

生态统计学

沈国春、李勤

2025-09-27

Contents

前言	5
0.1 课程在线资源	5
0.2 学习方式	6
1 统计编程基础	7
1.1 引言	7
1.2 核心能力培养框架	8
1.3 模块一：通用编程思维基础	18
1.4 模块二：LLM 协同编程技能	52
1.5 总结	59
2 附录 1 R 语言编程基础	61
2.1 R 语言介绍	61
2.2 数值向量创建与基本统计计算	74
2.3 字符型数据处理与向量索引操作	78
2.4 数据框结构理解与多类型数据管理	81
2.5 列表数据结构与数据分组管理	86
2.6 外部数据导入与条件筛选分析	91
2.7 缺失值和异常值的识别处理	97
2.8 描述性统计分析与基础数据可视化	103
2.9 条件判断、循环结构与函数编程	111
2.10 现代数据科学工具包应用	120
2.11 图形语法与科学绘图	131
2.12 综合练习	143

前言

这是一本关于生态统计学的教材，旨在为生态学研究者提供实用的数据分析方法。本书结合 R 语言，介绍生态学研究中常用的统计方法。

本书特点：

- 面向生态学研究实际问题
- 基于 R 语言实现
- 包含大量生态数据案例
- 循序渐进的教学安排

本书使用以下 R 包：

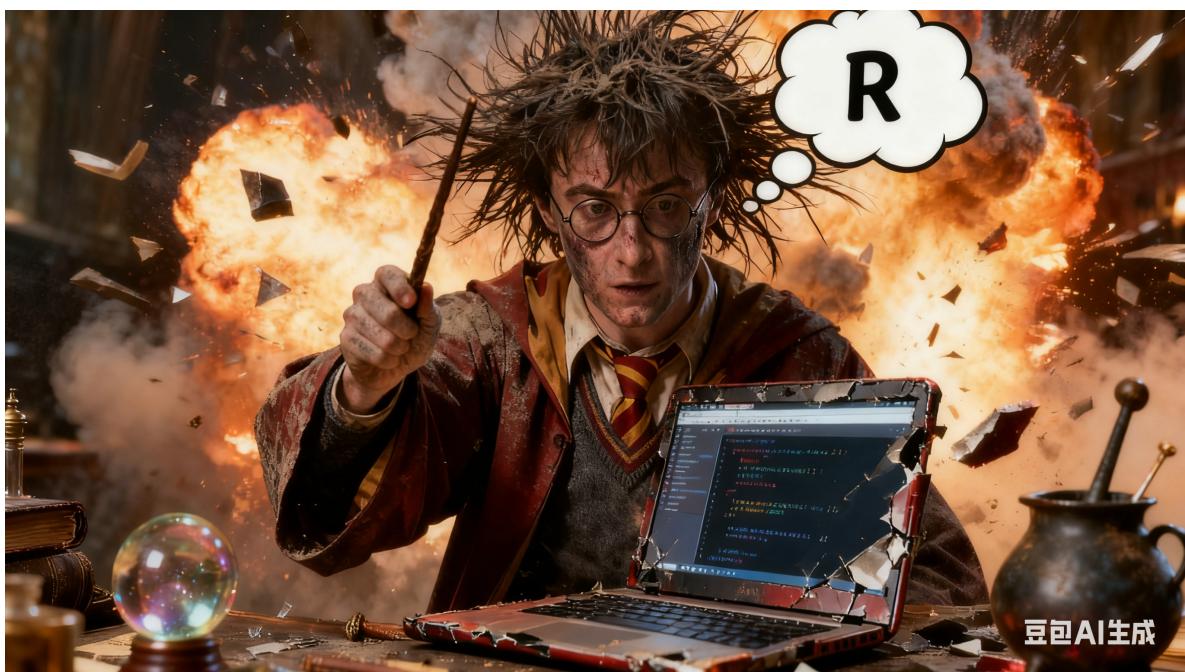
```
install.packages(c(  
  "tidyverse", "vegan", "lme4", "ggplot2",  
  "bookdown", "knitr", "rmarkdown", "DiagrammeR"  
) )
```

0.1 课程在线资源

课程简明手册

- 网页版 <https://guochunshen.github.io/ecological-statistics>
- PDF 版 <https://gitee.com/gcshen/ecological-statistics/blob/master/docs/ecological-statistics.pdf>
- 原代码 <https://gitee.com/gcshen/ecological-statistics>

0.2 学习方式



Chapter 1

统计编程基础

1.1 引言

在大语言模型（LLM）成为强大编程助手的今天，编程教育的重心正在发生根本性的转移。死记硬背语法和 API 细节的价值确实在大幅降低。这一变革标志着编程教育从”技能导向”向”思维导向”的深刻转型。过去，编程教学往往过分强调记忆各种语言的语法规则、函数库的 API 细节，以及特定框架的使用方法，学生需要花费大量时间在机械记忆上。然而，随着 Deepseek 等 AI 编程助手的普及，这些原本需要记忆的知识点现在可以通过简单的自然语言查询即时获得。这并不意味着编程变得不重要，恰恰相反，它意味着编程教育的价值需要重新定位。

在新的 AI 时代，编程教育的核心价值不再体现在”知道多少语法”，而是体现在”能够解决什么问题”和”如何设计分析方案”。学生需要培养的是更高层次的思维能力：如何将一个复杂的现实问题分解成可计算的分析步骤，如何选择合适的统计方法和数据处理技术，如何设计清晰的分析流程，如何与 AI 进行有效协作以验证和优化分析结果。这些能力构成了 AI 时代统计编程人员的核心竞争力。

具体而言，统计编程教育应该着重培养以下几个关键能力：首先是问题分解与抽象建模能力，这要求学生能够将复杂的统计问题转化为可计算的数学模型；其次是算法思维，理解不同统计方法的计算原理和适用条件；再次是数据处理能力，从数据收集、清洗到分析的完整流程设计；最后是与 AI 协作的能力，包括精准提问、代码审查和迭代优化。这些能力的培养需要项目驱动的教学方法，让学生在解决真实统计问题的过程中逐步建立编程思维框架。

这种教育重心的转移对学生提出了新的要求，也带来了新的机遇。学生不再需要为记忆琐碎的语法细节而苦恼，可以将更多精力投入到统计建模和数据分析中。教师的教学方法也需要相应调整，从传统的”语法讲解 + 练习题”模式转向”数据分析项目实践 + 统计思维训练”模式。通过这种转变，统计编程教育将更好地服务于培养学生的数据科学思维和统计分析能力这一根本目标，为他们在 AI 时代的科研和数据分析工作奠定坚实基础。

现在，对学生来说，最重要的不再是”如何写代码”，而是”解决什么问题”和”为何这样解决”。

本章将介绍 AI 时代的编程思维框架，帮助学生培养与 LLM 协同工作的核心能力，并通过 R 语言实践掌握现代数据分析方法。

1.2 核心能力培养框架

以下是学生在学习编程课程时最需要培养的核心技能，我将其分为三大类：

1.2.1 高阶思维与问题解决能力

最核心、最根本的能力是驾驭 LLM 的”方向盘”能力。在 AI 时代，数据分析师不再需要记忆繁琐的编程语法细节，但必须具备更高层次的思维框架来指导 AI 工具完成复杂的数据分析任务。这种高阶思维框架主要包括三个方面：首先是问题分解与抽象建模能力，即将复杂的生态学问题转化为清晰可执行的分析流程；其次是算法与数据结构思维，即对计算效率和数据处理优化的深刻理解；最后是数据分析流程设计与规划能力，即从宏观视角系统性地设计整个数据分析生命周期的能力。这三种能力共同构成了 AI 时代生态学数据分析师的核心竞争力，确保研究者能够站在战略高度设计分析方案，而不仅仅是执行具体的编程任务。

1.2.1.1 问题分解与抽象建模能力

是什么：将一个复杂的、模糊的现实世界问题，分解成一个个清晰的、可执行的分析步骤的能力。这种能力不仅涉及技术层面的分解，更包含对问题本质的深刻理解和抽象思维。在生态学研究中，这意味着能够将复杂的生态系统现象转化为可计算、可分析的统计模型和分析流程。

为什么重要：LLM 可以帮你写数据处理代码，但它无法替你决定”整个分析流程应该分成哪几个关键步骤”或”这个统计问题应该采用哪种分析方法”。这是人类分析师最核心的价值。在 AI 时代，这种能力变得更加关键，因为 LLM 擅长执行具体任务，但缺乏对复杂研

究问题的整体把握和统计规划能力。学生需要学会如何将模糊的研究问题转化为清晰的分析需求，这样才能有效指导 LLM 完成具体实现。

生态学案例：分析森林生态系统的物种多样性变化，需要分解为：数据收集、数据清洗、多样性指数计算、统计分析、结果可视化等步骤。具体而言，这个过程可以进一步细化为：首先确定研究目标和数据需求，包括样地选择标准、调查方法和数据格式；然后设计数据收集方案，考虑野外调查的可行性和数据质量控制；接着制定数据清洗流程，处理缺失值、异常值和数据标准化问题；再选择合适的多样性指数计算方法，如 Shannon-Wiener 指数、Simpson 指数等，并考虑其生态学意义；最后设计统计分析框架和可视化方案，确保结果能够清晰反映生态学规律。

这种问题分解能力在生态学研究中尤为重要，因为生态系统的复杂性往往超出直觉理解。通过系统性的分解和抽象，研究者能够将看似混沌的自然现象转化为有序的分析流程。例如，在研究气候变化对物种分布的影响时，需要将问题分解为气候数据获取、物种分布数据整合、生态位模型构建、未来情景预测等多个分析环节，每个环节都有其特定的技术要求和生态学考量。

培养这种能力的关键在于实践和反思。学生应该通过具体的生态学项目，学习如何识别问题的核心要素，如何设计合理的分析流程，以及如何在技术实现和生态学意义之间找到平衡。随着经验的积累，这种问题分解和抽象建模能力将成为学生在 AI 时代进行生态学研究的核心竞争力。

1.2.1.2 算法与数据结构思维

是什么：算法与数据结构思维是数据分析师的核心能力之一，它不仅仅是理解不同算法和数据结构的适用场景，更重要的是培养一种“计算效率意识”。这种思维要求我们理解不同统计方法的时间/空间复杂度 (Big O Notation)，知道在什么情况下应该选择哈希表而不是数组来快速查找数据，何时应该采用动态规划而不是暴力破解来处理复杂的优化问题。在生态学数据分析中，这种思维体现在对数据处理流程的优化意识上——比如知道在什么情况下应该使用向量化操作而不是循环，何时应该对数据进行预处理以提高后续分析的效率。这种思维还延伸到对统计方法计算复杂度的理解，比如知道某些复杂的生态位模型可能需要数小时甚至数天才能完成计算，而简单的线性回归可能只需要几秒钟。

为什么重要：在 AI 时代，LLM 可以根据你的要求实现一个统计函数，但你必须具备足够的专业知识来告诉它具体需要什么。比如，你不能简单地说“帮我做个 t 检验”，而应该明确说明“我需要一个能够处理缺失值、输出置信区间、并且可以进行方差齐性检验的 t 检验函数”。这种精确的需求描述能力来自于对统计方法内在逻辑的深刻理解。更重要的是，当

LLM 给出解决方案时，你需要具备评判能力——这个方案真的最优吗？有没有考虑边界情况？这个统计方法是否适用于我的数据类型？比如，LLM 可能会推荐使用 Pearson 相关系数来分析你的数据，但如果你知道自己的数据不符合正态分布，就应该选择 Spearman 或 Kendall 相关系数。这种批判性思维是 AI 难以替代的，它确保了分析结果的科学性和可靠性。

生态学案例：处理大规模的物种分布数据时，算法与数据结构思维显得尤为重要。假设你正在分析全国范围的鸟类分布数据，包含数百万条观测记录。如果你简单地使用线性搜索来查找特定物种的记录，可能需要数小时才能完成。但如果你具备数据结构思维，就会想到使用哈希表或数据库索引来建立快速查找机制，将查询时间从小时级降低到秒级。另一个例子是生态位模型的构建：如果你要使用 MaxEnt 模型分析某个濒危物种的分布规律，需要理解这个算法的计算复杂度，知道在什么情况下应该对数据进行降采样，什么情况下可以使用并行计算来加速模型训练。在处理时间序列的生态监测数据时，你需要知道何时应该使用滑动窗口分析而不是对整个数据集进行全局分析。这种思维还体现在对数据存储格式的选择上——知道在什么情况下应该使用 CSV 格式便于人工查看，什么情况下应该使用 Parquet 或 Feather 格式来提高读写效率。通过培养这种算法与数据结构思维，生态学研究者能够在面对海量生态数据时做出明智的技术决策，确保分析工作既高效又准确。

1.2.1.3 数据分析流程设计与规划能力

是什么：数据分析流程设计与规划能力是 AI 时代生态学研究者的核心竞争力，它要求从宏观视角系统性地设计整个数据分析的生命周期。这种能力不仅仅是知道如何使用各种统计工具，更重要的是能够站在研究问题的高度，设计出科学、高效、可复现的分析流程。具体包括：数据收集阶段的方案设计——如何确保数据的代表性和质量；数据清洗阶段的策略制定——如何处理缺失值、异常值和数据标准化；分析方法的选择与组合——如何根据研究问题和数据类型选择最合适的统计模型；结果验证与敏感性分析——如何确保分析结果的稳健性和可靠性；以及最终的可视化与报告呈现——如何将复杂的数据分析结果转化为清晰易懂的科学发现。这种能力还体现在对技术工具链的整体规划上，比如知道在什么情况下应该使用 R 而不是 Python，何时应该选择传统的统计方法而非机器学习算法，以及如何设计可扩展的分析框架来应对未来数据量的增长。

为什么重要：在 AI 协作的时代，LLM 确实可以高效执行具体的编程任务，但整个数据分析的战略规划必须由人类分析师来完成。LLM 是优秀的“执行者”，能够快速实现你指定的数据处理步骤，但它无法替代你在研究设计、方法选择和结果解释方面的专业判断。作为分析师，你需要负责确定分析的整体方向——比如是要探索数据的内在规律还是要验证特定

的科学假设？是要进行描述性统计还是要建立预测模型？这些战略决策直接影响着后续所有分析步骤的设计。更重要的是，只有你才能理解研究问题的生态学背景，知道哪些统计方法在生态学领域是公认有效的，哪些结果具有实际的生态学意义。**LLM** 可能会给出技术上的最优解，但你需要判断这个解是否科学合理、是否符合生态学理论。比如，**LLM** 可能会推荐使用复杂的深度学习模型来分析物种分布数据，但如果你知道传统的广义线性模型已经足够且更容易解释，就应该坚持使用更简单的方法。

生态学案例：在生态学研究中，数据分析流程设计与规划能力体现在对复杂研究项目的整体把控上。以研究气候变化对森林生态系统影响为例，一个完整的数据分析流程需要精心设计：首先，在数据收集阶段，你需要规划如何整合多源数据——包括气象站的长期观测数据、遥感影像的植被指数、野外调查的物种组成数据等，确保数据的时间序列和空间尺度相匹配。在数据清洗阶段，你需要制定统一的质量控制标准，比如如何处理不同来源数据的单位差异、如何填补缺失的气候数据、如何校正遥感数据的几何畸变。在分析方法选择上，你需要综合考虑研究目标——如果是要分析气候因子的相对重要性，可能会选择方差分解或随机森林；如果是要建立预测模型，可能会使用广义可加模型或机器学习算法。在整个流程中，你还需要规划结果的验证方法，比如使用交叉验证来评估模型的预测性能，或者使用独立的数据集来验证模型的泛化能力。最后，在结果呈现阶段，你需要设计清晰的可视化方案，确保复杂的统计分析结果能够被同行理解和接受。这种全方位的规划能力确保了生态学研究的科学性和可重复性，是 AI 时代生态学研究者不可或缺的核心素养。

1.2.2 与 LLM 协同工作的能力

在 AI 时代，与 **LLM** 的有效协作已经成为生态学数据分析师的核心技能。这种协作不是简单的命令与执行关系，而是一种需要精心设计的对话式工作流程。**LLM** 可以看作是一个知识渊博但缺乏专业判断的助手，它能够快速实现具体的技术任务，但需要人类分析师提供清晰的指导、专业的判断和严格的质量控制。这种协作关系要求分析师具备三个关键能力：首先是精确提问与 **Prompt** 工程能力，能够将复杂的生态学分析需求转化为 **LLM** 可以理解的明确指令；其次是代码审查与批判性验证能力，确保 **LLM** 输出的技术方案符合科学标准和实际需求；最后是迭代优化能力，通过多轮对话逐步完善分析方案。这三种能力共同构成了 AI 时代生态学研究者与智能工具协同工作的核心框架，确保研究者能够站在战略高度指导 AI 完成技术实现，同时保持对分析质量和科学性的全面把控。

1.2.2.1 精确提问与 Prompt 工程能力

是什么：精确提问与 Prompt 工程能力是 AI 时代生态学研究者与 LLM 有效协作的基础，它要求能够将复杂的生态学分析需求转化为清晰、无歧义的技术指令。这种能力不仅仅是简单的命令传达，更是一种需要专业知识和沟通技巧的对话艺术。具体包括：明确指定分析目标——是要探索数据模式还是要验证特定假设；清晰描述数据特征——包括数据结构、变量类型、数据质量等；设定技术约束——如使用的统计包、可视化要求、性能标准等；提供生态学背景——帮助 LLM 理解分析的科学意义和实际应用场景。这种能力还体现在对 LLM 输出格式的精确控制上，比如要求生成可复现的分析代码、添加详细的注释说明、确保代码符合生态学研究的最佳实践。

为什么重要：在生态学数据分析中，模糊的提问往往导致 LLM 生成不适用甚至错误的解决方案。比如，简单地说“帮我分析物种多样性”，LLM 可能会使用通用的多样性计算方法，而忽略生态学研究中需要考虑的特定约束条件，如样地面积标准化、稀有物种处理、空间自相关等问题。精确的提问能力确保 LLM 能够理解你的具体需求，生成符合生态学标准的分析代码。更重要的是，这种能力体现了研究者对分析问题的深刻理解——只有当你清楚地知道需要什么统计方法、什么数据预处理步骤、什么结果验证标准时，你才能向 LLM 提出精确的要求。在 AI 协作时代，这种精确描述需求的能力比记忆具体编程语法更加重要，它决定了你能否有效利用 AI 工具解决复杂的生态学问题。

生态学案例：在研究森林群落构建机制时，精确的提问能力显得尤为重要。假设你要分析环境过滤和生物相互作用对物种共现模式的影响，一个模糊的提问可能是：“帮我分析物种共现模式”。而精确的提问应该包括：明确分析目标——“使用零模型分析检验天童 20 个样地中木本植物的物种共现模式，检验环境过滤和竞争排斥的相对重要性”；数据约束——“数据包含每个样地的物种多度矩阵和环境因子（海拔、坡度、土壤 pH 值），需要排除 DBH<1cm 的个体”；方法要求——“使用 vegan 包计算 C-score 和 Checkerboard 指数，采用固定行列和固定物种丰富度的零模型，进行 999 次随机化，输出统计显著性和效应大小”；结果格式——“生成包含观察值、期望值、标准效应大小和 p 值的表格，以及物种对共现模式的可视化图表”。通过这种精确的提问，LLM 能够生成专业、可复现的生态学分析代码，大大提高了研究效率和分析质量。

1.2.2.2 代码审查与批判性验证能力

是什么：代码审查与批判性验证能力是确保 LLM 生成代码质量的关键保障，它要求生态学研究者具备专业的判断力来评估和验证 AI 输出的技术方案。这种能力不仅仅是检查语法

错误，更重要的是从多个维度进行综合评估：首先是功能正确性验证——确保代码逻辑符合生态学分析要求，统计方法选择恰当，计算结果准确可靠；其次是代码质量审查——检查代码的可读性、可维护性、性能效率，确保符合编程最佳实践；再次是生态学适用性判断——评估统计方法是否适合特定的生态数据类型和研究问题，比如时间序列数据是否需要考虑自相关，空间数据是否需要考虑空间依赖性；最后是科学合理性检验——确保分析结果具有生态学意义，统计推断符合科学标准。这种能力还体现在对边界情况的敏感度上，比如数据缺失、异常值处理、模型假设检验等关键环节的验证。

为什么重要：在生态学研究中，盲目接受 LLM 的输出可能导致严重的科学错误。LLM 虽然能够生成技术正确的代码，但它缺乏对生态学背景的深刻理解，可能会推荐不合适的统计方法或忽略重要的生态学约束条件。比如，LLM 可能会使用普通的线性回归来分析物种丰富度与环境因子的关系，而忽略了生态学中常用的广义线性模型或广义可加模型更适合处理计数数据和非线性关系。更重要的是，LLM 无法判断分析结果的实际生态学意义——一个统计上显著的相关性是否具有生物学重要性？模型预测是否超出了数据的合理范围？这些专业判断必须由人类研究者来完成。在 AI 协作时代，代码审查能力确保了分析结果的科学性和可靠性，是防止“垃圾进，垃圾出”现象的关键防线。

生态学案例：在分析气候变化对物种分布影响的研究中，代码审查能力尤为重要。假设 LLM 生成了一个使用 MaxEnt 模型预测物种分布变化的代码，研究者需要进行全面的审查：首先验证数据预处理——是否对气候变量进行了适当的标准化？是否考虑了变量间的多重共线性？是否使用了正确的投影坐标系？其次检查模型设置——是否设置了合理的正则化参数？是否进行了充分的模型调优？是否使用了适当的背景点采样策略？然后评估结果解释——模型预测的分布变化是否在生态学上合理？是否考虑了物种的扩散能力限制？预测的不确定性是否得到了充分评估？最后检查可复现性——代码是否包含了完整的随机种子设置？是否保存了中间结果以便后续验证？是否提供了清晰的文档说明？通过这种全面的代码审查，研究者能够确保分析结果的科学质量，避免因技术错误导致的研究结论偏差。

1.2.2.3 迭代与优化能力

是什么：迭代与优化能力是 AI 时代生态学研究者与 LLM 协同工作的核心流程，它体现了从初步方案到最终成果的渐进式完善过程。这种能力要求研究者具备系统性的反馈思维，能够基于 LLM 的初始输出进行多轮的精炼和优化。具体包括：问题诊断与反馈——准确识别 LLM 输出中的不足，如功能缺失、性能问题、代码风格不一致等，并提供具体的改进建议；方案调整与优化——根据实际需求调整分析方案，如改变统计方法、优化算法效率、改进可视化效果等；边界条件完善——补充 LLM 可能忽略的特殊情况处理，如数据缺失、异

常值、模型假设检验等；生态学细节补充——添加符合生态学研究标准的特定要求，如数据标准化、结果解释、不确定性评估等。这种能力还体现在对 LLM 学习曲线的把握上，通过持续的对话让 LLM 更好地理解研究者的分析习惯和偏好。

为什么重要：在复杂的生态学数据分析中，很少有分析方案能够一次性完美实现所有需求。LLM 的初始输出往往是一个基础框架，需要通过多轮迭代来完善细节、优化性能、增强鲁棒性。这种迭代过程不仅仅是技术修正，更重要的是科学思维的体现——通过反复的质疑、验证和优化，确保分析方案既技术正确又科学合理。比如，LLM 可能首先生成一个基本的物种多样性分析代码，但研究者需要通过多轮对话来添加样地面积标准化、稀有物种处理、统计检验等生态学分析必需的细节。更重要的是，迭代过程本身就是一个深度学习的机会——通过观察 LLM 如何响应不同的反馈，研究者能够更好地理解统计方法的实现细节，提升自己的编程和数据分析能力。在 AI 协作时代，这种迭代优化能力确保了分析方案的质量和适用性，是高效利用 AI 工具解决复杂生态学问题的关键。

生态学案例：在构建森林碳储量预测模型时，迭代优化能力发挥着关键作用。假设研究者首先向 LLM 提出需求：“帮我用 R 构建一个预测森林碳储量的模型”。LLM 可能首先生成一个简单的线性回归模型。研究者通过第一轮迭代反馈：“这个模型太简单了，森林碳储量与树高、胸径的关系可能是非线性的，请改用广义可加模型，并考虑树种差异的影响”。LLM 生成改进版本后，研究者进行第二轮迭代：“模型还需要考虑样地间的空间自相关性，请添加空间随机效应，使用混合效应模型框架”。第三轮迭代可能关注模型验证：“请添加交叉验证来评估模型预测性能，并生成残差分析图表检查模型假设”。第四轮迭代可能关注实际应用：“模型需要能够处理新的样地数据，请编写一个预测函数，并添加不确定性估计”。通过这种多轮迭代，研究者能够逐步完善分析方案，从基础模型发展到符合生态学研究标准的复杂预测系统，确保模型既技术先进又科学实用。

1.2.3 传统但愈发重要的软技能

在 AI 技术快速发展的背景下，一些传统的软技能不仅没有失去价值，反而在技术工具的辅助下变得更加关键。这些软技能构成了生态学研究者区别于 AI 的核心优势，是确保科学研究质量和影响力的根本保障。与 LLM 等技术工具不同，这些能力无法通过算法训练获得，而是需要通过长期的学术实践和人文素养培养来建立。具体包括：数据分析调试与问题排查能力——当复杂的生态学分析流程出现异常时，人类研究者的逻辑推理和经验判断是不可替代的；沟通与协作能力——向不同背景的受众解释分析结果、与团队成员协同完成研究项目的能力；持续学习与好奇心——在技术快速迭代的时代保持前沿知识更新的动力。这些软技能共同构成了 AI 时代生态学研究者的核心竞争力，确保研究者能够在技术工具的辅助下，依

