统计学讲义

战立侃

2023-12-12

目录

前言

统计为研究人员提供用于描述和解释其研究结果的了客观和系统的方法。科学研究是我们用来收集信息的系统，而统计学是我们用来将信息提炼成合理和可信结论的工具。本课程的目标不仅教授统计方法，还要传达科学所必需的客观性原则和基本逻辑，这对于科学和日常生活中的决策都具有价值。这门课程是《行为科学统计学》的入门课程。本课程的主要目标之一是尽可能地使学习统计变得简单和轻松。本课程内容包括统计学简介、频率分布、集中量数、变异量数、正态分布和 z 分数、概率、样本和概率、假设检验、t 检验（t 检验简介、独立样本 t 检验和相关样本 t 检验）、方差分析（方差分析简介、重复测量的方差分析和独立样本的双因素方差分析）、相关、回归分析、卡方检验和符号检验等。在课程结束时，您应该能够阅读使用基本统计方法的行为研究；进行基本的数据分析；以及有能力进一步学习更高级的统计课程。更具体地说，您将学会：

- 总结统计数据并发现数据中的模式；
- 用样本数据对相关整体的特征进行推论；
- 使用编程语言如 `Julia` 来分析统计数据。

本讲义的主要参考书为：

- Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the Behavioral Sciences* (10 ed.): Wadsworth Publishing.

第一部分

描述统计

第一章　统计学简介

1.1　本章 Julia 包

```julia
using Pkg; Pkg.activate(@__DIR__)
using Random, StatsBase, Distributions, Statistics
using CairoMakie
using DataFrames, StatsReIntro
```

1.2　统计、科学和观测

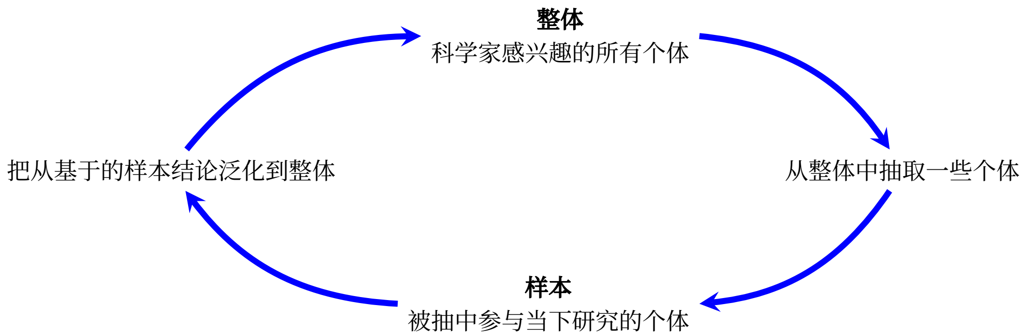
1.2.1　统计学的定义

统计学（Statistics）指一组用于组织、总结和解释信息的数学程序。

- 统计学被用于组织和总结信息，以便研究人员能够了解研究中发生了什么，并能够将结果传达给其他人。
- 统计学能通过获得的具体结果确定哪些一般结论是合理的，从而帮助回答触发当下具体研究的理论问题。
- 统计程序有助于确保信息或观察以准确和信息丰富的方式呈现和解释。

1.2.2　整体和样本

行为科学研究通常从一个关于特定群体(或多个群体)的一般问题开始。*总体*(Population)是指在特定研究中科学家感兴趣的所有个体的集合。由于总体通常非常庞大，科学家通常不可能研究总体中的每个个体。因此，科学家通常从总体中选择一个较小、更易处理的群体，并把其研究对象限定为所选群体中的个体。*样本*（Sample）是从总体中选择的一组个体，用于代表研究的全部个体。



1.2.3 变量和数据

通常，研究人员对总体中（或样本中）个体的特定特征感兴趣，或者对可能影响这些个体的外部因素感兴趣。*变量*（Variable）是指在不同个体间发生变化或具有不同值的特征或条件。为了展示变量的变化，有必要对所研究的变量进行测量。*数据*（复数）（Data）是测量或观察的结果。数据集（Data set）是一组测量或观察结果的集合。*数据点*（单数）（Datum）是单个测量或观察结果，通常称为分数或原始分数。

1.2.4 参数和统计量

在描述数据时，有必要区分数据是来自总体还是样本。*参数*（Parameter）是一个数量值，用于描述一个整体。参数通常是通过对整体中个体的测量得到的。*统计量*（Statistic）通常也是一个数量值，用于描述一个样本。统计量通常是通过对样本中个体的测量得到的。

1.2.5 描述统计和推论统计

研究者开发了不同的统计程序来组织和解释数据，这些程序可分为两个类别：*描述统计*（Descriptive statistics）用于总结、组织和简化数据；*推论统计*（Inferential statistics）用于研究样本，然后把结论泛化到样本来自的整体。

1.2.6 取样误差

由于总体通常非常庞大，通常不可能对总体中的每个个体进行测量。因此，选择一个样本来代表总体。通常情况下，样本统计量和相应的总体参数之间会存在一些差异。*取样误差*（Sampling error）是样本统计量和相应的总体参数之间存在的自然差异或误差。

如有如下十个人为整体的智商分数：

```
10-element Vector{Int64}:
 101
```

```
 104
  91
 101
 116
  76
 103
 113
 115
 123
```

其平均值是 104.3。从中选取一个样本：

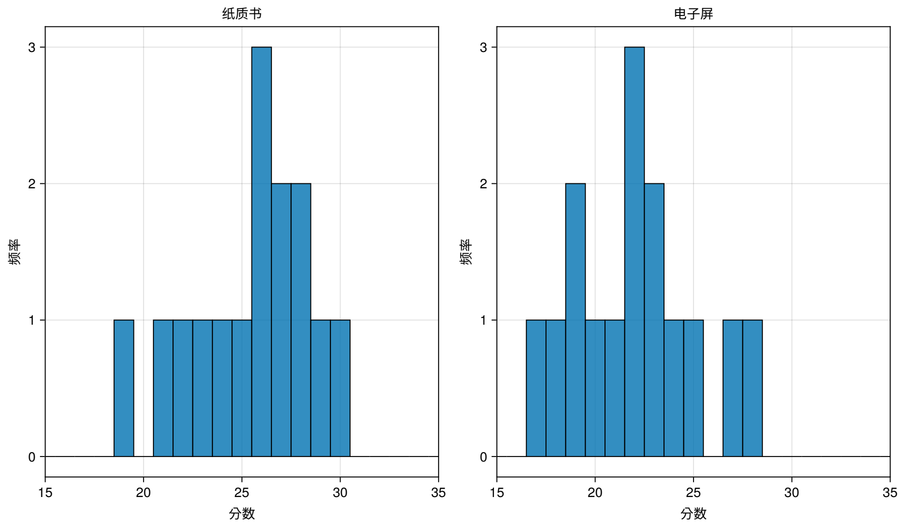
```
5-element Vector{Int64}:
 101
 116
 113
  91
 123
```

样本平均值是 108.8。

1.2.7 研究过程中的统计

```
纸质书 = [25, 26, 28, 27, 21, 27, 30, 28, 24, 19, 23, 26, 29, 26, 22]
电子屏 = [20, 22, 27, 23, 17, 23, 25, 28, 21, 22, 19, 22, 18, 24, 19]
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



样本数据显示两种学习方法之间存在 4 分差异，但有两种解释结果的方式。

- 一种解释是，两种学习方法实际上没有差异，样本差异是由于偶然因素（抽样误差）导致的。
- 另一种解释是，两种方法之间确实存在差异，样本数据准确反映了这种差异。

推断性统计的目标是帮助研究人员在这两种解释之间做出决策。

1.3 研究方法和统计

1.3.1 研究方法分类

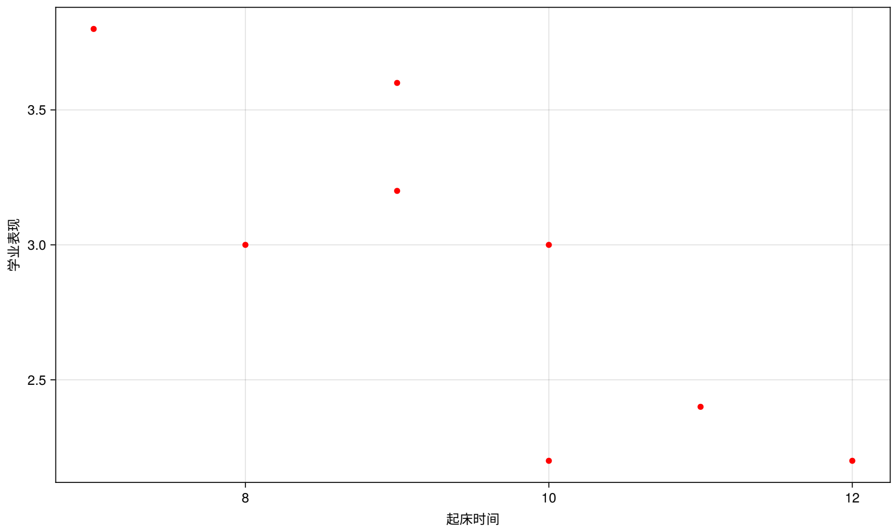
- 个体变量：描述性研究
- 变量之间的关系
  - 一个组中的每个个体都有两个变量的测量：相关方法
  - 比较两个（或更多）分组的分数：实验和非实验方法
- 描述性研究：有些研究仅仅是为了描述自然存在的个体变量。
- 变量之间的关系：大多数研究旨在考察两个或更多变量之间的关系。

1.3.2 变量之间的关系

要建立关系的存在，研究人员必须进行观察 - 即对两个变量进行测量。由此产生的测量可以分为两种不同的数据结构，这也有助于分类不同的研究方法和不同的统计技术。

在*相关方法*中（Correlational Method），观察两个不同的变量，以确定它们之间是否存在关系。

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



有时，用于相关研究的测量过程仅将个体分类为不对应数值的类别。例如，男女在微信和电话方面的使用偏好可以做如下总结：

	短信	电话	
男	30	20	50
女	25	25	50

相关性研究的结果可以证明两个变量之间存在关系，但它们并不能提供关系的解释。特别是，相关性研究不能证明因果关系。

1.3.3 实验和非实验方法

检验两个变量之间关系的第二种方法涉及比较两个或多个分数组。在这种情况下，通过使用其中一个变量来定义不同组别，然后测量第二个变量以获得每个组的分数来检查变量之间的关系。有两种明显不同的研究方法都能产生要进行比较的分数组：实验和非实验策略。这两种研究方法使用完全相同的统计方法，它们都展示了两个变量之间的关系。这两种研究策略之间的区别在于如何解释关系。实验的结果允许因果关系的解释。非实验研究不允许因果关系的解释。

实验研究的目标是展示两个变量之间的因果关系。具体来说，实验试图表明改变一个变量的值会导致第二个变量发生变化。为了实现这个目标，实验方法具有两个特点，这些特点将实验与其他类型的研究区分开来：*操纵*（Manipulation）：研究人员通过改变一个变量的值从一个水平到另一个水平来操纵一个变量。观察（测量）第二个变量以确定操纵是否导致变化发生。*控制*（Control）：研究人员必须对研究情境进行控制，以确保其他外部变量不会影响所研究的关系。

研究人员必须考虑两种一般类别的变量：*被试变量*（Participant Variables）：这些是如年龄、性别和智力等因素，它们在不同个体之间有所不同。*环境变量*（Environmental

序号	暴力	不暴力
1	7	8
2	8	4
3	10	8
4	7	3
5	9	6
6	8	5
7	6	3
8	10	4
9	9	4
10	6	5

Variables）：这些是环境的特征，如照明、时间和天气条件等。

研究人员通常使用三种基本技巧来控制其他变量：*随机指派*（Random assignment）：每个参与者有相等的机会被分配到每个处理条件中。控制变量的第二种技术是使用*匹配*（Matching）来确保等效的组或等效的环境。最后，研究人员可以通过保持控制变量*恒定*（constant）来控制变量。在*实验方法*中，一个变量被操纵，同时观察和测量另一个变量。为了建立两个变量之间的因果关系，实验试图控制所有其他变量，以防止它们对结果产生影响。

1.3.4 实验方法中的术语

实验研究通过操纵一个变量（自变量）并测量一个变量（因变量）来评估两个变量之间的关系。*自变量*（Independent variable）是研究人员操纵的变量。在行为研究中，自变量通常包括受试者暴露的两个（或更多）处理条件。自变量由在观察因变量之前操纵的前置条件组成。*因变量*（Dependent variable）是用于评估处理效应的变量。*控制条件*（Control condition）中的个体不接受实验处理。相反，他们要么不接受任何处理，要么接受中性的安慰剂处理。控制条件的目的是提供与实验条件进行比较的基线。*实验条件*（Experimental condition）中的个体接受实验处理。

1.3.5 非实验方法

真正的实验必须包括对自变量的操纵和对其他外部变量的严格控制。因此，还有许多其他研究设计不是真正的实验，但仍通过比较分数组来检查变量之间的关系。在非实验性研究中，用于创建不同分组的"自变量"通常被称为*伪自变量*（quasi-independent variable）。下面两个例子是涉及比较两组分数的非实验研究（现存或时间分组）

序号	男孩 (治疗前)	女孩 (治疗后)
1	17	12
2	19	10
3	16	14
4	12	15
5	17	13

序号	男孩 (治疗前)	女孩 (治疗后)
6	18	12
7	15	11
8	16	13

1.4 变量和测量

1.4.1 构念与操作定义

构念（Constructs）是无法直接观察但有助于描述和解释行为的内在属性或特征。*操作定义*（Operational definition）确定了一种测量过程（一组操作）来测量外部行为，并使用由测量得到的结果作为假设构念的定义和测量。请注意，操作定义具有两个组成部分。首先，它描述了一组测量构念的操作。其次，它根据所得的测量结果定义了构念。

1.4.2 离散变量和连续变量

研究中的变量可以根据可以分配给它们的值的类型来表征。*离散变量*（Discrete variable）由分离且不可分割的类别组成。在相邻两个类别之间不存在值。对于*连续变量*（Continuous variable），在任意两个观察值之间存在无限多个可能的值。连续变量可分割成无限多个分数部分。在测量连续变量时，对于两个不同个体获得相同的测量值应该非常罕见。在测量连续变量时，每个测量类别实际上都是一个由边界定义的间隔。

1.4.3 连续变量和实际限度

实际限度（Real limits）是表示在连续数字线上的分数间隔的边界。分开两个相邻分数的实际限度恰好位于分数之间的中间位置。每个分数都有两个实际限度。*上实际限度*（Upper real limit）位于间隔的顶部，*下实际限度*（Lower real limit）位于底部。

1.4.4 测量尺度

连续和离散这两个术语适用于正在测量的变量，而不适用于从测量中获得的分数。数据收集要求我们对观察结果进行测量。*测量*（Measurement）涉及将个体或事件分配到不同的类别中。用于测量变量的类别构成了*测量尺度*（Scale of measurement），而类别之间的关系确定了不同类型的测量尺度。

- *称名尺度*（Nominal scale）包括一组具有不同名称的类别。称名尺度上的测量用于标记和分类观察结果，但不会在观察结果之间进行任何数量上的区分。
- *顺序尺度*（Ordinal scale）包括一组按顺序排列的类别。顺序尺度上的测量将观察结果按大小或程度排名。

- *等距尺度*（Interval scale）包括按精确相同大小的间隔排列的类别。尺度上数字之间的相等差异反映了大小上的相等差异。但是，等距尺度上的零点是任意的，不表示所测量的变量的零量。
- *等比尺度*（Ratio scale）是具有绝对零点的等距尺度的附加特征。使用比率尺度时，数字的比率反映了大小的比率。

测量尺度很重要，因为它们有助于确定用于评估数据的统计方法。具体来说，有一些统计程序适用于来自等距或比率尺度的数值分数，而另一些统计程序适用于来自名义或顺序尺度的非数值分数。

1.5 统计符号

1.5.1 分数

在研究中测量一个变量会为每个个体产生一个值或分数。*原始分数*（Raw scores）是研究中获得的未经修改的原始分数。特定变量的分数通常用字母 X 表示。一组分数可以呈现在由 X 标头的列中。当为两个变量进行观察时，每个个体将有两个分数。数据可以呈现为两个标记为 X 和 Y 的列表，分别表示两个变量。字母 N 用于指定一组分数中有多少个分数。大写字母 N 表示群体中的分数数量，小写字母 n 表示样本中的分数数量。

1.5.2 求和符号

求和符号 Σ 可以读作“的总和”。求和符号后面总是跟着一个符号或数学表达式。求和过程通常包括其他几个数学运算，如乘法或平方运算。

1.5.3 数学运算的顺序

任何包含在*括号*内的计算首先进行。*平方*（或提高到其他指数）其次进行。*乘法*和/或 *除法*接下来进行。一系列乘法和/或除法操作应按从左到右的顺序进行。使用 Σ 符号的*求和*紧随其后进行。最后，进行其他*加法*和/或 *减法*。

```
X = [3, 1, 7, 4]
sum(X), sum(X.^2), sum(X)^2, sum(X .- 1), sum(X .- 1)^2, sum((X .- 1).^2)
```

```
(15, 75, 225, 11, 121, 49)
```

第二章 频率分布

2.1 本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie, DataFrames, StatsReIntro
using Distributions: Normal
```

2.2 频率分布

描述性统计的目标是简化数据的组织和呈现。对一组数据进行组织的最常见方法之一是将分数放入频率分布中。一个**频率分布**是一个有序的表格，列出了测量尺度上每个类别中的个体数量。频率分布可以结构化为表格或图形，但无论哪种情况，分布都呈现相同的两个元素：- a. 原始测量尺度中的一组类别。- b. 每个类别中的频率或个体数量的记录。

2.2.1 频率分布表

以下一组 N = 20 个分数是从一份包含 10 分统计学测验中获得的。

```
quiz = [8, 9, 8, 7, 10, 9, 6, 4, 9, 8, 7, 8, 10, 9, 8, 6, 9, 7, 8, 8]
```

化成频率分布表可得

```
ft = FreqTable(quiz)
```

根据频率分布表可以做一些基本的运算，如

- 总个体数，N

```
sum(ft.频率)
```

20

类别	频率
10	2
9	5
8	7
7	3
6	2
5	0
4	1

- 个体分数之和，$\sum X$

```
sum(ft.频率 .* ft.类别)
```

```
158
```

- 个体分数的平方和，$\sum X^2$

```
sum(ft.频率 .^2 .* ft.类别)
```

```
748
```

2.2.2 比例和百分比

- 表格可能包括一个显示每个类别相对频率的比例列

$$\text{比例} = p = \frac{f}{N}$$

- 表格可能包括一个显示与每个 X 值相关联的百分比列

$$\text{百分比} = p(100) = \frac{f}{N}(100)$$

```
FreqTable(quiz, prp = true)
```

类别	频率	比率	百分比
10	2	0.1	10.0%
9	5	0.25	25.0%
8	7	0.35	35.0%
7	3	0.15	15.0%
6	2	0.1	10.0%
5	0	0.0	0.0%
4	1	0.05	5.0%

2.2.3 频率分布表：实践

```
DF2 = DataFrame(类别 = 5:-1:1, 频率 = [1, 2, 3, 3, 1])
```

类别	频率
5	1
4	2
3	3
2	3
1	1

计算比率和百分比

```
FreqTable(DF2, prp = true)
```

类别	频率	比率	百分比
5	1	0.1	10.0%
4	2	0.2	20.0%
3	3	0.3	30.0%
2	3	0.3	30.0%
1	1	0.1	10.0%

2.3 分组频率分布表

2.3.1 分组频率分布表

建议频率分布表最多包含 10 到 15 行，以保持简洁。如果分数覆盖的范围比建议的最大范围要宽，通常习惯性地将范围划分为称为**类间隔**的部分。然后，这些间隔会与频率分布表一起列出，以及每个间隔中具有分数的个体数量或频率。结果称为**分组频率分布**（grouped frequency distribution）。

2.3.2 构建这种分布的指南

- 分组频率分布表应该有大约 10 个类间隔。
- 每个间隔的宽度应该是相对简单的数字，例如 2、5、10 或 20。
- 每个类间隔的底部分数应该是宽度的倍数。
- 所有间隔的宽度应该相同。

2.3.3 分组频率分布表

- 假设有一组 $N = 25$ 的考试分数

```
data = [82, 75, 88, 93, 53, 84, 87, 58,
        72, 94, 69, 84, 61, 91, 64, 87,
        84, 70, 76, 89, 75, 80, 73, 78, 60]
```

- 数据的范围是

```
extrema(data)
```

(53, 94)

- 三种频率分布表

- 宽度：2

```
FreqTable(data, 52:
```

类别	频率
[94, 96)	1
[92, 94)	1
[90, 92)	1
[88, 90)	2
[86, 88)	2
[84, 86)	3
[82, 84)	1
[80, 82)	1
[78, 80)	1
[76, 78)	1
[74, 76)	2
[72, 74)	2
[70, 72)	1
[68, 70)	1
[66, 68)	0
[64, 66)	1
[62, 64)	0
[60, 62)	2
[58, 60)	1
[56, 58)	0
[54, 56)	0
[52, 54)	1

- 宽度：10

```
FreqTable(data, 50:
```

类别	频率
[90, 100)	3
[80, 90)	9
[70, 80)	7
[60, 70)	4
[50, 60)	2

- 宽度：5

```
FreqTable(data, 50:5:95)
```

类别	频率
[90, 95)	3
[85, 90)	4
[80, 85)	5
[75, 80)	4
[70, 75)	3
[65, 70)	1
[60, 65)	3
[55, 60)	1
[50, 55)	1

2.3.4 实际上限和频率分布

当测量连续变量时，得到的测量值对应于数轴上的间隔，而不是单个点。

2.4 频率分布图

2.4.1 频率分布图

所有频率分布都以两条垂直的线条开始，称为坐标轴。水平线是 **X 轴**或横坐标（ab-SIS-uh）。垂直线是 **Y 轴**或纵坐标。测量刻度（一组 X 值）沿着 X 轴列出，值从左到右递增。频率沿着 Y 轴列出，值从下到上递增。

作为一个一般规则，两个坐标轴相交的点应该在分数和频率两方面都为零。最后一个一般规则是，图表应该构建成其高度（Y 轴）大约是其长度（X 轴）的 2/3 到 3/4。

频率分布图将分数列在水平轴上，将频率列在垂直轴上。显示分布的图表类型取决于所使用的测量刻度。对于间隔或比例刻度，应该使用**直方图**或**多边形图**。对于名义或顺序刻度，应该使用**条形图**。

2.4.2 直方图

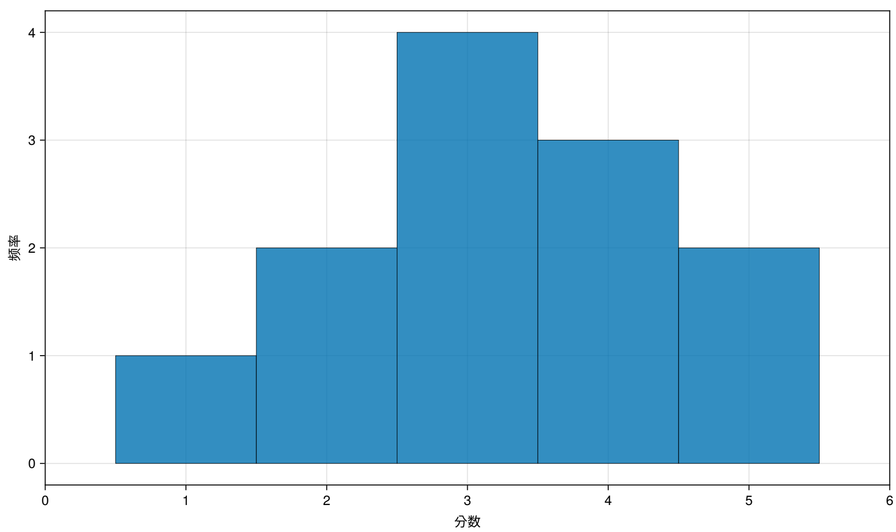
要构建直方图，首先在 X 轴上列出数值分数（测量的类别）。然后，为每个 X 值绘制一个条形，以便：

- 条形的高度对应于该类别的频率。
- 每个条形都延伸到分数的真正极限，以便相邻的条形相接触。

给定下面一组测验分数

```
quiz = [5, 5, 4, 4, 4, 3, 3, 3, 3, 2, 2, 1]
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



```julia
FreqTable(quiz)
```

类别	频率
5	2
4	3
3	4
2	2
1	1

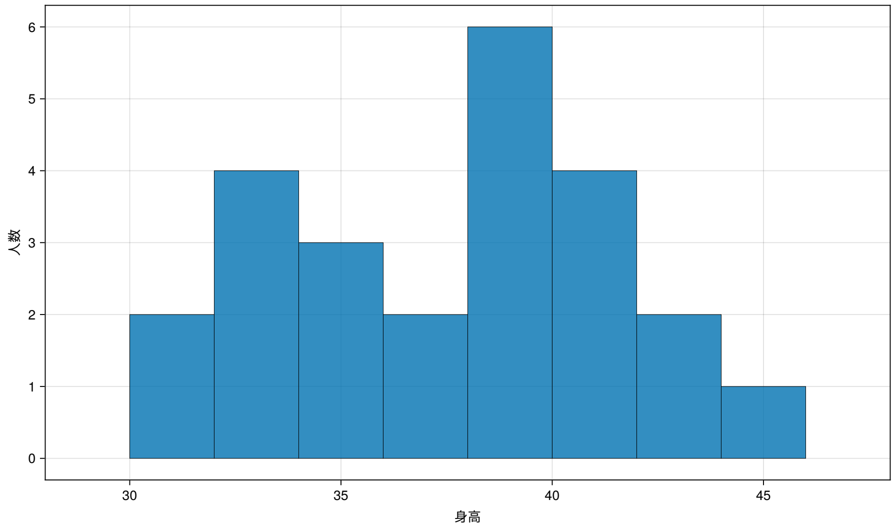
2.4.3 直方图：一组身高

```julia
身高 = [
    44.5, 42.2, 42.8, 40.1, 40.2, 40.3, 40.4, 38.4, 38.5, 38.6, 38.7,
    38.8, 38.9, 36.6, 36.8, 34.2, 34.3, 34.5, 33, 32.1, 32.2, 32.3, 30.1, 30.5]
```

