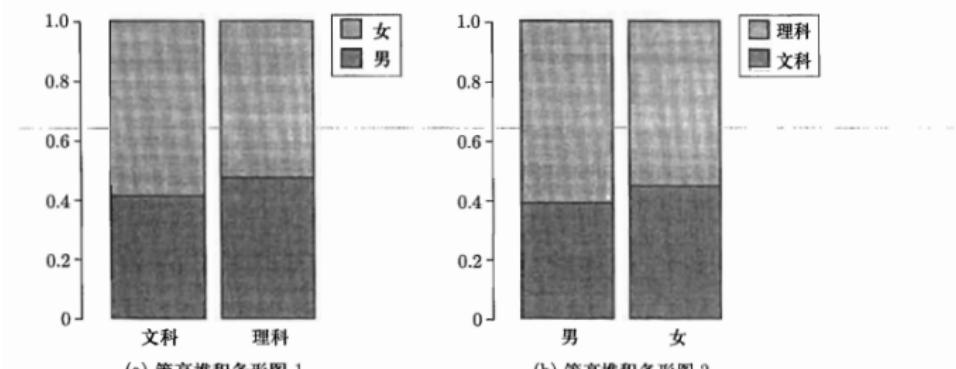


图 4.13 等高堆积条形图

图：等高堆积条形图示例

## 数据准备：创建列联表

首先将数据存储为矩阵或 table 对象：

```
1 # 创建矩阵数据
2 x <- matrix(c(12, 19, 17, 21), nrow=2, ncol=2)
3 colnames(x) <- c("男", "女")          # 列名：性别
4 rownames(x) <- c("文科", "理科")      # 行名：专业
5
6 y <- as.table(x)    # 转换为 table 对象
7 print(y)
```

	男	女
文科	12	17
理科	19	21

表：变量 y 的内容

## 绘制等高堆积条形图

等高堆积条形图主要用于比较结构比例（即条件概率）。

图 A：给定学科时的性别比例

```
1 # prop.table(..., margin=2) 表示按列计算比例(列和为1)
2 barplot(prop.table(t(y), margin=2),
3         xlim = c(0, 3.5),
4         legend.text = colnames(y),      # 图例
5         args.legend = list(x="topright"))
```

图 B：给定性别时的学科比例

```
1 barplot(prop.table(y, margin=2),
2         xlim = c(0, 3.5),
3         legend.text = rownames(y),
4         args.legend = list(x="topright"))
```

## 条件概率计算详解

函数 `prop.table(data, margin)` 用于计算比例：

- `margin=1`: 按行计算 (行和为 1)。
- `margin=2`: 按列计算 (列和为 1)。

示例结果：

`prop.table(t(y), margin=2)`

	文科	理科
男	0.41	0.48
女	0.59	0.52

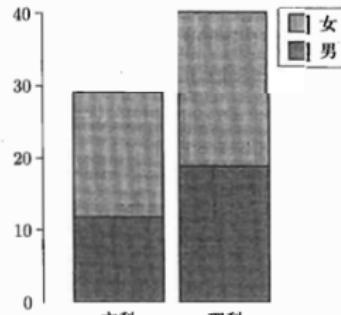
`prop.table(y, margin=2)`

	男	女
文科	0.39	0.45
理科	0.61	0.55

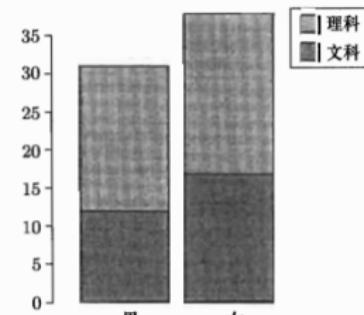
注：通过 `help(prop.table)` 可查看更多用法。

## 普通堆积条形图 (Stacked)

普通堆积条形图既能展示各组内部的比例结构，也能展示样本总量的差异。



(a) 堆积条形图 1



(b) 堆积条形图 2

图 4.14 堆积条形图

图: 非等高堆积条形图

## 绘制堆积条形图

直接使用原始频数数据  $y$  进行绘制，无需计算比例。

绘制图 (a)：按专业堆积

```
1 barplot(t(y),  
2         xlim = c(0, 3.5),  
3         legend.text = colnames(y),  
4         args.legend = list(x="topright"))
```

绘制图 (b)：按性别堆积

```
1 barplot(y,  
2         xlim = c(0, 3.5),  
3         legend.text = rownames(y),  
4         args.legend = list(x="topright"))
```

## 并列条形图 (Dodged)

并列条形图将各类别的柱子并排显示，便于直接比较数值大小。

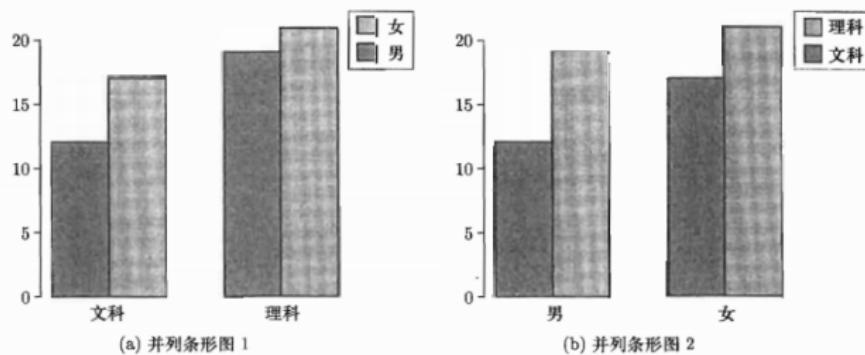


图 4.15 并列条形图

图: 并列条形图示例

## 绘制并列条形图

关键参数: `beside = TRUE`。

代码示例:

```
1 # 图 (a)
2 barplot(t(y),
3         beside = TRUE,      # 关键参数: 并列显示
4         legend.text = colnames(y),
5         args.legend = list(x="topleft"))
6
7 # 图 (b)
8 barplot(y,
9         beside = TRUE,
10        legend.text = rownames(y),
11        args.legend = list(x="topleft"))
```

## 课堂练习

---

练习题：某统计学院对学生性别与专业进行调查，数据如下：

专业	男生 (人)	女生 (人)	合计
文科	15	25	40
理科	20	30	50
工科	25	15	40

任务：请使用 R 语言完成以下图形绘制：

- ① 堆积条形图：展示每个专业中男、女生的数量构成。
- ② 等高堆积条形图：展示每个专业中男、女生的比例情况。
- ③ 并列条形图：分组对比不同专业男、女生的数量。

*Thanks!*