然保持对科学研究本质的深刻理解和主导地位。

1.2.3.1 数据分析调试与问题排查能力

是什么：数据分析调试与问题排查能力是生态学研究者面对复杂数据分析流程时必备的核心技能，它要求具备系统性的问题诊断思维和逻辑推理能力。这种能力不仅仅是解决技术错误，更重要的是从科学研究所的整体视角来识别和解决分析流程中的各种问题。具体包括：错误信息解读与定位——能够准确理解编程语言和统计软件的错误提示，快速定位问题根源；数据质量诊断——识别数据收集、录入、处理过程中的质量问题，如缺失值模式、异常值分布、数据一致性等；逻辑推理与假设检验——通过系统性的排除法确定问题原因，验证各种可能的假设；流程回溯与复现——能够重现问题发生的完整流程，确定问题出现的具体环节；解决方案设计与验证——提出针对性的解决措施，并验证解决方案的有效性。这种能力还体现在对问题严重性的判断上——区分技术性错误和科学性错误，确保问题解决不影响研究的科学完整性。

为什么重要：在生态学数据分析中，问题排查能力比单纯的编程技能更加重要。**LLM**虽然能够生成代码，但当分析流程出现复杂问题时，它往往难以理解问题的深层原因和科学背景。比如，当物种多样性分析结果出现异常值时，**LLM**可能只能提供技术性的检查建议，而人类研究者需要结合生态学知识来判断：是数据收集问题？是统计方法选择不当？还是生态系统本身的异常现象？更重要的是，许多数据分析问题涉及多个环节的交互影响，需要研究者具备全局思维来系统排查。在**AI**协作时代，这种调试能力确保了研究者能够主导分析过程，而不是被技术问题所困扰。它体现了研究者对数据分析流程的深刻理解和对科学问题的专业判断，是确保研究成果可靠性的关键保障。

生态学案例：在研究森林群落演替动态时，调试能力显得尤为重要。假设研究者使用**LLM**生成的代码分析 20 年长期监测数据，发现某些样地的物种丰富度变化模式异常——在理论上应该增加的情况下出现了下降。研究者需要进行系统性的问题排查：首先检查数据质量——验证野外调查记录是否完整，数据录入是否有误，样地边界是否发生变化；然后检查分析方法——确认多样性计算是否考虑了样地面积标准化，统计检验是否考虑了时间序列的自相关性；接着分析生态学背景——考察是否有干扰事件（如病虫害、火灾）影响，气候条件是否有异常变化；最后验证结果合理性——与其他独立数据源对比，咨询领域专家意见。通过这种系统性的调试过程，研究者可能发现问题是因为数据录入错误（某个年份的样地面积记录有误），或者是真实的生态现象（某种优势树种的大规模死亡导致多样性暂时下降）。这种深度的问题排查能力确保了研究结论的科学性，是**AI**工具难以替代的人类智慧。

1.2.3.2 沟通与协作能力

是什么：沟通与协作能力是生态学研究者将技术分析转化为科学影响力的关键桥梁，它要求具备多层次的交流技巧和团队合作素养。这种能力体现在多个维度：首先是跨学科沟通——能够向不同专业背景的研究者（如生态学家、统计学家、政策制定者）清晰解释复杂的数据分析结果，确保技术发现被正确理解和应用；其次是学术写作与报告——将数据分析过程、方法和结论转化为规范的学术论文、研究报告或政策建议，确保科学发现的传播和影响力；再次是团队协作与项目管理——在多人参与的研究项目中协调分工、统一研究思路、解决分歧，确保研究目标的顺利实现；最后是公众科普与政策沟通——将专业的生态学研究发现转化为公众易懂的语言，为环境保护决策提供科学依据。这种能力还体现在对不同受众需求的敏感度上——知道在什么场合使用什么语言，如何平衡专业性和可理解性。

为什么重要：在 AI 时代，技术工具可以高效完成数据分析任务，但科学的研究的最终价值需要通过有效的沟通和协作来实现。LLM 虽然能够生成技术报告，但它无法理解不同受众的知识背景、关注点和价值取向，也无法进行真正的情感共鸣和思想交流。比如，向政策制定者解释气候变化对生物多样性的影响时，需要将复杂的数据分析结果转化为具体的政策建议和风险评估，这要求研究者具备政策语言的理解和转化能力。更重要的是，科学的研究本质上是集体智慧的产物，需要研究者之间的深度协作——讨论研究设计、分享数据分析经验、批判性评价研究结论。这种协作过程不仅提高了研究质量，也促进了学术共同体的知识积累。在 AI 辅助的研究环境中，沟通协作能力确保了人类智慧的主导地位，是科学的研究社会价值实现的关键环节。

生态学案例：在开展跨学科的生态系统服务评估研究时，沟通协作能力发挥着核心作用。假设一个研究团队包括生态学家、经济学家、社会学家和政策专家，共同评估森林生态系统的碳汇功能和经济价值。生态学家需要向经济学家解释碳储量测算的科学原理和数据不确定性——“我们使用异速生长方程估算树木生物量，基于树种-specific 的转换系数计算碳储量，但这种方法在幼林和异龄林中的准确性需要谨慎评估”。经济学家则需要向生态学家说明价值评估的经济学方法——“我们采用影子价格法估算碳汇的市场价值，但需要考虑碳价格的时空变异性和平政策不确定性”。社会学家需要协调不同学科的观点，确保评估框架既科学严谨又社会相关——“我们需要平衡生态系统的长期服务功能与当地社区的短期生计需求”。政策专家则需要将研究成果转化为可操作的政策建议——“基于碳汇评估结果，我们建议建立生态补偿机制，但需要设计合理的补偿标准和监督体系”。通过这种深度的跨学科沟通和协作，研究团队能够产出既有科学价值又有政策影响力的综合研究成果，这是单纯的技术分析无法实现的。

1.2.3.3 持续学习与好奇心

是什么：持续学习与好奇心是生态学研究者在技术快速变革时代保持竞争力的根本动力，它体现为对知识的主动探索和对新技术的开放态度。这种能力不仅仅是被动接受信息，更是一种积极的知识建构过程。具体包括：技术前沿跟踪——主动关注统计学、生态学、数据科学等领域的最新进展，了解新的分析方法、软件工具和研究范式；批判性学习——不盲目追随技术潮流，而是基于科学需求评估新技术的适用性和局限性；跨学科知识整合——将其他领域的先进方法创造性应用于生态学问题解决，如机器学习、网络科学、复杂系统理论等；实践导向的学习——通过实际项目应用新技术，在解决具体问题的过程中深化理解；知识分享与传播——将学习成果转化为教学材料、技术文档或学术交流，促进学术共同体的知识更新。这种能力还体现在对未知问题的探索热情上——不满足于现有答案，始终保持对自然现象深层规律的好奇和追问。

为什么重要：在 AI 和数据分析技术日新月异的背景下，持续学习能力比掌握特定技术更加重要。**LLM** 等工具虽然能够提供当前的技术解决方案，但它们无法替代研究者对知识发展方向的判断和对新兴机遇的把握。比如，当新的空间统计方法出现时，研究者需要主动学习并评估其在生态学中的应用潜力，而不是等待 **LLM** 的推荐。更重要的是，生态学本身就是一个快速发展的学科，新的理论框架、研究方法和技术工具不断涌现。只有保持持续学习的研究者才能站在学科前沿，提出创新性的研究问题，设计先进的分析方案。在 AI 协作的研究环境中，持续学习能力确保了研究者对技术工具的主导地位——你知道什么时候应该采用新技术，什么时候应该坚持传统方法，如何将不同的技术工具组合使用来解决复杂的生态学问题。

生态学案例：在应对气候变化对生物多样性影响的研究中，持续学习能力发挥着关键作用。假设一位研究者十年前主要使用传统的物种分布模型（如 BIOCLIM、DOMAIN）来预测气候变化的影响。随着技术的发展，他需要主动学习新的建模方法——首先学习基于最大熵的 **MaxEnt** 模型，理解其相对于传统方法的优势；然后掌握集成建模方法（如 ensemble forecasting），学会组合多个模型的预测结果；接着探索机器学习方法（如随机森林、支持向量机）在生态学中的应用；最近可能需要学习深度学习技术（如卷积神经网络）处理高分辨率的遥感数据。在这个过程中，研究者不仅需要学习技术本身，还需要批判性评估每种方法的适用条件——**MaxEnt** 适合小样本数据，但可能过度依赖环境变量；机器学习方法预测性能好，但可解释性较差；深度学习方法能够捕捉复杂模式，但需要大量数据和计算资源。通过这种持续的学习过程，研究者能够根据具体的研究问题和数据条件，选择最合适的技术方法，确保研究成果既技术先进又科学可靠。更重要的是，这种学习过程本身会激发新的研究思路——比如将网络分析方法应用于物种互作研究，或将时间序列分析技术应用于长期生态监测数据，从而推动生态学研究的创新发展。

1.3 模块一：通用编程思维基础

在 AI 时代，编程教育的重点已从特定语言的语法细节转向通用的编程思维框架。通用编程思维基础之所以重要，是因为它提供了跨语言、跨工具的问题解决能力。无论使用 R、Python 还是其他编程语言，无论与哪种 AI 工具协作，扎实的编程思维都是有效沟通和高效解决问题的基石。这种思维框架确保学生能够理解计算的基本原理，而不仅仅是记忆特定 API 的使用方法。

接下来仅介绍任何编程语言都需要掌握的核心编程思维，包括变量与常量、数据结构、算法与数据结构思维、编程核心概念等。如果需要参考有关 R 语言的详细资料，可参考本书的附录 1 R 语言编程基础。

1.3.1 编程核心概念

1.3.1.1 计算机主要硬件与数据流

从统计分析编程的角度理解计算机硬件组成，对于优化数据分析效率和解决计算瓶颈至关重要。计算机系统主要由四大核心硬件组件构成：中央处理器（CPU）、图形处理器（GPU）、内存（RAM）和硬盘（存储设备），每个组件在统计分析中扮演着独特而关键的角色。

中央处理器（CPU） 是计算机的“大脑”，负责执行程序指令和进行逻辑运算。在统计分析中，CPU 的性能直接影响数据处理的效率。现代 CPU 通常包含多个核心，可以并行处理多个任务，这对于统计分析中的循环运算、矩阵计算等密集型操作尤为重要。例如，在执行蒙特卡洛模拟或 Bootstrap 重抽样时，多核 CPU 可以显著加速计算过程。CPU 的时钟频率决定了单线程任务的执行速度，而缓存大小则影响数据访问的效率。在 R 语言分析中，CPU 负责执行所有的统计函数调用、数据转换和模型拟合操作。

图形处理器（GPU） 最初设计用于图形渲染，但其高度并行的架构使其在特定类型的统计分析中表现出色。GPU 包含数千个小型处理核心，能够同时执行大量简单的计算任务。在统计分析中，GPU 特别适合处理大规模的矩阵运算、深度学习模型训练、以及需要大量并行计算的任务。例如，主成分分析（PCA）、奇异值分解（SVD）等线性代数运算在 GPU 上的执行速度可能比 CPU 快数十倍。然而，GPU 编程需要特定的库和框架支持，如 R 语言的 gpuR 包或 Python 的 CUDA 工具包。

内存（RAM） 是计算机的临时工作空间，用于存储当前正在处理的数据和程序。在统计分析中，内存容量直接决定了能够处理的数据集大小。当数据分析师读取一个 CSV 文件或

数据框时，整个数据集会被加载到内存中。如果数据集超过可用内存容量，系统将使用虚拟内存（硬盘空间），但这会显著降低性能。内存的速度也影响计算效率——更快的内存意味着 CPU 能够更快地访问数据。在 R 语言中，内存管理尤为重要，因为 R 通常将整个数据集保留在内存中进行分析。

硬盘（存储设备） 用于长期数据存储，包括原始数据文件、分析结果和程序代码。硬盘的性能影响数据读取和写入的速度。传统机械硬盘（HDD）速度较慢但容量大、成本低，适合存储大型历史数据集。固态硬盘（SSD）速度更快但价格较高，适合存储需要频繁访问的当前研究数据。在统计分析工作流中，合理的存储策略可以优化整体效率——将常用数据放在 SSD 上，将归档数据放在 HDD 上。

PCIe 总线 是连接 CPU、内存、硬盘和 GPU 等硬件组件的高速数据通道。PCIe (Peripheral Component Interconnect Express) 的带宽决定了不同硬件间数据传输的速度。现代 PCIe 4.0 x16 接口提供约 32GB/s 的带宽，而 PCIe 5.0 x16 接口带宽可达 64GB/s。在 GPU 加速计算中，PCIe 总线的性能直接影响 CPU 与 GPU 之间的数据传输效率，是决定整体计算性能的关键因素之一。

程序运行的基本流程 涉及这些硬件组件间的协同工作。当执行一个统计分析程序时：首先，程序代码从硬盘被加载到内存中；然后，CPU 从内存读取指令并逐条执行；在执行过程中，CPU 可能需要从内存读取数据，进行计算后将结果写回内存；最后，分析结果被保存到硬盘中。这个过程中，数据在不同硬件间流动：硬盘 \square 内存 \square CPU \square 内存 \square 硬盘。

GPU 加速计算的流程 与 CPU 有所不同。当程序需要利用 GPU 进行并行计算时：首先，CPU 将需要处理的数据从系统内存通过 PCIe 总线传输到 GPU 显存；然后，GPU 的数千个计算核心并行处理数据；计算结果暂存在 GPU 显存中；最后，CPU 将结果从 GPU 显存传回系统内存，再保存到硬盘。这个过程中，数据流动为：硬盘 \square 内存 \square GPU 显存 \square GPU 计算核心 \square GPU 显存 \square 内存 \square 硬盘。GPU 计算的关键在于减少 CPU 与 GPU 之间的数据传输次数，通过批处理和异步传输优化性能。

在统计分析的具体场景中，这种数据流转体现得更加明显。例如，当运行一个线性回归分析时：R 解释器从硬盘读取脚本文件到内存；数据文件从硬盘加载到内存的数据框中；CPU 执行 lm() 函数，在内存中进行矩阵运算；计算结果（系数、p 值等）存储在内存中的模型对象里；最终结果被写入硬盘的报告文件。如果分析涉及大规模数据，可能会出现内存瓶颈，此时需要采用分批处理或流式处理策略，让数据在硬盘和内存间分块流动。

理解硬件组成和程序运行流程有助于统计分析人员优化工作流程。例如，知道 CPU 多核特性可以指导使用并行计算包（如 parallel）来加速计算；了解 GPU 的并行能力可以指导选择适合 GPU 加速的算法；认识内存限制可以避免处理过大的数据集导致系统崩溃；理

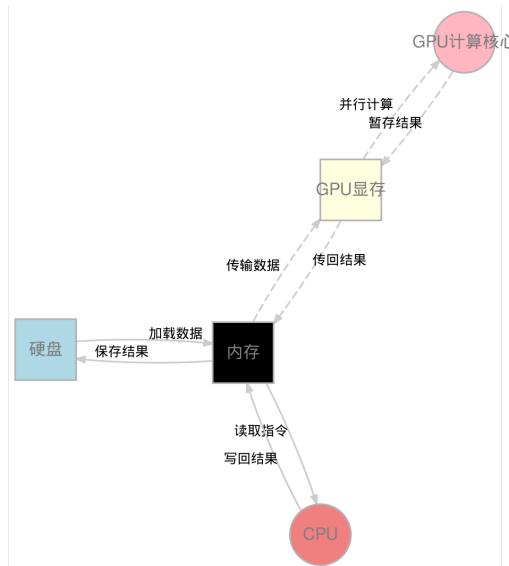


Figure 1.1 程序运行中的数据流动示意图

解硬盘性能差异可以优化数据存储策略。这种硬件意识是现代数据科学家必备的基础知识，它帮助研究者在技术约束下做出明智的决策，确保统计分析既高效又可靠。

1.3.1.2 变量与常量

```
# 变量就像可擦写的白板，可以随时修改
score <- 90
score <- 95 # 可以修改

# 常量就像刻在石头上的字，一旦设定就不能改变
PI <- 3.14159
# PI <- 3.14 # 不应该修改常量
```

为什么重要：变量和常量的区分体现了编程中的抽象思维和工程规范，是构建可维护、可复用代码的基础。在生态学数据分析中，这种区分尤为重要。变量就像生态学研究中的测量指标——它们会随着时间、空间或处理条件而变化，比如样地的温度读数、物种的个体数量、实验处理的效果值等。使用变量可以让代码适应不同的数据输入，实现分析的通用性和灵活性。而常量则代表那些在特定分析中保持不变的基础参数，比如圆周率 π 、重力加速度 g 、或者生态学中常用的转换系数（如生物量估算公式中的参数）。这些常量一旦设定就不应该被修改，因为它们代表了科学共识或物理规律。

从工程角度看，正确使用常量可以避免”魔法数字”问题——在代码中直接使用未经解释的数值，这不仅降低了代码的可读性，还增加了出错风险。比如，在计算森林碳储量时，如果直接将碳转换系数 0.5 写在计算公式中，其他研究者很难理解这个数字的含义，而且如果后续研究更新了这个系数，就需要在整个代码中搜索并修改所有出现 0.5 的地方。而如果将其定义为常量 CARBON_CONVERSION_FACTOR <- 0.5，代码就变得自文档化，修改也只需要在一个地方进行。

在生态学编程实践中，变量和常量的恰当使用还体现了对数据生命周期的理解。变量通常对应着分析过程中的中间结果或输入数据，它们的值会在分析流程中不断变化；而常量则对应着分析的基本假设和约束条件，它们定义了分析的边界和前提。这种区分有助于建立清晰的思维框架，让研究者能够更好地组织分析逻辑，确保代码的科学性和可复现性。更重要的是，在现代 AI 辅助编程环境中，明确的变量常量区分能够帮助 LLM 更好地理解代码意图，生成更符合生态学研究规范的解决方案。

1.3.1.3 基本数据类型

```
# 数字类型
temperature <- 25.5
count <- 100L

# 逻辑类型
is_raining <- TRUE
is_sunny <- FALSE

# 字符类型
species_name <- "Quercus acutissima"
habitat <- "deciduous forest"

# 空值
missing_data <- NULL
```

为什么重要：数据类型的正确理解和使用是生态学数据分析的基石，它直接影响分析的准确性、效率和可解释性。在生态学研究中，不同类型的数据对应着不同的统计方法和生态学意义，混淆数据类型可能导致严重的科学错误。比如，物种名称是分类数据（字符型），应该

使用频数统计和卡方检验；个体数量是计数数据（整数型），适合使用泊松回归或负二项回归；环境温度是连续数据（数值型），可以使用相关分析和回归模型；而存在/缺失数据（逻辑型）则需要使用二元响应模型。

从技术层面看，数据类型决定了可用的操作和函数。对数值型数据可以进行算术运算、统计检验和数学变换；对字符串型数据可以进行字符串处理、模式匹配和分类汇总；对逻辑型数据可以进行逻辑运算和条件筛选。如果混淆了数据类型，比如试图对物种名称进行算术平均，或者对温度数据进行字符串拼接，不仅会产生无意义的结果，还可能导致程序错误。更重要的是，在生态学数据分析中，数据类型的选择往往反映了对生态现象的深刻理解——比如将连续的环境梯度离散化为分类变量时，需要基于生态学理论来确定分类边界。

在 AI 辅助编程时代，数据类型知识变得更加重要。当向 LLM 描述分析需求时，明确的数据类型说明能够帮助 AI 生成更准确的代码。比如，“分析温度对物种丰富度的影响”这个需求中，如果明确指出温度是连续变量，物种丰富度是计数变量，LLM 就会推荐使用广义线性模型而不是普通的线性回归。此外，数据类型还关系到数据存储效率和计算性能——数值型数据比字符串型数据占用更少内存，整数运算比浮点运算更快。在处理大规模的生态监测数据时，这种效率差异可能决定分析是否可行。因此，掌握数据类型不仅是编程技术问题，更是生态学研究者科学素养的体现。

1.3.1.4 运算符

```
# 算术运算符
biomass <- dbh^2 * height * 0.6 # 幂运算和乘法

# 比较运算符
is_large_tree <- dbh > 30 # 大于比较
is_same_species <- species1 == species2 # 相等比较

# 逻辑运算符
suitable_habitat <- (temperature > 15) & (rainfall > 1000) # 与运算
rare_species <- (abundance < 10) | (distribution_area < 100) # 或运算

# 赋值运算符
species_count <- length(unique(species_list)) # 常规赋值
```

为什么重要：运算符是编程语言中执行基本操作的核心元素，它们将简单数据值组合成复杂的计算表达式。在生态学数据分析中，运算符的正确使用直接关系到分析结果的准确性和科学性。运算符可以分为几个主要类别：算术运算符（+、-、*、/、^）用于数值计算，如生物量估算、种群密度计算；比较运算符（>、<、==、!=）用于条件判断，如筛选特定大小的树木或特定温度范围的数据；逻辑运算符（&、|、!）用于组合多个条件，如同时满足温度和湿度要求的生态位分析。

运算符的优先级和结合性规则决定了复杂表达式的计算顺序，理解这些规则对于编写正确的代码至关重要。比如在表达式 `a + b * c` 中，乘法优先级高于加法，会先计算 `b * c` 再与 `a` 相加。如果不理解优先级，可能导致计算结果错误。在生态学建模中，这种精确性尤为重要——错误的运算符使用可能导致模型偏差或生态学意义的误解。

在 AI 协作环境中，明确的运算符使用能够显著提高与 LLM 的沟通效率。当向 AI 描述分析需求时，使用正确的运算符术语（如“使用逻辑与运算符组合温度和降水条件”）比模糊的描述（如“同时考虑温度和降水”）能生成更准确的代码。运算符还是连接数据与算法的桥梁，它们将原始生态数据转化为有意义的生态指标，是构建科学分析流程的基础构件。掌握运算符的使用不仅是一项编程技能，更是生态学研究者表达分析逻辑的重要工具。

1.3.1.5 集合数据类型

```
# 向量 - 同类型元素的集合
temperatures <- c(20, 22, 25, 18, 23)
species <- c("Oak", "Pine", "Maple", "Birch")

# 列表 - 可以包含不同类型的元素
forest_data <- list(
  name = "Tianmu Mountain Forest",
  area = 428,
  dominant_species = c("Cyclobalanopsis", "Castanopsis"),
  elevation_range = c(300, 1500)
)

# 数据框 - 表格形式的数据
forest_df <- data.frame(
```

```

plot_id = 1:5,
species = c("Quercus", "Pinus", "Acer", "Betula", "Fagus"),
dbh = c(25.3, 18.7, 12.4, 15.8, 22.1),
height = c(18.2, 15.6, 10.3, 12.7, 16.9)
)

```

为什么重要：集合数据类型的正确选择是生态学数据分析效率和质量的关键，它体现了对数据结构复杂性和分析需求的深刻理解。与基本数据类型（如数值、字符、逻辑值）处理单个数据元素不同，集合数据类型用于组织和存储多个相关数据，每种类型都有其独特的结构特性和适用场景。在生态学研究中，这种区分尤为重要——向量适合存储同类型的观测序列（如连续的温度读数），列表能够容纳复杂的嵌套结构（如包含样地信息、物种组成、环境因子的综合数据），而数据框则专门为表格型数据设计（如样地调查表）。

基本数据类型与集合数据类型的根本区别在于组织层次和操作粒度。基本数据类型关注单个数据点的属性和操作，比如数值的算术运算、字符的字符串处理；而集合数据类型关注数据之间的组织关系和整体操作，比如向量的元素索引、列表的嵌套访问、数据框的行列筛选。这种区别决定了它们的使用场景：当需要处理单一类型的序列数据时，向量提供了高效的内存存储和向量化运算；当数据结构复杂且异构时，列表的灵活性允许存储不同类型的数据对象；当数据呈现表格形式且需要同时处理多个变量时，数据框的结构化存储便于统计分析。

在生态学数据分析实践中，正确的集合类型选择直接影响分析效率和结果质量。比如，使用向量存储物种多样性指数序列可以实现快速的统计计算和可视化；使用列表组织不同样地的监测数据便于批量处理和比较分析；使用数据框管理样地调查表则可以直接应用各种统计函数和机器学习算法。更重要的是，集合数据类型的选择反映了对生态数据本质的理解——是时间序列、空间分布还是多变量关系？这种理解有助于设计更合理的分析流程，确保统计方法的适用性和结果的科学性。在 AI 协作环境中，明确的集合类型说明能够帮助 LLM 生成更符合生态学数据分析规范的代码，提高协作效率。

1.3.1.6 分支与循环

```

# 条件判断 - 根据条件选择不同路径
classify_tree_size <- function(dbh) {
  if (dbh < 10) {
    return("sapling")
  }
}

```

```
    } else if (dbh < 30) {
        return("medium tree")
    } else {
        return("large tree")
    }
}

# 循环 - 重复执行操作
# 计算每个样地的平均胸径
plot_dbh <- c(15.3, 22.7, 18.4, 25.1, 12.9)
average_dbh <- numeric(length(plot_dbh))

for (i in 1:length(plot_dbh)) {
    average_dbh[i] <- mean(plot_dbh[1:i])
}

# 更 R 风格的方式 - 使用向量化操作
average_dbh <- cumsum(plot_dbh) / 1:length(plot_dbh)
```