```julia
FreqTable(身高, 30:2:46)
```

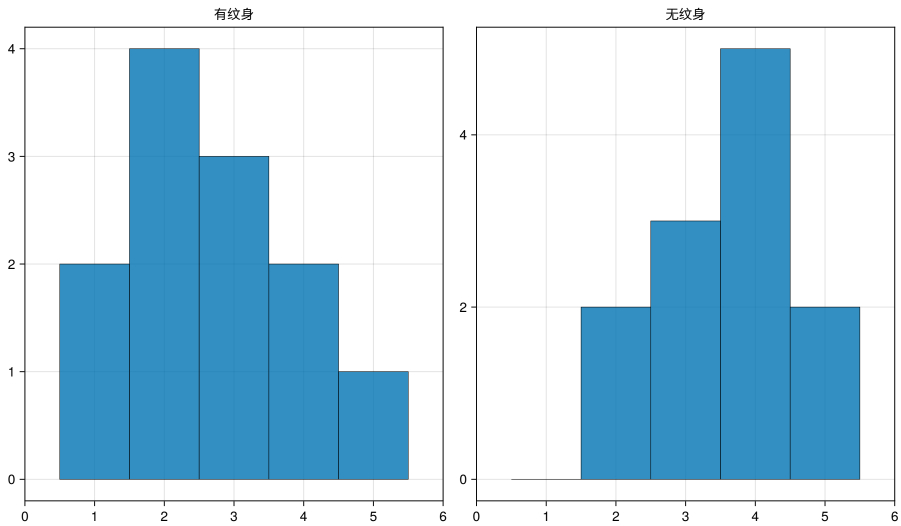
类别	频率
[44, 46)	1
[42, 44)	2
[40, 42)	4
[38, 40)	6
[36, 38)	2
[34, 36)	3
[32, 34)	4
[30, 32)	2

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



2.4.4 频率分布图：纹身和吸引力

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



2.4.5 频率多边形图

要构建多边形图，首先在 X 轴上列出数值分数（测量的类别）。然后，

- 在每个分数的正上方放置一个点，使得点的垂直位置与该类别的频率对应。

- 从一个点到另一个点画一条连续的线，以连接这些点。
- 在得分范围的两端，向下绘制一条线，直接指向 X 轴（零频率）。

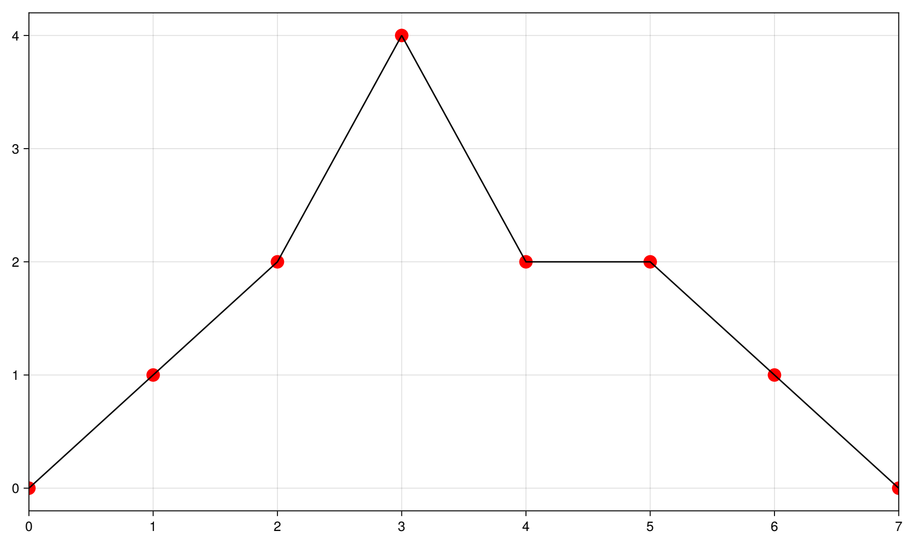
下面是一个例子

```
df = DataFrame(类别 = 6:-1:1, 频率 = [1, 2, 2, 4, 2, 1])
FreqTable(df)
```

类别	频率
6	1
5	2
4	2
3	4
2	2
1	1

```
pushfirst!(df, (maximum(df.类别) + 1, 0), promote = true)
push!(df, [minimum(df.类别) - 1, 0], promote = true)
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], limits = (extrema(df.类别), nothing))
scatter!(ax, df.类别, df.频率, markersize = 20, color = :red)
lines!(ax, df.类别, df.频率, color = :black)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



2.4.6 用于分组数据的频率分布多边形

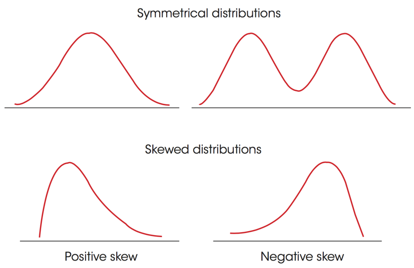


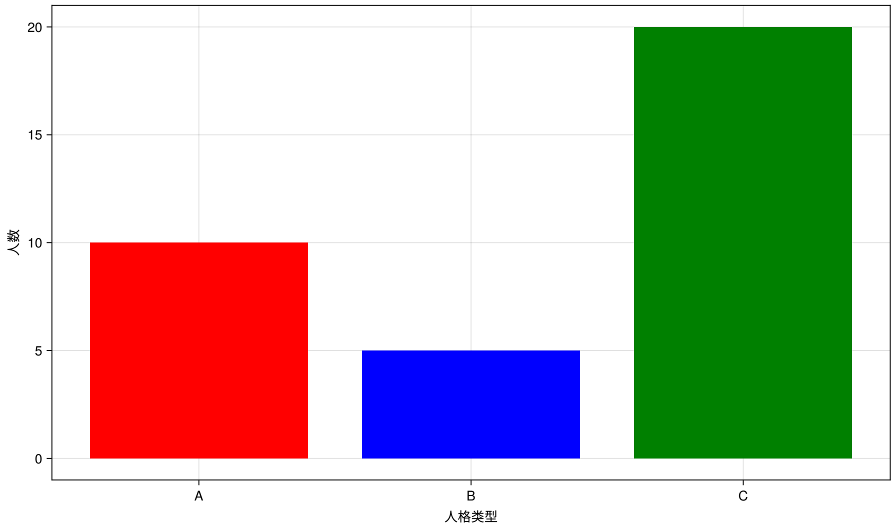
图 2.1: 形状

2.4.7 条形图

条形图与直方图基本相同，只是在相邻的条形之间留有间隔。对于名义刻度，间隔之间的空间强调了刻度由单独的、不同的类别组成。对于顺序刻度，使用单独的条形，因为不能假设所有类别都是相同大小的。

例如下面的条形图

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



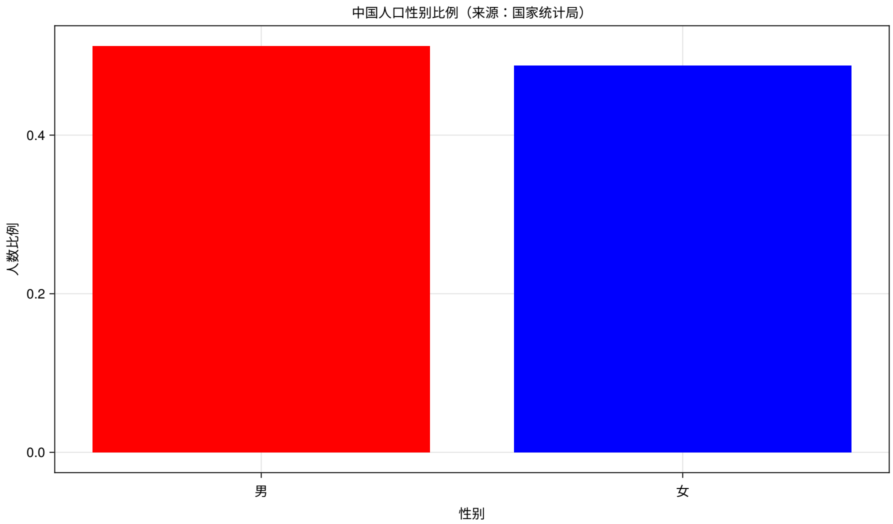
2.4.8 描述总体分布的图表

尽管仍然有可能构建显示极大总体的频率分布的图表，但这些图表通常涉及两个特殊的特征：**相对频率**（relative frequencies）和 **平滑曲线**（smooth curves）。虽然通常无法

找到每个分数的绝对频率，但往往可以获得相对频率。

2.4.9 相对频率

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



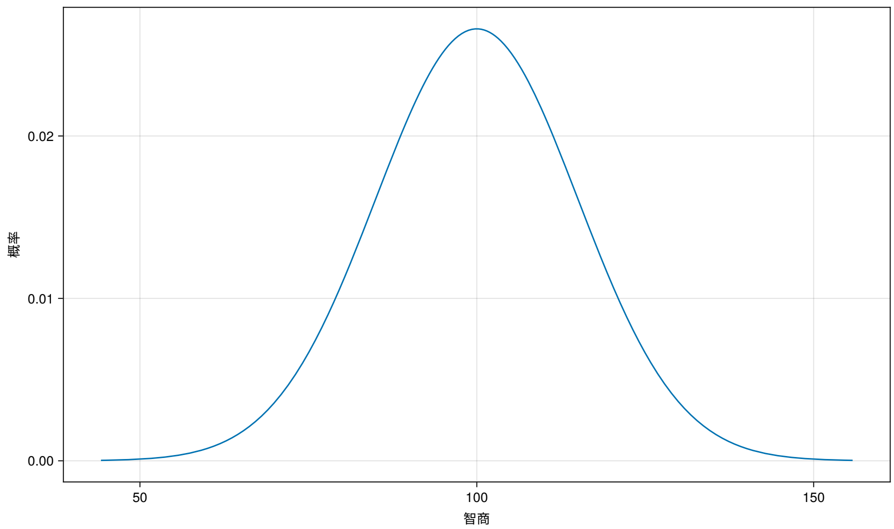
2.4.10 平滑曲线

当总体数据由等距或等比数值组成时，通常用平滑曲线，而不是用分段和锯齿状的直方图或多边形的来表示。平滑曲线说明不是连接一系列点（实际频率），而是从一个分数到下一个分数的相对变化。一个常见的总体分布是**正态曲线**。接下来我们将用**分布** (Distributions) 来指称平滑曲线。每当看到分布一词，您应该想到一个频率分布图。

2.4.11 正态曲线

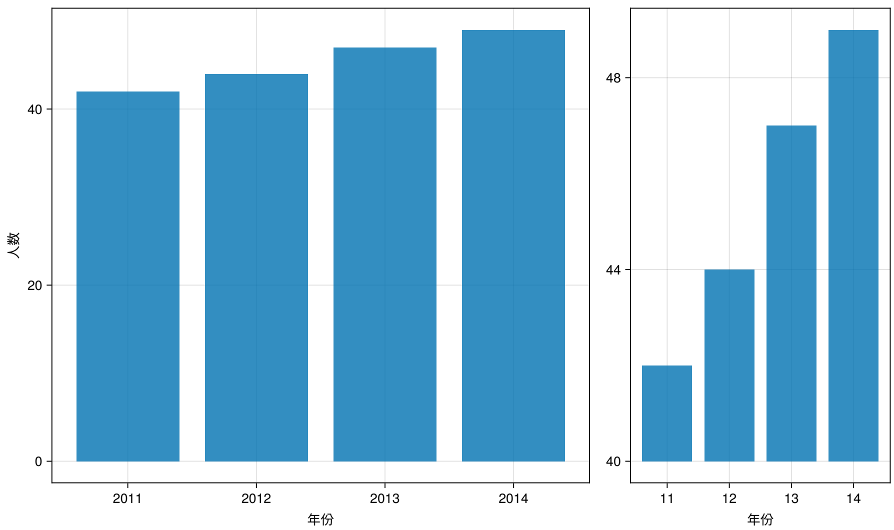
```
lines(Normal(100, 15), axis = (; xlabel = "智商", ylabel = "概率"))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



2.4.12 图表的使用与滥用

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



2.4.13 频率分布的形状

研究人员通常只需列出分布的特征，而不是绘制完整的频率分布图。有三个特征完全描述了任何分布：**形状**（shape）、**中心趋势**（central tendency）和**变异性**（variability）。

简单来说，中心趋势测量了分布的中心位置，变异性测量了分数在宽范围内的分散程度或集中程度。从技术上讲，分布的形状由一个方程来定义，该方程规定了图表上每个 X 和 Y 值之间的确切关系。但是，我们将依赖一些不太精确的术语来描述大多数分布的形状。

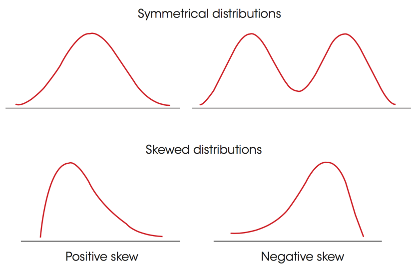


图 2.2: 形状

在**对称分布**中，可以通过分布中间绘制一条垂直线，以使分布的一侧成为另一侧的镜像。在**偏斜分布**中，分数倾向于堆积在刻度的一端，并在刻度的另一端逐渐减小。

在分布的一侧，分数逐渐减小的部分被称为**分布的尾部**（tail of the distribution）。一个右偏的分布尾部指向正（零上）端的 X 轴，因此被称为**正偏分布**（positively skewed）。如果尾部指向左侧，分布被称为**负偏分布**（negatively skewed）。

2.5 百分位数、百分位秩和插值

2.5.1 百分位数和百分位秩

百分位数和百分位秩用于描述个别分数在分布中的位置。特定分数的排名或**百分比等级**（percentile rank）被定义为分布中具有分数等于或低于特定值的个体百分比。当一个分数由其百分位秩来标识时，该分数称为**百分位数**（percentile）。请注意，百分比等级是指百分比，而百分位数是指分数。此外，请注意，百分比等级描述了分布中的确切位置。

2.5.2 累积频率和累积百分比

要确定百分位数或百分位秩，第一步是找到分布中每个点当下或下方的个体数量。这可以通过频率分布表最简单地完成，只需计算在刻度上的每个类别中或以下的分数数量。由此产生的值称为**累积频率**，因为它们表示在刻度上移动时个体的累积。

累积频率显示了每个分数当下或以下的个体数量。为了找到百分位数，我们必须将这些频率转换为百分比。由此产生的值称为**累积百分比**，因为它们显示随着在刻度上移动，个体的百分比如何积累。累积百分比值与相应分数或区间的上限相关联。

```
df = DataFrame(类别 = 5:-1:1, 频率 = [1, 5, 8, 4, 2])
FreqTable(df, cum = true)
```

类别	频率	累积频率	累积比率	累积百分比
5	1	20	1.0	100.0%
4	5	19	0.95	95.0%
3	8	14	0.7	70.0%
2	4	6	0.3	30.0%
1	2	2	0.1	10.0%

2.5.3 寻找百分位数

现在，我们可以使用百分位数的定义来找到分布中的任何特定百分位数。要找到百分位数，首先确定您感兴趣的分数在累积百分比中的位置。然后，找到与该位置相对应的分数。

例如

```
FreqTable(df, cum = true)
```

类别	频率	累积频率	累积比率	累积百分比
5	1	20	1.0	100.0%
4	5	19	0.95	95.0%
3	8	14	0.7	70.0%
2	4	6	0.3	30.0%
1	2	2	0.1	10.0%

- 什么是百分比等级为 95 百分位数？（答案：X = 4.5。）
- $X = 3.5$ 的百分比等级是多少？（答案：70%。）
- 什么数的百分比等级是 50？
- $X = 4$ 的百分位等级是多少？

2.5.4 插值

插值是一种估计在已知值之间的数值的方法。当我们需要估计特定百分位数的分数时，通常需要使用**插值** (Interpolation)。插值可以帮助我们找到分布中两个已知百分位数之间的分数。常见的插值方法之一是**线性插值**。线性插值假定在已知百分位数之间的数值均匀变化。通过连接这两个已知点并创建一条直线，可以估计百分位数之间的分数。

- 找 $X = 7.0$ 的百分比等级

```
df = DataFrame(
    类别 = 10:-1:5,
    频率 = [2, 8, 4, 6, 4, 1])
FreqTable(df, cum = true)
```

这些值可以用下表表示

类别	频率	累积频率	累积比率	累积百分比
10	2	25	1.0	100.0%
9	8	23	0.92	92.0%
8	4	15	0.6	60.0%
7	6	11	0.44	44.0%
6	4	5	0.2	20.0%
5	1	1	0.04	4.0%

	分数 (X)	百分比
上位值	7.5	44%
当下值	7.0	?
下位值	6.5	20%

其计算方式如下

$$\frac{7.5 - 7.0}{7.5 - 6.5} = \frac{44\% - ?}{44\% - 20\%}$$

解方程可得出其百分比等级为 32%.

- 找出百分比等级为 $40th$ 的百分位数

```
df = DataFrame(
    类别 = ["20-24", "15-19", "10-14", "05-09", "00-04"],
    频率 = [2, 3, 3, 10, 2])
FreqTable(df, cum = true)
```

类别	频率	累积频率	累积比率	累积百分比
20-24	2	20	1.0	100.0%
15-19	3	18	0.9	90.0%
10-14	3	15	0.75	75.0%
05-09	10	12	0.6	60.0%
00-04	2	2	0.1	10.0%

这些值可以用下表表示

	分数 (X	) 百分比
上位值	9.5	60%
当下值	?	40%
下位值	4.5	10%

其计算方式如下

$$\frac{9.5 - ?}{9.5 - 4.5} = \frac{60\% - 40\%}{60\% - 10\%}$$

解方程可得出其百分比等级为 40^{th} 的百分位数是 $X = 7.5$.

2.6 茎叶图

茎叶图 (stem and leaf display) 是一种组织数据的替代方法。每个分数被分成一个茎（前面的一个或一组数字）和一个叶子（最后的一个数字）。显示包括在一列中列出的茎，每

个分数的叶子写在其茎旁边。茎叶图类似于分组频率分布表，但茎叶图能够识别每个分数的确切值，而分组频率分布则不能。

```
tab = [83, 82, 63, 62, 93, 78,
         71, 68, 33, 76, 52, 97,
         85, 42, 46, 32, 57, 59,
         56, 73, 74, 74, 81, 76]
StemLeaf(tab)
```

茎	叶
9	37
8	1235
7	1344668
6	238
5	2679
4	26
3	23

第三章 集中趋势

3.1 本章 Julia 包

```julia
using Pkg; Pkg.activate(@__DIR__)
using Statistics: mean, median
using CairoMakie, StatsReIntro, DataFrames
```

3.2 集中趋势

描述性统计方法的一般目的是组织和总结一组分数。描述和总结分布最常见的方法之一是找到一个单一的值，用来定义平均分数并作为代表整个分布的典型示例。

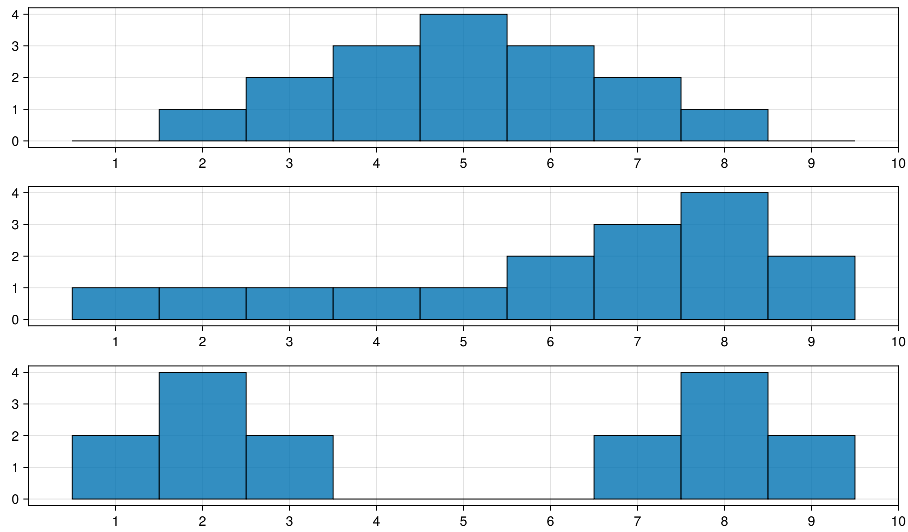
3.2.1 集中趋势

集中趋势是一种统计测量方法，用于确定定义分布中心的单一分数。集中趋势的目标是找到最典型或最具代表性的单一分数，以代表整个群体。

3.2.2 确定集中趋势

```julia
dt1 = [2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8]
dt2 = [1, 2, 3, 4, 5, 6, 6, 7, 7, 7, 8, 8, 8, 8, 9, 9]
dt3 = [1, 1, 2, 2, 2, 2, 3, 3, 7, 7, 8, 8, 8, 8, 9, 9]
fig = Figure(size = (1000, 600))
for (idx, dt) in enumerate([dt1, dt2, dt3])
    ax = Axis(fig[idx, 1], limits = ((0, 10), nothing), xticks = 1:10)
    hist!(ax, dt, bins = 0.5:1:9.5, strokewidth = 1, strokecolor = :black)
end
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



3.3 平均值

3.3.1 平均值

分布的**平均值**或**算术平均值**是分数总和除以分数的数量。

- *总体平均值*的公式是

$$\mu = \frac{\sum X}{N}$$

- *样本平均值*的公式是

$$\text{样本平均值} = M = \overline{X} = \frac{\sum X}{n}$$

Julia 语言中平均值的计算

```
import Statistics: mean
X = [3, 7, 4, 6]
sum(X) / length(X)
mean(X)
```

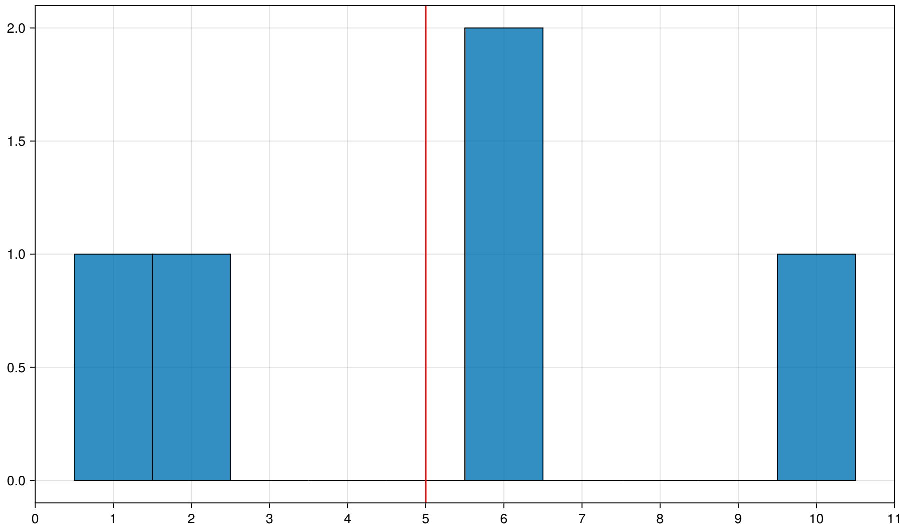
```
5.0
```

平均值的替代定义

- 等分总数。第一种替代方法是将平均值看作是在分布中的所有个体（N）之间平均分配总和（$\sum X$）时，每个个体收到的数量。
- *平均值作为平衡点*（重心）。平均值上方的距离与下方的距离完全平衡。

```
dt = [1, 2, 6, 6, 10]
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xticks = (0:11; ))
hist!(ax, dt, bins = 0.5:1:10.5, strokewidth = 1, strokecolor = :black)
vlines!(ax, [mean(dt)], color = :red)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



3.3.2 加权平均值

要计算两个合并样本的总体平均值，我们需要两个值：合并组的总分数（$\sum X$），以及合并组中数据的总个数（n）。有了这两个值，我们可以使用基本方程计算平均值：

$$\text{总体平均值} = \frac{\sum X}{n} = \frac{\sum X_1 + \sum X_2}{n_1 + n_2}$$

给定以下两个样本，其加权平均值是：

```
M1, M2 = 6, 7
n1, n2 = 12, 8
(M1 * n1 + M2 * n2) / (n1 + n2)
```

```
6.4
```

从逻辑上讲，当这两个样本合并时，较大的样本将对合并组产生比较小的样本更大的贡献。寻找加权平均值的替代方法

```
M1 * n1  / (n1 + n2) + M2 * n2  / (n1 + n2)
```

```
6.4
```

平均值也可以通频率分布表中计算。假设有一个频率分布表，

```
ft = FreqTable(DataFrame(类别 = 10:-1:6, 频率 = [1, 2, 4, 0, 1]))
```

类别	频率
10	1
9	2
8	4
7	0
6	1

其平均值为

```
sum(ft.类别 .* ft.频率) / sum(ft.频率)
```

```
8.25
```

3.3.3 平均值的特点

- 改变任何一个分数的值都会改变平均值。

```
X = [9, 8, 7, 5, 1]
println(mean(X))
X[5] = 20
println(mean(X))
```

```
6.0
9.8
```

- 向分布中添加新的分数或移除现有的分数通常会改变平均值。

```
M, n = 7, 5
X = 13
(n * M + X) / (n + 1)
```

```
8.0
```

- 如果在分布中的每个分数上都加上一个常数值，那么平均值也会增加相同的常数值。

```
X = [9, 8, 7, 5, 1]
println(mean(X))
X = X .+ 10
println(mean(X))
```

```
6.0
16.0
```

- 如果分布中的每个分数都乘以（或除以）一个常数值，平均值将以相同的方式改变。

```
X = [9, 8, 7, 5, 1]
println(mean(X))
X = X .* 10
println(mean(X))
```

```
6.0
60.0
```

3.4 中位数

3.4.1 中位数

如果分布中的分数按从小到大的顺序列出，**中位数**（Median）是列表的中点。更具体地说，中位数是分布中有 50% 的分数位于其下方的测量标度上的点。将中位数定义为分布的中点意味着将分数划分为两个大小相等的组。

要找到中位数，请按从小到大的顺序列出分数。从最小分数开始，数分数，随着列表的升序而数分数。中位数是您到达的第一个大于分布中 50 分数的点。中位数可以等于列表中的分数，也可以在两个分数之间。请注意，中位数没有代数定义（没有计算中位数的方程），这意味着：在确定确切值时存在一定程度的主观性。

3.4.2 大多数分布的中位数

```
function MyMedian(dt)
    len = length(dt)
    pos = isodd(len) ? (len+1)>>1 : [len>>1, len>>1+1]
    mean(sort(dt)[pos])
end
```

```
MyMedian (generic function with 1 method)
```

```
dt = [10, 11, 5, 3, 8]
MyMedian(dt), median(dt)
```

```
(8.0, 8.0)
```

```
dt = [4, 5, 7, 1, 1, 8]
MyMedian(dt), median(dt)
```

```
(4.5, 4.5)
```

3.4.3 连续变量的精确中位数

列出和计数分数的简单技巧足以确定大多数分布的中位数，并且始终适用于离散变量。但是，对于连续变量，可以精确地将分布分成两半，使得精确地 50% 的分布位于特定点的下方（和上方）。

```
X = [1; 2; 3; fill(4, 4); 6]
MyMedian(X), median(X)
```

```
(4.0, 4.0)
```

```
FreqTable(X, cum = true)
```

类别	频率	累积频率	累积比率	累积百分比
6	1	8	1.0	100.0%
5	0	7	0.875	87.5%
4	4	7	0.875	87.5%
3	1	3	0.375	37.5%
2	1	2	0.25	25.0%
1	1	1	0.125	12.5%

使用插值确定 $50th$ 百分位数。这些值显示在以下表格中:

	分数 (X)	百分比
上位值	4.5	87.5%
当下值	?	50%
下位值	3.5	37.5%

它可以计算为

$$\frac{4.5 - ?}{4.5 - 3.5} = \frac{87.5\% - 37.5\%}{50\% - 37.5\%}$$

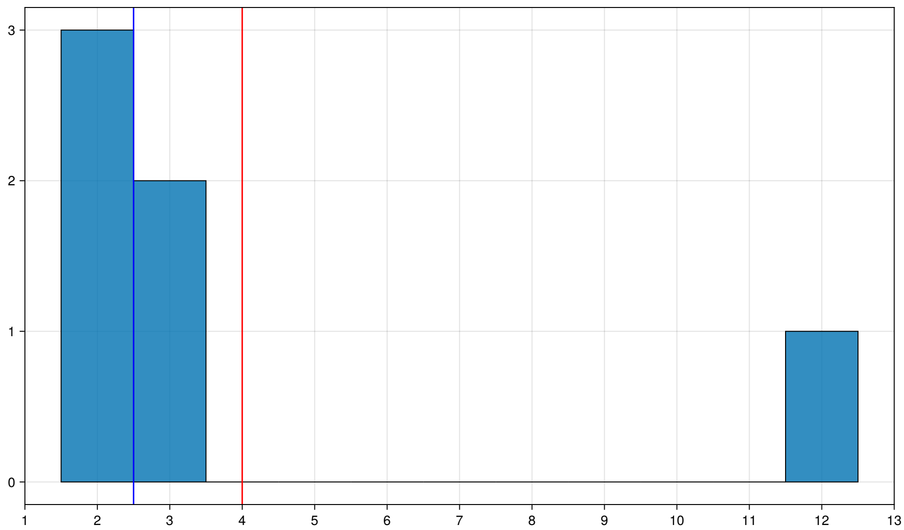
第 50 百分位数（中位数）是 $X = 3.75$ 。

3.4.4 中位数、平均数和中间值

如果用距离概念来定义"中间"，那么平均值位于分布的中间。中位数位于分布的中间，前提是"中间"由分数的数量定义。

```
X = [fill(2, 3); fill(3, 2); 12]
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1],
    limits = ((1, 13), nothing),
    xticks = (1:13; )
    )
hist!(ax, X, bins = 1.5:1:12.5, strokewidth = 1, strokecolor = :black)
vlines!(ax, [mean(X)], color = :red, label = "平均值")
vlines!(ax, [median(X)], color = :blue)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



3.5 众数

3.5.1 众数

在频率分布中，**众数**是具有最高频率的分数或类别。在频率分布图中，最高频率将显示为图形的最高部分。要找到众数，只需确定分布中最高点下方的分数。众数是与数据中的实际分数相对应的唯一的集中量数。

有两个众数的分布被称为双峰分布，有两个以上的众数的分布被称为多峰。具有几个相同高点的分布也被认为没有众数。

从技术上讲，众数是具有绝对最高频率的分数。然而，众数这个术语通常更随意地用来指代具有相对高频率的分数，即使这些高峰不是绝对最高点。当两个众数具有不同的频率时，更高的峰通常被称为主要众数，较短的众数被称为次要众数，以区分这两个值。

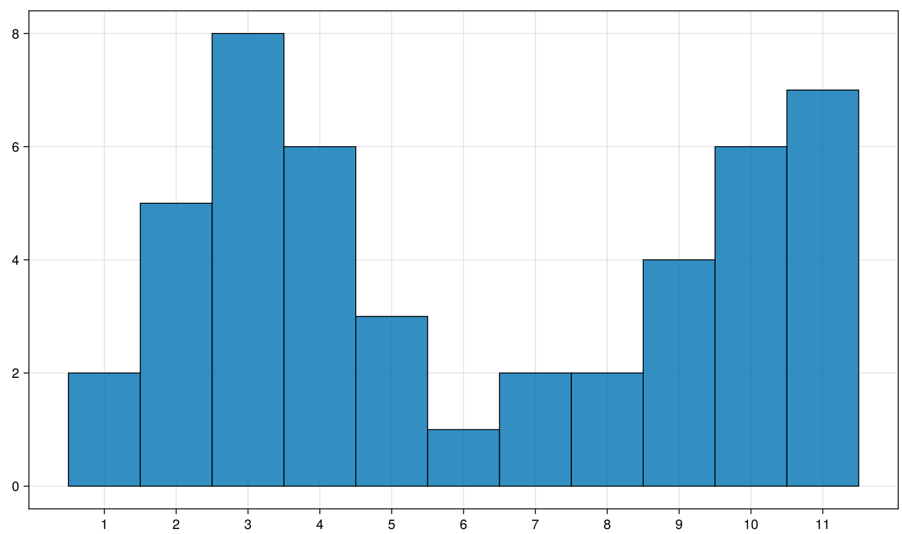
例如 $n = 100$ 名学生中选择的最喜欢的餐馆。

餐馆	频率
穆斯林	5
Hope	16
一心	42
韩国菜	18
南阳菜	7
食堂	7

3.5.2 频率分布

```
fq = [2, 5, 8, 6, 3, 1, 2, 2, 4, 6, 7]
X = vcat(fill.(1:length(fq), fq)...)
hist(X, bins = 0.5:1:11.5, strokewidth = 1, strokecolor = :black,
    axis = (; xticks = 1:11)
)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



3.6 集中趋势与分布形状

3.6.1 对称分布

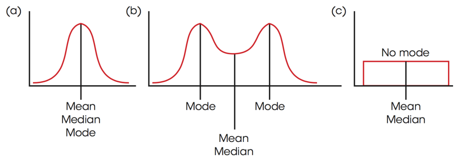


图 3.1: 对称

3.6.2 偏斜分布

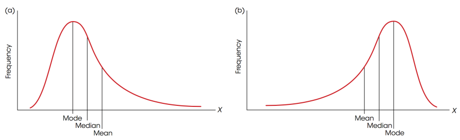


图 3.2: 偏态

3.7 选择集中趋势测量

3.7.1 何时使用平均值

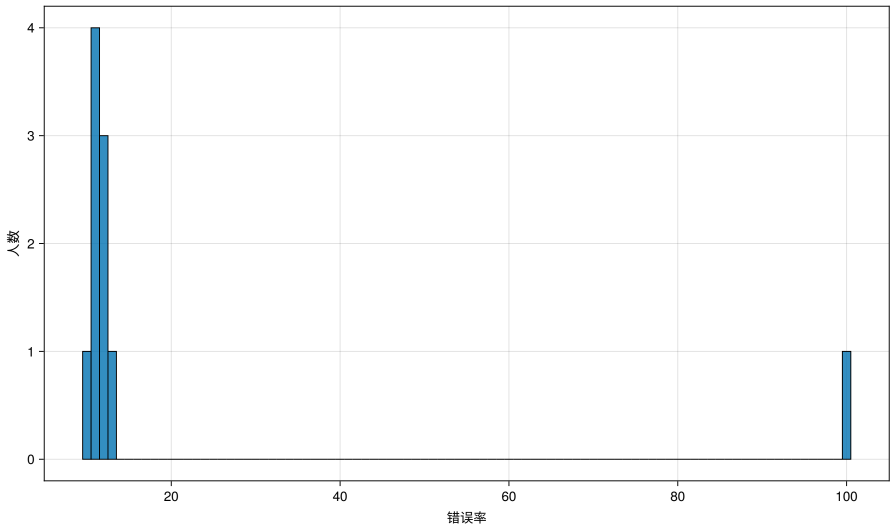
- 每当分数是数值值（区间或比例尺）时，平均值通常是首选的集中趋势测量。
- 因为平均值使用了分布中的每个分数，所以通常会产生一个良好的代表值。
- 除了是一个良好的代表值之外，平均值还有一个附加优势，即它与方差和标准差最常见的变异性测量相关。

3.7.2 何时使用中位数

- 极端分数或偏斜分布：当分布中有一些极端分数导致平均值的值被位移时，中位数是首选的集中趋势测量。

```
X = [10; fill(11, 4); fill(12, 3); 13; 100]
hist(X, bins = 9.5:1:100.5, strokewidth = 1, strokecolor = :black,
axis = (xlabel = "错误率", ylabel = "人数")
)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 未确定的值，开放性分布：当存在未确定的（无限，一个类别没有上限或下限）分数使得计算平均值成为不可能时，也可以使用中位数。

- 如组装木质拼图所需的分钟数

人物	时间（分钟）
1	8
2	11
3	12
4	13
5	17
6	从未完成

- 又如点披萨的数量

比萨的数量	频率
5 个或更多	3
4	2
3	2
2	3
1	6
0	4

- 有序刻度：最后，对于来自有序刻度的数据，中位数是首选的集中趋势测量。

3.7.3 何时使用众数

- 名义刻度：众数是描述名义数据集中趋势的唯一选择。
- 离散变量：在许多情况下，特别是在离散变量的情况下，人们更愿意使用众数产生的现实整数值。
- 描述形状：众数（或众数）的值还可以表示分布的形状，以及集中趋势的测量。

3.8 报告集中趋势测量

3.8.1 报告集中趋势测量

APA 风格使用字母 M 作为样本平均值的符号。因此，一项研究可能会陈述：*治疗组在任务上表现出较少的错误*（$M = 2.56$）*比对照组*（$M = 11.76$）*更少*。中位数可以使用缩写 Mdn 报告，如 $Mdn = 8.5$个错误，或者可以在叙述文本中报告，如下所示：*治疗组的错误中位数为 8.5，而对照组的错误中位数为 13*。

3.8.2 在图表中呈现平均值和中位数

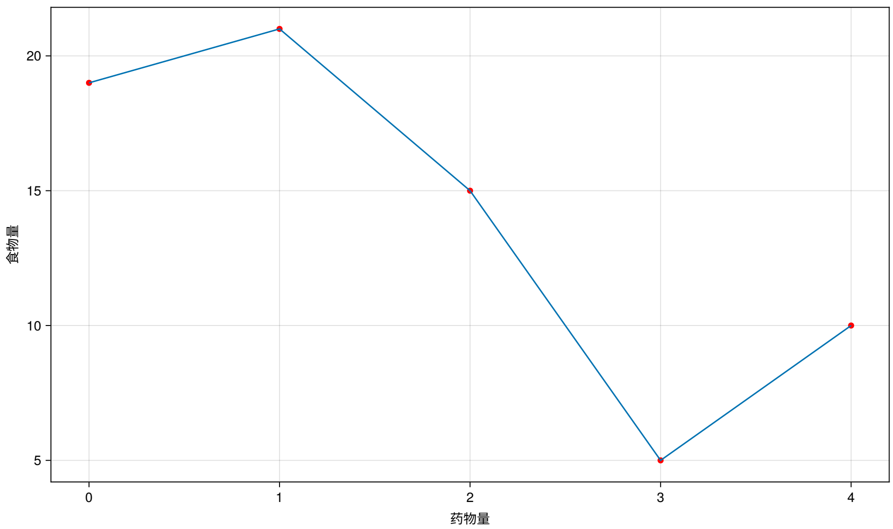
图表也可以用于报告和比较集中趋势测量。图表的价值在于它允许同时显示多个平均值（或中位数），因此可以快速比较不同组或处理条件。构建图表的规则有：

- 图表的高度应大约为其长度的三分之二到四分之三。
- 通常情况下，您应该从两个轴相交的地方开始编号 X 轴和 Y 轴。
- 然而，当数据中包含零值时，通常会将零点从交叉点移开，以便图表不会重叠轴。

3.8.3 在图表中呈现平均值和中位数

```
X, Y = 0:4, [19, 21, 15, 5, 10]
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xlabel = "药物量", ylabel = "食物量")
scatter!(ax, X, Y, color = :red)
lines!(ax, X, Y)
fig
```

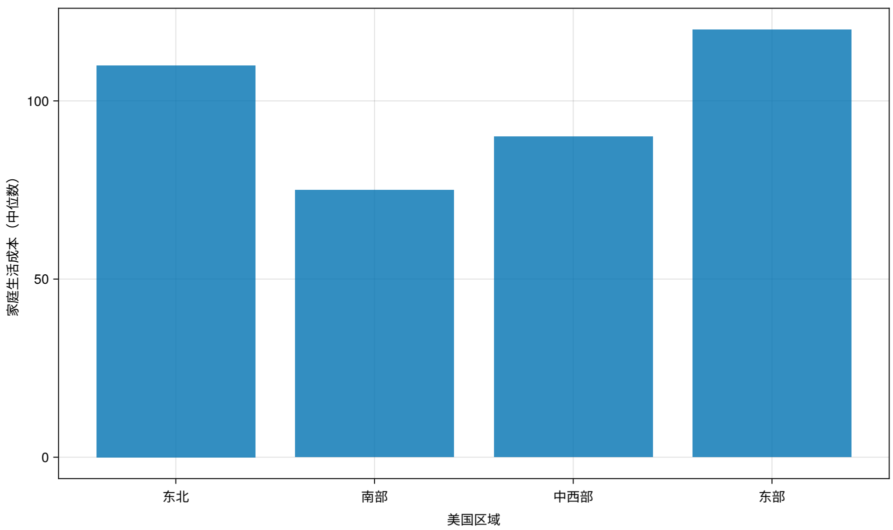
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



3.8.4 在图表中呈现平均值和中位数

```
Y = [110, 75, 90, 120]
barplot(Y, axis = (;
    xticks = (1:4, ["东北", "南部", "中西部", "东部"]),
    xlabel = "美国区域", ylabel = "家庭生活成本（中位数）"
))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



第四章　变异性

4.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using Distributions: Normal
using Statistics: mean, std, var, sqrt
using CairoMakie, DataFrames
using StatsReIntro
```

4.2　变异性介绍

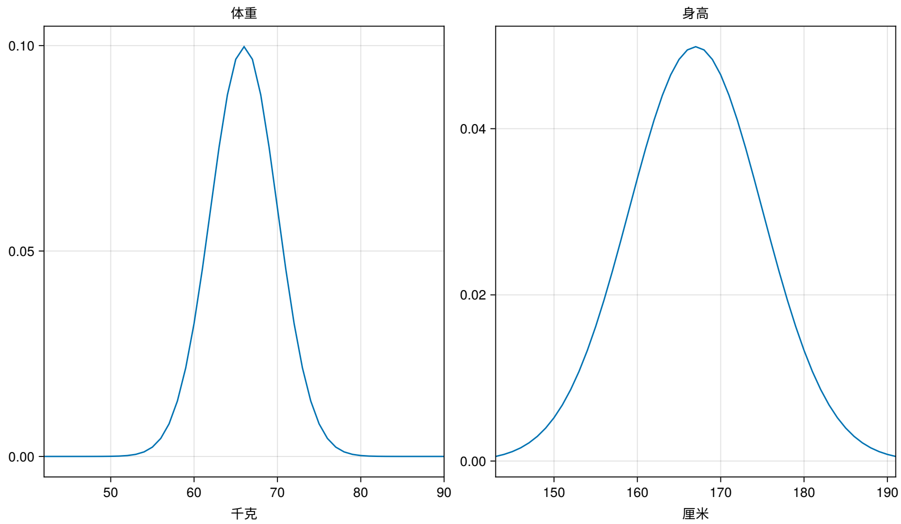
4.2.1　变异性介绍

变异性（Variability）度量了分布中分数间的差异数量，并描述了分布中分数分散或聚集在一起的程度。若分布中分数都相同，则就没有变异性。若分数间差异很小，则变异性较小。若分数间存差异很大，则变异性较大。

4.2.2　成年男性的身高和体重

```
fig = Figure(size = (1000, 600))
ax1 = Axis(fig[1, 1], limits = ((66 - 24, 66 + 24), nothing), title = "体重", xlabel = "千克"
ax2 = Axis(fig[1, 2], limits = ((167 - 24, 167 + 24), nothing), title = "身高", xlabel = "厘
lines!(ax1, (-24:24) .+  66, Normal(66, 4))
lines!(ax2, (-24:24) .+ 167, Normal(167, 8))

fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



4.2.3 变异量数的两个作用

1. 描述分布。变异性用“距离”来说明分数是聚集在一起还是分散在较大的距离上。
2. 变异性度量一个个体分数（或一组分数）代表整个分布的程度。变异性的这个方面在推论统计学中非常重要，推论统计学通常使用相对较小的样本来回答关于总体的问题。变异性提供了关于如果使用样本来代表总体时可以期望多少误差的信息。

4.2.4 常见的变异性度量

- 方差和标准差
- 全距
- 四分位距

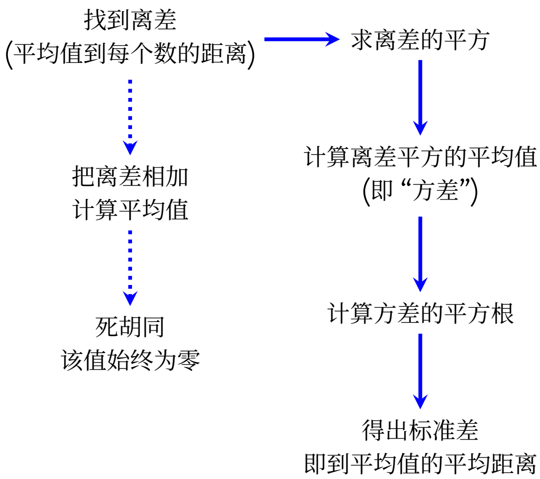
4.3 总体的方差和标准差

4.3.1 定义标准差和方差

标准差是最常用和最重要的变异性度量。标准差以分布的均值作为参考点，通过考虑每个分数与均值之间的距离来度量变异性。简单来说，标准差提供了对均值的标准或平均距离的度量，并描述了分数是否紧密聚集在均值周围或广泛分散。

4.3.2 方差和标准差的计算

尽管标准差的概念很简单，但实际方程往往更复杂。因此，我们首先看一下导致这些方程的逻辑。如果记得我们的目标是测量距离均值的标准或典型距离，那么这个逻辑和随后的方程应该更容易记住。**离差**（Deviation）是距离均值的距离：离差分数 $= X - \mu$



方差 (Variance) 等于平方偏差的均值。方差是距离均值的平均平方距离。**标准差**（Standard deviation）是方差的平方根，提供了对均值的标准或平均距离的度量。

$$\text{标准差} = \sqrt{\text{方差}}$$

4.3.3 方差和标准差的计算

给定以下 $N = 5$ 个分数

```
X = [1, 9, 5, 8, 7]
df = DataFrame(
    分数     = X,
    均值     = mean(X),
    离差     = X .- mean(X),
    离差平方 = (X .- mean(X)).^2
)
```

分数	均值	离差	离差平方
1	6.0	-5.0	25.0
9	6.0	3.0	9.0
5	6.0	-1.0	1.0
8	6.0	2.0	4.0
7	6.0	1.0	1.0

平方偏差的总和是:

$$\sum(X-\mu)^2 = 25+9+1+4+1 = 40.0$$

其方差是:

$$\text{方差} = \frac{\text{平方偏差总和}}{N} = \frac{40.0}{5} = 8.0$$

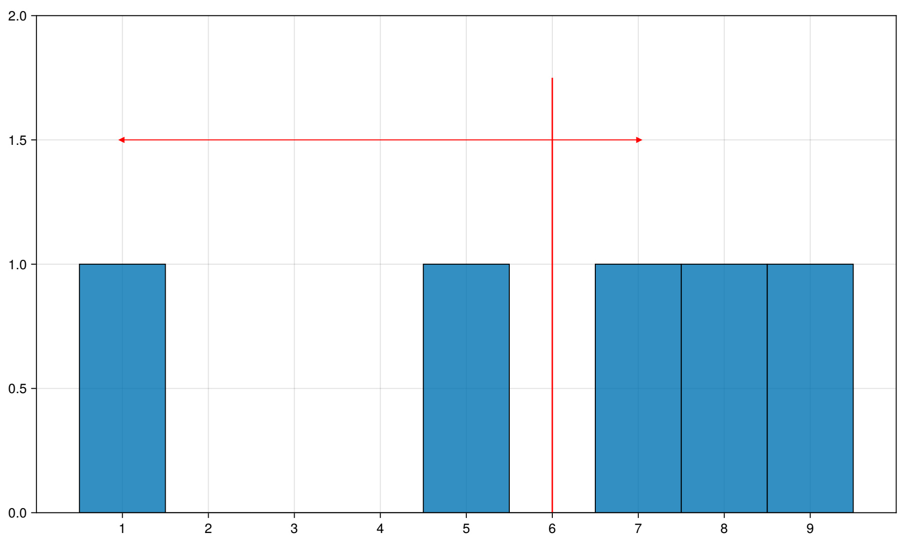
其标准差是:

$$\text{标准差} = \sqrt{\text{方差}} = \sqrt{8.0} = 2.8284271247461903$$

4.3.4 方差和标准差的计算

```
X = [1, 9, 5, 8, 7]
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xticks = 1:9, limits = ((0, 10), (0, 2)))
hist!(ax, X, bins = 0.5:1:9.5, strokewidth = 1)
vlines!(ax, [mean(X)], ymax = 1.75/2, color = :red)
arrows!(ax, [6, 6], [1.5, 1.5], [-5, 1], [0], color = :red)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



4.3.5 平方偏差总和（SS）

方差是平方偏差的均值。

$$\text{方差} = \text{平均平方偏差} = \frac{\text{平方偏差总和}}{\text{分数数量}}$$

离差平方和 (SS) 是平方偏差分数的总和。其定义公式为

$$SS = \sum (X-\mu)^2$$

根据其定义公式可做如下推导

$$\begin{aligned} SS &= \sum (X-\mu)^2 \\ &= \sum (X^2 - 2\cdot X \cdot \mu + \mu^2) \\ &= \sum X^2 - 2\cdot\mu\cdot\sum X + N\cdot\mu^2 \\ &= \sum X^2 - 2\cdot\frac{\sum X}{N}\cdot\sum X + N\cdot\frac{(\sum X)^2}{N^2} \\ &= \sum X^2 - 2\cdot\frac{(\sum X)^2}{N} + \frac{(\sum X)^2}{N} \\ &= \sum X^2 - \frac{(\sum X)^2}{N} \end{aligned}$$

从而得出离差平方和的计算公式：

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

在 Julia 语言中，离差平方和的定义和计算公式可做如下表示：

```
X = [1, 0, 6, 1]
SS = sum((X .- mean(X)).^2)
SS = sum(X.^2) .- (sum(X))^2 / length(X)
```

```
22.0
```

4.3.6 最终的公式和符号

总体方差（Population variance）由符号 σ^2 表示，等于距离均值的平方距离的平均值。总体方差通过将平方和除以 N 来获得。

$$\text{总体方差} = \sigma^2 = \frac{SS}{N}$$

总体标准差（Population standard deviation）由符号 σ 表示，等于总体方差的平方

根。

$$\text{总体标准差} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$$

4.4 样本的方差和标准差

推论统计学的目标是利用样本中的有限信息来推断总体的一般结论。这个过程的基本假设是样本应该代表它们来自的总体。这个假设对于变异性来说有一个特殊的问题，因为样本往往比它们的总体具有更小的变异性。

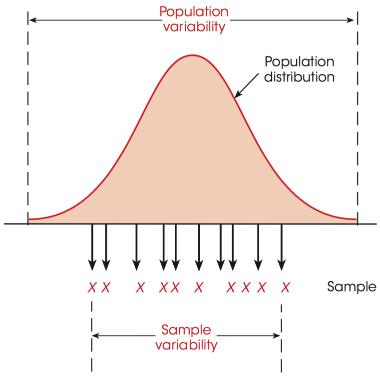


图 4.1: 样本变异性

总体中的一些极端分数往往会使总体的变异性相对较大。但是，当你选择一个样本时，很难获得这些极端值，这意味着样本的变异性相对较小。样本往往比其总体具有更小的变异性的事实意味着样本变异性给出了总体变异性的一个**有偏估计**。幸运的是，样本变异性中的偏差是一致且可预测的，这意味着它可以被校正。

4.4.1 样本离差平方和

样本 **离差平方和** (SS) 的定义公式是:

$$SS = \sum (X - M)^2$$

样本**离差平方和** (SS) 的计算公式是:

$$SS = \sum X^2 - \frac{(\sum X)^2}{n}$$

4.4.2 样本方差和标准差

样本方差（Sample variance）由符号 $\mathbf{s}^2$ 表示，等于距离均值的平方距离的平均值。样本方差通过将离差平方和除以 $n-1$ 来获得。

$$样本方差 = s^2 = \frac{SS}{n-1}$$

样本标准差（Sample standard deviation）由符号 $\mathbf{s}^2$ 表示，等于样本方差的平方根。

$$样本标准差 = s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

记住, 样本方差和标准差的公式的构建逻辑是使样本变异性能够很好地估计总体变异性。出于这个原因，样本方差通常被称为总体方差的估计（estimated population variance)，样本标准差被称为总体标准差的估计（estimated population standard deviation)。当只有一个样本能处理时，样本的方差和标准差提供了总体变异性的最佳估计。

4.4.3 样本变异性和自由度

尽管离差分数的概念和计算 SS 的方式在样本和总体中几乎完全相同，但符号上的微小差异实际上非常重要。

对于总体，你通过测量每个分数与总体均值之间的距离来找到每个分数的偏差。对于样本，总体均值 μ 是未知的，必须测量每个分数与样本均值的距离。由于样本均值的值在不同样本间会有所不同，所以在计算偏差前要先计算样本均值。样本均值 M 的计算对样本中的分数的变异性施加了限制。

假设我们选择了一个包含 $n = 3$ 个分数的样本并计算出其样本平均值为 $M = 5$。此时，如果样本的前两个数，如 $X = (2, 9, \_)$，则样本的第三个数就确定了。也就是说，对于一个包含 n 个分数的样本, 样本方差的**自由度** (Degrees of Freedom, df) 为 $df = n-1$。自由度决定了样本中独立和自由变化的分数的数量。

要计算样本方差，先求离差平方和，再除以可以自由变化的分数数量，即 $n-1 = df$。样本方差的公式是

$$s^2 = \frac{平方偏差总和}{可以自由变化的分数的数量} = \frac{SS}{df} = \frac{SS}{n-1}$$

4.4.4 样本方差和标准差

```
X = [4, 6, 5, 11, 7, 9, 7, 3]
n = length(X)
SS = sum((X .- mean(X)) .^2)
SS = sum(X.^2) - sum(X)^2 / n

v = SS / (n-1)
s = sqrt(v)
```

```
2.6186146828319083
```

Julia 语言中的两个函数

```
var(X), std(X)
```

```
(6.857142857142857, 2.6186146828319083)
```

4.4.5 有偏和无偏统计数据

如果样本统计量的平均值等于总体参数，则样本统计量是**无偏**（unbiased）的。统计量的平均值是从具有特定样本大小 n 的所有可能样本中获得的。如果统计量的平均值要么低估总体参数，要么高估总体参数，则统计量是**有偏**（biased）的。