为什么重要：分支与循环是构建复杂生态学数据分析逻辑的核心工具，它们将静态的数据处理转化为动态的、智能的分析流程。在生态学研究中，自然系统的复杂性和不确定性要求分析程序能够根据数据特征自动调整处理策略，这正是分支结构的价值所在。比如，在分析物种分布数据时，可能需要根据数据质量（完整性、准确性）选择不同的预处理方法；在处理环境梯度数据时，需要根据变量类型（连续型、分类型）应用不同的统计模型。这种条件判断能力使得分析程序能够适应真实世界的复杂性，而不是僵化地套用固定流程。

循环结构则解决了生态学数据分析中的规模化问题。生态学研究往往涉及大量的重复性操作——对数百个样地的数据执行相同的计算，对数十个环境变量进行相同的统计分析，对多年的监测数据进行相同的时间序列分析。手动重复这些操作不仅效率低下，还容易出错。循环结构通过自动化这些重复任务，确保了分析的一致性和可复现性。更重要的是，在 R 语言中，向量化操作往往比显式循环更高效，这体现了对计算效率的深入理解。

分支与循环的组合使用能够构建出真正智能的数据分析系统。比如，一个完整的生态数据分析流程可能包含：首先使用循环遍历所有样地，对每个样地使用分支结构检查数据质量，然后根据质量等级选择不同的清洗策略，接着使用嵌套循环分析不同时间尺度的变化模

式，最后根据统计显著性自动生成报告结论。这种复杂的逻辑结构正是现代生态学研究所需的——它能够处理大规模、多维度、异质性的生态数据，产生科学可靠的结论。在 AI 协作时代，理解这些控制结构有助于更好地指导 LLM 生成符合生态学分析逻辑的代码，而不是简单的脚本堆积。

向量化操作的重要性：在 R 语言中，向量化操作代表了更高级的编程思维，它通过将操作应用于整个数据向量而非单个元素，极大地简化了数据分析代码。对于生态学研究者而言，向量化不仅意味着代码简洁性的提升——比如用 `mean(temperature)` 替代繁琐的循环计算，更重要的是它体现了对数据整体性的理解。在处理生态监测数据时，向量化操作允许研究者一次性对整个时间序列或空间网格进行分析，而不是逐点处理，这大大提高了代码的可读性和可维护性。

从性能角度看，向量化操作通常比显式循环运行更快，因为 R 的内部优化能够利用底层 C/Fortran 代码的高效实现。这种性能优势在 R、Python 等解释型语言中尤为明显，因为这些语言中的循环通常较慢。相比之下，在 C++ 等编译型语言中，循环本身已经高度优化，向量化的性能优势相对较小，但向量化思维仍然有助于简化代码语法。在处理大规模的生态数据集（如遥感影像、长期监测记录）时，这种速度优势可能决定分析是否可行。然而，向量化操作也有其局限性：它要求数据具有相同的结构和类型，对于复杂的分支逻辑或条件处理可能不够灵活。此外，过度向量化可能降低代码的可调试性，因为错误可能隐藏在复杂的向量运算中。另一个重要缺陷是内存消耗问题：向量化操作通常需要将整个数据集加载到内存中进行批量处理，对于超大规模的生态数据集（如高分辨率遥感影像、全基因组序列），这可能超出计算机的内存容量，导致程序崩溃。相比之下，循环处理可以逐块读取数据，减少内存压力。因此，生态学研究者需要在向量化的简洁高效与循环的灵活可控之间找到平衡，根据具体分析需求选择最合适的编程范式。

1.3.1.7 表达式与语句

```
# 表达式 - 产生值的代码片段
total_trees <- 100 + 50 # 表达式，产生值 150
mean_dbh <- mean(c(25, 30, 35)) # 表达式，产生平均值

# 语句 - 执行动作的代码单元
if (temperature > 25) {
  cat(" 温度过高，需要调整实验条件\n") # 语句
```

```
}
```

```
for (plot in plots) {
```

```
  analyze_plot(plot) # 语句
```

```
}
```

为什么重要：表达式与语句的区分体现了编程中的两种基本思维模式——计算思维和流程控制思维。表达式（Expression）是能够产生值的代码片段，它们关注”计算什么”，通过运算符和函数调用来完成具体的数值计算或逻辑判断。比如 `dbh^2 * height * 0.6` 是一个表达式，它计算树木的生物量；`temperature > 25` 也是一个表达式，它产生逻辑值 TRUE 或 FALSE。表达式可以嵌套组合，形成复杂的计算逻辑，但最终都会归结为一个具体的值。

语句（Statement）则是执行动作的代码单元，它们关注”做什么”，控制程序的执行流程。语句不产生值（或者产生的值不是其主要目的），而是完成特定的操作任务。前面提到的分支（if-else）和循环（for/while）都是典型的语句类型——分支语句根据条件选择不同的执行路径，循环语句重复执行特定的代码块。其他常见的语句还包括赋值语句（`x <- 10`）、函数调用语句等。

这种区分在生态学数据分析中尤为重要：表达式用于构建统计模型和计算生态指标，如多样性指数计算、回归分析等；语句则用于控制分析流程，如根据数据质量选择不同的预处理方法，或者对多个样地执行相同的分析操作。理解这种区别有助于设计更清晰的分析架构，也便于与 LLM 有效协作——明确告诉 AI 需要计算什么（表达式）和需要执行什么操作（语句）。

1.3.1.8 函数/过程

```
# 定义计算物种多样性的函数
```

```
calculate_diversity <- function(species_list) {
```

```
  species_counts <- table(species_list)
```

```
  proportions <- species_counts / sum(species_counts)
```

```
  shannon <- -sum(proportions * log(proportions))
```

```
  return(shannon)
```

```
}
```

```

# 定义数据清洗函数
clean_forest_data <- function(raw_data) {
  cleaned <- raw_data %>%
    filter(!is.na(dbh) & dbh > 0) %>%
    mutate(species = str_trim(tolower(species)))
  return(cleaned)
}

# 使用函数
sample_species <- c("Oak", "Pine", "Oak", "Maple")
diversity_index <- calculate_diversity(sample_species)
cleaned_data <- clean_forest_data(raw_forest_data)

```

为什么重要：函数是编程中的抽象工具，它将复杂的操作封装成可重用的模块，体现了“一次编写，多次使用”的工程原则。在生态学数据分析中，函数的使用具有多重价值：首先是代码复用性——相同的分析逻辑可以在不同项目、不同数据集中重复使用，避免重复劳动。比如，一个计算 Shannon 多样性指数的函数可以在多个森林调查项目中重复使用，大大提高了分析效率。其次是可维护性——当分析逻辑需要修改时，只需修改函数定义，所有调用该函数的地方都会自动更新。例如，如果需要改进多样性指数的计算方法，只需修改 `calculate_diversity` 函数，而不需要在每个使用该计算的地方逐一修改。再次是模块化设计——通过将复杂分析流程分解为多个函数，使代码结构更清晰，便于理解和调试。在生态学研究中，一个完整的分析流程可能包含数据读取、清洗、多样性计算、统计检验、可视化等多个步骤，每个步骤都可以封装为独立的函数，使整体分析逻辑更加清晰。

从工程角度看，函数还促进了代码的标准化和规范化。在团队协作的生态学研究项目中，统一的函数接口可以确保不同研究者使用相同的分析方法，提高结果的可比性和可复现性。例如，定义标准的 `clean_forest_data` 函数可以确保所有参与者在数据清洗阶段采用相同的质量控制标准。

在 AI 协作环境中，函数思维变得更加重要。当向 LLM 描述分析需求时，明确的函数化架构能够帮助 AI 更好地理解分析逻辑，生成更模块化、可维护的代码。LLM 可以根据函数化的需求描述，分别生成数据读取、处理、分析、可视化等各个模块的代码，而不是生成一个冗长复杂的单一脚本。这种模块化的代码结构不仅便于人类理解，也便于后续的调试和优化。更重要的是，函数化的思维有助于建立清晰的测试框架——每个函数都可以独立测试，确保其功能的正确性，从而提高整个分析流程的可靠性。在生态学数据分析的复杂环境中，这

种函数化的设计思维是确保分析质量、提高协作效率的关键保障。

1.3.1.9 作用域

```
# 全局变量
global_species_count <- 0

analyze_forest <- function(plot_data) {
  # 局部变量 - 只在函数内部可见
  local_species <- unique(plot_data$species)
  local_count <- length(local_species)

  # 可以访问全局变量
  global_species_count <- global_species_count + local_count

  return(local_count)
}

# 在函数外部无法访问局部变量
# print(local_species) # 会报错

# 但可以访问全局变量
print(global_species_count)
```

为什么重要：作用域规则定义了变量的可见范围，是构建复杂、安全程序的基础机制。在生态学数据分析中，正确理解作用域具有多重重要意义：首先，作用域机制有效避免了命名冲突——不同的函数或模块可以使用相同的变量名而不会相互干扰。例如，在分析多个样地数据时，每个样地的分析函数都可以使用 `species_count` 作为局部变量，而不会影响其他样地的计算结果。这种隔离性大大简化了变量命名，降低了代码复杂度。

其次，作用域提供了精细的数据访问控制能力。在生态学研究中，某些敏感数据（如原始调查记录、物种分布坐标等）需要限制访问范围，防止意外修改或泄露。通过将敏感数据封装在特定作用域内，可以确保只有授权的函数能够访问和修改这些数据，提高了代码的安全性。

第三，作用域规则优化了内存管理效率。局部变量在函数执行结束时自动释放，避免了内存泄漏问题。在处理大规模的生态监测数据时，这种自动内存管理机制尤为重要，可以防止因变量积累导致的内存耗尽问题。例如，在循环处理多个年份的监测数据时，每次迭代的临时变量都会在迭代结束后自动清理，确保内存使用的高效性。

从软件工程角度看，作用域概念体现了信息隐藏原则，是构建模块化、健壮分析系统的关键。通过将内部实现细节隐藏在局部作用域中，只暴露必要的接口，可以降低模块间的耦合度，提高代码的可维护性和可扩展性。在团队协作的生态学研究项目中，明确的作用域规则可以防止意外的变量修改，确保不同开发者编写的代码能够安全地集成在一起。特别是在 AI 协作环境中，清晰的作用域设计有助于 LLM 生成更结构化的代码，避免全局变量污染和意外的副作用，从而提高生成代码的质量和可靠性。

1.3.1.10 错误与异常处理

```
# 基本的错误处理
safe_division <- function(numerator, denominator) {
  if (denominator == 0) {
    stop("分母不能为零")
  }
  return(numerator / denominator)
}

# 使用 tryCatch 进行异常处理
analyze_with_safety <- function(data_file) {
  result <- tryCatch({
    # 尝试执行可能出错的操作
    data <- read.csv(data_file)
    diversity <- calculate_diversity(data$species)
    return(diversity)
  }, error = function(e) {
    # 错误处理
    cat("分析失败:", e$message, "\n")
    return(NA)
  }, warning = function(w) {
```

```
# 警告处理
cat("警告:", w$message, "\n")
return(calculate_diversity(data$species)) # 继续执行
})

return(result)
}

# 使用示例
try_result <- analyze_with_safety("missing_file.csv")
```

为什么重要：错误与异常处理是构建健壮分析系统的关键机制，它确保程序在遇到意外情况时能够优雅地处理而不是崩溃。在生态学数据分析中，异常处理尤为重要，因为野外数据往往存在各种质量问题——文件缺失、格式错误、数据异常等。生态学研究的数据来源多样，包括野外调查记录、传感器监测、遥感影像等，这些数据在收集、传输和处理过程中容易出现各种问题。例如，野外调查可能因天气原因中断导致数据不完整，传感器可能因故障产生异常值，不同数据源可能使用不同的格式标准。

通过合理的错误处理，可以显著提高程序的稳定性。在复杂的生态数据分析流程中，一个环节的错误不应该导致整个分析流程的崩溃。例如，当处理多个样地的调查数据时，如果某个样地文件损坏或格式错误，异常处理机制可以捕获这个错误，记录问题并继续处理其他样地，而不是让整个批处理作业失败。这种容错能力对于长期生态监测项目尤为重要，因为数据收集往往跨越数年甚至数十年，期间难免会出现各种技术问题。

异常处理还能提供友好的用户体验。相比于直接显示晦涩的技术错误信息，精心设计的异常处理可以给出清晰、有指导意义的提示。例如，当数据文件缺失时，可以提示用户检查文件路径或提供替代数据源；当数据格式错误时，可以指出具体的问题所在并建议修正方法。这种用户友好的错误处理不仅提高了工具的易用性，也降低了非技术用户的使用门槛。

在自动化流程中，异常处理是确保连续运行的关键。生态学研究经常需要处理大规模数据集，如多年的气候监测数据或大范围的遥感影像。在这些场景下，手动干预每个错误是不现实的。通过异常处理机制，程序可以自动跳过问题数据、记录错误日志、尝试替代方案，确保分析流程能够持续运行。例如，在批量计算物种多样性指数时，如果某个样地的数据质量不合格，程序可以自动标记该样地并继续处理其他样地。

在 AI 生成代码的背景下，添加适当的错误处理是确保代码质量的重要环节。LLM 生

成的代码往往侧重于功能实现，可能忽略边界情况和异常处理。作为代码审查者，需要特别关注错误处理机制的完整性，确保生成的代码能够应对各种意外情况。同时，在向 LLM 描述需求时，明确要求包含完善的错误处理逻辑，可以显著提高生成代码的健壮性和实用性。

1.3.1.11 模块化与包管理

```
# 模块化代码组织

# data_processing.R - 数据处理模块
clean_data <- function(raw_data) { /* 数据清洗逻辑 */ }
normalize_data <- function(data) { /* 数据标准化逻辑 */ }

# analysis.R - 分析模块
calculate_diversity <- function(species) { /* 多样性计算 */ }
perform_stat_test <- function(data) { /* 统计检验 */ }

# visualization.R - 可视化模块
create_plots <- function(results) { /* 图表生成 */ }

# 主程序 - 协调各个模块
source("data_processing.R")
source("analysis.R")
source("visualization.R")

# 使用包管理
library(dplyr)      # 数据处理
library(ggplot2)    # 数据可视化
library(vegan)      # 生态学分析
```

为什么重要：模块化与包管理是构建可维护、可扩展分析系统的核心实践。模块化将复杂的分析流程分解为职责单一、接口清晰的代码单元，这种分解思维在生态学数据分析中具有深远的意义。从技术层面看，模块化显著提高了代码的可读性、可测试性和可维护性。一个典型的生态数据分析项目可能包含数据收集、清洗、统计分析、可视化等多个环节，将这些环节模块化后，每个模块都可以独立开发、测试和优化。例如，数据清洗模块可以专注于处理缺

失值和异常值，统计分析模块可以专注于算法实现，可视化模块可以专注于图表设计。这种职责分离使得代码结构更加清晰，便于理解和维护。

包管理则代表了现代编程的协作智慧，它充分利用社区资源，避免重复造轮子。在生态学领域，R 语言的包生态系统尤为丰富，提供了大量专业工具。`vegan` 包专门用于生态学多样性分析，`spatstat` 包提供了空间点模式分析的完整解决方案，`sp` 包处理空间数据，`lme4` 包实现混合效应模型等。这些经过社区验证的包不仅提供了可靠的功能实现，还包含了最佳实践和标准方法。

特别值得一提的是 `spatstat` 包，它本身就是模块化设计的典范。`spatstat` 将复杂的空间统计分析功能分解为多个子包：`spatstat.core` 处理核心的点模式分析功能，`spatstat.geom` 提供几何操作，`spatstat.model` 实现统计模型，`spatstat.explore` 支持探索性分析等。这种精细的模块化设计让用户可以根据具体需求选择性地加载所需功能，避免不必要的内存开销，同时也便于功能的独立开发和维护。

使用这些专业包，研究者可以快速构建复杂的分析流程，而不需要从零开始实现基础功能。例如，计算物种多样性指数时，直接使用 `vegan` 包的 `diversity()` 函数，既保证了计算准确性，又节省了开发时间。在进行空间点模式分析时，使用 `spatstat` 包可以轻松实现 Ripley's K 函数、对相关函数等复杂的空间统计方法。

在团队协作的生态学研究项目中，模块化架构特别重要。不同研究者可以负责不同模块的开发，如生态学家专注于分析逻辑的实现，程序员专注于技术架构的优化。清晰的模块接口确保了各个部分能够无缝集成，避免了因代码耦合度过高导致的协作困难。同时，模块化支持代码的渐进式改进——可以单独优化某个模块而不影响其他部分，这种灵活性对于长期的研究项目尤为重要。

包管理还促进了分析方法的标准化和可复现性。当整个研究领域都使用相同的分析包时，不同研究的结果具有更好的可比性。例如，如果所有森林生态学研究都使用 `vegan` 包计算多样性指数，那么不同研究的结果就可以进行有意义的比较和整合。这种标准化对于生态学知识的积累和科学共识的形成至关重要。

在 AI 时代，模块化思维变得更加重要。当向 LLM 描述分析需求时，明确的模块化架构能够帮助 AI 生成更结构化的代码。LLM 可以根据模块化的需求描述，分别生成数据读取、处理、分析、可视化等各个模块的代码，而不是生成一个冗长复杂的单一脚本。这种模块化的代码不仅便于人类理解，也便于后续的调试、优化和扩展。更重要的是，模块化思维有助于建立清晰的测试框架——每个模块都可以独立测试，确保其功能的正确性，从而提高整个分析流程的可靠性。

1.3.1.12 面向对象基础

```
# 简单的面向对象示例 - 使用 S3 系统
# 定义物种类
species <- function(name, abundance, habitat) {
  structure(list(
    name = name,
    abundance = abundance,
    habitat = habitat
  ), class = "species")
}

# 定义方法
print.species <- function(x) {
  cat(" 物种:", x$name, "\n")
  cat(" 多度:", x$abundance, "\n")
  cat(" 生境:", x$habitat, "\n")
}

# 使用示例
oak <- species("Quercus", 150, "deciduous_forest")
print(oak)

# 更现代的 R6 系统示例
library(R6)

ForestPlot <- R6Class("ForestPlot",
  public = list(
    plot_id = NULL,
    species_list = NULL,
    initialize = function(plot_id, species_list) {
      self$plot_id <- plot_id
```

```
self$species_list <- species_list
} ,

calculate_diversity = function() {
  table(self$species_list) %>% diversity()
} ,

print_info = function() {
  cat("样地", self$plot_id, "有", length(unique(self$species_li
}) ,
)

# 使用示例
plot1 <- ForestPlot$new(1, c("Oak", "Pine", "Oak"))
plot1/print_info()
diversity <- plot1$calculate_diversity()
```

为什么重要：面向对象编程（OOP）提供了一种更接近现实世界思维方式的编程范式，特别适合生态学这种研究复杂自然系统的学科。OOP 的核心优势在于其三大支柱：封装性、继承性和多态性，这些特性在生态学数据分析中具有独特的应用价值。

首先，封装性允许将数据和行为捆绑在一起，形成自包含的对象。在生态学研究中，这种封装思维非常自然——一个物种对象可以包含物种名称、生态特征、分布范围等属性，以及生长模型、竞争关系等方法。例如，可以创建一个 `Species` 类，包含 `name`、`habitat`、`growth_rate` 等属性，以及 `calculate_biomass()`、`predict_distribution()` 等方法。这种封装不仅使代码更加直观，还提高了数据的安全性，防止外部代码意外修改内部状态。

继承性通过类层次关系实现代码复用和扩展，这在生态学分类系统中表现得尤为明显。可以建立从 `Organism` 到 `Plant`、`Animal`，再到具体物种如 `Quercus`（栎属）的继承层次。每个层次都可以继承父类的通用属性和方法，同时添加特有的功能。例如，所有植物类都可以共享光合作用相关的计算方法，而木本植物可以在此基础上添加年轮分析等特有功能。这种继承结构大大减少了代码重复，提高了开发效率。

多态性允许同一操作在不同对象上产生不同行为，这为处理生态系统的复杂性提供了强大工具。例如，一个 `calculate_productivity()` 方法可以在不同的生态系统组件（如森林、草地、湿地）上产生不同的计算结果，但对外提供统一的接口。这种多态性使得代码更加灵活，能够适应生态系统中各种组件的差异性。

在生态学数据分析中，OOP 思维有助于建立更直观的模型。将现实世界的生态实体（如样地、物种、种群、群落）直接映射为程序中的对象，使得分析逻辑更加贴近研究者的思维模式。例如，可以创建 `ForestPlot` 类来表示森林样地，包含样地面积、物种组成、环境因子等属性，以及多样性计算、生物量估算等方法。这种对象化的建模方式不仅提高了代码的可读性，也使得模型更容易与生态学理论对接。

OOP 还显著提高了代码的可维护性。通过清晰的类接口隔离实现细节，当需要修改某个功能时，只需关注相关类的内部实现，而不影响其他部分的代码。例如，如果需要改进物种分布预测算法，只需修改 `Species` 类的相关方法，而不需要改动使用这些物种对象的其他代码。这种模块化的维护方式大大降低了代码修改的风险和成本。

在支持复杂系统模拟方面，OOP 表现出色。生态系统动态模型通常涉及多个相互作用的组件，如种群动态、资源竞争、环境变化等。使用 OOP 可以将这些组件建模为独立的对象，通过对对象间的消息传递来模拟生态过程。例如，可以构建一个包含 `Population`、`Resource`、`Environment` 等类的生态系统模型，通过对对象间的交互来模拟长期的生态演替过程。

虽然 R 语言传统上以函数式编程为主，但现代 R 开发已经广泛采用 OOP 概念。R6 包提供了完整的面向对象编程支持，许多重要的生态学包（如 `spatstat`、`lme4` 等）都采用了面向对象的设计。理解 OOP 概念不仅有助于更好地使用这些现代 R 包，还为与其他编程语言（如 Python、C++）的协作奠定了基础。在数据科学和生态建模日益跨学科的今天，这种多范式编程能力变得愈发重要。

在 AI 协作时代，OOP 思维同样具有重要价值。当向 LLM 描述复杂的生态分析需求时，使用面向对象的术语（如“创建一个 `Species` 类，包含以下属性和方法”）能够帮助 AI 生成更结构化、更易维护的代码。OOP 的抽象层次与人类对生态系统的认知层次更加匹配，这使得生成的代码更容易被研究者理解和验证。

1.3.1.13 内存管理基础

```
# 监控内存使用
memory_usage <- function() {
  cat("当前内存使用:", format(object.size(x = ls(envir = .GlobalEnv)))
}

# 大数据处理策略
# 策略 1: 分批处理
process_large_data <- function(data_file, chunk_size = 10000) {
  con <- file(data_file, "r")
  results <- list()

  while (TRUE) {
    chunk <- readLines(con, n = chunk_size)
    if (length(chunk) == 0) break

    # 处理当前块
    processed_chunk <- process_chunk(chunk)
    results <- c(results, list(processed_chunk))

    # 清理内存
    gc()
  }

  close(con)
  return(do.call(rbind, results))
}

# 策略 2: 使用高效数据结构
# 避免不必要的复制
large_vector <- 1:1e7 # 1000 万个元素
# 不好的做法: 创建多个副本
copy1 <- large_vector
copy2 <- large_vector
```

```
# 好的做法：使用引用或原地修改
large_vector[1] <- 100 # 原地修改
```

为什么重要：内存管理是处理大规模生态数据集时必须关注的关键问题。虽然 R 具有自动垃圾回收机制，但不合理的内存使用仍然会导致程序崩溃或性能下降。理解内存管理具有多重重要意义：首先，合理的内存使用可以显著优化程序性能。在生态数据分析中，避免不必要的数据复制和内存分配是提高效率的关键。例如，在处理大型物种分布矩阵时，使用原地修改而不是创建副本可以节省大量内存和时间。R 的向量化操作虽然高效，但如果注意内存使用，也可能导致意外的内存开销。

其次，内存管理能力决定了处理大数据集的能力。随着生态学研究规模的扩大，遥感数据、基因组数据、长期监测数据等大规模数据集的应用日益广泛。这些数据集往往超过单个计算机的内存容量。通过分批处理、流式处理、内存映射等技术，可以突破物理内存的限制，处理比可用内存大得多的数据集。例如，在处理高分辨率遥感影像时，可以分块读取和处理，避免一次性加载整个文件到内存。

第三，预防内存泄漏是确保程序稳定运行的关键。在长时间运行的生态模拟或批处理作业中，即使很小的内存泄漏也会逐渐累积，最终导致程序崩溃。理解 R 的垃圾回收机制，及时释放不再使用的对象，特别是大型数据对象，对于长期稳定性至关重要。例如，在循环处理多个年份的监测数据时，确保每次迭代后清理临时变量，防止内存占用持续增长。

在 AI 协作环境中，内存管理意识同样重要。LLM 生成的代码可能没有充分考虑内存使用效率，特别是处理大规模数据时。作为代码审查者，需要特别关注内存相关的优化，如避免不必要的数据复制、使用高效的数据结构、合理设置处理批次大小等。同时，在向 LLM 描述需求时，明确内存约束条件，可以引导 AI 生成更高效的代码解决方案。