假设有一个包含 $N = 3$ 个分数的总体：

```
X = [0, 3, 9]
```

```
3-element Vector{Int64}:
 0
 3
 9
```

这个总体的均值和方差分别为：

```
mean(X), var(X, corrected = false)
```

```
(4.0, 14.0)
```

接下来，我们从这个总体中选择 $n = 2$ 个分数的所有样本，即每一个可能的 $n = 2$ 个分数的样本。

```
samp = vcat([[x y] for x in X for y in X]...)
df = DataFrame(samp, [:数值1, :数值2])
transform!(df,
```

```
    AsTable(:) => ByRow(mean) => :均值,
    AsTable(:) => ByRow(x -> var(x, corrected=false)) => :有偏方差,
    AsTable(:) => ByRow(x -> var(x, corrected=true))  => :无偏方差
)
```

数值 1	数值 2	均值	有偏方差	无偏方差
0	0	0.0	0.0	0.0
0	3	1.5	2.25	4.5
0	9	4.5	20.25	40.5
3	0	1.5	2.25	4.5
3	3	3.0	0.0	0.0
3	9	6.0	9.0	18.0
9	0	4.5	20.25	40.5
9	3	6.0	9.0	18.0
9	9	9.0	0.0	0.0

样本均值和两个方差的平均值是

```
mean(df.均值), mean(df.有偏方差), mean(df.无偏方差)
```

```
(4.0, 7.0, 14.0)
```

4.5 补充内容

4.5.1 尺度变换

- 对每个分数添加一个常数不会改变标准差。
- 将每个分数乘以一个常数会导致标准差乘以相同的常数。

4.5.2 报告标准差

在报告研究结果时，研究人员通常会提供有关集中趋势和变异性的描述信息。心理学研究中的因变量通常是从等距或等比尺度上的测得的数值型数据。对于数值分数，最常见的描述性统计量是均值（集中趋势）和标准差（变异性），通常一起报告。

在许多期刊中，尤其是遵循 APA 样式的期刊，用符号 SD 表示样本标准差。例如

观看了暴力动画的儿童显示出更多的攻击性反应（$M = 12.45$，$SD = 3.7$），而观看了控制动画的儿童显示出更少的攻击性反应（$M = 4.22$，$SD = 1.04$）。

在报告多个群体的描述性统计量时，研究结果可以总结在一张表格中。

	视频游戏类型	
	暴力	非暴力
男性	M=7.72 SD=2.43	M=4.34 SD=2.16
女性	M=2.47 SD=0.92	M=1.61 SD=0.68

4.5.3 标准差和描述性统计

- **描述整个分布**。因为均值确定了分布的中心，标准差描述了距离均值的平均距离，所以这两个值应该让你能够创建一个相对准确的整个分布的图像。
- **描述个体分数的位置**。知道均值和标准差还应该让你能够描述分布中任何个体分数的相对位置。

4.5.4 方差和推论统计

从非常一般的角度来看，推论统计的目标是检测研究结果中的有意义和显著的模式。基本问题是观察到的样本数据中的模式是否反映了总体中存在的相应模式，还是只是偶然发生的随机波动。变异性在推论过程中起着重要作用，因为数据中的变异性影响着是否能够清晰地看到模式。一般来说，低变异性意味着现有模式可以清晰地看到，而高变异性倾向于掩盖可能存在的任何模式。

4.5.5 方差和推论统计

- 研究 A

方法 1	方法 2
35	39
34	40
36	41
35	40

- 研究 A 的平均值

方法	平均值
方法 1	35.0
方法 2	40.0

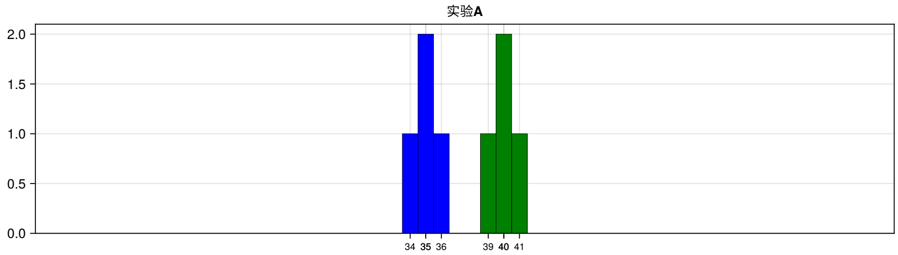
- 研究 B

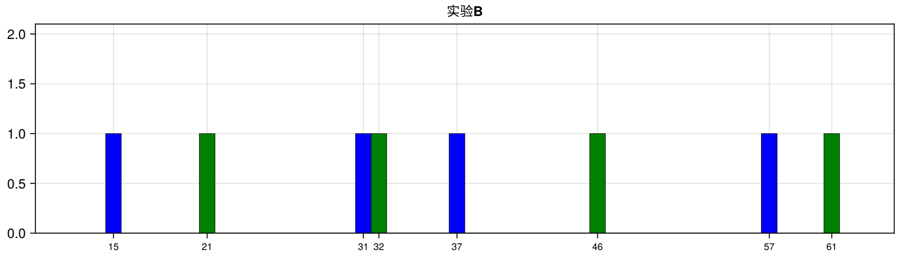
方法 1	方法 2
31	46
15	21
57	61
37	32

- 研究 B 的平均值

方法	平均值
方法 1	35.0
方法 2	40.0

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```





- 研究 A 的平均值和标准差

方法	平均值	标准差
方法 1	35.0	0.666667
方法 2	40.0	0.666667

- 研究 B 的平均值和标准差

方法	平均值	标准差
方法 1	35.0	301.333
方法 2	40.0	300.667

4.6 全距和四分位距

4.6.1 全距

全距（Range）是分布中分数的距离，从最小分数到最大分数。

$$\text{全距} = X_{max} - X_{min}$$

当分数是连续变量的测量时，全距可以定义为实际极限之间的差异。

$$\text{全距} = X_{max} - X_{min} + 1$$

4.6.2 Julia 中的全距函数

```
X = collect(2:3:17)
extrema(X),
minimum(X),
maximum(X),
min(X...),
max(X...),
```

```
diff([extrema(X)...])[]
```

```
((2, 17), 2, 17, 2, 17, 15)
```

4.6.3 四分位距

通常情况下，不是查看数据的整个全距，而是查看数据的中间 50 距离。为此，我们使用**四分位数**（Quartiles）。**四分位数**是将数据分为四个大致相等的部分的数值。第一个四分位数（Q1）是数据的下 25% 的分数。第二个四分位数（Q2）是数据的中间 50% 的分数，通常称为**中位数**。第三个四分位数（Q3）是数据的下 75% 的分数。四分位数的计算方式和中位数类似，只是它们将数据分为四个相等的部分，而不是两个。

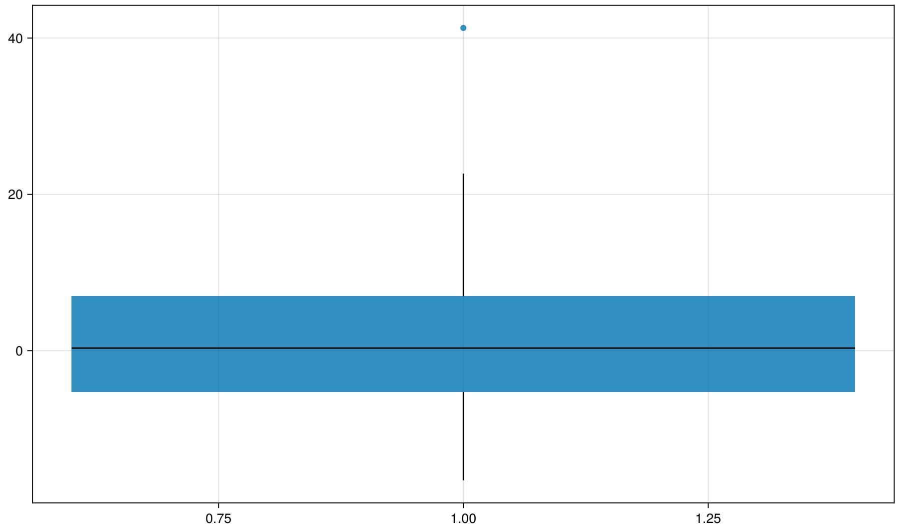
四分位距（Interquartile Range，IQR）是第三个四分位数（Q3）和第一个四分位数（Q1）之间的距离。四分位距是一个度量数据的变异性的方式，但它不受极端值的影响。计算四分位距的公式是

$$IQR = Q3 - Q1$$

四分位距提供了一个对中间 50% 数据的变异性的度量。IQR 提供了有关数据集的**分散程度**的信息。如果 IQR 较小，则数据分散较小，大多数数据分布在靠近中位数的地方。如果 IQR 较大，则数据分散较大，数据点分布在一个较宽的距离内。四分位距对于识别数据中的**异常值**非常有用。常见的异常值检测方法是使用四分位距的 1.5 倍或 3 倍来计算**内限**（Inner Fences）和**外限**（Outer Fences），然后识别落在外限之外的值作为异常值。一些箱线图将内限和外限的概念添加到箱线图中，以帮助可视化异常值。

```
X = randn(50) .* 10
boxplot(ones(length(X)), X)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



4.6.4 与标准差的比较

与标准差相比，四分位距具有一些优点和局限性。四分位距不受极端值的影响，因此对于含有极端值的数据集来说更稳健。标准差提供了更多关于数据的详细信息，但对于非正态分布的数据可能会产生不准确的度量。

4.6.5 结论

变异性是描述数据分布的重要特征，用于衡量数据点之间的差异程度。方差和标准差是衡量数据分散程度的常用统计指标，它们基于数据点与均值的距离。样本方差和标准差是用于估计总体方差和标准差的无偏估计。全距和四分位距是用于度量数据分散程度的另外两种方法，它们不受极端值的影响。四分位距还可以用于检测异常值。

第二部分

推论统计基础

第五章 z 分数介绍

5.1 本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using LinearAlgebra: inv
using Statistics: mean, std
using StatsReIntro
```

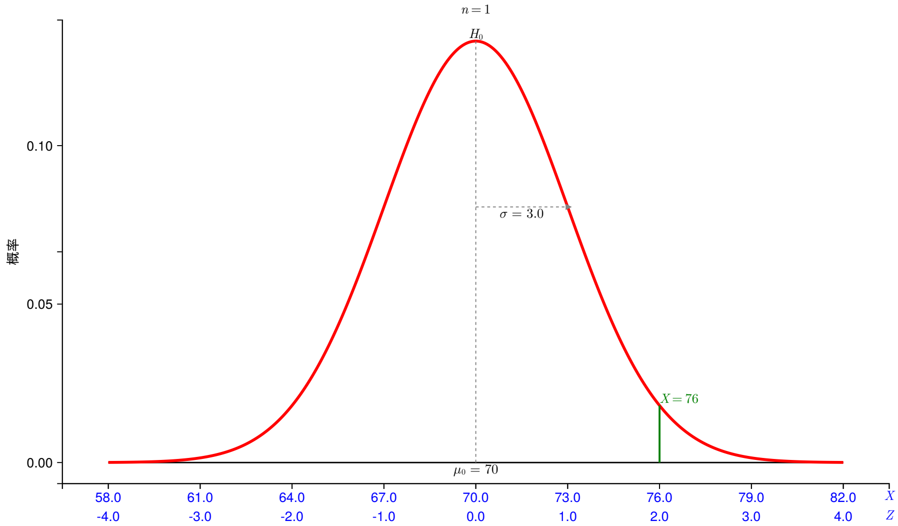
5.2 z 分数简介

在前两章中，我们介绍了均值和标准差的概念，作为描述整个分数分布的方法。现在我们将注意力转向分布内部的个体分数。在本章中，我们介绍一种使用均值和标准差将每个分数（X 值）转化为 z-分数或**标准分数** (standard score) 的统计技术。z-分数或标准分数的目的是确定和描述分布中每个分数的确切位置。

假设你在一次统计考试中得分 $\mathbf{X = 76}$。你学的好还是不好？

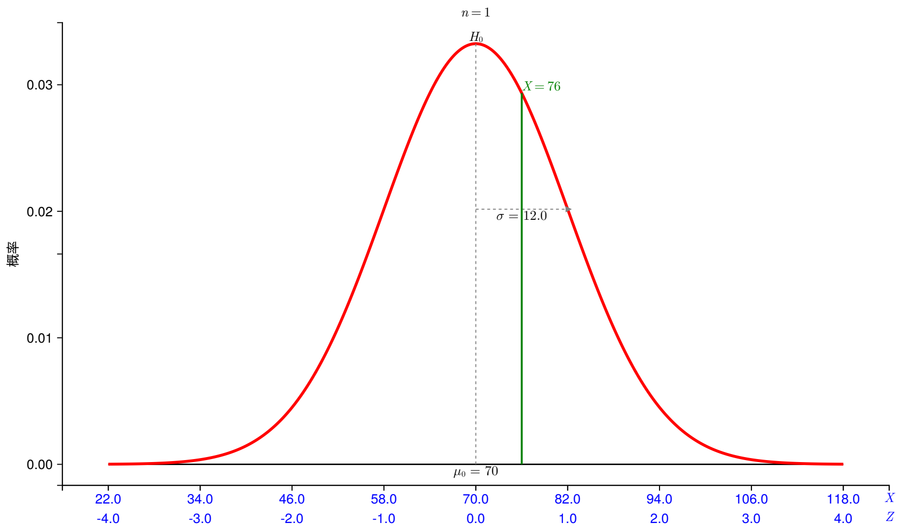
```
PlotZ(70, 3; XVal = 76)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```julia
PlotZ(70, 12; XVal = 76)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



这些原始、未改变的分数是测量的直接结果，被称为**原始分数** (raw scores)。为了使原始分数更有意义，它们通常被转化为包含更多信息的新值。这个转化是 z-分数的一个目的。特别地，我们将 X 值转化为 z-分数，以便得到的 z-分数精确地告诉原始分数的位置。z-分数的第二个目的是标准化整个分布。一个常见的标准化分布示例是智商分数的分布。尽管有几种不同的智商测量方法，但这些测量通常都被标准化，以使它们的均值为 100，标准差为 15。

z-分数的两个有用目的：(1) 每个 z-分数告诉原始 X 值在分布中的确切位置；(2) z-分数形成一个标准化分布，可以直接与其他已转化为 z-分数的分布进行比较。

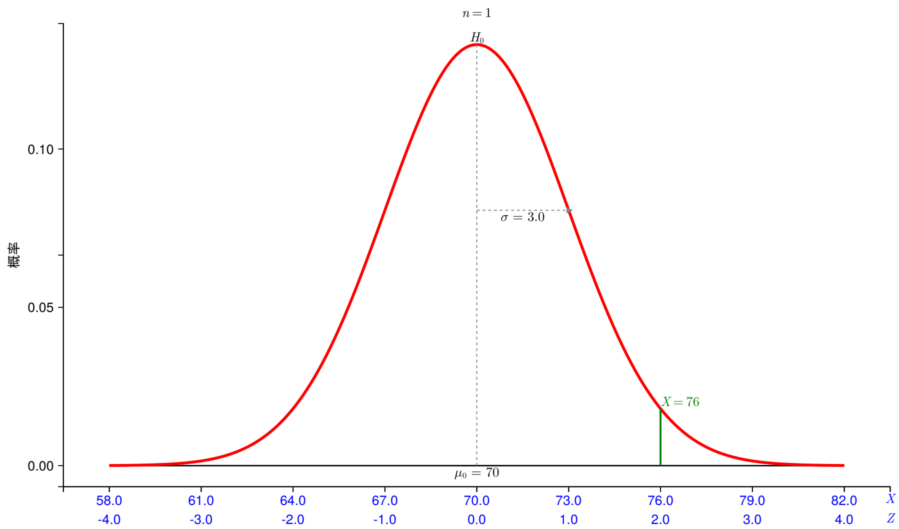
5.3 z-分数与分布中的位置

5.3.1 z-分数与分布中的位置

z-分数指定了分布中每个 X 值的精确位置。z-分数的符号（+ 或-）表示分数是否高于均值（正数）或低于均值（负数）。z-分数的数值表示从均值到 X 之间的标准差数量。由 z-分数确定的位置对于所有分布都是相同的，无论这些分布的均值或标准差是多少。

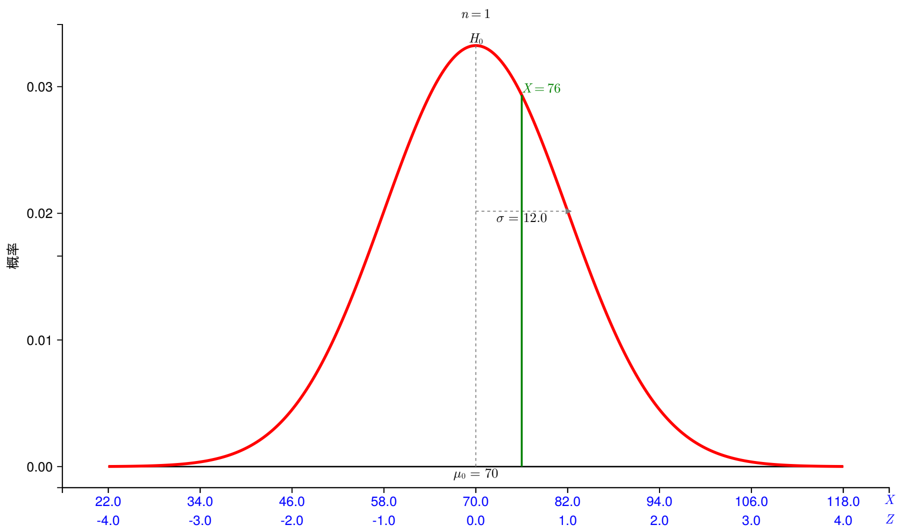
```
PlotZ(70, 3; XVal = 76, showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(70, 12; XVal = 76, showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



5.3.2 z-分数公式

将分数转化为 z-分数的公式为

$$z = \frac{X - \mu}{\sigma}$$

方程的分子，$X - \mu$，是一个离差分数；它表示 X 与 μ 之间的点之间的距离，并指示 X 是否位于均值之上或之下。然后将离差分数除以 σ，因为我们希望 z-分数以**标准差单位**（standard deviation units）来度量距离。

z-分数建立了**原始分数**（X）、**均值**（μ）和**标准差**（σ）之间的关系。知道这四个数中的任意三个，我们可以解出第四个。

$$X = z \times \sigma + \mu$$
$$\sigma = \frac{X - \mu}{z}$$
$$\mu = X - Z \times \sigma$$

5.3.3 z、X、μ 和 σ 之间的关系

- 一个分数分布的均值为 $\mu = 100$，标准差为 $\sigma = 10$。在这个分布中，$X = 130$ 的 z-分数是多少？

```
(130 - 100) / 10
```

```
3.0
```

- 一个分数分布的均值为 $\mu = 86$，标准差为 $\sigma = 7$。在这个分布中，$X = 95$ 的 z-分数是多少？

```
(95 - 86) / 7
```

```
1.2857142857142858
```

- 对于均值为 $\mu = 60$，标准差为 $\sigma = 8$ 的分布，一个 z-分数为 $z = -1.50$ 的 X 值是多少？

```
-1.50 * 8 + 60
```

```
48.0
```

- 在均值为 $\mu = 65$ 的总体中，分数 $X = 59$ 对应于 $z = -2.00$，这个总体的标准差是多少？

```
(59 - 65) / (-2)
```

```
3.0
```

- 在标准差为 $\sigma = 6$ 的总体中，分数 $X = 33$ 对应于 $z = +1.50$。这个总体的均值是多少？

```
33 - 1.5 * 6
```

```
24.0
```

$$z = \frac{X - \mu}{\sigma} \Leftrightarrow \sigma * z + \mu = X$$

- 在一个总体分布中，$X = 64$ 的 z-分数为 $z = 0.50$，$X = 72$ 的 z-分数为 $z = 1.50$，这个总体的均值和标准差是多少？

```
Cs = [0.5 1; 1.5 1]
Xs = [64; 72]
inv(Cs) * Xs
```

```
2-element Vector{Float64}:
  8.000000000000007
 59.99999999999999
```

- 在一个总体分布中，$X = 54$ 的 z-分数对应于 $z = +2.00$，$X = 42$ 的 z-分数对应于 $z = -1.00$。这个总体的均值和标准差是多少？

```
Cs = [2 1; -1 1]
Xs = [54; 42]
```

```
inv(Cs) * Xs
```

```
2-element Vector{Float64}:
  4.000000000000004
 46.0
```

5.4 z-分数标准化分布

5.4.1 z-分数分布的特性

可以将分布中的每个 X 值转化为相应的 z-分数。这个过程的结果是将所有 X 值的分布转化为 z-分数的分布。新的 z-分数分布具有使 z-分数转化成一个非常有用工具的特性。具体来说，如果每个 X 值都转化为 z-分数，那么 z-分数分布将具有以下特性:

- **形状**。z-分数的分布将与原始分数的分布形状完全相同。
- **均值**。z-分数分布的均值始终为零。
- **标准差**。z-分数分布的标准差始终为 1。

5.4.2 使用 z-分数标准化分布

当将任何分布（具有任何均值或标准差）转化为 z-分数时，结果分布将始终具有均值 $\mu = 0$ 和标准差 $\sigma = 1$。由于所有 z-分数分布都具有相同的均值和相同的标准差，所以 z-分数分布被称为**标准化分布** (standardized distribution)。

一个标准化分布由已转化以获得预定 μ 和 σ 值的分数组成。标准化分布用于使不同分布可比较。z-分数分布是一个均值为 $\mu = 0$ 和标准差为 $\sigma = 1$ 的标准化分布的例子。

5.4.3 使用 z-分数进行比较

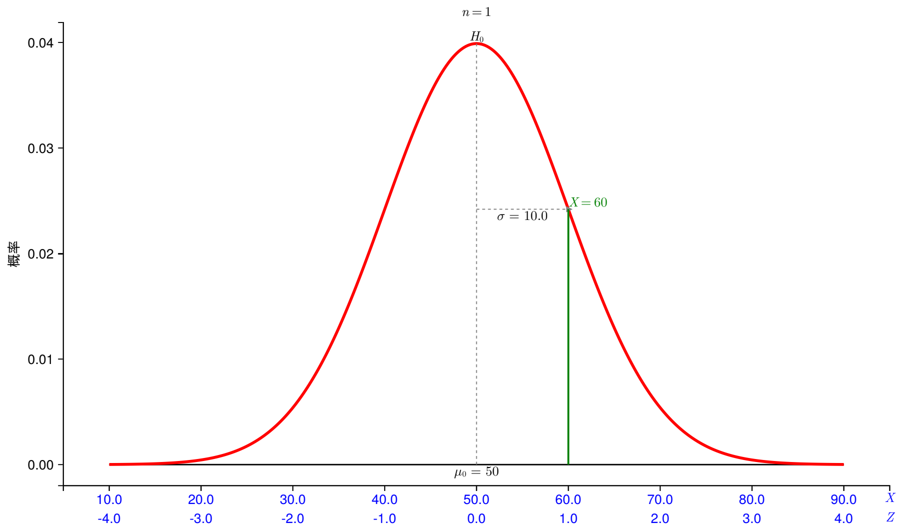
通常情况下，如果两个分数来自不同的分布，那么不可能直接比较它们。标准化分布的一个优势是，它可以使不同的分数或不同的个体进行比较，尽管它们来自完全不同的分布。

例如，假设张三在一次英语考试中得了 $X = 60$ 的分数，在一次法语考试中得了 $X = 56$ 的分数。他应该期望在哪门课程中获得更好的成绩？假设英语考试的分数均值为 $\mu = 50$，标准差为 $\sigma = 10$，而法语考试的分数均值为 $\mu = 48$，标准差为 $\sigma = 4$。

- 英语

```
PlotZ(50, 10; XVal = 60, showZ = true)
```

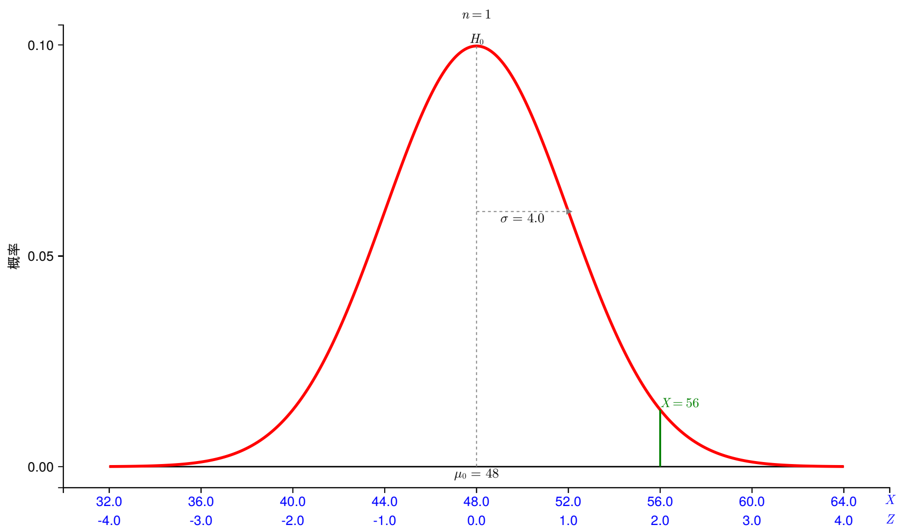
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 法语

```
PlotZ(48, 4; XVal = 56, showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



5.5 其他标准化分布

5.5.1 转化为具有预定 μ 和 σ 的分布

尽管 z-分数分布具有明显的优势，但许多人觉得它们很麻烦，因为它们包含负值和小数。因此，通常将分数转化为具有预定均值和标准差的新分布，这些均值和标准差是整数。目标是创建一个新的（标准化的）分布，具有“简单”的均值和标准差值，但不改变任何个体在分布中的位置。这种类型的标准化分数在心理或教育测试中经常使用。这种转化分两步：原始分数首先被被转化为 z-分数。然后，z-分数被转化为新的 X 值，以达到特定的 μ 和 σ。

例如，对于心理学考试，原始分数的分布均值为 $\mu = 57$，标准差为 $\sigma = 14$。我们希望通过将所有分数转化为一个新的标准化分布，该分布的均值为 $\mu = 50$，标准差为 $\sigma = 10$，来简化分布。假设有两名特定的学生：张三，在原始分布中得分 $X = 64$；李四，在原始分布中得分 $X = 43$。他们在新分布中的分数是多少？

```
function converter(X0, M0, SD0, M1, SD1)
    Z  = (X0 - M0) / SD0
    X1 = Z * SD1 + M1
    return X1
end
```

```
converter (generic function with 1 method)
```

```
converter(64, 57, 14, 50, 10),
converter(43, 57, 14, 50, 10)
```

```
(55.0, 40.0)
```

5.6 样本 z 分数

5.6.1 计算样本 z 分数

对于样本和总体，z-分数的定义是相同的，只要使用样本均值和样本标准差来指定每个 z-分数的位置即可。样本中的每个 X 值都可以转化为 z-分数，如下所示：

$$z = \frac{X - M}{s}$$

同样，每个 z-分数可以转化回 X 值，如下所示：

$$X = M + z \times s$$

- 在一个样本中，均值为 $M = 40$，标准差为 $s = 10$，与 $X = 35$ 对应的 z-分数是多少？与 $z = +2.00$ 对应的 X 值是多少？

```
(35 - 40) / 10,
2 * 10 + 40
```

```
(-0.5, 60)
```

5.6.2 标准化样本分布

转化后的 z-分数分布具有与将一组 X 值转化为 z-分数时存在的特性相同的特性。具体来说，(1) z-分数的样本将与原始分数的样本具有相同的形状；(2) z-分数的样本将具有均值 $M_z = 0$；(3) z-分数的样本将具有标准差 $s_z = 1$。

```
X = [0, 2, 4, 4, 5]
M, SD = mean(X), std(X)
Z = @. (X - M) / SD
```

```
5-element Vector{Float64}:
 -1.5
 -0.5
  0.5
  0.5
  1.0
```

```
mean(Z), std(Z)
```

```
(0.0, 1.0)
```

5.7 z-分数与推论统计

推论统计是使用样本的信息来回答关于总体的问题的技术。一个典型的研究从一个关于治疗如何影响总体中个体的问题开始。由于通常不可能研究整个总体，所以研究人员选择一个样本，并对样本中的个体进行治疗。为了评估治疗的效果，研究人员简单地比较接受治疗的样本与原始总体。如果样本中的个体与原始总体中的个体明显不同，那么研究人员有证据表明治疗产生了效果。另一方面，如果样本与原始总体没有明显不同，似乎治疗没有效果。请注意，研究结果的解释取决于样本是否**显著不同于**（noticeably different from）总体。

一个判断样本是否明显不同的技术是使用 z-分数。因此，我们可以使用 z-分数来帮助判断治疗是否产生了变化。具体地说，如果接受治疗的个体以极端的 z-分数结束研究，我们可以得出结论认为治疗似乎产生了效果。

第六章 概率

6.1 本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie
using Distributions: Normal, Binomial, pdf, cdf, ccdf, quantile, cquantile
using StatsReIntro
```

6.2 概率简介

6.2.1 推论统计和概率

研究开始时通常涉及整个总体的一般问题，但实际研究是使用样本进行的。推断统计的作用是使用样本数据来回答有关总体的问题。为了实现这个目标，推断过程通常围绕概率的概念构建。具体来说，样本和总体之间的关系通常是用*概率*来定义的。

例如，*假设你从一个装有 50 个黑色和 50 个白色弹珠的罐子中选择一个弹珠*。在这个例子中，弹珠罐是总体，要选择的单个弹珠是样本。尽管你不能保证样本的确切结果，但可以根据概率来讨论可能的结果。在这种情况下，*你有一半的机会获得任一颜色*。现在考虑*另一个罐子（总体），其中有 90 个黑色和只有 10 个白色弹珠*。再次，你无法指定样本的确切结果，但现在你知道*样本很可能是一个黑色弹珠*。通过了解总体的构成，我们可以确定获得特定样本的概率。通过这种方式，*概率*为我们提供了总体和样本之间的联系，这是接下来的章节中将要介绍的推断统计的基础。

你可能已经注意到，前面的例子都是从总体开始，然后使用概率来描述可能获得的样本。这与我们想要进行的推断统计正好*相反*。记住，推断统计的目标是从样本开始，然后回答有关总体的一般问题。我们通过两个阶段来实现这个目标。在第一阶段中，我们将概率作为从总体到样本的桥梁来发展。这个阶段涉及识别从特定总体中可能获得的样本类型。一旦建立了这个桥梁，我们只需反转概率规则，以允许我们从样本到总体的移动。

反转概率关系的过程可以通过再次考虑我们之前看过的两个弹珠罐（总体）来演示：

- *罐子 1 有 50 个黑色和 50 个白色弹珠*;
- *罐子 2 有 90 个黑色和只有 10 个白色弹珠*。

这次，假设在选择样本时你被蒙上了眼睛，所以你不知道使用了哪个罐子。你的任务是看你获得的样本，然后决定哪个罐子最有可能被使用。

如果你选择了一个由 4 颗弹珠组成的样本，而且全都是黑色，你会选择哪个罐子？很明显，从罐子 1 中获得这个样本的概率相对较低，因为在四次抽样中，你几乎肯定会得到至少 1 颗白色弹珠。另一方面，这个样本来自罐子 2 的概率很高，因为几乎所有的弹珠都是黑色的。因此，你的决定是这个样本很可能来自罐子 2。请注意，你现在是在使用样本来对总体做出推断。

6.2.2 概率的定义

概率是一个广泛的主题，远远超出了统计学入门的范围，我们不会在这里尝试全面讨论。相反，我们专注于入门推断统计所需的一些概念和定义。我们从概率的一个相对简单的定义开始。对于可能出现多种不同结果的情况，任何特定结果的**概率**（Probability）定义为所有可能结果的分数或*比率*（Proportion）。如果可能的结果被标识为 A、B、C、D 等，那么

$$P(A) = \frac{\text{被分类为 A 的结果数}}{\text{可能结果的总数}}$$

例如，如果你从一副完整的牌中选择一张牌，那么有 52 种可能的结果。选择红桃皇后的概率是：

$$P(\text{红桃皇后}) = \frac{1}{52}$$

选择一张 A 的概率是：

$$P(\text{A}) = \frac{4}{52}$$

选择一张黑桃牌的概率是：

$$P(\text{黑桃}) = \frac{13}{52} = \frac{1}{4} = 0.25 = 25\%$$

或者，如果你抛掷一枚硬币，得到正面的概率是

$$P(\text{正面}) = \frac{1}{2} = 0.5 = 50\%$$

6.2.3 概率值

你还应该注意，所有可能的概率值都包含在一个有限范围内。在一个极端情况下，当一个事件从不发生时，概率为零，或 0%。在另一个极端情况下，当一个事件总是发生时，概率为 1，或 100%。因此，所有概率值都包含在 0-1 的范围内。例如，假设你有一个罐子，里面有 10 颗白色弹珠。随机选择一个黑色弹珠的概率是

$$P(\text{黑色}) = \frac{0}{10} = 0$$

选择一个白色弹珠的概率是

$$P(\text{白色}) = \frac{10}{10} = 1\%$$

6.2.4 随机抽样

为了使概率的上述定义准确，必须通过一种称为随机抽样的过程来获得结果。**随机样本** (random sample) 要求总体中的每个个体都有*相等的机会*被选中。通过这个过程获得的样本被称为**简单随机样本** (simple random sample)。

第二个要求,对于许多统计公式来说是必要的,它规定如果选择的不止一个个体,从一次选择到下一次选择概率必须保持*不变*。添加这第二个要求会产生*独立随机抽样* (independent random sampling)。术语*独立*指的是选择任何特定个体的概率与已经选择的个体无关。例如，你被选中的概率是固定的，不会因为其他个体在你之前被选中而改变。**独立随机抽样**要求每个个体都有相等的机会被选中，并且如果选择不止一个个体，概率必须从一次选择到下一次选择保持不变。

由于独立随机样本通常是大多数统计应用的必要组成部分，我们将始终假定这是所使用的抽样方法。为了简化讨论，我们通常会省略"独立"这个词，只是将这种抽样技术简单地称为随机抽样。但你应该始终假设两个要求（*相等的机会*和*概率恒定*）都是这个过程的一部分。

6.2.4.1 随机抽样：相等机会

随机抽样的两个要求都有一些有趣的后果。第一个要求确保选择过程没有偏差。随机抽样的第一个要求禁止你将概率的定义应用于可能结果不均等的情况。例如，考虑从一副完整的牌中抽取 $n = 2$ 张牌的情况。对于第一次抽取，获得方块 J 的概率是

$$P(\text{方块 J}) = \frac{1}{52} = 2\%$$

在选择一张牌作为样本后，你准备好抽取第二张牌。这次获得方块 J 的概率是多少？

6.2.4.2 随机抽样：概率恒定

假设你仍然拿着第一张牌，有两种可能性：如果第一张牌不是方块 J，则获得方块 J 的概率为

$$P(\text{方块 J}) = \frac{1}{51} = 2\%$$

如果第一张牌是方块 J，则获得方块 J 的概率是

$$P(\text{方块 J}) = \frac{0}{51} = 0\%$$

在任一情况下，概率都与第一次抽取的概率不同。这违反了随机抽样的要求，该要求规定概率必须保持恒定。

为了使概率不会从一次选择到下一次选择发生变化，必须在每次选择之前将每个个体返回到总体中。这个过程称为*放回取样* (sampling with replacement)。随机样本的第二个要求（概率恒定）要求你进行放回取样。

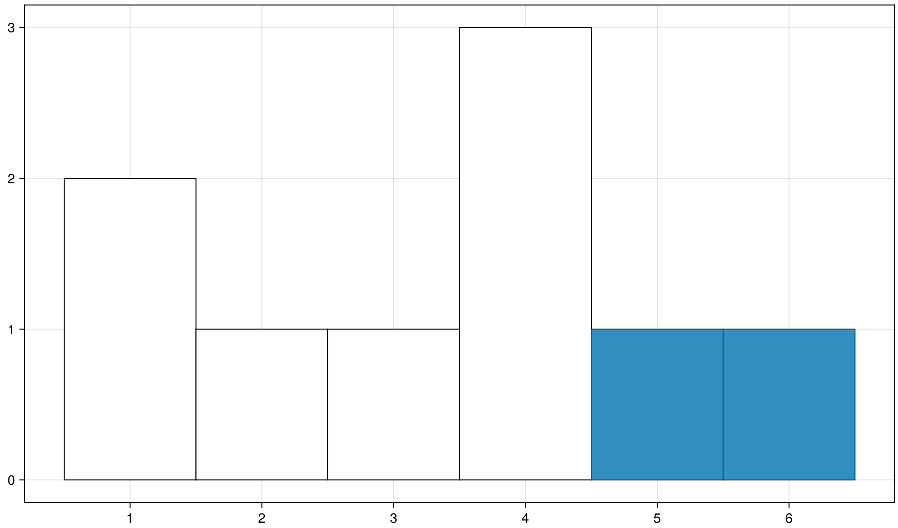
6.2.5 概率和频率分布

我们将使用一个非常简单的总体，其中包含 10 个分数，分别是 1、1、2、3、3、4、4、4、5、6。如果你从这个总体中随机选择一个 $n = 1$ 分数样本，那么获得大于 4 的个体的概率是多少？用概率符号表示，$P(X > 4) = ?$

我们通常关注的概率问题涉及到一个可以在频率分布图中显示的分数（X 值）总体。如果将图形视为表示整个总体，那么图形的不同部分表示总体的不同部分。由于概率和比率是等价的，图形的特定部分对应于总体中的特定比率。因此，每当将总体表现为频率分布图时，将可能将概率表示为图形的比率。

```
X = [1, 1, 2, 3, 4, 4, 4, 5, 6]
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xticks = unique(X))
hist!(ax, X,
  bins = (minimum(X)-0.5):1:(maximum(X)+0.5),
  strokewidth = 1, strokecolor = :black, color = :white,
)
hist!(ax, X[X.>4], bins = (minimum(X)-0.5):1:(maximum(X)+0.5))

fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



6.3 概率和正态分布

6.3.1 正态分布

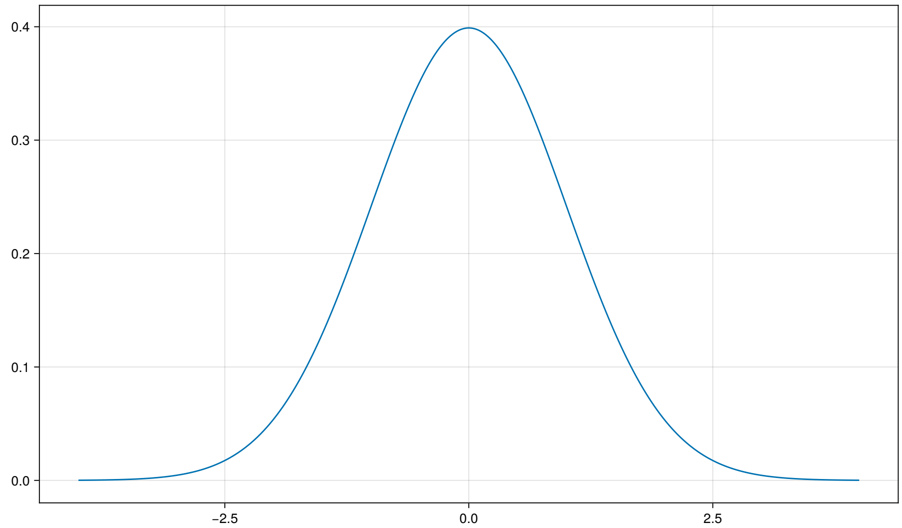
*正态分布*是总体分布中常见形状的一个例子。正态分布的确切形状由一个方程来指定，该方程将每个 X 值（分数）与每个 Y 值（频率）相关联。该方程是

$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

（π 和 e 是数学常数）

```
MyNorm(X; μ = 0, σ = 1) = 1/sqrt(2π*σ^2)*e^(-(X-μ)^2/(2σ^2))
lines(-4:0.01:4, MyNorm)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



简单地说，正态分布是对称的，并在中间有一个模（Mode）。离中间越远，频率就越低。正态分布的形状也可以用各个部分中包含的面积比率来描述。只有正态分布的所有比率才是正确的。

还有两点要注意。首先，你应该意识到分布左侧的部分与右侧的部分具有相同的面积，因为正态分布是对称的。第二，由于分布中的位置是由 z 分数确定的，图中显示的百分比适用于任何正态分布，而不管均值和标准差的值如何。

由于正态分布是许多自然分布的良好模型，而且在某些情况下这种形状是有保证的，所以我们对这种特定分布给予了相当多的关注。

SAT 分数的总体分布是正态的，均值为 $\mu = 500$，标准差为 $\sigma = 100$。从这个总体中随机选择一个个体，其 SAT 分数大于 700 的概率是多少？

6.3.2 单元正态分布

(A) z	(B) Proportion in body	(C) Proportion in tail	(D) Proportion between mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

图 6.1: 单元正态分布表

无论右侧还是左侧，主体总是对应于分布的较大部分。类似地，尾部总是较小的部分，无论是在右侧还是在左侧。由于正态分布是对称的，右侧的比率与左侧的相应比率完全相同。虽然 z 分数值从一侧变成了另一侧（+ 和-），但比率始终是正数。

6.3.3 概率和比率和 z 分数

单位正态表列出了 z 分数位置和正态分布中的比率之间的关系。对于任何 z 分数位置，你可以使用表格查找相应的比率。同样，如果你知道比率，可以使用表格找到特定的 z 分数位置。因为我们已经定义了概率等于比率，所以你也可以使用单位正态表来查找正态分布的概率。

6.3.3.1 找到 z 分数对应的比率

- 正态分布的比率对应于 z 分数值大于 $z = +1.00$ 的区域是多少?
- 对于正态分布，获得 z 分数小于 $z = 1.50$ 的概率是多少？用符号表示，$P(z < 1.50) = ?$
- $z > 0.80$
- $z > -0.75$

```
ccdf(Normal(0, 1), 1),
cdf(Normal(0, 1), 1.5),
ccdf(Normal(0, 1), 0.8),
ccdf(Normal(0, 1), -0.75)
```

```
(0.15865525393145702, 0.9331927987311419, 0.21185539858339664, 0.7733726476231318)
```

- 对于正态分布，把前面 10% 和剩余部分分离开来的 z 分数是多少？
- 对于正态分布，哪些 z 分数值将中间 60% 的分布与其余的分数分开？

```
cquantile(Normal(0, 1), 0.1),
quantile(Normal(0, 1), (1-0.6)/2),
cquantile(Normal(0, 1), (1-0.6)/2)
```

```
(1.2815515655446004, -0.8416212335729143, 0.8416212335729143)
```

6.3.4 从正态分布中获取概率

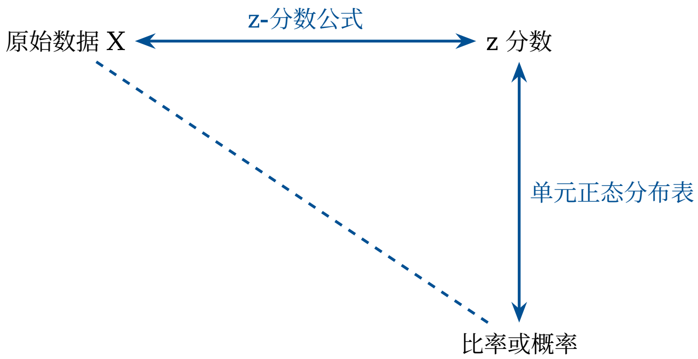
在前面的部分中，我们使用单位正态表查找了与特定 z 分数值相关的概率和比率。然而，在大多数情况下，需要找到具体 X 值的概率。要回答有关正态分布中的分数（X 值）的概率问题，你必须使用以下两个步骤：(1) 将 X 值转换为 z 分数; (2) 使用 z 分数来查找概率或比率。

- 已知智商分数服从均值为 $\mu = 100$，标准差为 $\sigma = 15$ 的正态分布。在这个信息的基础上，随机选择一个智商分数小于 120 的个体的概率是多少？
- 公路部门进行了一项研究，测量了一段州际公路的行驶速度。他们发现平均速度为 $\mu = 58$ 英里每小时，标准差为 $\sigma = 10$。分布大致呈正态分布。在这个信息的基础上，
  (1) 有多少比率的汽车的行驶速度在 55 到 65 英里每小时之间？
  (2) 有多少比率的汽车的行驶速度在 65 到 75 英里每小时之间？
- 对于均值 $\mu = 60$，标准差 $\sigma = 12$ 的正态分布，找到每个请求的概率。
  （1）$p(X > 66) = ?$；（2）$p(48 < X < 72) = ?$
- 根据美国人口普查局（2005）的报告，美国人平均每天通勤时间为 $\mu = 24.3$ 分钟。假设通勤时间的分布是正态分布，标准差为 $\sigma = 10$ 分钟，（1）要想在全国范围内属于最高的 10%，每天需要花费多少时间通勤？（2）定义分布的中间 90% 范围的数值范围是多少？

```
cdf(Normal(100, 15), 120),
diff([cdf(Normal(58, 10), X) for X in [55, 65]]),
diff([cdf(Normal(58, 10), X) for X in [65, 75]]),
ccdf(Normal(60, 12), 66),
diff([cdf(Normal(60, 12), X) for X in [48, 72]]),
cquantile(Normal(24.3, 10), 0.1),
diff([quantile(Normal(24.3, 10), X) for X in [0.05, 0.95]])
```

```
(0.9087887802741321, [0.3759477699658796], [0.19739818946453003], 0.3085375387259869, [0.68268949
```

6.3.4.1 z 分数，概率和比率的总结



z 分数是正态分布中的分数的标准测量单位。正态分布是对称的，z 分数告诉你一个分数在分布中的位置。单位正态表可以用来查找 z 分数值对应的概率或比率。要找到正态分布中特定分数的概率，需要将该分数转换为 z 分数，然后使用单位正态表查找。此外，你还可以使用 z 分数来找到将分布分成不同部分的 z 分数值。正态分布中 z 分数和概率之间的关系是非常有用的，因为它允许你在分布中找到特定分数的位置和概率。这对于许多统计应用非常重要，因为它可以帮助你回答有关概率和比率的问题。

6.4 概率和二项分布

6.4.1 二项分布

6.4.1.1 二项数据与二项分布

当一个变量在一个刻度上仅包含*两个类别*时 (two categories)，产生的数据称为二项数据 (binomial)。“二项” 一词可以宽松地翻译为 “两个名称”，指的是测量刻度上的两个类别。二项数据可以在一个变量自然存在仅有两个类别的情况下出现。例如，人可以被分类为男性或女性，抛硬币的结果要么是正面要么是反面。研究人员也常常通过将分数合并为两个类别来简化数据。例如，心理学家可以使用个性分数来将人们分类为攻击性高或攻击性低。

6.4.1.2 概率和二项分布

在二项情况下，研究人员通常知道与两个类别中的每一个相关的概率。例如，在使用一枚平衡硬币时，

$$p(\text{正面}) = p(\text{反面}) = \frac{1}{2}$$

感兴趣的问题是在一系列试验或一组个体样本中每个类别出现的次数。例如：

- 在抛 20 次平衡硬币的情况下，获得 15 次正面的概率是多少？
- 在抽取 50 名大一新生的样本中获得超过 40 名内向者的概率是多少？

6.4.1.3 二项分布

为了回答关于二项数据的概率问题，我们需要研究二项分布。为了定义和描述这个分布，首先引入一些符号。两个类别被标识为 A 和 B。与每个类别相关的概率（或比率）被标识为

$$p = p(A) = \text{A 的概率}$$
$$q = p(B) = \text{B 的概率}$$

注意，$p + q = 1.00$，因为 A 和 B 是唯一的两个可能的结果。样本中的个体或观察次数被标识为 n。变量 X 表示样本中类别 A 出现的次数.

注意，X 可以具有从 0（样本中没有类别 A）到 n（样本中所有类别都是 A）的任何值。使用此处呈现的符号，**二项分布**显示了从 $X = 0$ 到 $X = n$ 与每个 X 值相关联的概率。

6.4.1.4 二项分布：一个示例

对于这个示例，我们考虑的事件是抛硬币。有两种可能的结果：正面和反面。我们假设硬币是平衡的，所以 $p = p(\text{正面}) = \frac{1}{2}$，$q = p(\text{反面}) = \frac{1}{2}$

我们正在观察 $n = 2$ 次抛硬币的样本，感兴趣的变量是：$X = \text{正面的次数}$。为了构建二项分布，我们将考虑抛硬币 2 次的所有可能结果。

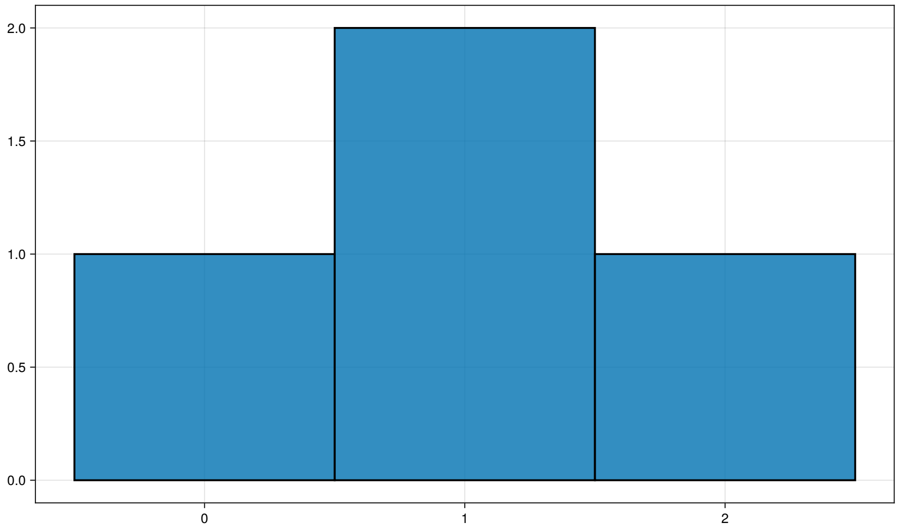
- 完整的 4 种结果集如下所示。

```
df = CombinationTable([0, 1], 2)
```

第 1 个	第 2 个	求和	均值	概率
0	0	0	0//1	1//4
1	0	1	1//2	1//4
0	1	1	1//2	1//4
1	1	2	1//1	1//4

```
hist(df.求和, bins = -0.5:1:2.5, strokewidth = 2)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



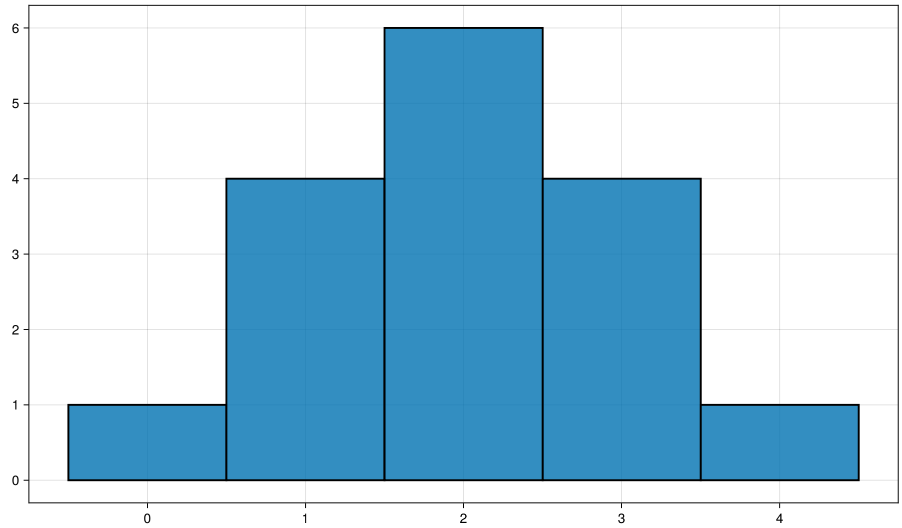
- 如果抽取四次，其结果如下所示

```
df = CombinationTable([0, 1], 4)
```

第 1 个	第 2 个	第 3 个	第 4 个	求和	均值	概率
0	0	0	0	0	0//1	1//16
1	0	0	0	1	1//4	1//16
0	1	0	0	1	1//4	1//16
1	1	0	0	2	1//2	1//16
0	0	1	0	1	1//4	1//16
1	0	1	0	2	1//2	1//16
0	1	1	0	2	1//2	1//16
1	1	1	0	3	3//4	1//16
0	0	0	1	1	1//4	1//16
1	0	0	1	2	1//2	1//16
0	1	0	1	2	1//2	1//16
1	1	0	1	3	3//4	1//16
0	0	1	1	2	1//2	1//16
1	0	1	1	3	3//4	1//16
0	1	1	1	3	3//4	1//16
1	1	1	1	4	1//1	1//16

```
hist(df.求和, bins = -0.5:1:4.5, strokewidth = 2)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



6.4.1.5 Julia 中的二项分布函数

```
cdf(Binomial(2, 0.5), 0),
cdf(Binomial(2, 0.5), 1),
cdf(Binomial(2, 0.5), 2)
```

```
(0.25, 0.75, 1.0)
```

```
cdf(Binomial(4, 0.5), 0),
cdf(Binomial(4, 0.5), 1),
cdf(Binomial(4, 0.5), 2),
cdf(Binomial(4, 0.5), 3),
cdf(Binomial(4, 0.5), 4)
```

```
(0.06250000000000001, 0.3125, 0.6875, 0.9375, 1.0)
```