1.3.1.14 测试基础

```
# 单元测试示例
test_diversity_calculation <- function() {
  # 测试用例 1：单一物种
  test1 <- calculate_diversity(rep("Oak", 10))
  stopifnot(abs(test1 - 0) < 1e-10) # 单一物种多样性应为 0
```

```
# 测试用例 2：两个物种各占一半
test2 <- calculate_diversity(rep(c("Oak", "Pine"), each = 5))
expected <- log(2) # 两个物种各占一半的理论值
stopifnot(abs(test2 - expected) < 1e-10)

cat(" 所有测试通过!\n")
}

# 使用 testthat 包进行更专业的测试
library(testthat)

test_that(" 多样性计算正确", {
  # 测试边界情况
  expect_equal(calculate_diversity(character(0)), 0) # 空向量
  expect_equal(calculate_diversity("Oak"), 0) # 单一物种

  # 测试已知结果
  species <- c("A", "B", "C")
  expect_true(calculate_diversity(species) > 0)
})

# 数据验证函数
validate_forest_data <- function(data) {
  errors <- c()

  if (any(data$dbh <= 0)) {
    errors <- c(errors, " 存在非正胸径值")
  }

  if (any(is.na(data$species))) {
    errors <- c(errors, " 存在缺失的物种名称")
  }
}
```

```

if (length(errors) > 0) {
  stop(paste(errors, collapse = "; "))
}

return(TRUE)
}

```

为什么重要：测试是确保代码质量和分析结果可靠性的关键实践。在生态学研究中，错误的分析代码可能导致严重的科学结论偏差，因此测试尤为重要。生态学数据分析往往涉及复杂的统计模型和算法，任何细微的编程错误都可能放大为显著的科学结论差异。例如，一个错误的多样性指数计算公式可能导致对生态系统健康状况的错误评估，进而影响保护决策的制定。

完善的测试体系具有多重价值。首先，测试验证功能正确性，确保代码在各种情况下都能产生预期结果。这包括正常情况测试、边界情况测试和异常情况测试。在生态学数据分析中，这意味着不仅要测试常规的数据输入，还要测试极端值、缺失值、异常数据等特殊情况。例如，测试多样性计算函数时，需要验证它对单一物种群落、均匀分布群落、以及包含稀有物种的群落都能正确计算。

其次，测试防止回归错误，在修改代码时确保原有功能不受影响。生态学分析代码往往需要长期维护和迭代改进，随着研究深入或新方法的出现，代码需要不断更新。如果没有完善的测试套件，修改一个功能可能会意外破坏其他相关功能。例如，在优化生物量估算算法时，测试可以确保新的实现不会影响已有的多样性分析功能。

第三，测试提高代码可信度，通过测试的代码更值得信赖。在科学的研究中，可复现性是基本原则。完善的测试不仅证明了代码在当前条件下的正确性，也为其他研究者验证和复现结果提供了基础。当研究论文附有经过充分测试的分析代码时，其科学结论的可信度会显著提高。

第四，测试支持重构优化，有了测试保障，可以放心地改进代码结构。随着分析需求的复杂化，代码可能需要重构以提高性能、可读性或可维护性。测试套件作为安全网，确保重构过程中不会引入新的错误。例如，可以将一个复杂的分析函数拆分为多个小函数，通过测试验证拆分后的功能完整性。

在 AI 生成代码的背景下，测试能力变得更加重要。LLM 生成的代码虽然功能上可能正确，但往往缺乏对边界情况的充分考虑。作为代码使用者，需要建立系统的测试策略来验证 AI 输出的代码：验证功能正确性——确保代码正确实现了分析需求；测试边界情况——检查代码对异常输入、极端值的处理能力；性能测试——评估代码在处理大规模数据时的效率；

兼容性测试——确保代码与现有分析框架的集成性。

更重要的是，测试思维应该贯穿整个 AI 协作过程。在向 LLM 描述需求时，可以同时要求生成相应的测试用例；在审查 LLM 输出时，测试是验证代码质量的重要手段；在迭代优化过程中，测试确保每次改进都不会破坏已有功能。这种测试驱动的 AI 协作模式，可以显著提高生成代码的可靠性和实用性。

1.3.1.15 代码风格与规范

良好的代码风格示例

```
# 变量命名 - 使用有意义的名称
tree_diameter <- 25.3 # 好的命名
td <- 25.3           # 不好的命名

# 函数命名 - 使用动词短语
calculate_tree_volume <- function(dbh, height) {
  # 函数体
}

get_tree_volume <- function(dbh, height) { # 也可以接受
  # 函数体
}

# 代码格式 - 一致的缩进和空格
if (dbh > 30) {
  tree_size <- "large"
} else if (dbh > 10) {
  tree_size <- "medium"
} else {
  tree_size <- "small"
}

# 注释规范
```

```

# 计算 Shannon-Wiener 多样性指数
# 参数: species_vector - 物种名称向量
# 返回: 多样性指数值

calculate_shannon_diversity <- function(species_vector) {
  species_counts <- table(species_vector) # 统计每个物种的频数
  proportions <- species_counts / sum(species_counts) # 计算比例
  -sum(proportions * log(proportions)) # 计算 Shannon 指数
}

# 使用 lintr 检查代码风格
# install.packages("lintr")
# lintr::lint("your_script.R")

```

为什么重要: 代码风格与规范是编程中的“礼仪”，它虽然不影响程序功能，但直接影响代码的可读性、可维护性和协作效率。一致的代码风格具有多重重要意义：首先，良好的代码风格显著提高可读性，让其他研究者（包括未来的自己）能够快速理解代码逻辑。在生态学研究中，分析代码往往需要被同行评审、复现或扩展，清晰的代码结构就像一篇组织良好的论文，便于他人理解和验证。例如，使用有意义的变量名（如 `species_richness` 而不是 `s_rich`）、一致的缩进和空格，都能大大降低理解成本。

其次，规范的代码风格有助于减少错误。清晰的格式使潜在的逻辑问题更容易被发现，比如不匹配的括号、错误的缩进层次等。在复杂的生态数据分析中，一个微小的格式错误可能隐藏着严重的逻辑问题。使用 `lintr` 等工具自动检查代码风格，可以在早期发现这些问题，避免它们演变为难以调试的 `bug`。

第三，统一的代码规范支持团队协作。在多人参与的生态研究项目中，不同的编码风格会导致理解困难和集成冲突。制定并遵守统一的编码规范，就像使用共同的语言交流，确保团队成员能够顺畅协作。例如，约定使用蛇形命名法、特定的注释格式、一致的文件组织结构等，都可以提高协作效率。

在 AI 时代，代码规范的重要性进一步提升。LLM 生成的代码质量很大程度上取决于输入提示的规范性。当向 AI 描述需求时，使用规范的术语和结构化的描述，有助于生成更符合标准的代码。同时，规范的代码也更容易被 AI 理解和改进——当需要优化或扩展 AI 生成的代码时，规范的代码结构降低了理解难度。此外，在代码审查环节，规范的代码使人类审査者能够更专注于逻辑和功能问题，而不是纠结于格式不一致。这种人与 AI 的高效协作，正

是现代生态学研究所需的能力。

1.3.2 算法复杂度

算法是计算机科学的核心概念，它代表解决特定问题的明确、有限的步骤序列。在生态学数据分析中，算法思维尤为重要——无论是计算物种多样性指数、拟合生态位模型，还是分析时间序列数据，本质上都是在执行特定的算法。一个优秀的算法应当具备正确性（能够准确解决问题）、效率性（在合理时间内完成计算）、可读性（便于理解和维护）和鲁棒性（能够处理各种边界情况）等特征。

算法复杂度分析正是评估算法效率性的核心工具。它帮助我们理解算法性能如何随数据规模的变化而变化，这种理解对于生态学研究至关重要。例如，当处理小样本的野外调查数据时，简单的双重循环可能足够高效；但当分析全国范围的遥感数据时，只有具备良好复杂度特征的算法才能胜任。复杂度分析不仅关注时间效率（时间复杂度），也关注空间效率（空间复杂度），这两者在处理大规模生态数据集时都极为重要。

掌握算法复杂度分析，意味着能够从本质上理解不同统计方法的计算代价，为数据驱动的生态学研究提供坚实的技术基础。这种能力使研究者能够在方法选择、实验设计和结果解释中做出更加明智的决策，确保科学研究既高效又可靠。

1.3.2.1 为什么需要复杂度分析？

当解决一个问题时，通常有多种算法可供选择。我们如何评判哪个算法更“好”？这个看似简单的问题背后涉及深刻的计算科学原理。在生态学数据分析中，选择合适的算法不仅影响计算效率，更关系到研究结果的可靠性和可复现性。

方法 1：实际运行时间是一种直观但存在严重局限性的评估方式。通过在特定计算机上运行不同算法并比较执行时间，这种方法看似客观，实则受到多重外部因素的干扰。硬件配置的差异（CPU 性能、内存容量、硬盘速度）、编程语言的选择（解释型语言如 R/Python 与编译型语言如 C++ 的性能差异）、编译器优化程度、操作系统调度策略、甚至运行时的系统负载都会显著影响测试结果。更重要的是，这种测试结果具有高度的情境依赖性——在某台机器上表现优异的算法，在另一台配置不同的机器上可能表现平平。对于生态学研究而言，这种不确定性是难以接受的，因为科学分析需要可预测和可复现的性能表现。

方法 2：复杂度分析则提供了一种更加科学和根本的评估框架。这种方法不依赖于具体的运行环境，而是从算法本身的逻辑结构出发，通过数学建模来估算其资源消耗随数据规模增长的变化趋势。复杂度分析的核心优势在于其理论性和普适性——它关注的是算法内在的

效率特征，而非外在的执行环境。通过大 O 表示法等数学工具，我们可以量化分析算法的时间复杂度（执行时间增长趋势）和空间复杂度（内存占用增长趋势）。这种分析方法使得我们能够在算法设计阶段就预判其性能特征，为不同规模的数据集选择最合适的解决方案。

对于生态学数据分析师而言，掌握复杂度分析具有双重意义。从技术层面看，它帮助我们避免在大规模数据处理中陷入性能陷阱——一个在小型数据集上运行良好的 $O(n^2)$ 算法，在处理百万级生态监测记录时可能变得完全不可用。从科学层面看，复杂度分析确保了分析方法的可扩展性和可复现性，这是现代生态学研究的基本要求。通过理解算法的本质效率特征，我们能够构建既高效又可靠的生态数据分析流程，为科学研究提供坚实的技术支撑。

1.3.2.2 时间复杂度和空间复杂度的定义

1. 时间复杂度

- **定义：**全称是”渐进时间复杂度”，它表示算法的执行时间随数据规模增长的增长趋势。
- **理解：**它描述的并不是具体的执行时间（比如多少秒），而是当输入数据量 n 变得非常大时，执行时间的一个”量级”。比如，是线性增长？指数增长？还是对数增长？

2. 空间复杂度

- **定义：**全称是”渐进空间复杂度”，它表示算法的存储空间随数据规模增长的增长趋势。
- **理解：**它评估的是算法运行过程中临时占用的内存空间大小。同样，关注的是增长趋势，而不是具体的字节数。

1.3.2.3 大 O 表示法

我们使用 **大 O 表示法**来描述这种复杂度。它表示的是最坏情况下的复杂度上界，即”运行时间/占用空间最多会增长多快”。这种表示法的数学本质是描述函数增长率的渐近行为，重点关注当输入规模 n 趋向于无穷大时的主导趋势。大 O 表示法之所以选择最坏情况分析，是因为在生态学研究中，我们往往需要确保算法在最不利的条件下仍然能够完成计算任务，这对于长期监测和预测分析尤为重要。

核心思想：抓住主要矛盾体现了复杂度分析的精髓。在生态学数据分析中，我们面对的计算问题往往包含多个组成部分，但真正决定算法性能的是其中增长最快的部分。这种抓大放小的思维方式与生态学研究中的主导因子分析有着异曲同工之妙——正如在生态系统分析中我们关注关键物种和主导环境因子，在算法分析中我们关注决定性能的主导项。

在具体计算复杂度时，我们遵循几个关键原则：
* 只关注循环次数最多的那部分代码（最高阶项），因为当数据规模足够大时，低阶项的影响可以忽略不计。
* 忽略常数项。例如， $O(2n)$ 和 $O(3n)$ 都记为 $O(n)$ ，因为常数因子在不同硬件和实现中的差异很大，而大 O 表示法关注的是算法本身的本质特征。
* 忽略低阶项。例如， $O(n^2 + n)$ 记为 $O(n^2)$ ， $O(n + \log n)$ 记为 $O(n)$ ，因为随着 n 的增大，高阶项的增长速度会远远超过低阶项。

这些简化原则使得复杂度分析既实用又具有理论深度，为算法选择和优化提供了清晰的指导框架。

1.3.2.4 常见复杂度等级与计算示例

我们从低到高介绍常见的复杂度，这是面试和实际工作中最常被问到的。

1.3.2.4.1 $O(1)$ - 常数阶

- 描述：算法的执行时间/空间不随输入数据规模 n 的变化而变化。
- R 示例：

```
# 常数阶算法示例
constant_time_algorithm <- function(arr) {
  return(arr[1]) # 无论数组多大，只取第一个元素
}

# 测试
test_vector <- 1:1000
result <- constant_time_algorithm(test_vector)
print(result) # 输出：1
```

– 计算：该操作只执行一次，与 arr 的长度 n 无关。

1.3.2.4.2 $O(\log n)$ - 对数阶

- 描述：增长非常缓慢，是仅次于常数阶的高效复杂度。通常出现在“分而治之”的算法中。
- R 示例：二分查找

```

# 二分查找算法
binary_search <- function(arr, target) {
  low <- 1
  high <- length(arr)

  while (low <= high) {
    mid <- floor((low + high) / 2) # 每次都将搜索范围减半

    if (arr[mid] == target) {
      return(mid)
    } else if (arr[mid] < target) {
      low <- mid + 1
    } else {
      high <- mid - 1
    }
  }

  return(-1) # 未找到
}

# 测试
sorted_vector <- c(1, 3, 5, 7, 9, 11, 13, 15)
position <- binary_search(sorted_vector, 7)
print(position) # 输出: 4

```

- 计算：每次循环都将数据规模 n 减半。最坏情况下，需要减半多少次直到范围为空？即求解 $2^k = n$ ，得到 $k = \log_2 n$ 。所以复杂度是 $O(\log n)$ 。

1.3.2.4.3 $O(n)$ - 线性阶

- 描述：性能与数据规模 n 成正比。
- R 示例：遍历向量

```
# 线性阶算法示例
linear_time_algorithm <- function(arr) {
  total <- 0
  for (num in arr) { # 这个循环会执行 n 次
    total <- total + num
  }
  return(total)
}

# 测试
test_vector <- 1:100
result <- linear_time_algorithm(test_vector)
print(result) # 输出: 5050
```

– 计算：循环体内的操作是 $O(1)$ ，循环执行了 n 次，所以总复杂度是 $O(n)$ 。

1.3.2.4.4 $O(n \log n)$ - 线性对数阶

- 描述：性能较好，是许多高效排序算法的复杂度。
- R 示例：快速排序

```
# 快速排序算法
quick_sort <- function(arr) {
  if (length(arr) <= 1) {
    return(arr)
  }

  pivot <- arr[1]
  left <- arr[arr < pivot]
  middle <- arr[arr == pivot]
  right <- arr[arr > pivot]

  return(c(quick_sort(left), middle, quick_sort(right)))
}
```

```

}

# 测试
unsorted_vector <- c(5, 2, 8, 1, 9, 3)
sorted_vector <- quick_sort(unsorted_vector)
print(sorted_vector) # 输出: 1 2 3 5 8 9

```

- 计算：快速排序将数组层层对半分开（类似二叉树），深度是 $O(\log n)$ 。在每一层，都需要进行 $O(n)$ 级别的分区操作。因此总复杂度是 $O(n \log n)$ 。

1.3.2.4.5 $O(n^2)$ - 平方阶

- 描述：性能较差，通常出现在嵌套循环中。
- R 示例：冒泡排序

```

# 平方阶算法示例
quadratic_time_algorithm <- function(arr) {
  n <- length(arr)
  for (i in 1:n) {           # 外层循环 n 次
    for (j in 1:n) {         # 内层循环 n 次
      # O(1) 的操作
      cat(arr[i], arr[j], "\n")
    }
  }
}

# 测试（使用小数据集避免过多输出）
small_vector <- c(1, 2, 3)
quadratic_time_algorithm(small_vector)

```

- 计算：内层循环执行 n 次，外层循环执行 n 次，总操作次数是 $n * n = n^2$ ，所以复杂度是 $O(n^2)$ 。

1.3.2.4.6 $O(2^n)$ - 指数阶

- **描述：**性能极差，通常出现在暴力求解或递归未优化的场景。
- **R示例：**斐波那契数列（朴素递归）

```
# 指数阶算法示例 - 斐波那契数列（低效版本）
fibonacci_inefficient <- function(n) {
  if (n <= 1) {
    return(n)
  }
  return(fibonacci_inefficient(n-1) + fibonacci_inefficient(n-2))
}

# 测试（注意：n 不能太大，否则会非常慢）
result <- fibonacci_inefficient(10)
print(result) # 输出: 55
```

– 计算：这会产生一棵深度为 n 的递归树，节点数约为 2^n ，因此复杂度为 $O(2^n)$ 。

1.3.2.4.7 $O(n!)$ - 阶乘阶

- **描述：**性能最差，几乎不可用。通常出现在求解全排列、旅行商问题等暴力算法中。
- **R示例：**生成全排列

```
# 阶乘阶算法示例 - 生成全排列
generate_permutations <- function(elements) {
  if (length(elements) == 1) {
    return(list(elements))
  }

  permutations <- list()
  for (i in 1:length(elements)) {
    first <- elements[i]
```

```

rest <- elements[-i]

for (p in generate_permutations(rest)) {
  permutations <- c(permutations, list(c(first, p)))
}

return(permutations)
}

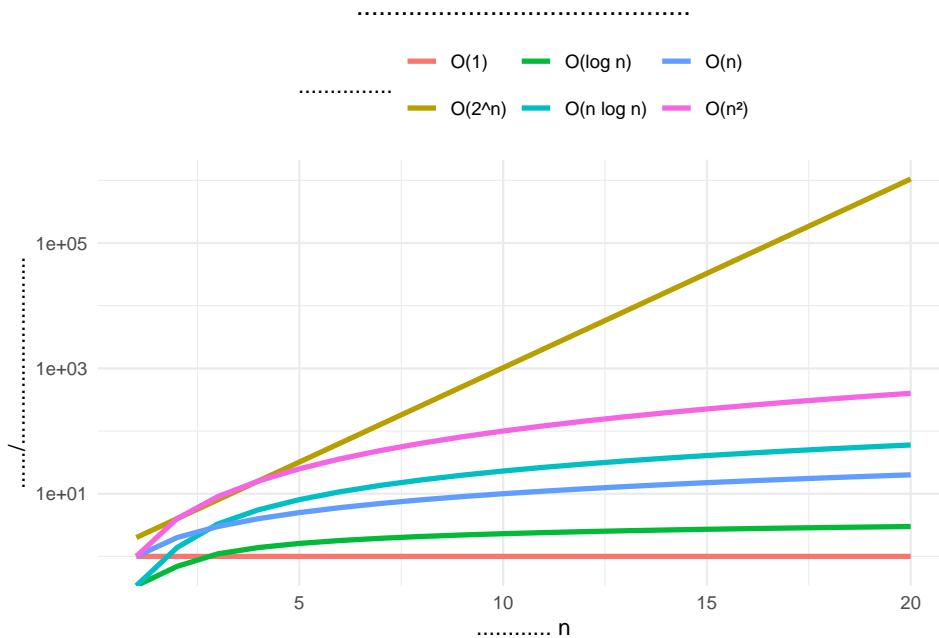
# 测试（使用小数据集）
small_set <- c("A", "B", "C")
perms <- generate_permutations(small_set)
print(length(perms)) # 输出: 6 (3! = 6)

```

- 计算： n 个元素的全排列有 $n!$ 种可能，因此复杂度为 $O(n!)$ 。

1.3.2.5 复杂度曲线图

下面这张图直观地展示了不同复杂度随数据量增长的趋势。Y 轴可以理解为时间或空间消耗。



结论： $O(1)$ 和 $O(\log n)$ 是极其高效的， $O(n)$ 和 $O(n \log n)$ 是优秀的， $O(n^2)$ 在 n 较小时可以接受，而 $O(2^n)$ 和 $O(n!)$ 应尽量避免。

1.3.2.6 给数据分析师的启示

数据规模是关键这一认知在生态学数据分析中具有决定性意义。当处理小规模数据集时，算法选择的差异可能并不明显——一个 $O(n^2)$ 的算法在几百条记录上运行可能只需要几毫秒。然而，当数据规模从 1 万条增长到 100 万条时，复杂度差异的威力就会充分显现。一个 $O(n^2)$ 的算法（如双重循环进行物种匹配）耗时将增加 1 万倍，从原本的 1 秒延长到近 3 小时；而一个 $O(n \log n)$ 的高效排序算法可能只增加不到 20 倍，从 1 秒延长到 20 秒左右。这种指数级的性能差异决定了某些分析方法在大数据场景下的可行性。在生态学研究中，这意味着我们需要根据预期的数据规模来前瞻性地选择分析方法，而不是等到数据积累到一定规模后再被动调整。

理解 R 语言操作的代价是生态学数据分析师的核心能力。许多看似简单的 R 操作背后都隐藏着复杂的算法实现。例如，`table()` 函数用于统计物种频数，其时间复杂度通常是 $O(n)$ ，但当处理大规模数据时仍需注意内存使用；`merge()` 函数进行数据框合并，其复杂度取决于合并策略，可能达到 $O(n \log n)$ 或更高；`sort()` 函数的性能差异更为明显——R 默认使用快速排序（平均 $O(n \log n)$ ），但在最坏情况下可能退化到 $O(n^2)$ 。理解这些操作的复杂度特征，能够帮助我们在设计分析流程时做出明智的决策。比如，在处理大型物种分布矩阵时，应该避免在循环内部重复调用 `table()`，而是应该预先计算好统计结果。

空间换时间是数据分析中经典的优化策略，在生态学研究中尤为实用。这种策略的核心思想是利用额外的内存存储来避免重复计算，从而显著降低时间复杂度。一个典型的例子是物种多样性分析：如果我们使用双重循环计算所有物种对之间的共存关系，时间复杂度为 $O(n^2)$ ；但如果我们将物种信息存储在一个哈希表中，然后通过单次遍历完成计算，时间复杂度可以降低到 $O(n)$ 。虽然这会增加 $O(n)$ 的空间复杂度，但在现代计算机内存充足的情况下，这种权衡通常是值得的。另一个例子是生态位模型预测：通过预先计算和缓存环境变量的响应曲线，可以避免在预测阶段重复进行复杂的数学运算，从而大幅提升模型运行效率。这种优化思维不仅适用于编程实现，也适用于分析流程设计——通过合理的设计和中间结果存储，我们可以构建既高效又可靠的大规模生态数据分析系统。

练习：尝试分析你写过的一些数据处理脚本，找出其中循环和操作，估算其时间复杂度和空间复杂度。这将极大地提升你的代码质量和性能优化能力。

1.4 模块二：LLM 协同编程技能

1.4.1 AI 编程模块安装

大语言模型辅助编程（AI-assisted programming）是当前最热门的 AI 技术之一，它通过大规模预训练语言模型（如 GPT-4、Claude、Qwen 等）来协助程序员编写代码。这种技术革命性地改变了编程工作的本质，将程序员从繁琐的语法记忆和重复性编码任务中解放出来，使其能够更专注于算法设计、系统架构和问题解决等高层次思维活动。

然而，大模型训练和调用都需要大量计算资源，这使得 AI 协作编程的可行性在早期受到限制。直到 2025 年初 DeepSeek 等开源模型的出现，普通人 AI 协作编程的可行性才得到极大提升。随后这种 AI 辅助编程工具呈现爆发式增长，从最初的 GitHub Copilot、DeepSeek，到后来的 Claude Code、Codex、Cursor 等，几乎每隔几天就有新的工具问世。这种快速迭代反映了 AI 技术在编程领域的巨大潜力和激烈竞争。

当前 AI 协作编程工具的种类已经非常丰富，涵盖了从代码补全、错误修复、代码重构到完整功能实现的各个层面。但技术的快速演进也意味着任何具体工具的详细介绍都可能很快过时。因此，本章不追求对特定工具的详尽介绍，而是聚焦于通用的 AI 协作编程思维框架和核心技能培养。无论使用哪种具体工具，研究者都需要掌握精确提问、代码审查、迭代优化等基本能力。

作为代表性工具，Qwen Code 是一个基于大模型的 AI 协作编程工具，它借鉴了 Claude Code 的设计理念，支持通过 Qwen Coder 或 Claude Code 等大模型来协助编程工作。Qwen Code 的特点包括命令行工作流设计、OAuth 一键登录、会话管理等功能，为生态学研究者提供了便捷的 AI 编程协作体验。但更重要的是，通过学习和使用这类工具，研究者能够建立起与 AI 有效协作的思维模式，这种能力将超越具体工具的局限，成为 AI 时代生态学数据分析的核心竞争力。



Qwen Code (阿里通义: **qwen** 命令的 CLI)

明确对标 Claude Code 的命令行工作流工具，对 **Qwen3-Coder** 优化；支持 **OAuth** 一键登录（国际/国内均有渠道），也支持 OpenAI-兼容 API；提供 `/compress`、`/stats` 等会话管理命令。([GitHub][8])

安装

```
# 需 Node.js 20+
npm install -g @qwen-code/qwen-code@latest
qwen --version
# 或 Homebrew (macOS/Linux)
brew install qwen-code
```

([GitHub][8])

登录 (两种模式)

```
# 方式 1: Qwen OAuth (零配置, 推荐)
qwen      # 会自动弹浏览器登录 qwen.ai 账户并缓存凭证

# 方式 2: OpenAI 兼容 API (适合私有部署/跨地区)
export OPENAI_API_KEY=...
```

```
export OPENAI_BASE_URL=...      # 例: DashScope/ModelScope/OpenRouter
export OPENAI_MODEL=qwen3-coder-plus
```

上手 & 常用命令

```
cd your-project
qwen
# 交互里可以直接自然语言:
> Explain the codebase structure
> Refactor this function
> Generate unit tests

# 会话管理
/clear  /compress  /stats  /exit
```

VS Code 中集成

在插件市场吗，搜索 **Qwen Code** 安装即可。

1.4.2 精准提问技巧

1.4.2.1 Prompt 工程基本原则

生态学数据分析的 **Prompt** 示例：

请用 R 语言帮我分析森林样地数据：

数据描述：

- 数据框包含以下列：plot_id（样地编号）、species（物种名称）、dbh（胸径）
- 数据已保存在CSV文件中，路径为"data/forest_survey.csv"

分析要求：

1. 读取数据并检查数据质量（缺失值、异常值）
2. 计算每个样地的物种丰富度（物种数）

3. 计算每个样地的平均胸径和平均树高
4. 绘制物种丰富度与平均胸径的关系散点图
5. 使用ggplot2进行可视化，添加适当的标题和标签

请为关键步骤添加注释，并确保代码具有良好的可读性。

1.4.2.2 上下文提供与约束条件设定

改进的 **Prompt** 示例:

我正在分析天童森林动态监测样地的数据，需要计算生物多样性指数。

约束条件：

- 数据包含200个固定样地的调查结果
- 每个样地面积为20m×20m
- 只统计DBH≥1cm的木本植物
- 需要排除外来物种和栽培物种

具体要求：

1. 计算每个样地的Shannon-Wiener多样性指数
2. 计算每个样地的Simpson多样性指数
3. 分析多样性指数与海拔的相关性
4. 生成专业的研究报告图表

请使用vegan包进行多样性计算，确保代码符合生态学研究的标准做法。

1.4.3 代码审查能力

1.4.3.1 LLM 输出代码的常见错误类型

```
# LLM 可能生成的有问题的代码示例
# 问题 1: 缺乏错误处理
calculate_density <- function(area, count) {
```

```

density <- count / area # 如果 area 为 0 会出错
return(density)
}

# 改进版本
calculate_density_safe <- function(area, count) {
  if (area <= 0) {
    stop(" 面积必须大于 0")
  }
  density <- count / area
  return(density)
}

# 问题 2: 使用过时的函数
# LLM 可能推荐使用旧的函数版本
old_way <- read.table("data.csv") # 较老的函数

# 改进: 使用更现代的 tidyverse 方法
library(tidyverse)
modern_way <- read_csv("data.csv") # 更简洁的语法

```

1.4.3.2 功能正确性验证方法

```

# 创建测试用例验证函数正确性
test_diversity_calculation <- function() {
  # 测试用例 1: 单一物种
  single_species <- rep("Oak", 10)
  result1 <- calculate_diversity(single_species)

  # 单一物种的 Shannon 指数应该为 0
  if (abs(result1$shannon - 0) > 1e-10) {
    stop(" 单一物种测试失败")
  }
}

```

```
}

# 测试用例 2: 两个物种各占一半
two_species <- rep(c("Oak", "Pine"), each = 5)
result2 <- calculate_diversity(two_species)

# 两个物种各占一半的 Shannon 指数应该为 log(2)
expected <- log(2)
if (abs(result2$shannon - expected) > 1e-10) {
  stop(" 两个物种测试失败")
}

cat(" 所有测试通过! \n")
}

# 运行测试
test_diversity_calculation()
```

1.4.4 调试与错误处理

1.4.4.1 错误信息解读与定位

```
# 常见的 R 错误信息及解决方法

# 错误 1: 对象未找到
# Error: object 'x' not found
# 解决方法: 检查变量名拼写, 确保变量已赋值

# 错误 2: 函数参数不匹配
# Error in mean(x) : 参数不是数值也不是逻辑值: 回传 NA
# 解决方法: 检查数据类型, 确保输入是数值型
```

```
# 错误 3: 下标越界
# Error in x[5] : 下标出界
# 解决方法: 检查向量长度, 确保索引在有效范围内

# 实用的调试技巧
debug_calculation <- function(data) {
  # 使用 browser() 进行交互式调试
  browser()

  result <- complex_calculation(data)
  return(result)
}

# 使用 tryCatch 处理错误
safe_calculation <- function(data) {
  result <- tryCatch({
    # 尝试执行可能出错的操作
    calculate_diversity(data)
  }, error = function(e) {
    # 错误处理
    cat(" 计算失败:", e$message, "\n")
    return(NULL)
  }, warning = function(w) {
    # 警告处理
    cat(" 警告:", w$message, "\n")
    return(calculate_diversity(data)) # 继续执行
  })

  return(result)
}
```

1.5 总结

本章系统性地构建了 AI 时代生态学统计编程的全新教育框架，标志着编程教育从”技能导向”向”思维导向”的根本性转变。在大语言模型成为强大编程助手的今天，编程教育的核心价值已不再体现在语法记忆和 API 细节掌握上，而是转向更高层次的思维能力培养。

本章重点培养了两大核心能力体系：首先是**高阶思维与问题解决能力**，包括问题分解与抽象建模能力、算法与数据结构思维、数据分析流程设计与规划能力。这些能力构成了生态学研究者驾驭 AI 工具的”方向盘”，确保研究者能够站在战略高度设计分析方案，而不仅仅是执行具体的编程任务。其次是**与 LLM 协同工作的能力**，涵盖精确提问与 Prompt 工程、代码审查与批判性验证、迭代与优化等关键技能，这些能力使研究者能够有效指导 AI 完成复杂的数据分析任务。

在技术层面，本章通过模块化的学习路径，系统介绍了通用编程思维基础，包括变量与常量、数据类型、运算符、集合数据类型、分支与循环、函数、作用域、错误处理、模块化、面向对象、内存管理、测试和代码规范等核心概念。这些知识为生态学数据分析提供了坚实的技术基础，确保研究者能够理解计算的基本原理，而不仅仅是记忆特定工具的使用方法。

特别值得强调的是，本章提出的”分析方案设计师 + AI 指令员 + 质量保证官”三位一体的角色定位，精准地捕捉了 AI 时代生态学研究者的核心竞争力。研究者不再需要为琐碎的编程细节所困扰，而是将精力集中在更具价值的分析设计、方法选择和结果解释上。这种角色转变不仅提高了研究效率，更提升了研究的科学性和创新性。

通过本章的学习，学生将建立起现代数据分析的思维框架，能够将复杂的生态学问题转化为清晰可执行的分析流程，并利用 AI 工具高效实现技术方案。这种能力框架具有高度的通用性和适应性，不仅适用于当前的 R 语言生态，也为未来学习其他编程语言和分析工具奠定了坚实基础。

在后续章节中，我们将基于本章建立的编程思维框架，深入探讨更专业的生态统计方法。但无论技术工具如何发展，本章所强调的分析思维、问题解决能力和 AI 协作素养，都将成为生态学研究者应对技术变革、推动学科发展的核心竞争优势。这种以思维为导向的编程教育，正是培养未来生态学创新人才的关键路径。

Chapter 2

附录 1 R 语言编程基础

2.1 R 语言介绍

- 理解为什么生态学专业需要学习 R 语言
- 掌握 R 和 RStudio 的安装和配置
- 建立良好的项目文件组织习惯
- 了解 R 包管理的基本方法

2.1.1 为什么生态学专业需要学习 R?

2.1.1.1 数据分析能力的必要性

现代生态学研究产生海量数据：野外调查数据、实验室测量数据、遥感影像数据、基因序列数据等。传统的 Excel 已无法满足复杂的统计分析需求，而 R 语言提供了完整的数据科学工具链。

2.1.1.2 可重现研究的科学要求

- **重现性危机：**越来越多的科学研究无法被重现，影响科学可信度
- **R 脚本的优势：**
 - 每一步分析都有记录，任何人都可以重现你的分析过程
 - 错误检查：代码可以被审查，减少人为错误

- 版本控制：分析过程的每次修改都有记录

2.1.1.3 职业发展的核心竞争力

- 学术界要求：顶级期刊越来越要求提供数据和代码
- 就业市场需求：环保部门、研究院所、咨询公司都需要数据分析能力
- 跨学科合作：与计算机科学、统计学等领域合作的桥梁
- 终身学习：编程思维有助于快速学习新的分析方法

2.1.1.4 开源免费的经济优势

- 成本优势：SPSS 单机版数万元，SAS 更昂贵，R 完全免费
- 功能更新：商业软件更新缓慢，R 社区每天都有新功能
- 全球社区：遇到问题可以在全球社区寻求帮助
- 未来保障：开源软件不会因为公司倒闭而消失

2.1.1.5 学术发表的必要工具

- 期刊要求：Nature、Science 等顶级期刊要求提供分析代码
- 同行评议：审稿人可以检查你的分析方法是否正确
- 引用优势：提供代码的论文被引用次数更高
- 学术诚信：透明的分析过程展现严谨的科学态度

2.1.1.6 国际交流的通用语言

- 国际会议：生态学国际会议上，R 是数据分析的主流工具
- 合作研究：与国外学者合作时，R 是共同的工作语言
- 在线学习：全球最优秀的生态学分析教程都使用 R
- 职业流动：掌握 R 可以在全球范围内寻求工作机会

2.1.2 R 语言简介与趣味应用

2.1.2.1 什么是 R 语言？

- 统计算语言：专门为统计分析和数据可视化设计的编程语言

- **开源免费**: 由全球统计学家共同维护发展
- **交互式环境**: 可以立即看到代码执行结果
- **扩展性强**: 超过 18,000 个专业扩展包

2.1.2.2 如何获取帮助?

```
# 查看函数帮助文档
?plot
help("plot")

# 搜索帮助文档
?? "regression"

# 示例代码演示
example(plot)
example(lm)
demo(persp)
demo(graphics)
demo(Hershey)
demo(plotmath)

# 在线资源:
# - R 官方文档: https://cran.r-project.org/manuals.html
# - RStudio 社区: https://community.rstudio.com
```

2.1.2.3 R 语言的趣味应用示例

```
# 安装必要包 (首次需要)
install.packages(c("ggplot2", "ggridge", "gapminder"))
library(ggplot2)
```

```

library(gganimate)
library(gapminder)

# 使用 gapminder 数据集 (包含各国多年生态经济数据)
# 绘制动态变化图
fig <- ggplot(gapminder, aes(gdpPercap, lifeExp, size = pop, color =
  geom_point() +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  labs(title = '年份: {frame_time}', x = '人均 GDP', y = '预期寿命') +
  transition_time(year) +
  ease_aes('linear')

# 渲染并保存动画
# 需要先安装 gifski 包: install.packages("gifske")
animate(fig, renderer = gifske_renderer("./imgs/gapminder_animation.gif"),
        width = 800, height = 600, fps = 10, duration = 15)

# 提示: 运行后会生成展示生态经济数据随时间变化的动画
# 可以清楚地看到不同大陆国家生态经济指标的变化趋势

```

2.1.2.3.1 生态数据动态可视化

2.1.2.4 电子水母

```

# 加载必要的包
library(ggplot2)
library(gganimate) # 用于创建动画
library(dplyr)      # 用于数据处理

# 设置参数 - 减少点数提高性能
i <- 0:2000 # 从 10000 减少到 2000 个点
x <- i %% 200      # R 中的模运算

```

```
y <- i / 43

# 预计算固定变量 (不依赖于 t)
k <- 5 * cos(x/14) * cos(y/30)
e <- y/8 - 13
d <- (k^2 + e^2)/59 + 4
a <- atan2(k, e)    # 注意 R 中 atan2 的参数顺序是 (y,x)

# 预计算所有帧数据 (向量化操作)
t_values <- seq(0, 10 * pi, by = pi / 10) # 减少帧数: 从 20π 到 10π, 步长为 π/10

# 使用矩阵运算代替循环
c_val_matrix <- outer(d / 2 + e / 99, -t_values / 18, "+")
q_matrix <- 60 - 3 * sin(a * e) + outer(k * (3 + 4 / d), sin(d ^ 2 - t_values ^ 2), "-")

x_matrix <- q_matrix * sin(c_val_matrix) + 200
y_matrix <- (q_matrix + 9 * d) * cos(c_val_matrix) + 200

# 创建数据框 (更高效的方式)
all_frames <- data.frame(
  t = rep(t_values, each = length(i)),
  X = as.vector(x_matrix),
  Y = as.vector(y_matrix)
)

# 创建动画
anim <- ggplot(all_frames, aes(X, Y)) +
  geom_point(size = 0.2, alpha = 0.2, color = "white") +
  theme_void() +
  theme(plot.background = element_rect(fill = "black", color = NA),
        panel.background = element_rect(fill = "black", color = NA)) +
  coord_cartesian(xlim = c(0, 400), ylim = c(0, 400)) +
  transition_time(t) +
```

```

shadow_mark(past = FALSE, alpha = 0.05) # 保留轨迹痕迹

# 渲染动画（可能需要几分钟时间）
animate(anim,
        fps = 20,
        duration = 30,
        width = 900,
        height = 900,
        renderer = gifski_renderer())

# 保存动画（可选）
# anim_save("animation.gif", animation = last_animation())

```

```

install.packages("audio")
install.packages("dplyr")
library(audio)
library(dplyr)
notes <- c(A = 0, B = 2, C = 3, D = 5, E = 7, F = 8, G = 10)
pitch <- "D D E D G F# D D E D A G D D D5 B G F# E C5 C5 B G A G"
duration <- c(rep(c(0.75, 0.25, 1, 1, 1, 2), 2),
               0.75, 0.25, 1, 1, 1, 1, 0.75, 0.25, 1, 1, 1, 1, 2)
bday <- data_frame(pitch = strsplit(pitch, " ")[[1]],
                     duration = duration)

bday <-
  bday %>%
  mutate(octave = substring(pitch, nchar(pitch)) %>%
    {suppressWarnings(as.numeric(.))} %>%
    ifelse(is.na(.), 4, .),
    note = notes[substr(pitch, 1, 1)],
    note = note + grepl("#", pitch) -

```

```

grepl("b", pitch) + octave * 12 +
  12 * (note < 3),
freq = 2 ^ ((note - 60) / 12) * 440

tempo <- 120
sample_rate <- 44100

make_sine <- function(freq, duration) {
  wave <- sin(seq(0, duration / tempo * 60, 1 / sample_rate) *
    freq * 2 * pi)
  fade <- seq(0, 1, 50 / sample_rate)
  wave * c(fade, rep(1, length(wave) - 2 * length(fade)), rev(fade))
}

bday_wave <-
  mapply(make_sine, bday$freq, bday$duration) %>%
  do.call("c", .)

play(bday_wave)

```

2.1.2.4.1 生成音乐

2.1.2.4.2 更多有趣功能

- 动态报告：用 R Markdown 生成可交互报告
- 网络爬虫：抓取生态监测站数据
- 地图绘制：可视化物种分布
- 机器学习：预测生态变化趋势

2.1.3 环境设置和配置

2.1.3.1 R 语言安装

2.1.3.1.1 下载和安装 R Windows 系统：

1. 访问 <https://cran.r-project.org/bin/windows/base/>
2. 下载最新版 R 安装包 (.exe)
3. 右键以管理员身份运行安装程序
4. 安装路径不要包含中文或空格
5. 勾选”创建桌面快捷方式”

macOS 系统:

1. 访问 <https://cran.r-project.org/bin/macosx/>
2. 下载最新版 R 安装包 (.pkg)
3. 双击安装, 可能需要右键”打开” 绕过 Gatekeeper 限制
4. 或通过 Homebrew 安装: `brew install --cask r`

Linux 系统:

- Ubuntu/Debian: `sudo apt-get install r-base`
- CentOS/RHEL: `sudo yum install R`

2.1.3.1.2 下载和安装 RStudio

1. 访问 <https://www.rstudio.com/products/rstudio/download/>
2. 选择适合你系统的 RStudio Desktop 免费版
3. 安装注意事项:
 - Windows: 确保已安装 R 后再安装 RStudio
 - macOS: 可能需要允许来自”未识别开发者”的应用
 - Linux: 可能需要安装依赖库 `libssl-dev` 等

2.1.3.1.3 常见安装问题解决

- 中文路径问题: 安装路径和用户名不要包含中文
- 防火墙拦截: 临时关闭防火墙或添加 R/RStudio 为例外
- 镜像源设置: 安装后运行:

```
options(repos = c(CRAN="https://mirrors.tuna.tsinghua.edu.cn/CRAN"))
```

- 依赖缺失:
 - Windows: 安装 Rtools
 - macOS: 安装 Xcode 命令行工具
 - Linux: 安装开发工具链

```
# 在 RStudio 中运行这行代码，应该显示 R 的版本信息
R.version.string

# 检查基本功能是否正常
1+1
plot(1:10)
```

2.1.3.1.4 验证安装

2.1.3.1.5 在 VSCode 中配置 R 环境 为什么选择 VSCode?

- 轻量高效: 启动速度快, 占用资源少
- 扩展性强: 丰富的扩展生态系统
- 多语言支持: 同时支持 R、Python、Markdown 等多种语言
- 版本控制集成: 内置 Git 支持, 便于代码管理
- 远程开发: 支持 SSH、容器、WSL 等远程开发环境

VSCode 安装和配置步骤:

1. 安装 VSCode

- 访问 <https://code.visualstudio.com/>
- 下载适合你操作系统的版本并安装

2. 安装 R 扩展

- 打开 VSCode，点击左侧扩展图标（或按 Ctrl+Shift+X）
- 搜索“R”并安装“R”扩展（由 REditorSupport 开发）
- 搜索“R Debugger”并安装调试器扩展

3. 基本配置

- 打开设置 (Ctrl+,)
- 搜索“rrterm”设置 R 执行路径
- 搜索“rrpath”设置 R 语言服务器路径
- 推荐设置：

```
{
  "r.rterm.windows": "C:\\Program Files\\R\\R-4.3.1\\bin\\x64\\",
  "r.rterm.mac": "/usr/local/bin/R",
  "r.rterm.linux": "/usr/bin/R",
  "r.lsp.path": "/usr/lib/R/bin/R",
  "r.sessionWatcher": true,
  "r.bracketedPaste": true
}
```

4. 常用功能

- 代码执行：Ctrl+Enter 执行当前行或选中代码
- 代码补全：自动提供函数和参数建议
- 语法高亮：彩色显示代码结构
- 调试功能：设置断点，逐步调试代码
- 绘图查看：内置绘图查看器
- **Markdown 支持**：直接编写和预览 R Markdown 文档

5. 推荐扩展

- **Rainbow CSV**：彩色显示 CSV 文件

- **GitLens**: 增强的 Git 功能
- **Bracket Pair Colorizer**: 彩色匹配括号
- **Project Manager**: 项目管理工具
- **Live Share**: 实时协作编程

6. 优势对比

- 相比 **RStudio**: 更轻量, 启动更快, 扩展更多
- 相比纯文本编辑器: 提供完整的 IDE 功能
- 适合场景: 大型项目、多语言开发、远程开发

注意事项:

- 确保已安装 R 后再配置 VSCode
- 路径设置需要根据实际安装位置调整
- 首次使用可能需要安装相关依赖包

2.1.4 项目文件夹结构设置

建立良好的文件组织习惯是数据分析的基础:

生态学R语言课程 /

```
|   └── 第01课 - 森林调查 /  
|       ├── 数据 /  
|       ├── 脚本 /  
|       └── 结果 /  
|   └── 第02课 - 物种名称 /  
|       ├── 数据 /  
|       ├── 脚本 /  
|       └── 结果 /  
└── ...  
└── 参考资料 /
```

2.1.5 基本包管理

2.1.5.1 R 包的概念

- **什么是 R 包:** R 包是扩展 R 功能的代码、数据和文档集合
- **包的作用:** 提供专业统计方法、可视化工具、数据导入等功能
- **生态学常用包:** vegan(群落分析)、ggplot2(可视化)、dplyr(数据处理) 等

2.1.5.2 Rtools 的作用

- **Windows 专用:** 用于编译需要 C/C++/Fortran 代码的 R 包
- **适用场景:**
 - 安装需要编译的包 (如部分生态学模型包)
 - 开发自己的 R 包
 - 使用某些高性能计算功能
- **安装方法:** 从 CRAN 下载对应 R 版本的 Rtools 安装包

2.1.5.3 包管理基础

```
# 检查已安装的包
installed.packages()

# 安装新包 (从 CRAN)
install.packages("ggplot2")

# 从 GitHub 安装开发版包
# install.packages("devtools")
# devtools::install_github(" 作者/包名")

# 加载包
library(ggplot2)

# 更新所有包
```

```
update.packages()
```

```
# 查看包帮助文档  
help(package="ggplot2")
```

2.1.5.4 如何寻找合适的包

- CRAN 任务视图: <https://cran.r-project.org/web/views/>
 - 如 Environmetrics 任务视图包含生态学相关包
- RStudio 包推荐: 通过 RStudio 的 Packages 面板浏览
- 学术文献: 参考领域内论文使用的方法和包
- 社区推荐: R-bloggers、Stack Overflow 等平台

2.1.6 工作目录设置

2.1.6.1 什么是工作目录?

- 简单理解: 就像你办公桌上的文件夹, R 会默认从这个文件夹里找文件
- 作用: 告诉 R 在哪里读取数据和保存结果
- 类比: 就像在图书馆找书需要知道书架位置一样

2.1.6.2 为什么要设置工作目录?

- 方便管理: 所有课程文件可以分类存放
- 避免错误: R 能准确找到你的数据文件
- 提高效率: 不用每次都输入完整文件路径

2.1.6.3 基本操作

```
# 查看当前工作目录 (就像查看你现在在哪个文件夹)  
getwd()
```

```
# 设置工作目录 (告诉 R 使用哪个文件夹)
# 注意: 路径中的斜杠方向
setwd("C:/Users/你的用户名/生态学 R 语言课程")

# 列出当前目录下的文件 (看看这个文件夹里有什么)
list.files()

# 小技巧: 在 RStudio 中可以通过菜单设置工作目录更简单:
# Session → Set Working Directory → Choose Directory
```

2.1.6.4 注意事项

- 路径中不要有中文
- Windows 系统使用正斜杠"/"或双反斜杠"\"
- 建议为每节课创建单独的子文件夹

2.1.7 课前准备检查清单

- R 软件安装成功
- 能够运行简单的 R 代码
- 创建了课程文件夹结构
- 理解了学习 R 语言的重要性
- 准备好投入时间学习编程思维

2.2 数值向量创建与基本统计计算

2.2.1 生态学背景

在野外森林调查中，测量树木胸径（DBH，距地面 1.3 米处的直径）是评估森林生长状况的基本方法。我们需要计算样地内树木的平均胸径来了解林分特征。

2.2.2 演示数据

```
# 某样地内 10 棵马尾松的胸径测量值（单位：厘米）
tree_dbh <- c(15.2, 18.7, 22.1, 19.5, 16.8, 20.3, 17.9, 21.4, 19.2,
```

2.2.3 课堂演示过程

```
# 1. 创建胸径数据向量
```

```
tree_dbh <- c(15.2, 18.7, 22.1, 19.5, 16.8, 20.3, 17.9, 21.4, 19.2,
```

```
# 2. 查看数据
```

```
tree_dbh
length(tree_dbh) # 查看测量了多少棵树
```

```
# 3. 计算平均胸径
```

```
mean_dbh <- mean(tree_dbh)
mean_dbh
```

```
# 4. 计算总树数
```

```
tree_count <- length(tree_dbh)
tree_count
```

```
# 5. 简单的数学运算
```

```
max_dbh <- max(tree_dbh) # 最大胸径
min_dbh <- min(tree_dbh) # 最小胸径
total_dbh <- sum(tree_dbh) # 胸径总和
```

2.2.4 R 语言知识点详解

2.2.4.1 向量创建函数 **c()**

- 是什么: **c()** 是 **combine** 的缩写, 用于将多个值组合成一个向量

- 语法: `c(值 1, 值 2, 值 3, ...)`
- 重要特点:
 - 向量中的所有元素必须是同一类型（数值、字符或逻辑）
 - 如果混合不同类型, R 会自动转换为最通用的类型
 - 向量是 R 中最基本的数据结构
- 常见错误: 忘记加逗号分隔值, 如 `c(1 2 3)` 是错误的
- 最佳实践: 给向量起有意义的名字, 如 `tree_dbh` 而不是 `x`

2.2.4.2 变量赋值操作符 `<-`

- 是什么: 将右边的值赋给左边的变量名
- 语法: 变量名 `<- 值`
- 为什么用 `<-` 而不是 `=`:
 - `<-` 是 R 的传统赋值符号, 更清晰地表示赋值方向
 - `=` 也可以用, 但在某些情况下可能引起混淆
 - 建议统一使用 `<-` 保持代码风格一致
- 变量命名规则:
 - 只能包含字母、数字、点(.) 和下划线(_)
 - 不能以数字开头
 - 区分大小写
 - 建议使用有意义的名称

2.2.4.3 基本统计函数

2.2.4.3.1 `mean()` - 计算平均值

- 语法: `mean(x, na.rm = FALSE)`
- 参数说明:
 - `x`: 数值向量
 - `na.rm`: 是否移除缺失值, 默认 `FALSE`
- 返回值: 数值, 向量的算术平均数
- 注意事项: 如果向量包含 NA 值, 结果会是 NA, 除非设置 `na.rm = TRUE`