6.4.2 二项分布的正态近似

当 n 较大时，二项分布倾向于近似正态分布。具体来说，当 pn 和 qn 都大于或等于 10 时，二项分布将几乎完美地近似为正态分布。在这些情况下，二项分布将以以下参数**近似**正态分布：

$$\text{均值：} \mu = pn$$
$$\text{标准差：} \sigma = \sqrt{npq}$$

在这个正态分布内，每个 X 值都有一个对应的 z 分数，

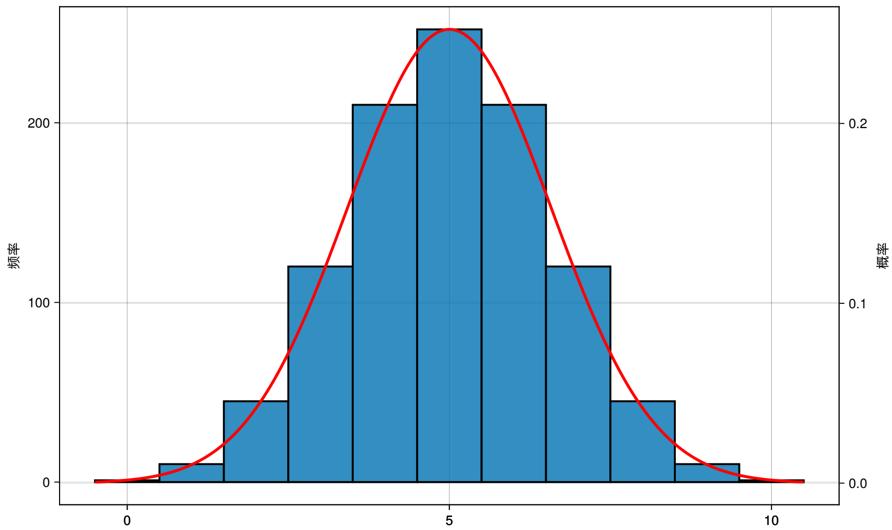
$$z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}}$$

二项分布倾向于呈正态形状意味着我们可以直接从 z 分数和单位正态表中计算概率值。

重要的是要记住，正态分布仅是真正的二项分布的近似。二项值，例如抛硬币一系列次数中正面次数，是*离散的*。例如，如果你记录了硬币抛掷 4 次中正面次数，你可能观察到 2 次正面或 3 次正面，但在 2 和 3 之间没有其他值。另一方面，正态分布是*连续的*。但是，在许多情况下，正态近似提供了计算二项概率的非常准确的模型。

```
n, p = 10, 0.5
q = 1-p
mu, sigma = p*n, sqrt(n*p*q)
df = CombinationTable([0, 1], n)
fig = Figure(size = (1000, 600))
ax1 =  Axis(fig[1, 1], ylabel = "频率")
ax2 =  Axis(fig[1, 1], ylabel = "概率", yaxisposition = :right)
hist!(ax1, df.求和, bins = -0.5:1:(n+0.5), strokewidth = 2)
lines!(ax2, -0.5:.001:(n+0.5),
  X -> pdf(Normal(mu, sigma), X), linewidth = 3, color = :red)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



在使用正态近似时，为了获得最大的准确性，您必须记住，二项分布中的每个 X 值实际上对应于直方图中的一根条。例如，直方图中的得分 $X = 6$ 由 5.5 和 6.5 的实际限制界定。

$X = 6$ 的实际概率由包含在此条中的面积确定。

要使用正态分布近似来估算这个概率，您应该找到包含在两个实际限制之间的面积。同样，如果您使用正态近似来找到得分大于 $X = 6$ 的概率，您应该使用超出 6.5 的实际限制边界的区域。

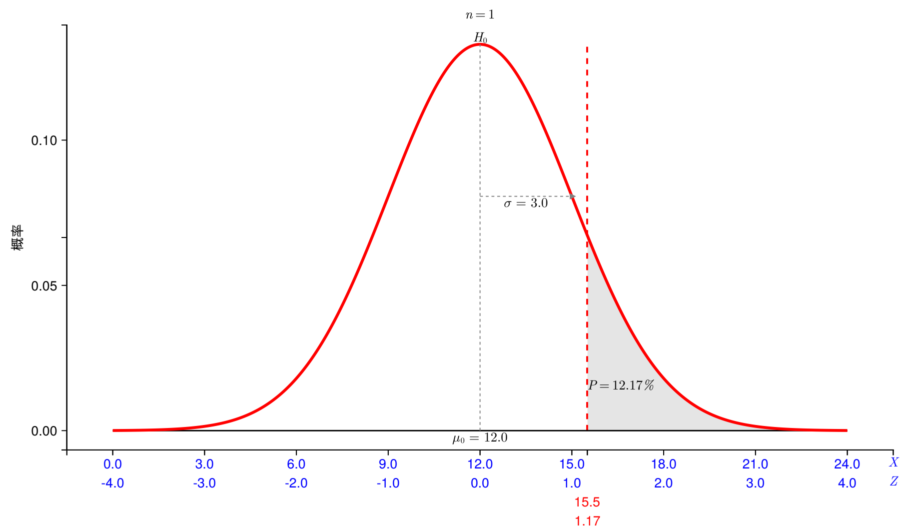
例如，假设您计划通过要求人们预测从一副完整的牌中随机选择的卡的花色来测试超感知（超感觉知觉）。对于没有超感知的个体来说，在 48 次试验中正确猜测超过 15 次的概率是多少。

```
n, p = 48, 1/4
q = 1 - p
M = n * p
S = sqrt(n * p * q)
ccdf(Binomial(n, p), 15),
ccdf(Normal(M, S), 15),
ccdf(Normal(M, S), 15.5)
```

```
(0.12317776308777559, 0.15865525393145702, 0.12167250457438122)
```

```
PlotZ(M, S; XBand = [15.5, Inf], showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



6.5 推断统计展望

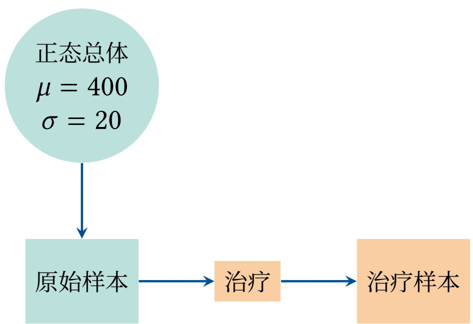
概率在样本和它们来自的总体之间建立了直接联系。这个联系是未来章节中推断统计学的基础。

要确定治疗是否有效，研究人员只需比较受治疗的样本与原始总体。如果样本中的个体得分约为 400（原始总体均值），则必须得出结论治疗似乎没有效果。另一方面，如果接受治疗的个体得分与 400 明显不同，那么研究人员有证据表明治疗确实有效。请注意，这项研究使用样本来回答有关总体的问题；这是推断统计学的本质。

对于研究人员来说，问题是确定“与 400 明显不同”的确切含义是什么。在第 5 章中，我们建议 z 分数提供了解决这个问题的一种方法。具体来说，我们建议 z 分数的值超过 $z = 2.00$（或 -2.00）是一个极端值，因此明显不同。但是，选择 $z = \pm 2.00$ 是纯粹任意的。

现在我们有了另一个工具，*概率*，来帮助我们确定确切的边界。中间 95% 高概率值（接近 $\mu = 400$ 的分数）表明治疗没有效果。极端 5% 的分数几乎不可能来自原始总体，因此提供了治疗效果的证据。

使用单位正态表的列 C，右尾和左尾的 z 分数边界分别为 $z = +1.96$ 和 $z = -1.96$。如果我们从未接受治疗的原始总体中选择一个个体，那么我们很难获得超出 $z = \pm 1.96$ 边界的分数。在 $z = \pm 1.96$ 处设置的边界为决定我们的样本是否提供了治疗效果证据的客观标准。



第七章　样本均值的分布

7.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie, StatsReIntro
using Statistics, Distributions, DataFrames
```

7.2　样本、总体和样本均值

7.2.1　z-分数和概率

在前面的两章中，介绍了 z-分数和概率的主题。每当从总体中选择一个分数时，您应该能够计算一个 z-分数，精确描述分数在分布中的位置。如果总体是正态分布的，您还应该能够确定获得任何个体分数的概率值。然而，到目前为止，我们考虑的 z-分数和概率仅限于样本由单个分数组成的情况。在本章中，我们将扩展 z-分数和概率的概念，以涵盖更大样本的情况。具体而言，我们将介绍一种将样本均值转化为 z-分数的方法。

7.2.2　抽样误差

抽样误差（Sampling error）是样本统计量与其对应的总体参数之间的自然差异或误差量。正如前面提到的，即使从同一总体中取出两个独立的样本，它们也可能不同。这些样本将包含不同的个体、不同的分数、不同的均值等等。在大多数情况下，可以从一个总体中获得成千上万个不同的样本。幸运的是，可能样本的庞大集合形成了一个相对简单和有序的模式，使得能够相对准确地预测样本的特征。预测样本特征的能力基于样本均值的分布。

7.2.3　样本均值的分布

到目前为止，我们总是讨论分数的分布；现在分布中的值不再是分数，而是统计量（样本均值）。由于统计量是从样本中获得的，所以统计量的分布被称为抽样分布。**抽样分布**是通

过从总体中选择特定大小的所有可能样本来获得的统计量的分布。样本均值的分布是抽样分布的一个示例。**样本均值的分布**（distribution of sample means ）是所有可能的特定大小（n）的随机样本的样本均值的集合，这些样本可以从总体中获得。

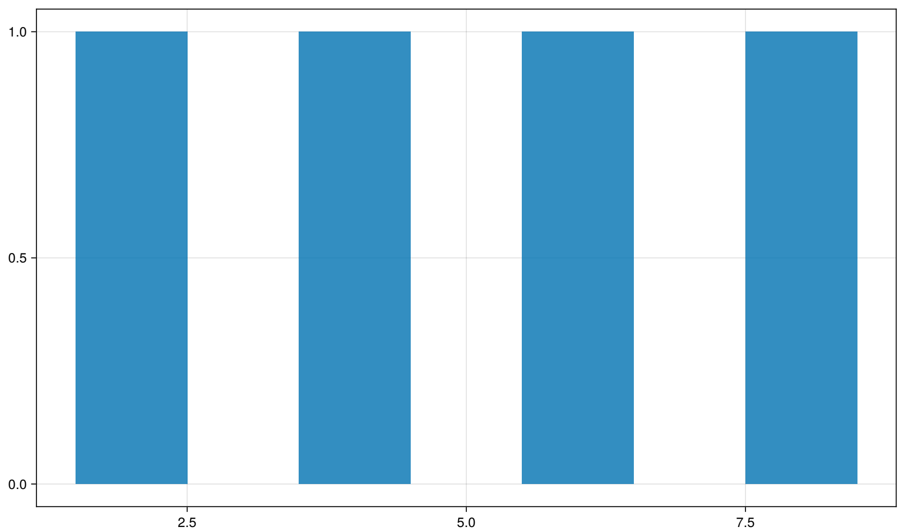
7.2.4 样本均值分布的特点

- 样本均值应该在总体均值附近堆积。不要期望样本是完美的，但它们应该代表总体。因此，大多数样本均值应该与总体均值相对接近。
- 样本均值的堆积应该形成一个近似正态分布。从逻辑上讲，大多数样本的均值应该接近 μ，发现与 μ 显著不同的样本均值应该相对较少。因此，样本均值应该在分布中心（μ 周围）堆积，随着 M 和 μ 之间的距离增加，频率应该逐渐减小。这描述了一个近似正态分布。
- 通常情况下，样本大小越大，样本均值应该越接近总体均值 μ。从逻辑上讲，大样本应该更好地代表小样本。因此，大样本大小获得的样本均值应该相对接近总体均值；从小样本中获得的均值应该更分散。

```
X = [2, 4, 6, 8]
```

```
hist(X, bins = 1.5:1:8.5, figiure = (; size = (800, 400)))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```

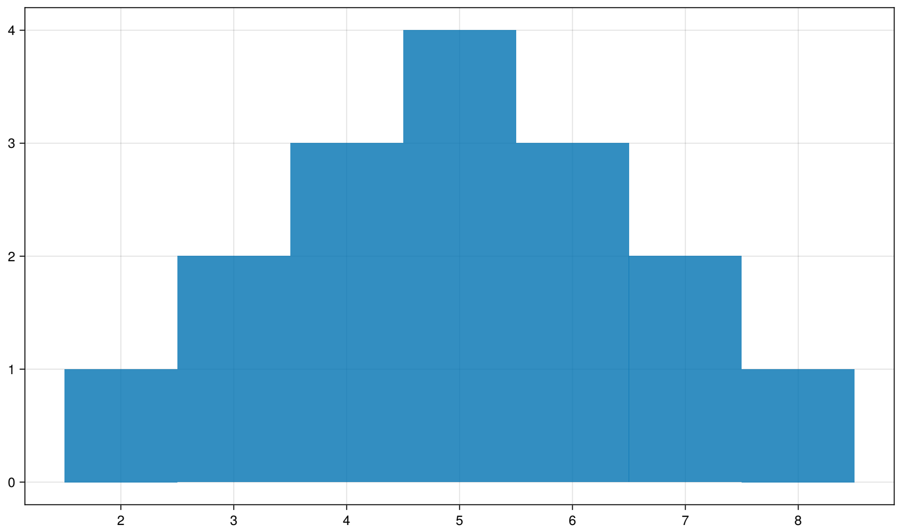



```
XM = CombinationTable(X, 2)
```

第 1 个	第 2 个	求和	均值	概率
2	2	4	2//1	1//16
4	2	6	3//1	1//16
6	2	8	4//1	1//16
8	2	10	5//1	1//16
2	4	6	3//1	1//16
4	4	8	4//1	1//16
6	4	10	5//1	1//16
8	4	12	6//1	1//16
2	6	8	4//1	1//16
4	6	10	5//1	1//16
6	6	12	6//1	1//16
8	6	14	7//1	1//16
2	8	10	5//1	1//16
4	8	12	6//1	1//16
6	8	14	7//1	1//16
8	8	16	8//1	1//16

```julia
hist(XM.均值,
    bins = 1.5:1:8.5,
    axis=(; xticks = 2:8)
)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



问：如果您从原始总体中抽取 $n = 2$ 个分数的样本，获得均值大于 7 的样本的概率是多少？

$$p(M > 7) = \frac{1}{16}$$

7.3 样本均值分布

7.3.1 样本均值分布的形状

如果满足以下两个条件之一，样本均值分布几乎是完全正态的：（1）从中选择样本的总体是正态分布。（2）每个样本中的分数数量（n）相对较大，约为 30 或更多。

样本均值分布的均值等于分数总体的均值 μ，称为 **M 的期望值**（expected value of M）。样本均值分布的标准差 σ_M 称为 **M 的标准误差**（standard error of M）。标准误差提供了样本均值（M）与总体均值（μ）之间平均距离的度量。

与标准差和原始分数的关系类似，标准误差对于样本均值分布有相同的两个作用：（1）标准误差描述样本均值分布。（2）标准误差衡量个体样本均值代表整个分布的程度。

标准误差的大小由两个因素确定：（1）样本大小。（2）从中选择样本的总体的标准差。

大数定理（law of large numbers）认为，样本大小（n）越大，样本均值接近总体均值的可能性越大。

标准误差的公式表达了标准差与样本大小（n）之间的关系。

$$\text{标准误} = \sigma_M = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

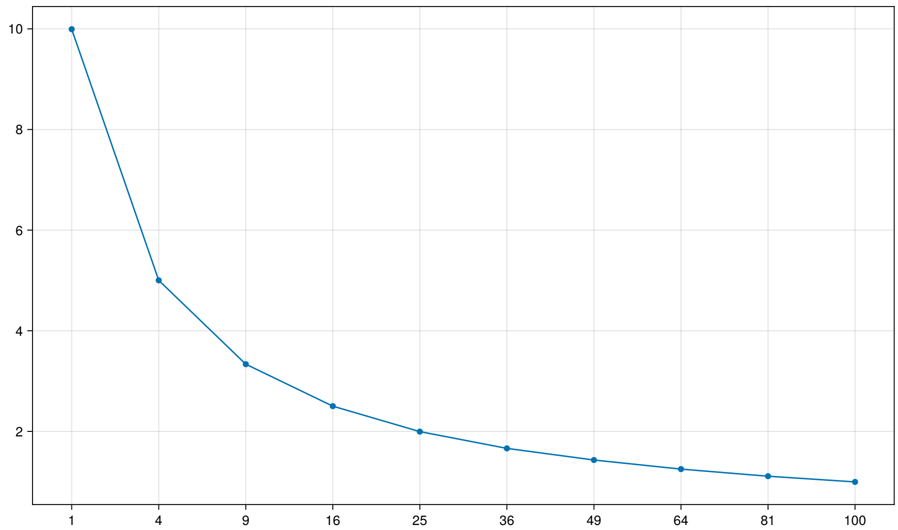
当样本只包含一个分数（$n = 1$）时，标准误差与标准差相同（$\sigma_M = \sigma$）。随着样本大小（n）的增加，标准误差的大小减小。（更大的样本更准确。）

```
sstd(n) = std([mean(rand(Normal(0, 10), n)) for _ in 1:1e5])
df = vcat([DataFrame(样本容量 = n, 标准误 = sstd(n)) for n in (1:10).^2]...)
```

样本容量	标准误
1	9.99305
4	5.00229
9	3.33572
16	2.50153
25	1.99556
36	1.6624
49	1.4313
64	1.25075
81	1.10941
100	0.99608

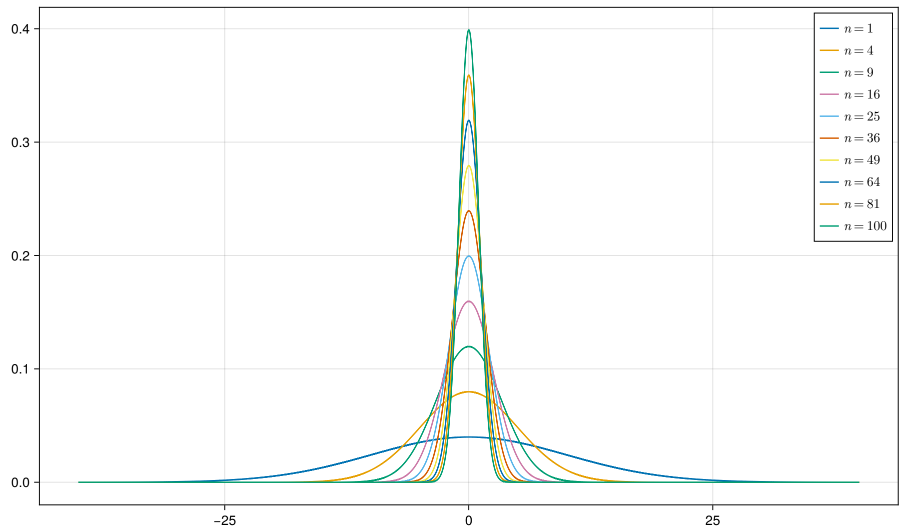
```
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xticks = df.样本容量, xscale = sqrt)
scatter!(ax, df.样本容量, df.标准误)
lines!(ax, df.样本容量, df.标准误)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1])
for n in (1:10).^2
lines!(ax, -10*4:0.001:10*4,
    x -> pdf(Normal(0, 10/sqrt(n)), x),
    label = L"n=%$n")
end
axislegend()
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



7.3.2 中心极限定理

中心极限定理（Central Limit Theorem）：对于任何具有均值 μ 和标准差 σ 的总体，样本大小为 n 的样本均值分布将具有均值 μ 和标准差 $\sigma/\sqrt{n}$，并且当 n 趋近无穷大时，将趋近于正态分布。

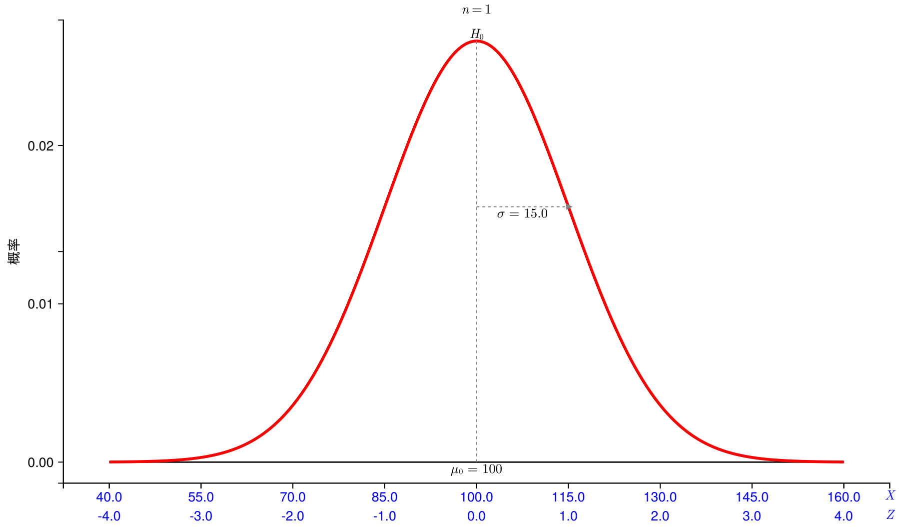
这个定理的价值来自两个简单的事实：（1）首先，它描述了*任何总体*的样本均值分布，不管其形状、均值或标准差如何；（2）其次，样本均值的分布会非常迅速地“趋近”于正态分布；（3）一旦样本大小达到 $n = 30$，分布几乎完美地符合正态分布。

7.3.3 三种不同的分布

- 整体分布

```
# lines((100-4*15):0.01:(100+4*15), X -> pdf(Normal(100, 15), X))
PlotZ(100, 15)
```

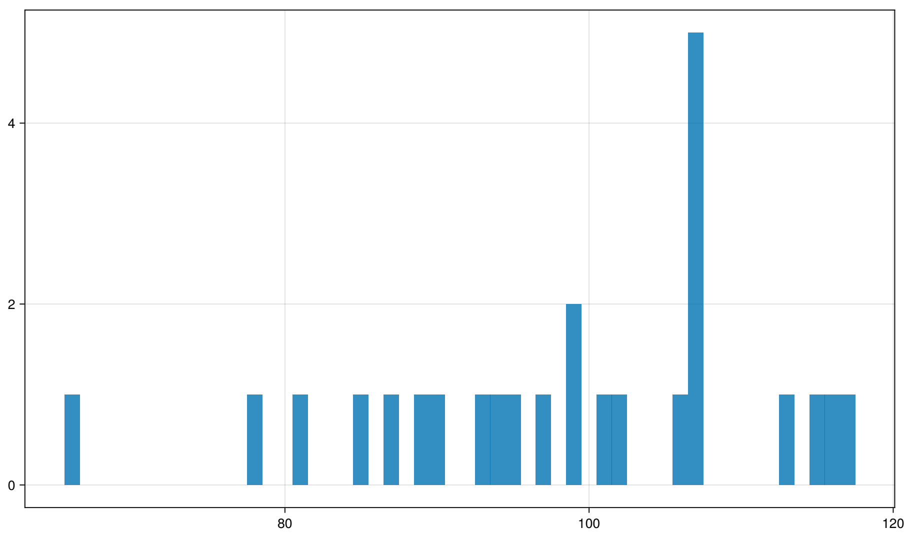
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 样本分布

```
X = round.(rand(Normal(100, 15), 25), digits = 0)
hist(X, bins = minimum(X)-0.5:1:maximum(X)+0.5)
```

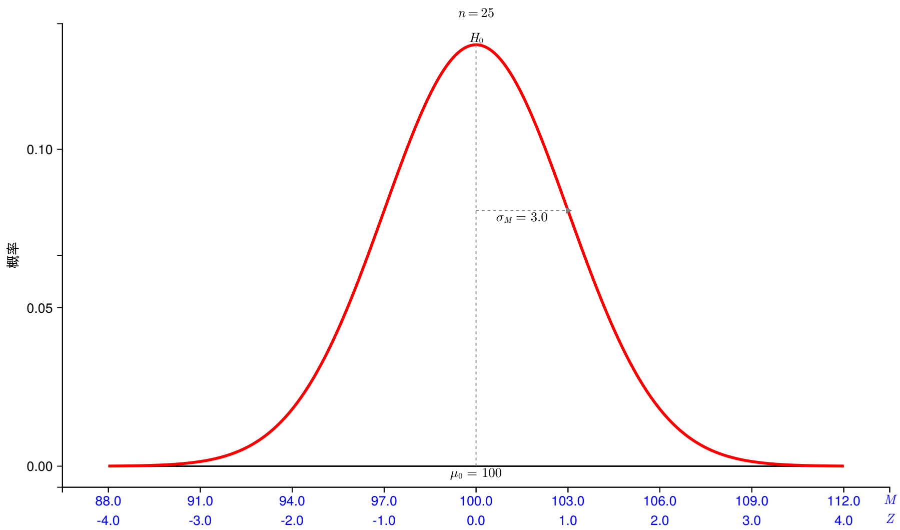
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 样本均值分布

```
PlotZ(100, 15, n = 25)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



7.4 概率与样本均值分布

SAT 分数的总体构成一个均值为 $\mu = 500$ 和 $\sigma = 100$ 的正态分布。如果您随机选择了 $n = 16$ 名学生的样本，样本均值大于 $M = 525$ 的概率是多少?

样本均值的分布具有以下特点：- 该分布是正态的，因为 SAT 分数的总体是正态的。- 该分布的均值为 500，因为总体均值是 $\mu = 500$。- 对于 $n = 16$，该分布的标准误差为：

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{16}} = 125.0$$

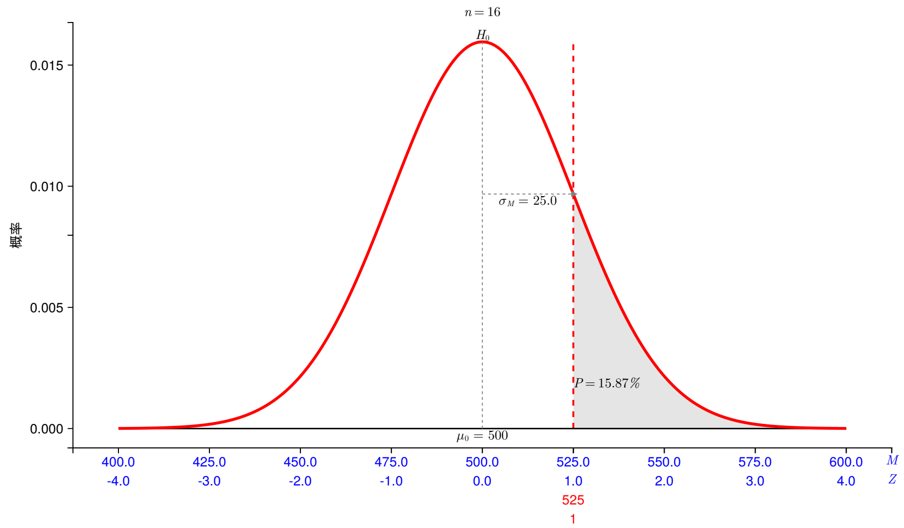
- 样本均值大于 $M = 525$ 的概率是