2.2.4.3.2 `length()` - 计算向量长度

- 语法: `length(x)`
- 作用: 返回向量中元素的个数
- 应用场景: 统计样本数量、检查数据完整性

2.2.4.3.3 `max()` 和 `min()` - 最大值和最小值

- 语法: `max(x, na.rm = FALSE)`, `min(x, na.rm = FALSE)`
- 作用: 找出向量中的最大值和最小值
- 参数: 与 `mean()` 相同, 可以设置 `na.rm` 参数

2.2.4.3.4 `sum()` - 求和

- 语法: `sum(x, na.rm = FALSE)`
- 作用: 计算向量所有元素的总和
- 应用: 计算总量、累计值等

2.2.4.4 数据查看

- 直接输入变量名: 最简单的查看方式, 直接显示变量内容
- 自动打印机制: R 会自动显示表达式的结果
- 向量显示格式: 会显示 [1] 表示第一个元素的位置

2.2.5 课后练习

题目: 某湿地调查中测量了 8 棵柳树的树高 (单位: 米) : `tree_height <- c(4.2, 5.1, 3.8, 4.7, 5.3, 4.9, 4.1, 4.6)`

- 请完成 (仅使用本课学过的向量和基本统计函数):
1. 创建树高向量并查看数据
 2. 计算平均树高
 3. 找出最高的树有多高
 4. 统计总共测量了多少棵树

5. 计算所有树的总高度
6. 计算树高的标准差（提示：使用 `sd()` 函数）

2.3 字符型数据处理与向量索引操作

2.3.1 生态学背景

在野外鸟类观察中，需要记录观察到的鸟类物种名单，这是生物多样性调查的基础工作。我们要学会如何在 R 中管理物种名称数据。

2.3.2 演示数据

```
# 某公园早晨观察到的鸟类物种
bird_species <- c("白头鵙", "麻雀", "喜鹊", "乌鸦", "红嘴蓝鹊", "大
```

2.3.3 课堂演示过程

```
# 1. 创建鸟类物种名单
bird_species <- c("白头鵙", "麻雀", "喜鹊", "乌鸦", "红嘴蓝鹊", "大"

# 2. 查看物种名单
bird_species
print(bird_species) # 另一种显示方法

# 3. 统计观察到的物种数量
species_count <- length(bird_species)
species_count

# 4. 访问特定位置的物种
bird_species[1] # 第一个物种
bird_species[3] # 第三个物种
```

```
# 5. 添加新观察到的物种  
bird_species <- c(bird_species, "燕子")  
bird_species  
  
# 6. 字符串操作  
paste("今天观察到", length(bird_species), "种鸟类")
```

2.3.4 R 语言知识点详解

2.3.4.1 字符型数据 (Character Data)

- 是什么：用引号包围的文本数据，R 中的基本数据类型之一
- 语法："文本内容" 或 '文本内容'
- 重要特点：
 - 单引号和双引号都可以，但要成对使用
 - 如果文本中包含引号，需要转义或使用另一种引号包围
 - 字符型数据在 R 中以向量形式存储
- 与数值的区别：
 - 字符型："123" - 这是文本，不能进行数学运算
 - 数值型：123 - 这是数字，可以进行数学运算
- 编码注意：中文字符需要确保 R 的编码设置正确

2.3.4.2 向量索引 (Vector Indexing)

- 是什么：通过位置编号访问向量中特定元素的方法
- 语法：向量名 [位置]
- 索引规则：
 - R 的索引从 1 开始（不是 0!）
 - 可以使用负数排除特定位置：bird_species[-1]（排除第一个）
 - 可以一次访问多个位置：bird_species[c(1,3,5)]
- 超出范围：如果索引超出向量长度，返回 NA
- 实际应用：在生态学中用于提取特定样本、物种等

2.3.4.3 数据显示函数

2.3.4.3.1 `print()` 函数

- 语法: `print(x)`
- 与直接输入变量名的区别:
 - 直接输入: 仅在交互模式下显示
 - `print()`: 在脚本和函数中也会显示, 更可控
- 应用场景: 在循环、函数中需要显示结果时

2.3.4.4 向量合并和扩展

- 添加元素: `c(原向量, 新元素)`
- 重要概念: R 中的向量是不可变的, 每次”添加”实际上是创建新向量
- 效率考虑: 频繁添加元素效率较低, 大量数据建议预先分配空间
- 实际应用: 野外调查中动态添加新发现的物种

2.3.4.5 字符串连接函数 `paste()`

- 语法: `paste(..., sep = " ", collapse = NULL)`
- 参数详解:
 - ...: 要连接的多个元素
 - `sep`: 分隔符, 默认是空格
 - `collapse`: 如果提供, 将结果向量合并为单个字符串
- 相关函数:
 - `paste0()`: 等同于 `paste(..., sep = "")`, 不使用分隔符
 - `sprintf()`: 格式化字符串, 类似其他语言的 `printf`
- 实际应用: 生成报告文本、标签、文件名等

2.3.5 课后练习

题目: 某湿地调查中记录的水生植物: `water_plants <- c("荷花", "芦苇", "菖蒲", "水葫芦", "睡莲")`

请完成（使用向量、字符串操作、索引等已学内容）：

1. 创建植物名称向量并显示所有植物名称

2. 计算记录了多少种植物（使用 `length()` 函数）
3. 显示第 2 种和第 4 种植物的名称（使用向量索引）
4. 添加“慈姑”到植物名单中（使用 `c()` 函数合并）
5. 用 `paste()` 函数创建一句完整的调查报告
6. 尝试查找“芦苇”在向量中的位置（提示：使用 `which()` 函数和 `==` 运算符）

2.4 数据框结构理解与多类型数据管理

2.4.1 生态学背景

在植物群落调查中，需要同时记录多种信息：样方编号、物种名称、株高、是否存活等。这些不同类型的数据需要组织在一个表格中，这就需要用到数据框。

2.4.2 演示数据

```
# 某山坡 5 个样方的植物调查数据
plot_data <- data.frame(
  plot_id = c("S001", "S002", "S003", "S004", "S005"),
  species = c("马尾松", "杉木", "樟树", "栎树", "枫香"),
  height_m = c(12.5, 8.3, 15.2, 10.7, 9.8),
  diameter_cm = c(18.2, 12.5, 22.1, 16.8, 14.3),
  alive = c(TRUE, TRUE, FALSE, TRUE, TRUE)
)
```

2.4.3 课堂演示过程

```
# 1. 创建植物调查数据框
plot_data <- data.frame(
  plot_id = c("S001", "S002", "S003", "S004", "S005"),
  species = c(" 马尾松", " 杉木", " 樟树", " 栎树", " 枫香"),
  height_m = c(12.5, 8.3, 15.2, 10.7, 9.8),
  diameter_cm = c(18.2, 12.5, 22.1, 16.8, 14.3),
  alive = c(TRUE, TRUE, FALSE, TRUE, TRUE)
)

# 2. 查看数据框
plot_data
print(plot_data)

# 3. 查看数据框结构
str(plot_data) # 显示每列的数据类型

# 4. 查看数据框维度
nrow(plot_data) # 行数
ncol(plot_data) # 列数
dim(plot_data) # 行数和列数

# 5. 查看前几行和后几行
head(plot_data, 3) # 前 3 行
tail(plot_data, 2) # 后 2 行

# 6. 访问特定列
plot_data$species      # 物种列
plot_data$height_m     # 高度列

# 7. 计算统计量
mean(plot_data$height_m) # 平均高度
sum(plot_data$alive)     # 存活植物数量
```

2.4.4 R 语言知识点详解

2.4.4.1 数据框 (Data Frame)

- 是什么: R 中最重要的数据结构, 类似于 Excel 表格或数据库表
- 语法: `data.frame(列名 1 = 向量 1, 列名 2 = 向量 2, ...)`
- 核心特点:
 - 每列可以是不同的数据类型 (数值、字符、逻辑)
 - 但每列内部必须是相同类型
 - 所有列必须有相同的长度 (行数)
 - 每行代表一个观察单位, 每列代表一个变量
- 与矩阵的区别:
 - 矩阵: 所有元素必须是相同类型
 - 数据框: 不同列可以是不同类型
- 生态学应用: 完美适合存储野外调查数据

2.4.4.2 R 中的数据类型系统

2.4.4.2.1 数值型 (Numeric/Double)

- 特点: 包含小数点的数字
- 示例: `12.5, 8.3, 15.2`
- 用途: 测量值、计数、比例等

2.4.4.2.2 字符型 (Character)

- 特点: 文本数据, 用引号包围
- 示例: `"S001", " 马尾松", " 杉木"`
- 用途: 名称、标识符、分类标签

2.4.4.2.3 逻辑型 (Logical)

- 特点: 只有两个值: TRUE 和 FALSE

- 示例: TRUE, FALSE
- 用途: 是/否判断、条件标记
- 运算: 可以进行数学运算 (TRUE=1, FALSE=0)

2.4.4.3 数据框结构查看函数

2.4.4.3.1 **str()** - Structure 函数

- 作用: 显示数据框的完整结构信息
- 显示内容:
 - 数据框类型和维度
 - 每列的数据类型
 - 前几个值的预览
- 读法技巧:
 - 'data.frame': 5 obs. of 5 variables - 5 行 5 列的数据框
 - \$ plot_id : chr - plot_id 列是字符型
 - \$ height_m: num - height_m 列是数值型

2.4.4.3.2 **head()** 和 **tail()**

- 语法: head(x, n = 6), tail(x, n = 6)
- 作用: 查看数据框的开头或结尾几行
- 参数: n 指定显示的行数, 默认 6 行
- 应用场景:
 - 快速了解数据格式
 - 检查数据导入是否正确
 - 大数据集的初步查看

2.4.4.3.3 维度函数

- **nrow()**: 返回行数 (观察数量)
- **ncol()**: 返回列数 (变量数量)
- **dim()**: 返回维度向量 c(行数, 列数)
- **names()** 或 **colnames()**: 返回列名

2.4.4.4 数据框列访问

2.4.4.4.1 美元符号 \$ 操作符

- 语法: 数据框名 \$ 列名
- 特点:
 - 返回该列的向量
 - 支持名称自动补全 (在 RStudio 中)
 - 最常用的列访问方式
- 注意: 列名不需要引号

2.4.4.4.2 双括号 [[]] 操作符

- 语法: 数据框名 [["列名"]] 或 数据框名 [[列位置]]
- 与 \$ 的区别:
 - 支持变量名作为列名: df[[var_name]]
 - 可以使用数字索引: df[[1]]

2.4.4.4.3 单括号 [] 操作符

- 语法: 数据框名 [行, 列]
- 特点: 返回数据框 (保持原结构)
- 示例: plot_data[, "species"] - 选择 species 列但保持数据框格式

2.4.5 课后练习

题目: 某河流生态调查数据:

```
river_survey <- data.frame(  
  site_id = c("R01", "R02", "R03", "R04"),  
  fish_species = c("草鱼", "鲤鱼", "鲫鱼", "青鱼"),  
  length_cm = c(25.3, 18.7, 12.4, 31.2),  
  weight_g = c(680, 420, 180, 1200),
```

```

mature = c(TRUE, FALSE, FALSE, TRUE)
)

```

请完成（使用数据框操作、向量计算等已学内容）：

1. 创建数据框并显示整个数据框
2. 使用 `str()` 函数查看数据框的结构
3. 计算鱼类的平均长度和平均重量（使用 `mean()` 和 `$` 操作符）
4. 统计有多少条成熟的鱼（使用 `sum()` 和逻辑值运算）
5. 显示所有鱼类的名称（使用 `$` 操作符）
6. 计算最大和最小的鱼重量（使用 `max()` 和 `min()` 函数）
7. 创建一个包含调查总结的字符串（使用 `paste()` 函数）

2.5 列表数据结构与数据分组管理

2.5.1 生态学背景

不同栖息地类型的物种多样性差异很大。我们需要比较森林、草地、湿地三种生境中的物种数量，这种多组数据的管理需要用到列表结构。

2.5.2 演示数据

```

# 不同栖息地的物种数量调查（每个生境调查了 4 个样点）
forest_species <- c(25, 30, 28, 32)
grassland_species <- c(15, 18, 20, 16)
wetland_species <- c(12, 14, 11, 13)

```

2.5.3 课堂演示过程

```
# 1. 创建各栖息地物种数据
forest_species <- c(25, 30, 28, 32)
grassland_species <- c(15, 18, 20, 16)
wetland_species <- c(12, 14, 11, 13)

# 2. 创建栖息地数据列表
habitats <- list(
  forest = forest_species,
  grassland = grassland_species,
  wetland = wetland_species
)

# 3. 查看列表内容
habitats
str(habitats) # 查看列表结构

# 4. 访问列表中的元素
habitats$forest      # 使用 $ 访问
habitats[["forest"]] # 使用 [[]] 访问
habitats[[1]]         # 使用位置索引

# 5. 计算各生境的平均物种数
forest_mean <- mean(habitats$forest)
grassland_mean <- mean(habitats$grassland)
wetland_mean <- mean(habitats$wetland)

# 6. 创建结果向量
habitat_means <- c(forest_mean, grassland_mean, wetland_mean)
names(habitat_means) <- c("森林", "草地", "湿地")

# 7. 比较结果
```

```
habitat_means
max(habitat_means) # 哪个生境物种最多
which.max(habitat_means) # 物种最多的生境位置
```

2.5.4 R 语言知识点详解

2.5.4.1 列表 (List) 数据结构

- 是什么: R 中最灵活的数据结构, 可以存储不同类型、不同长度的数据
- 语法: `list(名称 1 = 数据 1, 名称 2 = 数据 2, ...)`
- 核心特点:
 - 每个元素可以是不同的数据类型 (向量、数据框、甚至其他列表)
 - 每个元素可以有不同的长度
 - 元素可以有名称, 也可以没有
 - 是递归数据结构 (可以包含其他列表)
- 与向量、数据框的区别:
 - 向量: 同类型, 一维
 - 数据框: 不同列可以不同类型, 但同列必须同类型, 二维表格
 - 列表: 最灵活, 可以存储任何类型的 R 对象

2.5.4.2 列表元素访问方法

2.5.4.2.1 美元符号 \$ 访问 (推荐)

- 语法: `列表名 $ 元素名`
- 特点:
 - 最直观、最常用的方法
 - 只能用于有名称的元素
 - 支持 RStudio 中的自动补全
 - 返回元素的原始类型

2.5.4.2.2 双括号 [[]] 访问

- 语法: 列表名 [[" 元素名 "]] 或 列表名 [[位置]]

- 特点:

- 更灵活, 可以使用变量作为索引
- 可以使用数字位置索引
- 返回元素的原始类型

- 示例:

```
element_name <- "forest"
habitats[[element_name]]    # 使用变量
habitats[[1]]                # 使用位置
```

2.5.4.2.3 单括号 [] 访问

- 语法: 列表名 [元素名或位置]
- 特点: 返回包含该元素的子列表 (仍然是列表类型)
- 与 [[]] 的区别:
 - `habitats[1]` 返回包含第一个元素的列表
 - `habitats[[1]]` 返回第一个元素本身 (向量)

2.5.4.3 向量命名系统

2.5.4.3.1 `names()` 函数

- 作用: 为向量的每个元素分配名称
- 语法: `names(向量) <- c(" 名称 1", " 名称 2", ...)`
- 好处:
 - 增加数据的可读性
 - 便于后续的数据访问和处理

- 在图表中自动显示有意义的标签

- 应用:

```
# 创建时命名
scores <- c(数学 = 95, 英语 = 88, 物理 = 92)

# 后续命名
scores <- c(95, 88, 92)
names(scores) <- c("数学", "英语", "物理")
```

2.5.4.4 比较和查找函数

2.5.4.4.1 max() 和 min()

- 作用: 找到向量中的最大值或最小值
- 语法: `max(x, na.rm = FALSE)`
- 参数: `na.rm` 控制是否忽略缺失值

2.5.4.4.2 which.max() 和 which.min()

- 作用: 返回最大值或最小值的位置索引
- 语法: `which.max(x)`
- 返回值: 整数, 表示最大值在向量中的位置
- 应用场景: 找到最优样地、最佳条件等
- 注意: 如果有多个相同的最大值, 只返回第一个的位置

2.5.4.5 数据组织策略

- 何时使用列表:
 - 存储相关但结构不同的数据
 - 分组存储实验数据
 - 存储分析结果的不同组成部分
- 命名的重要性:

- 提高代码可读性
- 便于数据访问
- 减少错误发生

- **最佳实践:**

- 使用有意义的名称
- 保持命名风格一致
- 适当添加注释说明数据来源

2.5.5 课后练习

题目：某保护区三个监测站的哺乳动物目击次数：

```
station_a <- c(8, 12, 6, 10)
station_b <- c(15, 18, 14, 16)
station_c <- c(3, 5, 2, 4)
```

请完成（使用向量、列表、命名等已学内容）：

1. 将三个监测站的数据组织成一个列表（使用 `list()` 函数）

2. 计算每个监测站的平均目击次数（使用 `mean()` 和列表访问）
3. 创建一个命名向量显示三个站点的平均值（使用 `names()` 函数）
4. 找出哪个监测站的平均目击次数最高（使用 `which.max()` 函数）
5. 计算所有监测站的总目击次数（使用 `sum()` 和向量合并）
6. 创建一个数据框，包含站点名称和对应的平均目击次数
7. 比较站点 A 和站点 B 的数据变异程度（使用 `sd()` 函数计算标准差）

2.6 外部数据导入与条件筛选分析

2.6.1 生态学背景

长期鸟类监测项目通常将数据保存在 Excel 或 CSV 文件中。我们需要学会将这些外部数据导入 R 中进行分析，并根据研究需要筛选特定时间段的数据。

2.6.2 演示数据文件 (**bird_monitoring.csv**)

```
date,month,species,count,observer
2023-03-15,3, 白头鵙,12, 张三
2023-03-15,3, 麻雀,25, 张三
2023-04-20,4, 喜鹊,8, 李四
2023-04-20,4, 白头鵙,15, 李四
2023-05-10,5, 燕子,20, 王五
2023-06-05,6, 麻雀,18, 张三
2023-06-05,6, 白头鵙,22, 张三
2023-07-12,7, 喜鹊,10, 李四
```

```
bird_data <- data.frame(
  date = c("2023-03-15", "2023-03-15", "2023-04-20",
  "2023-04-20",
  "2023-05-10", "2023-06-05", "2023-06-05",
  "2023-07-12"),
  month = c(3, 3, 4, 4, 5, 6, 6, 7),
  species = c("白头鵙", "麻雀", "喜鹊", "白头鵙", "燕子",
  "麻雀", "白头鵙", "喜鹊"),
  count = c(12, 25, 8, 15, 20, 18, 22, 10),
  observer = c("张三", "张三", "李四", "李四", "王五", "张三",
  "张三", "李四")
)

# 保存为 CSV 文件
write.csv(bird_data, "bird_observation_data.csv", row.names =
  FALSE, fileEncoding = "UTF-8")
# 注意, 这里的 fileEcoding 参数要依据 OS 系统编码设定
```

2.6.3 课堂演示过程

```
# 1. 读取鸟类监测数据
bird_data <- read.csv("bird_monitoring.csv", stringsAsFactors = FALSE)

# 2. 查看数据概况
head(bird_data)      # 前几行
tail(bird_data)      # 后几行
str(bird_data)        # 数据结构
summary(bird_data)   # 数据摘要

# 3. 查看数据维度
nrow(bird_data)     # 有多少条记录
ncol(bird_data)      # 有多少个变量

# 4. 查看具体列的信息
unique(bird_data$species)    # 观察到哪些物种
unique(bird_data$observer)   # 有哪些调查员
range(bird_data$month)       # 调查的月份范围

# 5. 筛选春季数据 (3-5 月)
spring_birds <- subset(bird_data, month %in% c(3, 4, 5))
spring_birds

# 6. 筛选特定物种
baijitou_data <- subset(bird_data, species == "白头鹀")
baijitou_data

# 7. 条件组合筛选
spring_baijitou <- subset(bird_data, month %in% c(3, 4, 5) & species == "白头鹀")
spring_baijitou

# 8. 计算统计量
```

```
total_count <- sum(bird_data$count)
mean_count <- mean(bird_data$count)
paste(" 总观察个体数:", total_count, " 平均每次观察:", round(mean_count,
```

2.6.4 R 语言知识点详解

2.6.4.1 数据导入函数 `read.csv()`

- 作用：从 CSV（逗号分隔值）文件中读取数据，创建数据框
- 语法：`read.csv(file, header = TRUE, sep = ", ", stringsAsFactors = FALSE, ...)`
- 重要参数详解：
 - `file`: 文件路径，可以是本地文件或网络 URL
 - `header = TRUE`: 第一行是否为列名，默认 `TRUE`
 - `sep = ", "`: 字段分隔符，CSV 默认逗号
 - `stringsAsFactors = FALSE`: 是否将字符串转换为因子，建议设为 `FALSE`
 - `encoding`: 文件编码，中文文件可能需要设置为”UTF-8”
- 文件路径注意事项：
 - Windows 系统使用反斜杠\，但 R 中需要转义\\或使用正斜杠/
 - 使用相对路径时，基于当前工作目录
 - 可用 `getwd()` 查看当前工作目录，`setwd()` 设置工作目录

2.6.4.2 数据概览函数集

2.6.4.2.1 `summary()` - 数据摘要

- 作用：提供每列数据的统计摘要
- 不同数据类型的摘要：
 - 数值型：最小值、第一四分位数、中位数、均值、第三四分位数、最大值
 - 字符型：长度、类别、模式
 - 因子型：各水平的频数
- 应用价值：快速了解数据分布、发现异常值

2.6.4.2.2 `unique()` - 唯一值

- 作用：返回向量中的所有不重复值
- 语法：`unique(x)`
- 应用场景：
 - 查看分类变量的所有类别
 - 检查数据录入是否有错误（如拼写错误）
 - 了解数据的多样性
- 相关函数：`duplicated()` 检查重复值

2.6.4.2.3 `range()` - 值域范围

- 作用：返回向量的最小值和最大值
- 语法：`range(x, na.rm = FALSE)`
- 返回值：长度为 2 的向量，`c(最小值, 最大值)`
- 应用：快速了解数据的取值范围

2.6.4.3 数据筛选系统

2.6.4.3.1 `subset()` 函数（推荐方式）

- 作用：根据条件筛选数据框的行
- 语法：`subset(x, subset, select)`
- 参数详解：
 - `x`: 要筛选的数据框
 - `subset`: 逻辑条件表达式
 - `select`: 选择的列（可选）
- 优势：语法简洁，不需要重复写数据框名称

2.6.4.3.2 逻辑操作符详解

2.6.4.3.2.1 %in% 操作符

- 作用：检查左边的值是否在右边的向量中
- 语法：`x %in% y`
- 示例：`month %in% c(3, 4, 5)` 检查月份是否是 3、4、5 中的一个
- 与 `==` 的区别：
 - `==` 只能比较单个值
 - `%in%` 可以同时比较多个值

2.6.4.3.2.2 == 等于操作符

- 作用：检查两个值是否相等
- 注意事项：
 - 区分大小写：`"A" == "a"` 为 FALSE
 - 精确匹配：`"cat" == "cats"` 为 FALSE
 - 用于字符串时必须完全匹配

2.6.4.3.2.3 & 逻辑与操作符

- 作用：连接多个条件，所有条件都必须为 TRUE
- 语法：条件 1 & 条件 2 & ...
- 相关操作符：
 - `|`：逻辑或，任一条件为 TRUE 即可
 - `!`：逻辑非，取反

2.6.4.4 4. 数值处理函数

2.6.4.4.1 round() - 四舍五入

- 语法：`round(x, digits = 0)`
- 参数：
 - `x`：要舍入的数值
 - `digits`：保留的小数位数
- 应用：美化输出结果，控制精度

2.6.4.5 数据导入最佳实践

- **文件检查**: 导入前先用文本编辑器查看文件格式
- **编码处理**: 中文数据注意编码问题
- **数据验证**: 导入后立即检查数据结构和内容
- **备份原始数据**: 避免在原始数据上直接修改
- **路径管理**: 使用项目文件夹, 保持文件路径的一致性

2.6.5 课后练习

题目: 假设有一个植被监测数据文件包含以下列:

- date: 调查日期
- season: 季节 (春、夏、秋、冬)
- plot: 样地编号
- coverage: 植被覆盖度 (%)
- height: 平均高度 (cm)

请完成 (使用数据导入、数据框操作、条件筛选等已学内容): 1. 创建模拟数据或读取数据文件

2. 查看数据的基本信息 (使用 nrow(), ncol(), str(), summary())
3. 筛选夏季的数据 (使用 subset() 函数)
4. 筛选植被覆盖度大于 80% 的记录 (使用 subset() 函数和条件)
5. 计算所有样地的平均植被覆盖度和平均高度 (使用 mean() 函数)
6. 找出覆盖度最高的样地 (使用 which.max() 函数)
7. 创建一个汇总报告 (使用 paste() 函数)

2.7 缺失值和异常值的识别处理

2.7.1 生态学背景

在水质监测中, 由于仪器故障、人为记录错误等原因, 经常出现缺失值和异常值。数据清理是生态学数据分析的重要步骤, 需要识别和处理这些问题数据。

2.7.2 演示数据

```
# 湖泊水质监测数据（包含缺失值和异常值）
water_quality <- data.frame(
  site_id = c("湖心", "入水口", "出水口", "湖心", "入水口", "出水口",
  date = c("2023-05-01", "2023-05-01", "2023-05-01", "2023-05-15", "2023-05-15"),
  pH = c(7.2, 6.8, NA, 7.5, 6.9, 7.1),
  temperature_C = c(18.5, 19.2, 20.1, 999, 19.8, 20.3), # 999 为仪器
  dissolved_oxygen = c(8.2, 7.5, 8.8, 8.1, NA, 8.4)
)
```