```
ccdf(Normal(500, 100/sqrt(16)), 525)
```

```
0.15865525393145702
```

```
PlotZ(500, 100, n = 16, XBand = [525, Inf], showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



7.4.1 样本均值的 z-分数

用于定位样本均值的 z-分数公式是

$$z = \frac{M - \mu}{\sigma_M}$$

就像每个分数（X）都有一个描述其在分数分布中位置的 z-分数一样，每个样本均值（M）都有一个描述其在样本均值分布中位置的 z-分数。具体而言，z-分数的符号告诉我们样本均值是在 μ 之上（+）还是在 μ 之下（-），z-分数的数值是样本均值和 μ 之间的标准误差数目。当样本均值分布是正态的时，可以使用 z-分数和单位正态分布表找到与任何特定样本均值相关联的概率。

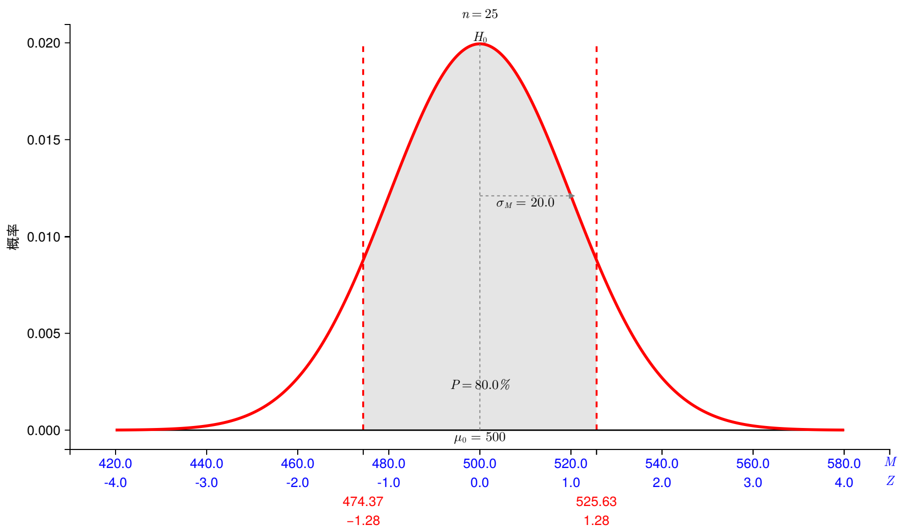
再次提醒，SAT 分数的分布形成一个均值为 $\mu = 500$ 和标准差为 $\sigma = 100$ 的正态分布。为了这个例子，我们将确定在随机选择 $n = 25$ 名学生的样本中，平均 SAT 分数有望获得哪种样本均值，这种情况下，我们将确定 80 的时间内预期的值范围。

```
[quantile(Normal(500, 100/sqrt(25)), P) for P in [0.1, 0.9] ]
```

```
2-element Vector{Float64}:
 474.368968689108
 525.631031310892
```

```
PlotZ(500, 100, n = 25, QBand = [0.1, 0.9])
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



7.5　标准误的更多内容

- 两个调整：（1）现在我们使用样本均值分布而不是分数分布。（2）现在我们使用标准误差而不是标准差。
- 一个单一的规则：每当您使用样本均值时，必须使用标准误差。

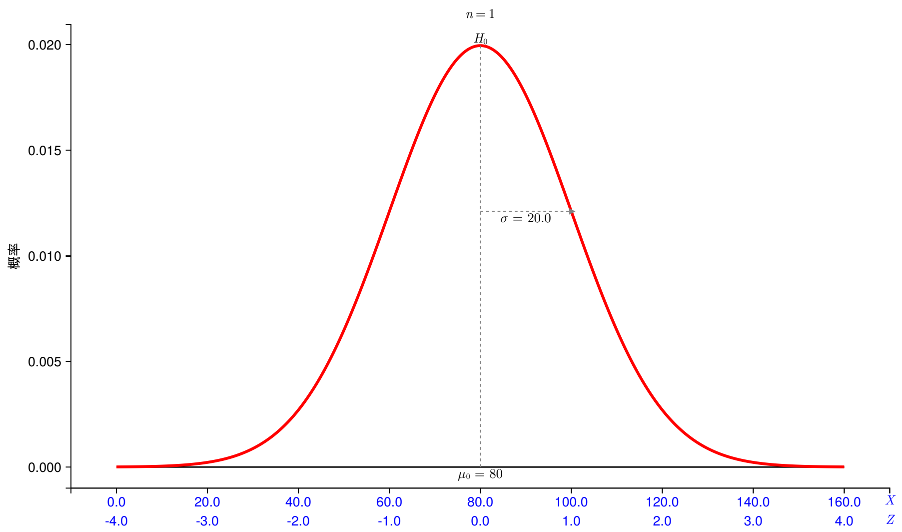
7.5.1　抽样误差和标准误差

抽样误差（Sampling Error）的一般概念是，样本通常不会提供总体的完美准确代表。标准误差（Standard Error）提供了定义和衡量抽样误差的方法。知道标准误差可以让研究人员很好地了解他们的样本数据如何准确地代表他们正在研究的总体。

每天您花多少时间观看电子视频（在线、电视、手机、平板电脑等）？平均响应是 $\mu = 80$ 分钟，观看时间的分布大致呈正态分布，标准差为 $\sigma = 20$。接下来，我们从这个总体中取一个样本，检查样本均值如何准确地代表总体均值。具体而言，我们将考虑三种不同的样本：一个有 $n = 1$ 名学生的样本，一个有 $n = 4$ 名学生的样本和一个有 $n = 100$ 名学生的样本，以考虑样本大小如何影响准确性。

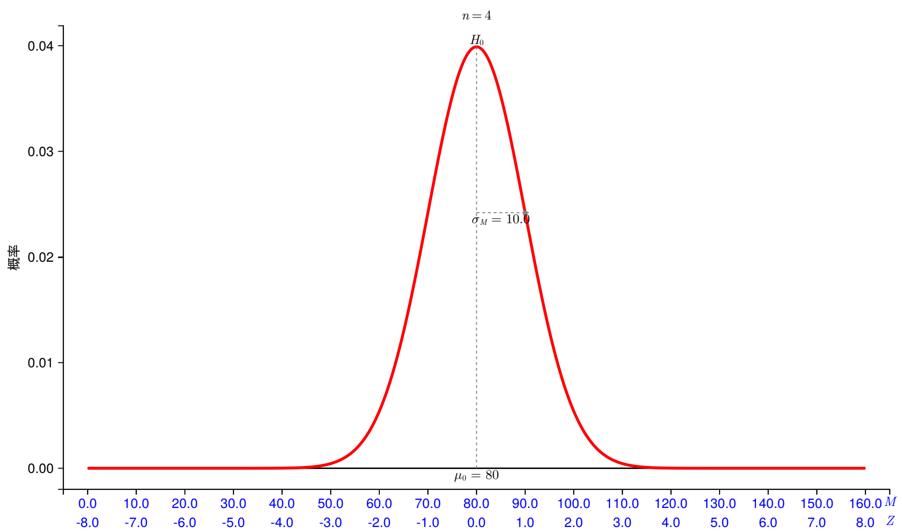
```
PlotZ(80, 20, n = 1; XExtrema = [0, 160])
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



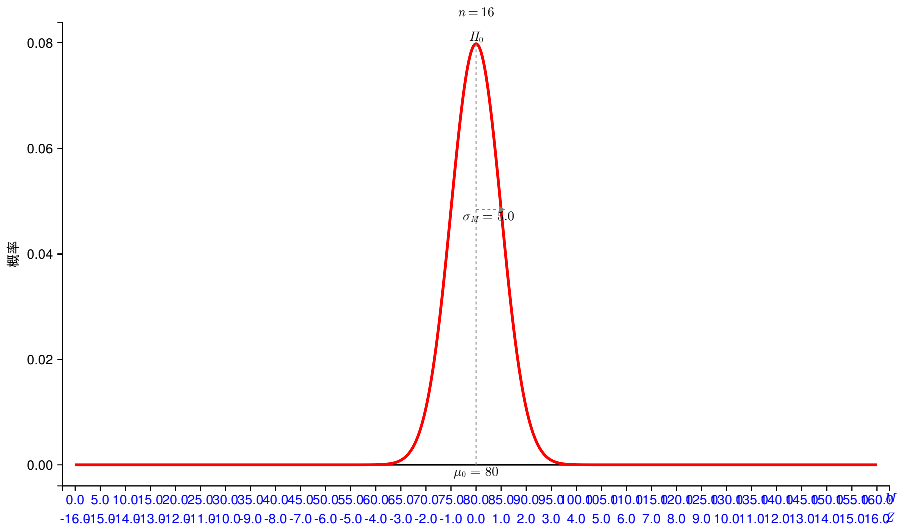

```
PlotZ(80, 20, n = 4; XExtrema = [0, 160])
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(80, 20, n = 16; XExtrema = [0, 160])
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



7.6 报告标准误差

正如我们将在后面的章节中看到的，标准误差在推断统计学中扮演着非常重要的角色。由于其关键作用，科学论文通常报告样本均值的标准误差，而不是样本标准差。科学期刊在如何提到标准误差方面存在差异，但经常使用符号 SE 和 SEM（均值的标准误差）。标准误差有两种报告方式：与样本均值一起，它可以在表中报告；或者，标准误差可以在图表中报告。

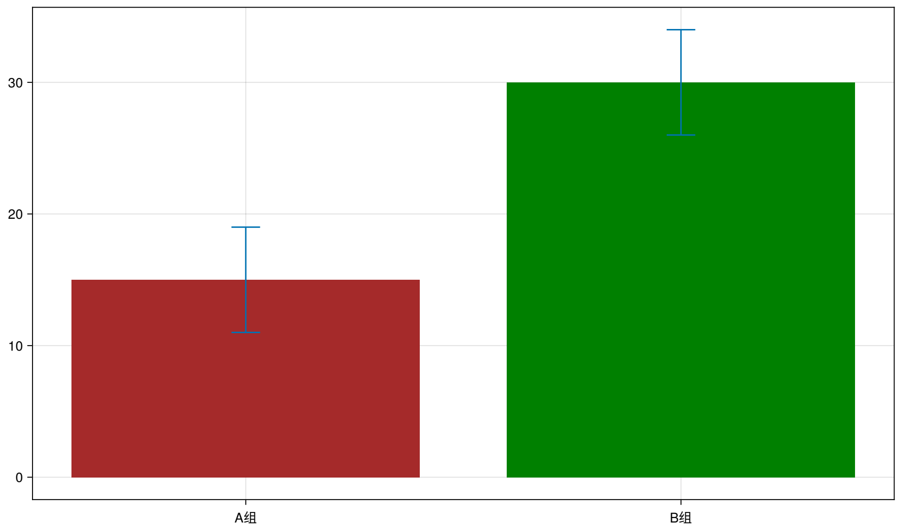
- 报告标准误差：表格

	n	均值	SE
控制组	17	32.23	2.31
照相机组	15	45.17	2.78

- 报告标准误差：条形图

```julia
Mean, SE = [15, 30], 4
fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xticks = (1:2, ["A组", "B组"]))
barplot!(ax, Mean, color = [:brown, :green])
errorbars!(ax, [1, 2], Mean, SE, whiskerwidth = 30)

fig
```

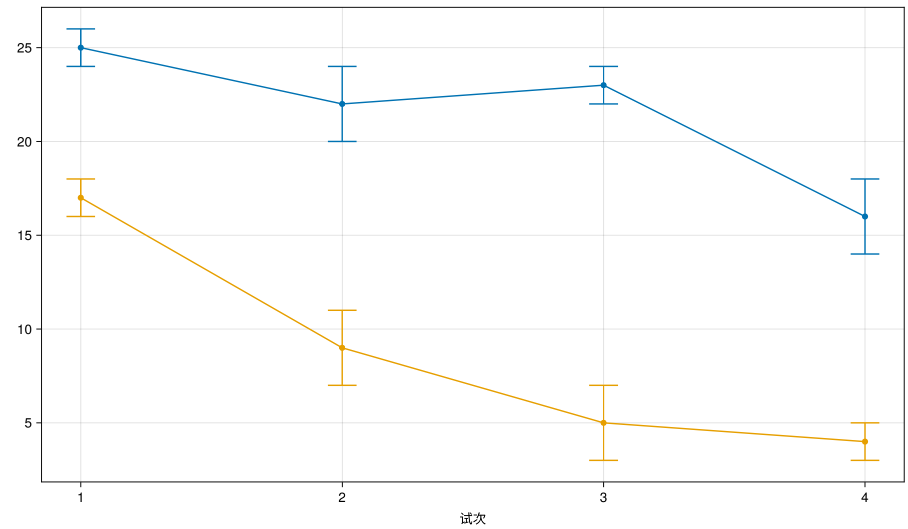
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 报告标准误差：折线图

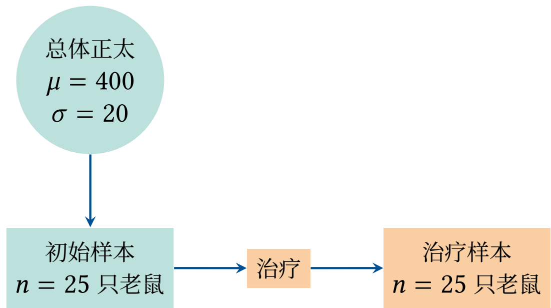
```
MA  = [25, 22, 23, 16]
SEA = [1, 2, 1, 2]
MB  = [17, 9, 5, 4]
SEB = [1, 2, 2, 1]

fig = Figure(size = (1000, 600))
ax = Axis(fig[1, 1], xlabel = "试次", ylabel = "")
[lines!(ax, X) for X in [MA, MB]]
[scatter!(ax, X) for X in [MA, MB]]
[errorbars!(ax, 1:4, X, Y, whiskerwidth = 30) for (X, Y) in zip([MA, MB], [SEA, SEB])]
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



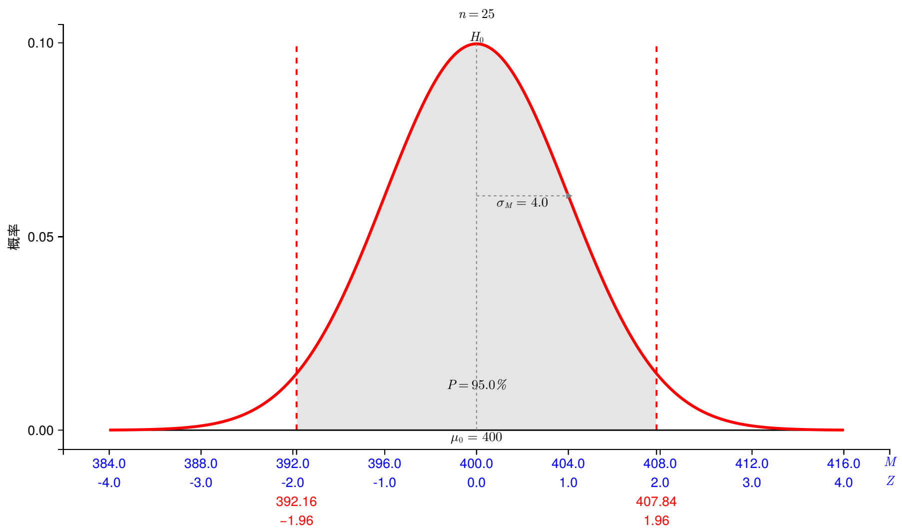
7.7　展望推断统计学

假设一位心理学家正在计划一项评估新生长激素效果的研究。已知普通成年大鼠（没有激素）的平均体重为 $\mu = 400g$ 。当然，并不是所有的大鼠都一样大，它们的体重分布是均匀的，标准差为 $\sigma = 20$。心理学家计划选择一组 $n = 25$ 只新生大鼠的样本，注射激素，然后在它们成年后测量它们的体重。




```
PlotZ(400, 20, n = 25; QBand = [0.025, 0.975], showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyw
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



换句话说，如果治疗没有效果，那么获得均值大于 407.84 或小于 392.16 的样本均值是非常不可能的。如果我们的样本产生了这些极端均值之一，那么我们就有了治疗效果的证据。具体来说，没有治疗效果的情况下很不可能发生这样的极端结果。我们通过确定受治疗样本与未受治疗样本是否明显不同来评估治疗的效果。

在大多数研究中，研究人员必须依赖一个样本来提供对正在研究的总体的准确代表。在这种情况下，标准误差可以看作是样本均值可靠性的度量。可靠性这个词是指对同一事物的不同测量的一致性。

第八章　假设检验简介

8.1　本章 Julia 包

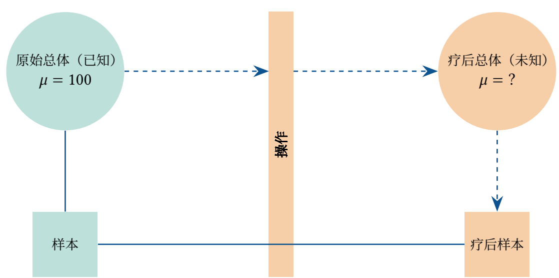
```julia
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie, StatsReIntro
```

8.2　假设检验的逻辑

8.2.1　假设检验的逻辑

假设检验（hypothesis test）是一种使用样本数据来评估关于总体的假设的统计方法。首先，我们提出有关总体的假设。在选择样本之前，我们使用假设来预测样本应具备的特征。接下来，我们从总体中获取一个随机样本。最后，我们将获得的样本数据与从假设中得出的预测进行比较。如果样本均值与预测一致，我们会得出假设是合理的结论。但是，如果数据与预测之间存在较大差异，我们将决定假设是错误的。

根据研究类型和数据类型的不同，假设检验的细节会在不同的研究情境中发生变化。在后续章节中，我们将研究用于不同类型研究的不同版本的假设检验。然而，现在，我们关注所有假设检验都共有的基本元素。为了实现这一总体目标，我们将研究假设检验作为它适用于最简单情境的一个例子：使用样本均值来检验有关总体均值的假设。



8.2.2 假设检验的四个步骤

- 第一步：陈述假设
- 第二步：设定决策标准
- 第三步：收集数据并计算样本统计量
- 第四步：做出决策

8.2.2.1 第一步：陈述假设

假设检验的过程始于对未知总体提出假设。实际上，我们提出了两种相反的假设：

零假设（H_0）(Null Hypothesis）声明在总体中没有变化、没有差异或没有关系。在实验的背景下，H_0 预测自变量（处理）对总体中的因变量（分数）没有影响。

备择假设（H_1）(Alternative Hypothesis）声明总体中存在变化、差异或关系。在实验的背景下，H_1 预测自变量（处理）对因变量有影响。

注意：零假设和备择假设是互斥且穷尽的（exclusive and exhaustive)。它们不能同时成立。数据将决定哪一个应该被拒绝。

8.2.2.2 第二步：设定决策标准

最终，研究者将使用样本数据来评估零假设的可信度。数据将要么支持零假设，要么倾向于推翻零假设。特别是，如果数据与假设之间存在较大差异，我们将得出假设错误的结论。为了形式化决策过程，我们使用零假设来预测应该获得的样本均值类型。具体来说，我们确定哪些样本均值与零假设一致，哪些样本均值与零假设不一致。

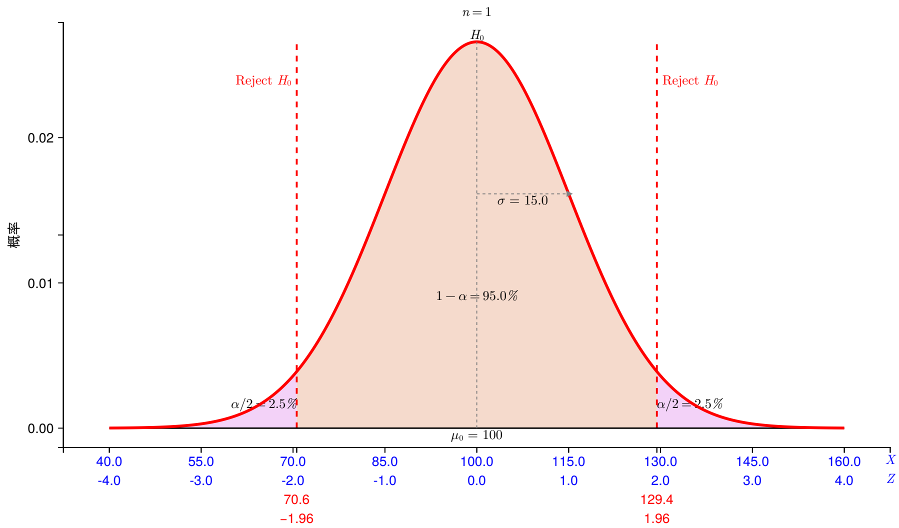
根据零假设，样本均值的分布分为两个部分：如果 H_0 成立，即接近零假设的样本均值，那么有可能获得的样本均值。如果 H_1 成立，即与零假设差异很大，那么不太可能获得的样本均值。

为了找到分隔高概率样本和低概率样本的边界，我们必须确切地定义"低"概率和"高"概率的含义。**alpha 水平**（α 水平），或**显著性水平**，是用于定义在假设检验中"非常不可能"概念的概率值。根据惯例，常用的 alpha 水平是 $\alpha = .05$，$\alpha = .01$，和 $\alpha = .001$。

临界区域由极端样本值组成，如果零假设成立，则这些值非常不可能（根据 alpha 水平的定义）被获得。临界区域的边界由 alpha 水平决定。如果样本数据位于临界区域内，那么拒绝零假设。技术上来说，临界区域是由样本结果定义的，如果操作没有效果（即零假设成立）时，非常不可能发生的结果。从另一角度来看，我们也可以将临界区域定义为能够提供确凿证据表明操作确实具有效果的样本值。为了确定定义临界区域边界的确切位置，我们使用 alpha 水平的概率和标准正态分布表。

```julia
PlotZ(100, 15, PVal = 0.05, showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.2.2.3 第三步：收集数据并计算样本统计量

注意，在研究者陈述假设并建立决策标准之后，才收集数据。这一事件序列有助于确保研究者对数据进行了诚实客观的评估，并在了解实验结果之后不会篡改决策标准。接下来，对样本的原始数据进行总结，使用适当的统计量，常见统计量为样本均值。

现在，研究者可以比较样本均值（数据）与零假设。这是假设检验的核心：将数据与假设进行比较。通过计算一个 z 分数来实现比较，该分数描述了样本均值相对于假设的总体均值的确切位置。在第二步中，我们构建了如果零假设成立（即处理没有效应）将预期的样本均值分布，即如果处理没有效应，可以获得的整个样本均值集合。现在，我们计算一个 z 分数，用于确定我们的样本均值位于这个假设分布中的位置。

$$z = \frac{M-\mu}{\sigma_M} = \frac{\text{样本均值} - \text{假定的整体平均值}}{M\text{和}\mu\text{间的标准误}}$$

8.2.2.4 第四步：做出决策

在最后一步中，研究者使用第三步得到的 z 分数值根据第二步建立的标准对零假设做出决策。有两种可能的结果。

- 样本数据位于临界区域内。

按定义，临界区域内的样本值如果零假设成立则非常不可能发生。因此，我们*拒绝零假设* (reject the null hypothesis)。请记住，零假设表明没有处理效应，所以拒绝 H_0 意味着我们得出结论，处理确实具有效应。

- 样本数据不在临界区域内。

在这种情况下，样本均值与零假设中指定的总体均值相当接近（位于分布中心）。由于数据没有提供零假设错误的强有力证据，我们 *无法拒绝零假设* (fail to reject the null hypothesis)。这个结论意味着处理似乎没有效应。

假设检验的类比：假设检验过程类似于陪审团审判的过程。*假定无罪，直到被证明有罪*。（Innocent until proven guilty）

8.2.3 更详细地看 z 分数统计量

用于假设检验中的 z 分数统计量是被称为*检验统计量*（test statistic）的第一个具体示例。术语“检验统计量”只是指将样本数据转化为用于检验假设的一个具体统计量。在后续章节将介绍用于不同研究情境的其他检验统计量。但是，大多数新检验统计量都与 z 分数有相同的基本结构和相同的目的。

z 分数方程是比较样本数据和总体假设的正式方法。

$$z = \frac{M - \mu}{\sigma_M}$$

第一、z 分数方程，像任何方程一样，可以看作是一个食谱。如果按照说明并使用所有正确的成分，该方程将产生一个 z 分数。但是，在假设检验情况下，您没有所有必要的成分。具体来说，不知道总体均值（μ）的值，这是方程中的一个组成部分。

第二、将 z 分数方程视为比值

$$\begin{aligned} z &= \frac{M - \mu}{\sigma_M} \\ &= \frac{\text{样本均值} - \text{假设的总体均值}}{\text{样本均值（M）和总体均值（}\mu\text{）之间的标准误差}} \\ &= \frac{\text{样本（M）和假设（}\mu\text{）之间的实际差异}}{\text{没有处理时样本（M）和 }\mu\text{ 之间的标准差}} \end{aligned}$$

8.2.4 例一：穿红色衣服与魅力

先前研究表明, 女性穿红色衣服时, 男性会认为她们更有吸引力 (Elliot & Niesta, 2008)。Gueguen 和 Jacob（2012）认为，同样的现象可能会影响男性对穿红色衣服的女服务员的反应。餐厅的女服务员 (和男服务员) 通常穿着白色衬衫和黑色裤子，餐厅记录显示，男性顾客给女服务员的小费平均为 $\mu = 15.8$ %，标准差为 $\sigma = 2.4$ 个百分点。小费金额的分布大致正态。研究期间女服务员穿红色衬衫，研究者计划记录 $n = 36$ 名男性顾客的小费。

如果样本的平均小费与基准平均值明显不同，研究者可以得出穿红色衬衫似乎对小费有影响的结论。另一方面，如果样本均值仍然约为 15.8%(与基准相同)，研究者必须得出红色衬衫似乎没有任何影响的结论。