2.7.3 课堂演示过程

```
# 1. 创建包含问题的水质数据
water_quality <- data.frame(
  site_id = c("湖心", "入水口", "出水口", "湖心", "入水口", "出水口",
  date = c("2023-05-01", "2023-05-01", "2023-05-01", "2023-05-15", "2023-05-15"),
  pH = c(7.2, 6.8, NA, 7.5, 6.9, 7.1),
  temperature_C = c(18.5, 19.2, 20.1, 999, 19.8, 20.3),
  dissolved_oxygen = c(8.2, 7.5, 8.8, 8.1, NA, 8.4)
)

# 2. 查看原始数据
print(water_quality)
str(water_quality)

# 3. 检查缺失值
is.na(water_quality) # 显示所有缺失值位置
sum(is.na(water_quality$pH)) # pH 缺失值个数
sum(is.na(water_quality$dissolved_oxygen)) # 溶解氧缺失值个数

# 4. 识别异常值
```

```
summary(water_quality$temperature_C) # 查看温度的统计摘要
water_quality$temperature_C > 50 # 找出不合理的高温值

# 5. 处理缺失值 - 用均值填补
ph_mean <- mean(water_quality$pH, na.rm = TRUE) # 计算 pH 均值 (忽略 NA)
water_quality$pH[is.na(water_quality$pH)] <- ph_mean

# 6. 处理异常值 - 替换为 NA
water_quality$temperature_C[water_quality$temperature_C > 50] <- NA

# 7. 查看清理后的数据
print(water_quality)

# 8. 删除包含 NA 的整行 (如果需要)
clean_data <- na.omit(water_quality)
print(clean_data)

# 9. 计算清理后的统计量
mean(clean_data$temperature_C)
mean(clean_data$pH)
mean(clean_data$dissolved_oxygen)
```

2.7.4 R 语言知识点详解

2.7.4.1 缺失值 (Missing Values) 处理系统

2.7.4.1.1 缺失值的概念

- 什么是 NA: Not Available 的缩写, 表示缺失或不可用的数据
- NA 的特点:
 - 任何包含 NA 的运算结果都是 NA
 - NA 具有传染性: $1 + \text{NA} = \text{NA}$
 - NA 不等于任何值, 包括它自己: $\text{NA} == \text{NA}$ 返回 NA 而不是 TRUE

2.7.4.1.2 `is.na()` 函数

- 作用：检测缺失值的位置
- 语法：`is.na(x)`
- 返回值：与输入同样结构的逻辑向量/矩阵，`TRUE` 表示缺失
- 应用方式：
 - 检查单个向量：`is.na(vector)`
 - 检查整个数据框：`is.na(data.frame)`
 - 统计缺失值数量：`sum(is.na(vector))`

2.7.4.1.3 `na.rm` 参数

- 作用：在统计计算中移除缺失值
- 语法：`function(x, na.rm = FALSE)`
- 适用函数：`mean()`, `sum()`, `max()`, `min()`, `sd()` 等
- 重要性：不设置 `na.rm = TRUE` 时，有 `NA` 的计算结果都是 `NA`
- 示例对比：

```
x <- c(1, 2, NA, 4)
mean(x)           # 返回 NA
mean(x, na.rm = TRUE) # 返回 2.33
```

2.7.4.1.4 `na.omit()` 函数

- 作用：删除包含任何缺失值的完整行
- 语法：`na.omit(x)`
- 返回值：不含任何 `NA` 的数据框
- 注意事项：
 - 可能导致大量数据丢失
 - 需要评估删除行对分析的影响
 - 适合缺失值较少且随机分布的情况

2.7.4.2 异常值 (Outliers) 识别与处理

2.7.4.2.1 异常值的识别方法

- **统计方法**: 使用 `summary()` 查看数据分布, 识别明显不合理的值
- **业务逻辑**: 基于专业知识判断, 如温度 999°C 明显错误
- **可视化方法**: 使用箱线图、散点图等发现异常值
- **统计阈值**: 如超出 3 倍标准差的值

2.7.4.2.2 异常值处理策略

1. **删除异常值**: 适用于明显的录入错误
2. **替换为 NA**: 保留数据结构, 标记为缺失
3. **替换为合理值**: 用中位数、均值等替换
4. **保留但标记**: 在分析中特殊处理

2.7.4.3 条件替换技术

2.7.4.3.1 逻辑索引替换

- **语法**: `data[condition] <- new_value`
- **原理**: 通过逻辑条件选择满足条件的元素进行替换
- **示例**:

```
# 将所有负值替换为 0
data[data < 0] <- 0

# 将异常高值替换为 NA
data[data > threshold] <- NA
```

2.7.4.3.2 `which()` 函数

- 作用：返回满足条件的元素位置索引
- 语法：`which(condition)`
- 与直接逻辑索引的区别：
 - 逻辑索引：返回 TRUE/FALSE 向量
 - `which()`：返回位置数字向量
- 应用：当需要知道具体位置时使用

2.7.4.4 数据清理流程和最佳实践

2.7.4.4.1 标准数据清理流程

1. 数据探索：使用 `str()`, `summary()`, `head()`, `tail()` 了解数据
2. 缺失值检查：使用 `is.na()`, `sum(is.na())` 统计缺失情况
3. 异常值识别：结合统计和专业知识识别异常值
4. 清理决策：选择合适的处理方法
5. 执行清理：应用处理方法
6. 验证结果：检查清理后的数据质量

2.7.4.4.2 数据清理的注意事项

- 保留原始数据：清理前备份原始数据
- 记录清理过程：文档化所有清理步骤和决策理由
- 验证合理性：确保清理后的数据符合业务逻辑
- 评估影响：分析清理对后续分析的影响

2.7.4.4.3 缺失值填补方法选择

- 均值填补：适用于数值变量，数据接近正态分布
- 中位数填补：适用于有偏斜的数值变量

- **众数填补**: 适用于分类变量
- **前向/后向填补**: 适用于时间序列数据
- **预测模型填补**: 基于其他变量预测缺失值

2.7.5 课后练习

题目: 某森林土壤调查数据:

```
soil_data <- data.frame(  
  plot = c("A1", "A2", "A3", "B1", "B2", "B3"),  
  organic_matter = c(3.2, NA, 2.8, 3.5, 2.9, 3.1),  
  nitrogen_mg = c(45, 52, -10, 48, 51, 49), # -10 为异常负值  
  moisture = c(25.5, 28.2, 22.1, NA, 26.8, 24.9)  
)
```

请完成（使用缺失值处理、条件判断、数据清理等已学内容）：

1. 检查每列的缺失值个数（使用 `is.na()` 和 `sum()` 函数）

2. 识别 `nitrogen_mg` 列中的异常值（使用逻辑判断和 `which()` 函数）
3. 用均值填补 `organic_matter` 的缺失值（使用 `mean()` 和 `na.rm` 参数）
4. 将 `nitrogen_mg` 中的异常值替换为 `NA`（使用条件赋值）
5. 创建一个完全没有缺失值的干净数据集（使用 `na.omit()`）
6. 计算清理后数据的各项平均值（使用 `mean()` 函数）
7. 对比清理前后数据的 `summary()` 结果

2.8 描述性统计分析与基础数据可视化

2.8.1 生态学背景

不同森林类型的物种多样性存在显著差异。通过比较松林、栎林、混交林的物种数量，我们可以了解森林结构对生物多样性的影响。这需要用到描述性统计和基础可视化。

2.8.2 演示数据

```
# 三种森林类型各 5 个样地的物种数量
pine_forest <- c(22, 25, 20, 28, 24)          # 松林
oak_forest <- c(35, 32, 38, 30, 34)           # 栎林
mixed_forest <- c(45, 42, 48, 40, 46)         # 混交林
```

2.8.3 课堂演示过程

```
# 1. 创建三种森林类型数据
pine_forest <- c(22, 25, 20, 28, 24)
oak_forest <- c(35, 32, 38, 30, 34)
mixed_forest <- c(45, 42, 48, 40, 46)

# 2. 计算描述性统计
# 平均值
pine_mean <- mean(pine_forest)
oak_mean <- mean(oak_forest)
mixed_mean <- mean(mixed_forest)

# 标准差
pine_sd <- sd(pine_forest)
oak_sd <- sd(oak_forest)
mixed_sd <- sd(mixed_forest)

# 最大值和最小值
range(pine_forest)
range(oak_forest)
range(mixed_forest)

# 3. 创建汇总表
forest_summary <- data.frame(
```

```
森林类型 = c("松林", "栎林", "混交林"),
平均物种数 = c(pine_mean, oak_mean, mixed_mean),
标准差 = c(pine_sd, oak_sd, mixed_sd),
最大值 = c(max(pine_forest), max(oak_forest), max(mixed_forest)),
最小值 = c(min(pine_forest), min(oak_forest), min(mixed_forest))
)
print(forest_summary)

# 4. 箱线图比较
boxplot(pine_forest, oak_forest, mixed_forest,
         names = c("松林", "栎林", "混交林"),
         ylab = "物种数量",
         main = "不同森林类型物种多样性比较",
         col = c("lightgreen", "lightblue", "lightyellow"))

# 5. 添加平均值点
points(1:3, c(pine_mean, oak_mean, mixed_mean),
        col = "red", pch = 19, cex = 1.5)

# 6. 条形图显示平均值
barplot(c(pine_mean, oak_mean, mixed_mean),
        names.arg = c("松林", "栎林", "混交林"),
        ylab = "平均物种数",
        main = "各森林类型平均物种数量",
        col = c("lightgreen", "lightblue", "lightyellow"))

# 7. 方差分析 (简单介绍)
all_data <- c(pine_forest, oak_forest, mixed_forest)
forest_type <- rep(c("松林", "栎林", "混交林"), each = 5)
forest_df <- data.frame(species_count = all_data, type = forest_type)
```

2.8.4 R 语言知识点详解

2.8.4.1 描述性统计函数深入解析

2.8.4.1.1 `sd()` - 标准差函数

- 作用：计算样本标准差，衡量数据的离散程度
- 语法：`sd(x, na.rm = FALSE)`
- 数学含义：
 - 标准差越大，数据越分散
 - 标准差越小，数据越集中在均值附近
 - 单位与原数据相同
- 与方差的关系：标准差 = $\sqrt{\text{方差}}$
- 相关函数：
 - `var()`：计算方差
 - `mad()`：计算中位数绝对偏差（对异常值更稳健）

2.8.4.1.2 `range()` - 值域函数

- 作用：返回最小值和最大值组成的向量
- 语法：`range(x, na.rm = FALSE)`
- 返回值：长度为 2 的数值向量 `c(min, max)`
- 应用：
 - 快速了解数据的取值范围
 - 检查数据是否在合理范围内
 - 设置图形的坐标轴范围

2.8.4.2 R 基础绘图系统详解

2.8.4.2.1 `boxplot()` - 箱线图函数

- 作用：绘制箱线图，显示数据的分布特征
- 语法：`boxplot(..., names, main, xlab, ylab, col)`

- 箱线图解读:

- 盒子: 第一四分位数 (Q1) 到第三四分位数 (Q3), 包含 50% 的数据
- 中线: 中位数 (Q2)
- 须线: 延伸到 1.5 倍四分位数间距的范围
- 点: 超出须线的异常值

- 参数详解:

- **names**: 各组的标签
- **main**: 图形标题
- **xlab, ylab**: x 轴和 y 轴标签
- **col**: 填充颜色
- **border**: 边框颜色
- **notch**: 是否显示置信区间缺口

2.8.4.2.2 **barplot()** - 条形图函数

- 作用: 绘制条形图, 比较不同组的数值

- 语法: `barplot(height, names.arg, main, xlab, ylab, col)`

- 参数详解:

- **height**: 条形的高度值
- **names.arg**: 条形的标签
- **beside**: 并排显示多组数据时设为 TRUE
- **horiz**: 是否绘制水平条形图

- 应用场景:

- 比较不同组的均值
- 显示分类数据的频数
- 展示比例或百分比

2.8.4.3 图形参数和美化

2.8.4.3.1 颜色参数 **col**

- 预定义颜色: "red", "blue", "green" 等

- 颜色名称: `colors()` 查看所有可用颜色名称
- 十六进制: "#FF0000" (红色)
- RGB 函数: `rgb(1, 0, 0)` (红色)
- 颜色向量: 为不同元素指定不同颜色

2.8.4.3.2 点的形状参数 `pch`

- 常用形状:
 - `pch = 1`: 空心圆 □
 - `pch = 19`: 实心圆 □
 - `pch = 2`: 空心三角形 □
 - `pch = 17`: 实心三角形 □
 - `pch = 15`: 实心方形 □
- 字符形状: `pch = "A"` 使用字符 A 作为点

2.8.4.3.3 大小参数 `cex`

- 作用: 控制图形元素的大小
- 默认值: 1.0
- 用法:
 - `cex = 1.5`: 放大 1.5 倍
 - `cex = 0.8`: 缩小为 0.8 倍
- 相关参数:
 - `cex.main`: 标题大小
 - `cex.lab`: 轴标签大小
 - `cex.axis`: 轴数字大小

2.8.4.4 图形叠加和增强

2.8.4.4.1 `points()` - 添加点

- 作用: 在现有图形上添加点

- 语法: `points(x, y, col, pch, cex)`
- 坐标系统: 使用与原图相同的坐标系统
- 应用: 在箱线图上标记均值、在散点图上突出特定点

2.8.4.4.2 图形叠加的原理

- 图层概念: R 绘图采用图层叠加的方式
- 顺序重要: 后绘制的元素会覆盖先绘制的元素
- 坐标统一: 所有叠加元素必须使用相同的坐标系统

2.8.4.5 数据重组和整理

2.8.4.5.1 `rep()` - 重复函数

- 作用: 重复向量元素
- 语法: `rep(x, times, each, length.out)`
- 参数说明:
 - `times`: 整个向量重复的次数
 - `each`: 每个元素重复的次数
 - `length.out`: 输出向量的长度
- 示例:

```
rep(c("A", "B"), times = 2)      # "A" "B" "A" "B"  
rep(c("A", "B"), each = 2)        # "A" "A" "B" "B"  
rep(c("A", "B"), length.out = 5)  # "A" "B" "A" "B" "A"
```

2.8.4.6 图形设计最佳实践

2.8.4.6.1 色彩选择原则

- 对比度: 确保不同组别容易区分
- 色盲友好: 避免仅依赖红绿色区分
- 一致性: 同一类型数据使用相同色系
- 专业性: 避免过于鲜艳的颜色

2.8.4.6.2 标签和标题

- **信息完整**: 包含变量名称和单位
- **简洁明了**: 避免过长的标题
- **中文支持**: 确保中文字符正确显示

2.8.4.6.3 图形尺寸和比例

- **合适的比例**: 避免图形过于压缩或拉伸
- **合理的尺寸**: 适合展示媒介的大小
- **留白空间**: 给图形元素足够的空间

2.8.5 课后练习

题目：某保护区三种植被类型的蝴蝶物种数调查：

```
shrubland <- c(12, 15, 10, 14, 13)      # 灌丛
meadow <- c(18, 22, 20, 19, 21)          # 草甸
riparian <- c(25, 28, 23, 27, 26)        # 河岸林
```

请完成（使用描述统计、基础绘图等已学内容）：

1. 计算三种植被类型的平均物种数和标准差（使用 `mean()` 和 `sd()` 函数）
2. 创建一个汇总表显示基本统计信息（使用 `data.frame()`）
3. 绘制箱线图比较三种植被类型（使用 `boxplot()` 函数）
4. 绘制条形图显示平均物种数（使用 `barplot()` 函数）
5. 在箱线图上添加平均值点（使用 `points()` 函数）
6. 判断哪种植被类型的蝴蝶多样性最高（使用 `which.max()` 和 `max()` 函数）
7. 计算每种植被类型的变异系数（标准差/平均值 ×100）

2.9 条件判断、循环结构与函数编程

2.9.1 生态学背景

在群落生态学研究中，经常需要根据不同条件对物种进行分类处理，或者对大量样地数据进行批量处理。这需要用到编程中的条件判断和循环结构，让 R 能够自动化完成重复性工作。

2.9.2 演示数据

```
# 某自然保护区不同海拔的物种调查数据
sites_data <- data.frame(
  site_id = paste0("S", 1:10),
  elevation = c(1200, 1450, 1800, 2100, 2350, 1650, 1900, 2200, 1750),
  species_count = c(45, 52, 38, 28, 22, 48, 35, 25, 42, 30),
  dominant_species = c("栎树", "栎树", "云杉", "冷杉", "高山杜鹃",
)})
```

2.9.3 课堂演示过程

2.9.3.1 条件判断基础

```
# 创建示例数据
sites_data <- data.frame(
  site_id = paste0("S", 1:10),
  elevation = c(1200, 1450, 1800, 2100, 2350, 1650, 1900, 2200, 1750),
  species_count = c(45, 52, 38, 28, 22, 48, 35, 25, 42, 30),
  dominant_species = c("栎树", "栎树", "云杉", "冷杉", "高山杜鹃",
))

# 简单的 if 语句
elevation_threshold <- 2000
```

```

if (sites_data$elevation[1] > elevation_threshold) {
  print("高海拔样地")
} else {
  print("低海拔样地")
}

# if-else 判断所有样地
for (i in 1:nrow(sites_data)) {
  if (sites_data$elevation[i] > 2000) {
    print(paste(sites_data$site_id[i], "是高海拔样地"))
  } else {
    print(paste(sites_data$site_id[i], "是低海拔样地"))
  }
}

```

2.9.3.2 向量化条件判断

```

# 使用 ifelse() 函数进行向量化判断
sites_data$elevation_zone <- ifelse(sites_data$elevation > 2000, "高海拔样地", "低海拔样地")
print(sites_data[, c("site_id", "elevation", "elevation_zone")])

# 多层条件判断
sites_data$vegetation_type <- ifelse(sites_data$elevation < 1500, "荒漠",
                                         ifelse(sites_data$elevation < 2000, "草原", "森林"))
print(sites_data[, c("site_id", "elevation", "vegetation_type")])

```

2.9.3.3 for 循环处理

```

# 计算每个样地的多样性指数类别
diversity_categories <- character(nrow(sites_data))

```

```
for (i in 1:nrow(sites_data)) {  
  species_num <- sites_data$species_count[i]  
  if (species_num >= 40) {  
    diversity_categories[i] <- "高多样性"  
  } else if (species_num >= 30) {  
    diversity_categories[i] <- "中等多样性"  
  } else {  
    diversity_categories[i] <- "低多样性"  
  }  
}  
  
sites_data$diversity_category <- diversity_categories  
print(sites_data[, c("site_id", "species_count", "diversity_category")])
```

2.9.3.4 自定义函数编写

```
# 编写海拔带判断函数  
classify_elevation_zone <- function(elevation) {  
  if (elevation < 1500) {  
    return("低山带")  
  } else if (elevation < 2000) {  
    return("中山带")  
  } else {  
    return("高山带")  
  }  
}  
  
# 测试函数  
classify_elevation_zone(1800)  
classify_elevation_zone(2200)  
  
# 批量应用函数
```

```
sites_data$elevation_belt <- sapply(sites_data$elevation, classify_elevation)
print(sites_data[, c("site_id", "elevation", "elevation_belt")])
```

2.9.3.5 复杂条件处理

```
# 编写综合评估函数
assess_conservation_value <- function(elevation, species_count, dominant_sp) {
  score <- 0

  # 海拔因子
  if (elevation > 2000) {
    score <- score + 2
  } else if (elevation > 1500) {
    score <- score + 1
  }

  # 物种多样性因子
  if (species_count > 40) {
    score <- score + 2
  } else if (species_count > 30) {
    score <- score + 1
  }

  # 优势种稀有性因子
  rare_species <- c("高山杜鹃", "冷杉")
  if (dominant_sp %in% rare_species) {
    score <- score + 1
  }

  # 返回保护价值等级
  if (score >= 4) {
    return("极高价值")
  }
}
```

```
    } else if (score >= 3) {
      return("高价值")
    } else if (score >= 2) {
      return("中等价值")
    } else {
      return("一般价值")
    }
}

# 应用综合评估
sites_data$conservation_value <- mapply(assess_conservation_value,
                                             sites_data$elevation,
                                             sites_data$species_count,
                                             sites_data$dominant_species)

print(sites_data[, c("site_id", "conservation_value")])
```

2.9.4 R 语言知识点详解

2.9.4.1 条件判断结构

2.9.4.1.1 **if** 语句

- 语法: **if** (条件) { 执行代码 }
- 条件: 必须是逻辑值 (TRUE/FALSE)
- 执行规则: 条件为 TRUE 时执行大括号内的代码
- 注意事项: 条件必须是长度为 1 的逻辑向量

2.9.4.1.2 **if-else** 语句

- 语法:

```
if (条件) {
  # 条件为 TRUE 时执行
} else {
  # 条件为 FALSE 时执行
}
```

- **多重条件:** `else if` 可以链式连接
- **最佳实践:** 始终使用大括号, 即使只有一行代码

2.9.4.1.3 `ifelse()` 函数 (向量化)

- **语法:** `ifelse(test, yes, no)`
- **优势:** 可以处理向量, 一次性判断多个元素
- **参数:**
 - `test`: 逻辑向量条件
 - `yes`: 条件为 TRUE 时返回的值
 - `no`: 条件为 FALSE 时返回的值
- **嵌套使用:** 可以嵌套实现多重条件判断

2.9.4.2 循环结构

2.9.4.2.1 `for` 循环

- **语法:** `for (变量 in 序列) { 循环体 }`
- **常见用法:**

```
# 按索引循环
for (i in 1:10) { }

# 按元素循环
for (item in vector) { }
```

```
# 按名称循环
for (name in names(list)) { }
```

- 循环控制:

- **break**: 跳出循环
- **next**: 跳过当前迭代

2.9.4.2.2 其他循环类型

- **while** 循环: while (条件) { 循环体 }
- **repeat** 循环: repeat { 循环体; if(条件) break }

2.9.4.3 函数定义

2.9.4.3.1 基本函数语法

- 语法:

```
函数名 <- function(参数 1, 参数 2 = 默认值) {
  # 函数体
  return(返回值)
}
```

- 参数:

- 必需参数: 调用时必须提供
- 可选参数: 有默认值, 可省略

- 返回值:

- 显式返回: 使用 **return()**
- 隐式返回: 函数最后一个表达式的值

2.9.4.3.2 函数设计原则

- 单一职责：一个函数只做一件事
- 参数验证：检查输入参数的有效性
- 错误处理：使用 `stop()`、`warning()` 处理异常
- 文档化：添加注释说明函数用途和参数

2.9.4.4 高级应用函数

2.9.4.4.1 `sapply()` - 简化的 `apply`

- 作用：对向量或列表的每个元素应用函数
- 语法：`sapply(X, FUN, ...)`
- 返回值：简化后的向量或矩阵
- 与 `lapply()` 的区别：
 - `lapply()` 总是返回列表
 - `sapply()` 尝试简化结果

2.9.4.4.2 `mapply()` - 多变量 `apply`

- 作用：同时对多个向量应用函数
- 语法：`mapply(FUN, ..., MoreArgs = NULL)`
- 应用场景：函数需要多个参数时使用
- 示例：`mapply(function(x, y) x + y, vector1, vector2)`

2.9.4.5 逻辑运算符

2.9.4.5.1 基本逻辑运算符

- `==`: 等于
- `!=`: 不等于
- `>、<、>=、<=**`: 比较运算符
- `&`: 与（向量化）
- `|`: 或（向量化）

- `!`: 非
- `&&`、`||**`: 短路逻辑运算符（只判断第一个元素）

2.9.4.5.2 成员测试

- `%in%`: 检查元素是否在向量中
- `is.na()`: 检查缺失值
- `is.null()`: 检查空值

2.9.4.6 编程最佳实践

2.9.4.6.1 代码组织

- **缩进**: 使用一致的缩进（建议 2 或 4 个空格）
- **命名**: 使用有意义的变量名和函数名
- **注释**: 解释复杂逻辑和算法思路
- **模块化**: 将复杂任务分解为简单函数

2.9.4.6.2 性能考虑

- **向量化**: 优先使用向量化操作而非循环
- **预分配**: 循环前预分配存储空间
- **避免增长**: 不要在循环中动态增长向量

2.9.4.6.3 调试技巧

- `print()`: 在关键位置输出变量值
- `browser()`: 设置断点进行交互式调试
- `traceback()`: 查看错误调用堆栈
- **分步测试**: 逐步测试函数的各个部分

2.9.5 课后练习

题目：某湿地鸟类监测数据包含以下信息：

```

bird_monitoring <- data.frame(
  site = c("A1", "A2", "B1", "B2", "C1", "C2"),
  water_depth = c(15, 25, 45, 35, 65, 55), # 水深 (cm)
  bird_abundance = c(8, 12, 20, 16, 5, 8), # 鸟类丰度
  season = c("春季", "春季", "夏季", "夏季", "秋季", "秋季")
)

```

请完成（使用 if-else、循环、函数等编程内容，结合之前学过的数据处理方法）：

1. 使用 `ifelse()` 函数，根据水深将栖息地分类 (<30cm 浅水区, 30-50cm 中等深度, >50cm 深水区)
2. 编写函数 `classify_habitat_quality()`, 综合水深和鸟类丰度评估栖息地质量
3. 使用 `for` 循环，计算每个季节的平均鸟类丰度
4. 创建一个新列，标记高丰度样地（丰度 >15 为高丰度，使用 `ifelse()`）
5. 编写函数处理整个数据集，输出每个样地的综合评估报告
6. 使用 `apply` 族函数重做第 3 题（比较循环和向量化方法的差异）

2.10 现代数据科学工具包应用

2.10.1 生态学背景

现代生态学研究产生的数据日益复杂，传统的基础 R 语法在处理复杂数据操作时略显繁琐。`tidyverse` 是 R 语言的现代数据处理工具包，提供了更直观、更高效的数据处理方法，特别适合处理多变量、多时间点的生态学数据。

2.10.2 演示数据

```
# 模拟某森林样地多年监测数据
library(tidyverse)

# 创建模拟数据
forest_monitoring <- data.frame(
  plot_id = rep(paste0("Plot_", 1:5), each = 12),
  year = rep(2018:2021, times = 15),
  season = rep(c(" 春", " 夏", " 秋"), times = 20),
  temperature = rnorm(60, mean = 15, sd = 5),
  humidity = rnorm(60, mean = 70, sd = 10),
  species_richness = rpois(60, lambda = 25),
  tree_height = rnorm(60, mean = 12, sd = 3),
  soil_ph = rnorm(60, mean = 6.5, sd = 0.5)
)
```

2.10.3 课堂演示过程

2.10.3.1 tidyverse 包的加载和数据查看

```
# 安装和加载 tidyverse
# install.packages("tidyverse")
library(tidyverse)

# 创建示例数据（简化版）
forest_data <- tibble(
  plot_id = rep(c("A", "B", "C", "D"), each = 6),
  year = rep(2020:2022, times = 8),
  season = rep(c(" 春", " 夏"), times = 12),
  temperature = c(12, 18, 15, 22, 10, 16, 14, 20, 13, 19, 11, 17,
                 16, 21, 14, 20, 12, 18, 15, 23, 13, 21, 11, 19),
  species_count = c(22, 28, 25, 32, 20, 24, 26, 30, 24, 28, 21, 25,
                   28, 34, 26, 32, 23, 27, 30, 36, 27, 31, 24, 28)
```

```
)  
  
# 查看数据结构  
glimpse(forest_data)  
head(forest_data)
```

2.10.3.2 数据筛选与选择

```
# 使用 filter() 筛选行  
summer_data <- forest_data %>%  
  filter(season == "夏")  
  
high_diversity <- forest_data %>%  
  filter(species_count > 30)  
  
recent_summer <- forest_data %>%  
  filter(year >= 2021 & season == "夏")  
  
# 使用 select() 选择列  
temp_species <- forest_data %>%  
  select(plot_id, temperature, species_count)  
  
# 选择特定范围的列  
core_variables <- forest_data %>%  
  select(plot_id:season, species_count)  
  
# 排除特定列  
without_year <- forest_data %>%  
  select(-year)
```

2.10.3.3 数据变换与新变量创建

```
# 使用 mutate() 创建新变量
forest_enhanced <- forest_data %>%
  mutate(
    temp_category = case_when(
      temperature < 15 ~ "低温",
      temperature < 20 ~ "中温",
      TRUE ~ "高温"
    ),
    diversity_index = species_count / 10, # 简化的多样性指数
    temp_celsius = temperature,
    temp_fahrenheit = temperature * 9/5 + 32
  )

# 查看结果
forest_enhanced %>%
  select(plot_id, temperature, temp_category, species_count, diversi
```

2.10.3.4 数据排序与分组汇总

```
# 使用 arrange() 排序
forest_data %>%
  arrange(desc(species_count)) %>%
  head()

forest_data %>%
  arrange(plot_id, year, season)

# 使用 group_by() 和 summarise() 进行分组统计
plot_summary <- forest_data %>%
  group_by(plot_id) %>%
```

```

summarise(
  n_observations = n(),
  mean_temperature = mean(temperature),
  mean_species = mean(species_count),
  max_species = max(species_count),
  sd_temperature = sd(temperature),
  .groups = 'drop'
)

print(plot_summary)

# 多变量分组
season_plot_summary <- forest_data %>%
  group_by(season, plot_id) %>%
  summarise(
    mean_temp = mean(temperature),
    mean_species = mean(species_count),
    .groups = 'drop'
)

print(season_plot_summary)

```

2.10.3.5 数据重塑：长宽格式转换

```

# 宽格式转长格式 (gather/pivot_longer)
forest_long <- forest_data %>%
  pivot_longer(
    cols = c(temperature, species_count),
    names_to = "variable",
    values_to = "value"
)

```

```
head(forest_long)

# 长格式转宽格式 (spread/pivot_wider)
forest_wide <- forest_long %>%
  pivot_wider(
    names_from = variable,
    values_from = value
  )

head(forest_wide)
```

2.10.3.6 数据连接

```
# 创建额外的样地信息
plot_info <- tibble(
  plot_id = c("A", "B", "C", "D"),
  elevation = c(1200, 1450, 1800, 1600),
  soil_type = c("壤土", "砂土", "黏土", "壤土"),
  management = c("保护", "管理", "保护", "管理")
)

# 左连接
forest_complete <- forest_data %>%
  left_join(plot_info, by = "plot_id")

head(forest_complete)

# 按管理类型分析
management_analysis <- forest_complete %>%
  group_by(management) %>%
  summarise(
    mean_temperature = mean(temperature),
```

```

mean_species = mean(species_count),
  .groups = 'drop'
)

print(management_analysis)

```

2.10.4 R 语言知识点详解

2.10.4.1 tidyverse 哲学与管道操作

2.10.4.1.1 管道操作符 %>%

- 作用：将左侧结果作为右侧函数的第一个参数
- 优势：
 - 代码更易读：从左到右，从上到下
 - 减少中间变量：避免创建临时对象
 - 链式操作：多个操作连续进行
- 语法: `data %>% function()`
- 等价写法: `function(data)`

2.10.4.1.2 tibble vs data.frame

- **tibble** 特点：
 - 更好的打印输出
 - 更严格的子集操作
 - 保持字符串为字符串（不自动转因子）
 - 支持列名包含空格和特殊字符

2.10.4.2 数据筛选与选择

2.10.4.2.1 filter() - 行筛选

- 语法: `filter(data, condition1, condition2, ...)`

- 多条件:

- 逗号分隔: 逻辑与 (AND)
- |: 逻辑或 (OR)
- !: 逻辑非 (NOT)

- 常用条件:

- ==、 !=: 等于、不等于
- >、<、>=: 大小比较
- %in%: 成员检查
- is.na(): 缺失值检查

2.10.4.2.2 `select()` - 列选择

- 基本选择: `select(data, col1, col2)`
- 范围选择: `select(data, col1:col3)`
- 排除选择: `select(data, -col1, -col2)`
- 辅助函数:
 - `starts_with()`: 以某字符开头
 - `ends_with()`: 以某字符结尾
 - `contains()`: 包含某字符
 - `matches()`: 正则表达式匹配

2.10.4.3 数据变换

2.10.4.3.1 `mutate()` - 新变量创建

- 基本用法: `mutate(data, new_col = expression)`
- 多变量: 可同时创建多个新变量
- 引用新建变量: 在同一个 `mutate()` 中可引用前面创建的变量
- 变量类型转换:
 - `as.numeric()`: 转数值
 - `as.character()`: 转字符
 - `as.factor()`: 转因子

2.10.4.3.2 `case_when()` - 多条件分类

- 语法:

```
case_when(
  condition1 ~ value1,
  condition2 ~ value2,
  TRUE ~ default_value
)
```

- 优势: 替代复杂的嵌套 `ifelse()`
- 注意: 条件从上到下评估, 满足即停止

2.10.4.4 数据排序与汇总

2.10.4.4.1 `arrange()` - 数据排序

- 基本排序: `arrange(data, col)`
- 降序排序: `arrange(data, desc(col))`
- 多列排序: `arrange(data, col1, col2)`

2.10.4.4.2 `group_by()` 与 `summarise()`

- 分组概念: 将数据按指定变量分组
- 汇总函数:
 - `n()`: 计数
 - `mean()`、`median()`: 均值、中位数
 - `sum()`、`min()`、`max()`: 求和、最小值、最大值
 - `sd()`、`var()`: 标准差、方差
- `.groups` 参数: 控制结果的分组状态

2.10.4.5 数据重塑

2.10.4.5.1 长宽格式概念

- 宽格式：每个变量一列，观察单位一行
- 长格式：变量名和变量值分别存储在不同列中
- 选择原则：
 - 分析时通常用长格式
 - 展示时通常用宽格式

2.10.4.5.2 `pivot_longer()` - 宽转长

- 语法: `pivot_longer(data, cols, names_to, values_to)`
- 参数：
 - `cols`: 要转换的列
 - `names_to`: 存储变量名的新列名
 - `values_to`: 存储变量值的新列名

2.10.4.5.3 `pivot_wider()` - 长转宽

- 语法: `pivot_wider(data, names_from, values_from)`
- 参数：
 - `names_from`: 提供新列名的列
 - `values_from`: 提供新列值的列

2.10.4.6 数据连接

2.10.4.6.1 连接类型

- `left_join()`: 保留左表所有行
- `right_join()`: 保留右表所有行
- `inner_join()`: 仅保留匹配行
- `full_join()`: 保留所有行

2.10.4.6.2 连接语法

- 基本语法: `left_join(x, y, by = "key")`
- 多键连接: `by = c("key1", "key2")`
- 不同列名: `by = c("x_key" = "y_key")`

2.10.5 课后练习

题目: 某湿地生物多样性调查数据:

```
wetland_survey <- tibble(
  site_id = rep(c("W1", "W2", "W3"), each = 8),
  date = rep(c("2022-05", "2022-08", "2022-05", "2022-08"), times =
  plant_species = c(15, 18, 12, 16, 20, 25, 18, 22, 8, 12, 6, 10),
  bird_species = c(8, 12, 6, 9, 15, 18, 12, 14, 4, 7, 3, 5),
  water_level = c(45, 38, 50, 42, 35, 28, 40, 33, 55, 48, 60, 53)
)

site_characteristics <- tibble(
  site_id = c("W1", "W2", "W3"),
  area_ha = c(12, 8, 15),
  protection_status = c("保护区", "缓冲区", "实验区")
)
```

请完成 (使用 tidyverse 工具链, 结合之前学过的统计和可视化方法):

1. 筛选出 5 月份的调查数据, 并计算植物和鸟类物种总数 (使用 `filter()` 和 `mutate()`)
2. 按站点分组, 计算各站点的平均物种数和水位变化范围 (使用 `group_by()` 和 `summarise()`)
3. 连接站点特征数据, 创建物种密度指标 (物种数/面积) (使用 `left_join()` 和 `mutate()`)
4. 将数据从宽格式转换为长格式, 便于后续统计分析 (使用 `pivot_longer()`)
5. 根据保护状态和季节, 分析不同组合下的生物多样性特征 (使用 `group_by()` 和统计函数)

6. 使用 `ggplot2` 创建专业的可视化图表展示分析结果

7. 与第 9 课的传统方法对比，体会 `tidyverse` 的优势

2.11 图形语法与科学绘图

2.11.1 生态学背景

数据可视化是生态学研究中传达发现和支持论证的关键工具。与基础 R 绘图相比，`ggplot2` 采用图形语法，能够创建更加专业、美观的科学图表，满足期刊发表和学术报告的高标准要求。

2.11.2 演示数据

```
# 某保护区多年生物多样性监测数据
library(ggplot2)
library(dplyr)

biodiversity_data <- data.frame(
  year = rep(2018:2022, each = 12),
  month = rep(1:12, times = 5),
  temperature = rnorm(60, mean = 15 + 5*sin(2*pi*(rep(1:12, times=5)),
  species_richness = rpois(60, lambda = 25 + 10*sin(2*pi*(rep(1:12,
  habitat = rep(c("森林", "草地", "湿地"), length.out = 60),
  elevation = rep(c(1200, 1000, 800), length.out = 60)
)
```

2.11.3 课堂演示过程

2.11.3.1 `ggplot2` 基础语法

```
library(ggplot2)
library(dplyr)

# 创建示例数据
bird_data <- data.frame(
  species = c("白头鹎", "麻雀", "喜鹊", "乌鸦", "燕子", "画眉"),
  abundance = c(45, 78, 32, 28, 56, 41),
  habitat = c("森林", "城市", "农田", "城市", "农田", "森林"),
  body_mass = c(25, 15, 180, 350, 18, 35)
)

# 基础散点图
ggplot(bird_data, aes(x = body_mass, y = abundance)) +
  geom_point()

# 添加颜色映射
ggplot(bird_data, aes(x = body_mass, y = abundance, color = habitat))
  geom_point(size = 3)

# 添加标题和标签
ggplot(bird_data, aes(x = body_mass, y = abundance, color = habitat))
  geom_point(size = 3) +
  labs(
    title = "鸟类体重与丰度关系",
    subtitle = "不同栖息地类型的比较",
    x = "体重 (g)",
    y = "丰度 (个体数)",
    color = "栖息地类型"
  )
```

2.11.3.2 不同类型的图表

```
# 创建时间序列数据
time_series_data <- data.frame(
  month = rep(1:12, 3),
  species_count = c(
    20, 25, 35, 45, 55, 60, 58, 52, 42, 32, 25, 22, # 2020 年
    22, 28, 38, 48, 58, 65, 62, 55, 45, 35, 28, 25, # 2021 年
    25, 30, 40, 50, 60, 68, 65, 58, 48, 38, 30, 28 # 2022 年
  ),
  year = rep(c("2020", "2021", "2022"), each = 12)
)

# 线图
ggplot(time_series_data, aes(x = month, y = species_count, color = y
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +
  labs(
    title = "月度物种数量变化",
    x = "月份",
    y = "物种数量",
    color = "年份"
  ) +
  theme_minimal()

# 柱状图
habitat_summary <- bird_data %>%
  group_by(habitat) %>%
  summarise(
    mean_abundance = mean(abundance),
    se_abundance = sd(abundance) / sqrt(n()),
    .groups = 'drop'
```

```
)  
  
ggplot(habitat_summary, aes(x = habitat, y = mean_abundance, fill =  
  geom_col() +  
  geom_errorbar(  
    aes(ymin = mean_abundance - se_abundance,  
        ymax = mean_abundance + se_abundance),  
    width = 0.2  
) +  
  labs(  
    title = "不同栖息地的鸟类平均丰度",  
    x = "栖息地类型",  
    y = "平均丰度",  
    fill = "栖息地"  
) +  
  theme_classic()  
  
# 箱线图  
ggplot(bird_data, aes(x = habitat, y = abundance, fill = habitat)) +  
  geom_boxplot() +  
  geom_jitter(width = 0.2, alpha = 0.6) +  
  labs(  
    title = "不同栖息地鸟类丰度分布",  
    x = "栖息地类型",  
    y = "丰度"  
) +  
  theme_minimal() +  
  theme(legend.position = "none")
```

2.11.3.3 多面板图形

```
# 创建多组数据
multi_species_data <- data.frame(
  species = rep(c("鸟类", "哺乳动物", "昆虫"), each = 24),
  month = rep(1:12, 6),
  year = rep(c("2021", "2022"), each = 12, times = 3),
  abundance = c(
    # 鸟类数据
    rnorm(24, mean = 30 + 15*sin(2*pi*(1:24-1)/12), sd = 5),
    # 哺乳动物数据
    rnorm(24, mean = 15 + 8*sin(2*pi*(1:24-1)/12), sd = 3),
    # 昆虫数据
    rnorm(24, mean = 80 + 40*sin(2*pi*(1:24-1)/12), sd = 15)
  )
)

# 分面图
ggplot(multi_species_data, aes(x = month, y = abundance, color = year)) +
  geom_line(size = 1) +
  geom_point() +
  facet_wrap(~ species, scales = "free_y") +
  scale_x_continuous(breaks = c(3, 6, 9, 12)) +
  labs(
    title = "不同类群动物的季节性变化模式",
    x = "月份",
    y = "丰度",
    color = "年份"
  ) +
  theme_bw()
```

2.11.3.4 专业主题和自定义

```
# 创建专业期刊风格的图表
publication_plot <- ggplot(bird_data, aes(x = body_mass, y = abundance))
  geom_point(aes(color = habitat), size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, color = "black", linetype =
  scale_color_manual(values = c("森林" = "#2E8B57", "城市" = "#DC143C"))
  labs(
    title = "鸟类体重与种群丰度的关系",
    x = "体重 (g)",
    y = "种群丰度 (个体数)",
    color = "栖息地类型",
    caption = "数据来源: 某自然保护区鸟类调查 (2022)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10),
    legend.title = element_text(size = 11),
    legend.text = element_text(size = 10),
    panel.grid.minor = element_blank(),
    plot.caption = element_text(size = 8, color = "gray50")
  )

print(publication_plot)

# 保存图片
ggsave("bird_analysis.png", publication_plot,
       width = 8, height = 6, dpi = 300)
```

2.11.3.5 复杂的生态学可视化

```
# 群落组成气泡图
community_data <- data.frame(
  site = rep(c("样地 A", "样地 B", "样地 C"), each = 6),
  species = rep(c("物种 1", "物种 2", "物种 3", "物种 4", "物种 5"),
  abundance = c(25, 15, 8, 32, 12, 6, # 样地 A
                18, 22, 12, 25, 8, 15, # 样地 B
                12, 8, 25, 18, 20, 17), # 样地 C
  biomass = c(2.5, 3.2, 1.8, 4.1, 2.0, 1.2,
             3.0, 2.8, 2.2, 3.5, 1.5, 2.5,
             2.2, 1.8, 4.0, 2.9, 3.8, 3.2)
)

ggplot(community_data, aes(x = species, y = site)) +
  geom_point(aes(size = abundance, color = biomass), alpha = 0.7) +
  scale_size_continuous(range = c(2, 12), name = "丰度") +
  scale_color_gradient(low = "lightblue", high = "darkred", name =
  labs(
    title = "群落物种组成与生物量分布",
    x = "物种",
    y = "样地"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid = element_line(color = "gray90", size = 0.3)
  )
```

2.11.4 R 语言知识点详解

2.11.4.1 ggplot2 的图形语法

2.11.4.1.1 基本概念

- **数据 (Data)**: 要可视化的数据集
- **美学映射 (Aesthetics)**: 数据变量到图形属性的映射
- **几何对象 (Geometries)**: 用来表示数据的图形元素
- **统计变换 (Statistics)**: 对原始数据的统计总结
- **坐标系统 (Coordinates)**: 数据如何映射到平面
- **分面 (Facets)**: 将数据分割成子集的方法
- **主题 (Themes)**: 控制图形整体外观

```
ggplot(data, aes(x = var1, y = var2)) +  
  geom_*( ) +  
  scale_*( ) +  
  labs( ) +  
  theme_*( )
```

2.11.4.1.2 基本语法结构

2.11.4.2 美学映射系统

2.11.4.2.1 aes() 函数

- 位置映射: x、y
- 颜色映射: color (边框)、fill (填充)
- 大小映射: size
- 形状映射: shape
- 透明度映射: alpha
- 线型映射: linetype

2.11.4.2.2 映射 vs 设定

- 映射: `aes(color = variable)`, 颜色根据变量值变化
- 设定: `geom_point(color = "red")`, 所有点都是红色

2.11.4.3 几何对象详解

2.11.4.3.1 点图相关

- `geom_point()`: 散点图
- `geom_jitter()`: 抖动散点图
- 参数: `size`、`shape`、`alpha`、`stroke`

2.11.4.3.2 线图相关

- `geom_line()`: 线图
- `geom_path()`: 路径图
- `geom_smooth()`: 拟合线
- 参数: `size`、`linetype`、`method`

2.11.4.3.3 柱状图相关

- `geom_col()`: 柱状图 (使用实际值)
- `geom_bar()`: 柱状图 (统计计数)
- `geom_histogram()`: 直方图
- 参数: `width`、`position`

2.11.4.3.4 分布图相关

- `geom_boxplot()`: 箱线图
- `geom_violin()`: 小提琴图
- `geom_density()`: 密度图

2.11.4.4 标度系统

2.11.4.4.1 颜色标度

- 连续型:

- `scale_color_gradient()`: 双色渐变
- `scale_color_gradient2()`: 三色渐变
- `scale_color_viridis_c()`: viridis 调色板

- 离散型:

- `scale_color_manual()`: 手动设置颜色
- `scale_color_brewer()`: ColorBrewer 调色板

2.11.4.4.2 坐标轴标度

- 连续型:

- `scale_x_continuous()`: 连续 x 轴
- `scale_y_log10()`: 对数 y 轴

- 离散型:

- `scale_x_discrete()`: 离散 x 轴

- 日期型:

- `scale_x_date()`: 日期 x 轴

2.11.4.5 分面系统

2.11.4.5.1 `facet_wrap()`

- 用途: 按一个变量分面, 排列成网格
- 语法: `facet_wrap(~ variable, ncol = 2)`
- 参数:
 - `ncol`、`nrow`: 列数和行数
 - `scales`: 坐标轴缩放方式

2.11.4.5.2 `facet_grid()`

- 用途：按两个变量分面，形成矩阵
- 语法：`facet_grid(rows ~ cols)`
- 特殊语法：
 - `facet_grid(. ~ variable)`: 仅按列分面
 - `facet_grid(variable ~ .)`: 仅按行分面

2.11.4.6 主题系统

2.11.4.6.1 预设主题

- `theme_minimal()`: 简洁主题
- `theme_classic()`: 经典主题
- `theme_bw()`: 黑白主题
- `theme_void()`: 空白主题

2.11.4.6.2 自定义主题元素

- 文本元素: `element_text()`
 - `size`: 字体大小
 - `color`: 字体颜色
 - `face`: 字体样式 (“bold”、“italic”)
 - `hjust`、`vjust`: 水平和垂直对齐
- 线条元素: `element_line()`
 - `color`: 线条颜色
 - `size`: 线条粗细
 - `linetype`: 线条类型
- 矩形元素: `element_rect()`
 - `fill`: 填充颜色
 - `color`: 边框颜色
- 移除元素: `element_blank()`

2.11.4.7 图片保存

2.11.4.7.1 `ggsave()` 函数

- 语法: `ggsave(filename, plot, width, height, dpi, units)`
- 支持格式:
 - 矢量格式: PDF、SVG、EPS
 - 位图格式: PNG、JPEG、TIFF
- 推荐设置:
 - 期刊投稿: 300-600 DPI
 - 演示文稿: 150-300 DPI
 - 网页使用: 72-150 DPI

2.11.4.8 色彩设计原则

2.11.4.8.1 科学可视化色彩指南

- 连续数据: 使用渐变色, 避免彩虹色
- 分类数据: 使用对比鲜明的颜色
- 色盲友好: 避免红绿组合, 推荐 viridis 调色板
- 发表要求: 考虑黑白印刷效果

2.11.4.8.2 推荐调色板

- **Viridis** 系列: 色盲友好, 打印友好
- **ColorBrewer**: 专业的制图调色板
- 自然色彩: 模仿自然界的颜色组合

2.11.5 课后练习

题目: 某国家公园植被多样性调查数据:

```
vegetation_survey <- data.frame(  
  transect = rep(c("山顶", "山腰", "山底"), each = 20),  
  species_richness = c(rnorm(20, 15, 3), rnorm(20, 25, 4), rnorm(20,  
  coverage_percent = c(rnorm(20, 60, 10), rnorm(20, 75, 8), rnorm(20,  
  slope_degree = c(rnorm(20, 25, 5), rnorm(20, 15, 3), rnorm(20, 5,  
  soil_depth = c(rnorm(20, 15, 3), rnorm(20, 25, 4), rnorm(20, 40, 6  
)
```

请完成（使用 `ggplot2` 高级功能，结合之前学过的所有内容）：

1. 创建物种丰富度与植被覆盖度的散点图，用颜色区分不同海拔带（使用 `geom_point()` 和 `aes()`）
2. 绘制三个海拔带物种丰富度的箱线图，添加个体数据点（使用 `geom_boxplot()` 和 `geom_jitter()`）
3. 创建多面板图，展示不同海拔带的各项指标分布（使用 `facet_wrap()`）
4. 设计一个期刊级别的综合图表，展示海拔梯度上的植被特征变化（使用多个 `geom` 层）
5. 自定义主题，确保图表符合学术发表标准（使用 `theme()` 函数）
6. 保存高质量图片用于论文发表（使用 `ggsave()` 函数）
7. 与第 7 课的基础绘图方法对比，总结 `ggplot2` 的优势
8. 尝试创建动态或交互式可视化（选做，可查阅相关资料）

2.12 综合练习

2.12.1 练习 1

请用上面所学产生一个斐波那契序列的前 46 个值，其中斐波那契序列的具体定义如下：

$$F_0 = 0, F_1 = 1, F_n = F_{n-1} + F_{n-2} \quad (n \geq 2, n \in \mathbb{N})$$

2.12.2 练习 2

针对一段给定的字符串: `ladjgl5ug6e6wjgl12saj98gd0la`, 请用 R 语言提取其中的数字并求和。要求, 给出详细的编码过程及最终结果。

2.12.3 练习 3

在项目文件夹的 `data` 目录下, 存在一个天童样地的数据 `Tiantong_Sample.csv`, 其中包含 7 列, 分别为: `spCN` (物种中文名)、`sp` (物种代码)、`tag` (个体编码)、`gx` (个体 x 坐标)、`gy` (个体 y 坐标)、`dbh` (胸径)、`height` (个体高度)。请读取该文件, 并统计总共有多少个物种、胸径大于 10 的个体有多少个、平均胸径是多少、平均高度是多少。最后统计每个物种的在天童样地的个体数, 并计算 `shannon` 多样性指数。