8.2.4.1 第一步、陈述假设并选择 alpha 水平

- 穿红色衬衫对男性顾客的小费行为没有影响。

$$H_0 \ : \ \mu_{\text{穿红衫}} = 15.8$$

  即使穿红色衬衫，均值仍为 15.8%。

- 穿红色衬衫会影响男性顾客的小费，并导致均值发生变化。

$$H_1 \ : \ \mu_{\text{穿红衫}} \neq 15.8$$

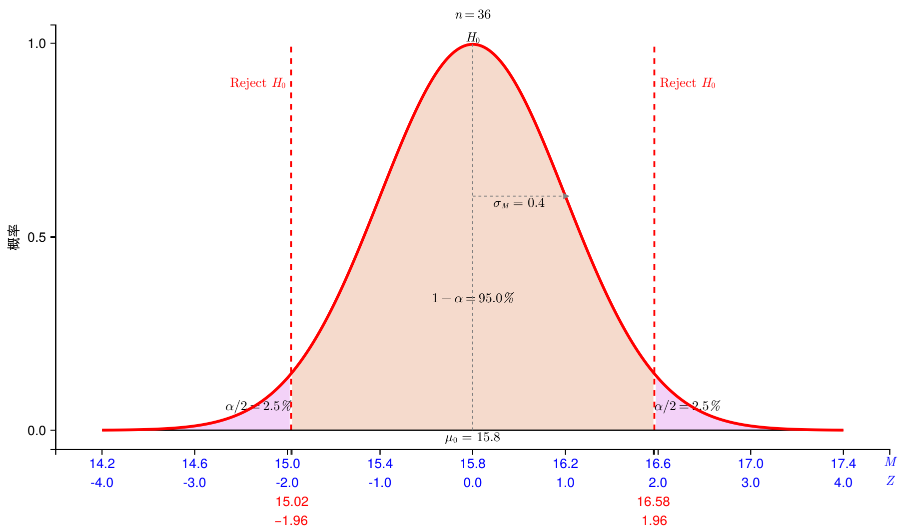
  穿红衫，小费均值将不同于 15.8%。

我们设置 $\alpha = .05$。对于 $\alpha = .05$ 的正态分布，临界区域包括产生极端尾部 z 分数的样本均值，超出 $z = \pm 1.96$。

8.2.4.2 第二步：设定决策标准

```
PlotZ(15.8, 2.4, n = 36, PVal = 0.05, showZ = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.2.4.3 第三步、计算检验统计量

男性顾客给女服务员的小费平均为 $\mu = 15.8\%$，标准差为 $\sigma = 2.4$ 个百分点。假设样本包含 $n = 36$ 名个体，平均小费为 $M = 16.7\%$。样本均值的标准误差为

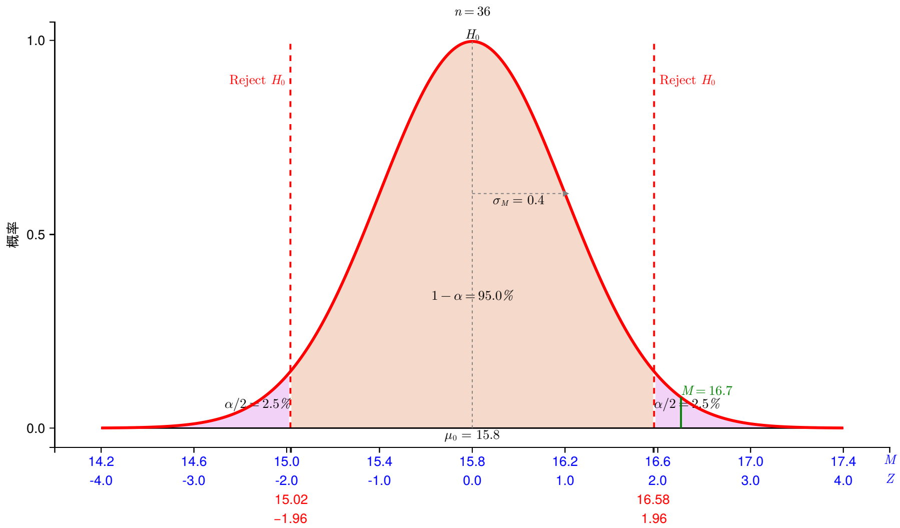
$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{2.4}{6} = 0.4$$

因此，样本均值 $M = 16.7$ 产生了 z 分数

$$z = \frac{M - \mu}{\sigma_M} = \frac{16.7 - 15.8}{0.4} = \frac{0.9}{0.4} = 2.25$$

```
PlotZ(15.8, 2.4, n = 36, PVal = 0.05, showZ = true, XVal = 16.7)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.2.4.4 第四步、做出决策

z 分数位于临界区域，这意味着如果零假设成立，这个样本均值非常不可能出现。因此，我们拒绝零假设，并得出结论：穿红色衬衫确实对男性的小费产生了影响。

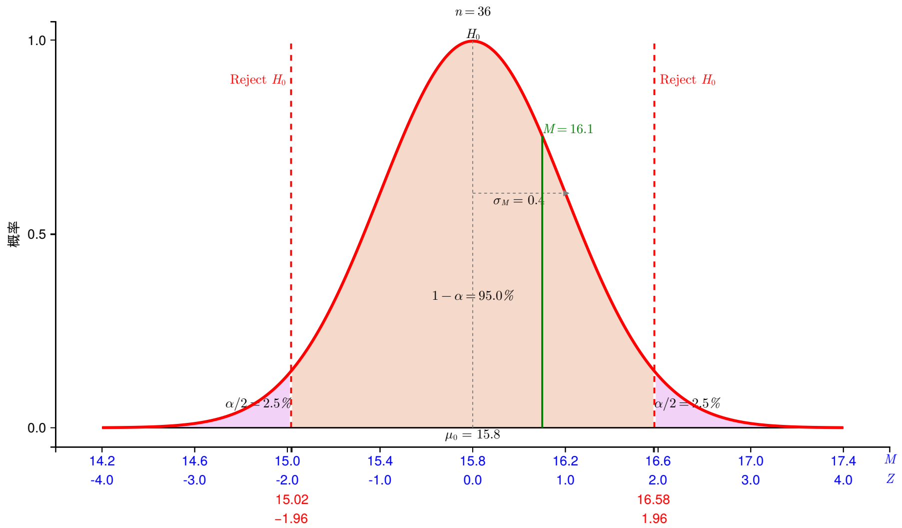
8.2.5 例二

假设样本包含 $n = 36$ 名个体，平均小费为 $M = 16.1\%$。这产生了一个 z 分数

$$z = \frac{M - \mu}{\sigma_M} = \frac{16.1 - 15.8}{0.4} = \frac{0.3}{0.4} = 0.75$$

```
PlotZ(15.8, 2.4, n = 36, PVal = 0.05, showZ = true, XVal = 16.1)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.2.6 例三

一个正态分布总体的均值 $\mu = 40$，标准差为 $\sigma = 8$。在从总体中抽取的一个由 $n = 16$ 个个体组成的样本接受处理后，发现样本均值 $M = 45$。在 $\alpha = .05$ 水平下，这个样本提供足够的证据来拒绝零假设并得出结论，存在显著的处理效应吗？

8.3 假设检验的错误和不确定性

8.3.1 假设检验中的错误

假设检验是一种*推论过程*，这意味着它使用有限的信息作为达成一般结论的基础。具体来说，样本只提供了关于整个总体的有限或不完整的信息，然而假设检验使用一个样本来对

总体做出结论。在这种情况下，总是存在着可能做出错误结论的可能性。在假设检验中，可以发生两种不同类型的错误。

8.3.2 第一类错误

第一类错误（也称为 I 类错误）发生在研究者拒绝一个实际上为真的零假设时。在典型的研究情境中，第一类错误意味着研究者得出结论，即一种处理具有效应，而实际上它没有效应。第一类错误发生在研究者无意中获得了一个极端的、不代表性的样本时。

假设检验的 **alpha 水平**（α level）是测试将导致第一类错误的概率。换句话说，alpha 水平确定了在零假设为真的情况下获得临界区域内的样本数据的概率。

8.3.3 第二类错误

第二类错误（也称为 II 类错误）发生在研究者未能拒绝一个实际上为假的零假设时。在典型的研究情境中，第二类错误意味着假设检验未能检测到真实的处理效应。第二类错误的后果通常不如第一类错误严重。类似于 *错误地确定了有罪*或 *未能找到有罪的人*。

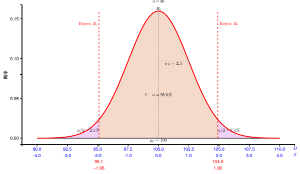
与第一类错误不同，不可能确定第二类错误的单一确切概率。相反，第二类错误的概率取决于各种因素，因此是一个函数，而不是一个具体的数字。尽管如此，第二类错误的概率由希腊字母 β 表示。

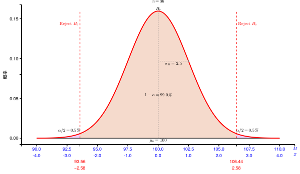
		实际情况	
		无效应 H_0 正确	有效应 H_0 错误
实验者 决策	拒绝 H_0	I 类错误	决策正确 (统计效力)
	无法拒绝 H_0	决策正确	II 类错误

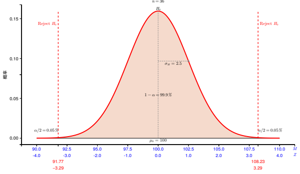
8.3.4 选择 α 水平

- $\alpha = .05$
- $\alpha = .01$
- $\alpha = .001$

```
┌ Warning: Found `resol
└ @ Makie ~/.julia/pack
```







8.4 更多关于假设检验

8.4.1 假设检验的摘要

1. 陈述假设并选择 alpha 水平
2. 定位临界区域
3. 计算检验统计量（z 分数）
4. 做出决策

8.4.2 报告统计检验的结果

穿红色衬衫对男性顾客所留小费的大小产生了显著影响， $z = 2.25$，$p < .05$。

Wearing a red shirt had a significant effect on the size of the tips left by male customers, z = 2.25, p < .05.

首先，什么是 “显著”？如果在零假设为真时发生的可能性非常小，那么结果被称为**显著**（significant）或**统计显著**（statistically significant）。也就是说，结果足以拒绝零假设。因此，如果假设检验的决策是拒绝 H_0，则处理有显著效应。接下来，$z = 2.25$ 是什么意思？z 表示使用 z 分数作为检验统计量来评估样本数据，并且其值为 2.25。最后，$p < .05$ 是什么意思？陈述的这一部分是指明假设检验中使用的 alpha 水平的传统方式。这也承认了（以及概率上存在）第一类错误的可能性。

在统计决策是不拒绝 H_0 的情况下，报告可能会陈述如下：

没有证据表明穿红色衬衫对男性顾客所留小费的大小产生了影响， $z = 0.75$，$p > .05$。

There was no evidence that the red shirt had an effect on the size of the tips left by male customers, z = 0.75, p > .05.

当使用计算机程序进行假设检验时，打印输出通常不仅包括 z 分数值，还包括 p 的精确值，即在没有任何处理效应的情况下发生结果的概率。

穿红色衬衫对男性顾客所留小费的大小产生了显著影响， $z = 2.25$，$p = .0244$。

Wearing a red shirt had a significant effect on the size of the tips left by male customers, z = 2.25, p = .0244.

然而，使用精确的 p 值时，仍然必须满足传统的显著性标准。

8.4.3 影响假设检验的因素

z 分数

$$z = \frac{M - \mu}{\sigma_M}$$

影响假设检验的因素（1）样本均值与零假设中的总体均值之间的差异；（2）标准误差。标准误差由分数的变异性（标准差或方差）和样本中的分数数量（n）决定。

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

8.4.4 假设检验的假设

- 随机抽样
- 独立观察：赌徒的谬误；无替代抽样。一位研究人员有兴趣研究儿童的电视偏好。为了获得一个包含 $n = 20$ 名儿童的样本，研究人员从家庭 A 选择了 4 名儿童，从家庭 B 选择了 3 名儿童，从家庭 C 选择了 5 名儿童，从家庭 D 选择了 2 名儿童，从家庭 E 选择了 6 名儿童。
- 操作不会改变 σ 的值。假设检验中 z 分数公式的一个关键部分是标准误差 σ_M 的值。为了计算标准误差的值，我们必须知道样本大小（n）和总体标准差（σ）。然而，在假设检验中，样本来自一个未知的总体。如果总体真的未知，那么它暗示着我们不知道标准差，因此无法计算标准误差。为了解决这个困境，我们进行了一种假设。具体来说，我们假设未知总体（在处理前）的标准差与处理后的总体标准差相同。实际上，这一假设是许多统计程序的一部分的更一般假设的结果。这个一般性假设陈述了*处理的效应是将一个常数增加（或减少）到总体中的每个分数*。
- 正态抽样分布为了评价 z 分数的假设，我们使用单位正态表来识别临界区域。只有在样本均值分布是正态分布时才能使用这个表。

8.5 方向性假设检验

8.5.1 方向性假设检验的逻辑

假设检验程序的标准格式是双尾测试。“双尾”这个术语来自于临界区域分布在分布的两个尾部。这个格式是迄今为止最广泛接受的假设检验程序。尽管如此，本节讨论的是另一种备选方法。

在**方向性假设检验**（directional hypothesis test）或**单尾检验**（one-tailed test）中，统计假设（H_0 和 H_1）指定总体均值的增加或减少。也就是说，它们对效应的方向进行了陈述。

8.5.2 方向性检验：一个例子

在研究中，对于 $n = 36$ 名参与者的样本，每个参与者由一名穿红衬衫的女服务员招待，小费的大小被记录下来。对于男性顾客的一般总体（女服务员穿着白衬衫），小费的分布大致正态，均值为 $\mu = 15.8\%$，标准差为 $\sigma = 2.4$ 个百分点。对于这个例子，预期效应是红色会增加小费。如果研究者获得 $n = 36$ 名参与者的样本均值为 $M = 16.5\%$，这个结果是否足以得出红衫确实增加小费的结论？

8.5.3 方向性测试的假设

- 零假设：操作没有效果。

H_0: 小费没有增加。

- 备择假设：操作效果与预期一致。

H_1: 小费有所增加。

要用符号表示方向性假设，通常从备择假设（H_1）开始更容易。

- 备择假设：穿红色衬衫时，平均小费大于 15.8%。

H_1: $\mu > 15.8$

- 零假设：穿红色衬衫时，平均小费不大于 15.8%。

H_0: $\mu \leq 15.8$

请再次注意，这两个假设是互斥的，并涵盖了所有可能性。

8.5.4 方向性测试的临界区域

如果预测是操作将导致分数的增加，那么临界区域完全位于分布的右尾。如果预测是操作将导致分数的减少，那么临界区域完全位于分布的左尾。由于临界区域包含在分布的一个尾部，方向性测试通常被称为单尾测试。还要注意，由 alpha 水平指定的比例不会分散在两个尾部，而是完全包含在一个尾部。

请注意，方向性（单尾）测试需要在假设检验过程的步骤中进行两个更改：（1）在假设检验的第一步中，将方向性预测纳入到假设的陈述中。（2）在过程的第二步中，临界区域完全位于分布的一个尾部。完成这两个更改后，单尾测试将与常规的双尾测试完全相同。具体来说，您计算 z 分数统计量，然后根据 z 分数是否在临界区域内来对 H_0 做出决策。

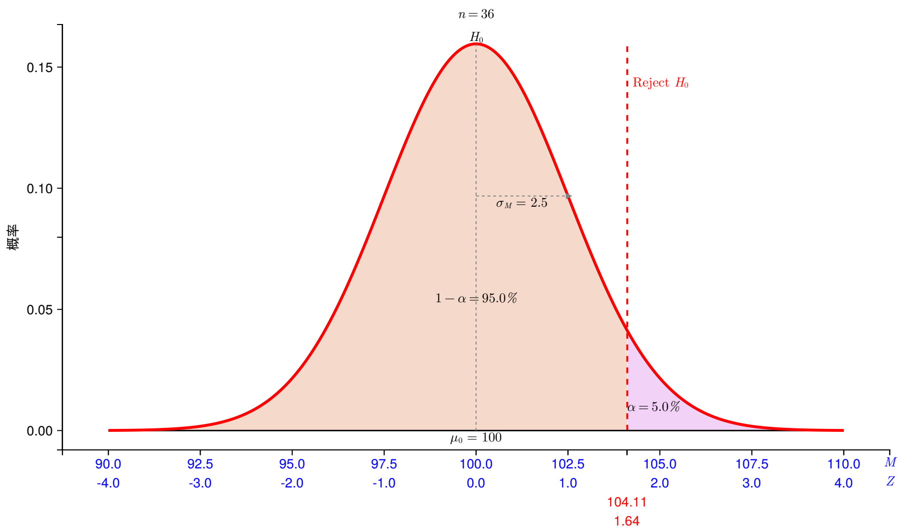
对于这个例子，研究人员获得了 36 名由穿红衬衫的女服务员招待的参与者的平均值为 $M = 16.5\%$。这个样本均值对应于一个 z 分数

$$z = \frac{M - \mu}{\sigma_M} = \frac{16.5 - 15.8}{0.4} = \frac{0.7}{0.4} = 1.75$$

z 分数 $z = 1.75$ 位于单尾测试的临界区域内。如果 H_0 为真，这是一个非常不可能的结果。因此，我们拒绝零假设，并得出结论：穿红衫会显著增加男性顾客的小费。

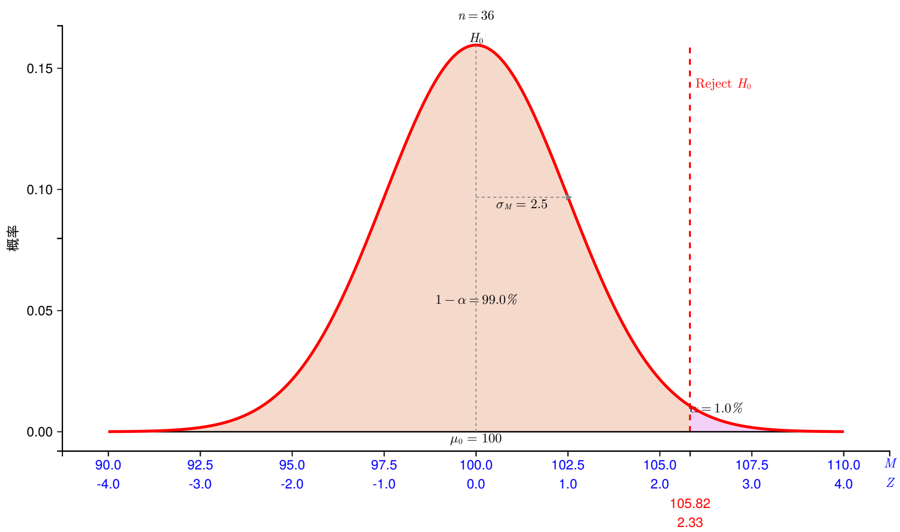
```
PlotZ(100, 15, n = 36, PVal = 0.05, showZ = true, Tail = "Right")
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



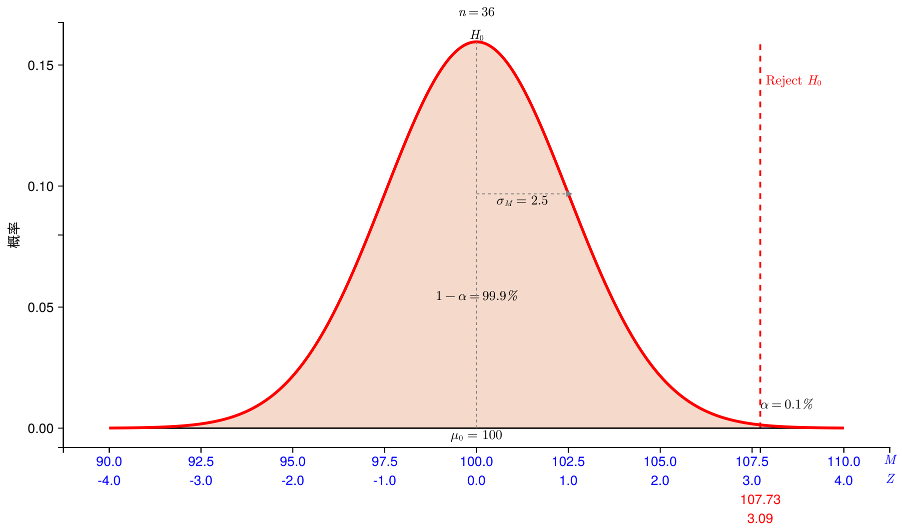

```
PlotZ(100, 15, n = 36, PVal = 0.01, showZ = true, Tail = "Right")
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(100, 15, n = 36, PVal = 0.001, showZ = true, Tail = "Right")
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.5.5 报告统计检验的结果

在文献中，这个结果将报告如下：

穿红色衬衫显著增加了小费， $z = 1.75$*，* $p < .05$*，单尾。*

Wearing a red shirt produced a significant increase in tips, z = 1.75, p < .05, one tailed.

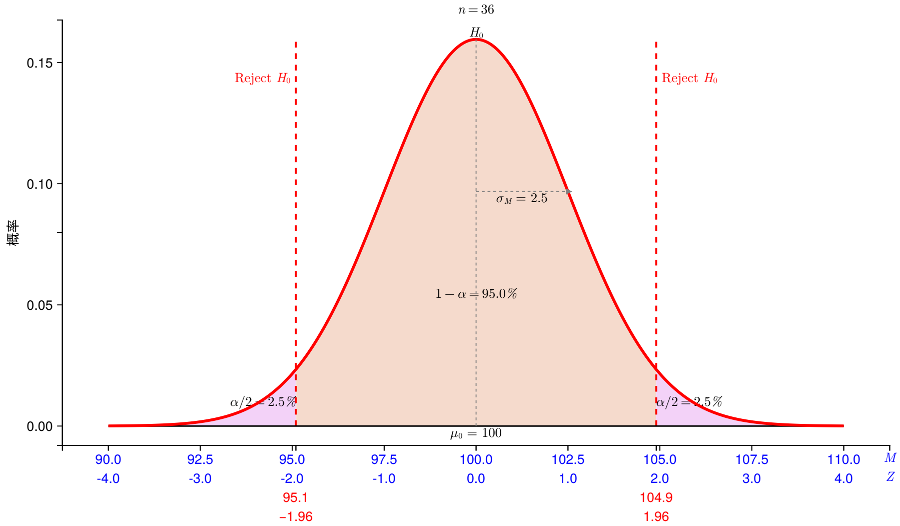
请注意，报告明确承认使用了单尾测试。

8.5.6 单尾测试与双尾测试的比较

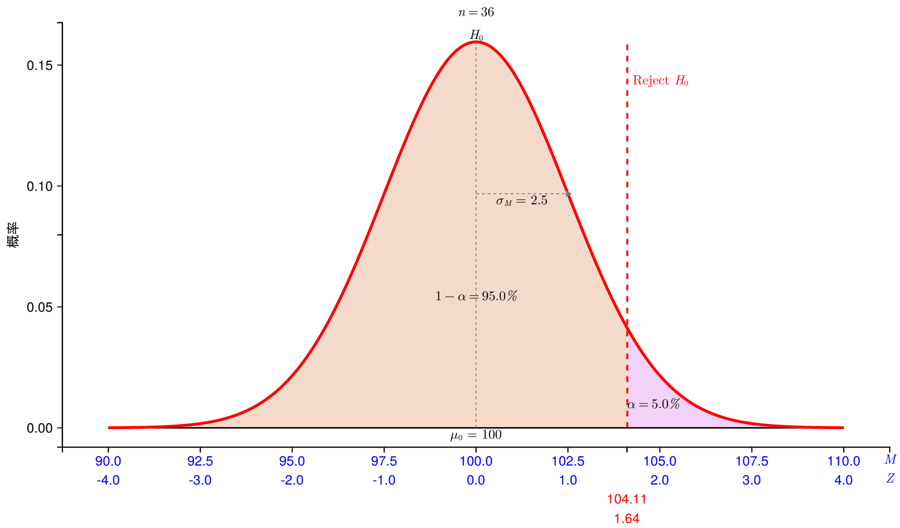
假设检验的一般目标是确定特定处理对总体是否有影响。测试是通过选择一个样本、对样本进行处理，然后将结果与原始总体进行比较来执行的。如果处理的样本与原始总体明显不同，那么我们就说处理产生了影响，并拒绝 H_0。另一方面，如果处理的样本仍然与原始总体相似，那么我们得出处理效应没有充分证据，并无法拒绝 H_0。做出这一决策的关键因素是处理的样本与原始总体之间的差异大小。大的差异表明处理有效；小的差异不足以说明处理有任何影响。

单尾测试和双尾测试之间的主要区别在于它们用于拒绝 H_0 的标准。单尾测试允许在样本与总体之间的差异相对较小的情况下拒绝零假设，只要差异是在指定的方向上。另一方面，双尾测试需要相对较大的差异，而不考虑方向。

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



一般来说，在没有明确方向性预期或存在两个竞争性预测的研究情境中应使用双尾测试。只有在进行研究之前进行了方向性预测并且有充分理由进行方向性预测的情况下，才应使用单尾测试。特别地，如果双尾测试未达到显著性水平，绝对不应在相同数据上进行第二次尝试以救回显著结果。

8.6 假设检验的担忧

8.6.1 两个严重的限制

使用假设检验来确定处理效应的显著性存在两个严重的限制。

第一个关注的是假设检验的重点是数据而不是假设。具体地说，当拒绝零假设时，我们实际上是在对样本数据而不是对零假设做出有力的概率陈述。一个显著的结果允许得出以下结论：

如果零假设为真，那么这个特定的样本均值非常不可能（$p < .05$）。

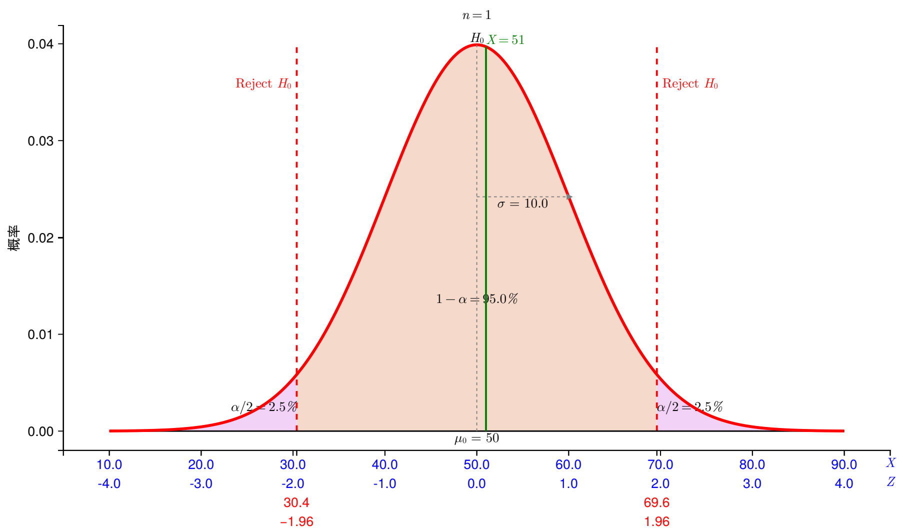
请注意，这个结论没有就零假设的真假提出明确的陈述。数据非常不可能表明零假设也非常不可能，但我们没有任何坚实的理由对零假设提出概率陈述。具体来说，您不能因为使用 $\alpha = .05$ 拒绝了零假设，就得出零假设真实性的概率小于 5% 的结论。

第二个担忧是，证明显著的处理效应并不一定意味着显著的处理效应。特别地，统计显著性并不能提供有关处理效应的绝对大小的任何真实信息。相反，假设检验仅仅是通过计算标准误差来确定样本均值与 μ 之间的差异是否合理，并展示所得的均值差异是否显著大于标准误差。

注意，该测试是进行相对比较的：处理效应的大小是相对于标准误差而言的。如果标准误差非常小，那么处理效应也可以非常小，但仍然足够大，以至于显著。因此，显著效应并不一定意味着效应很大。

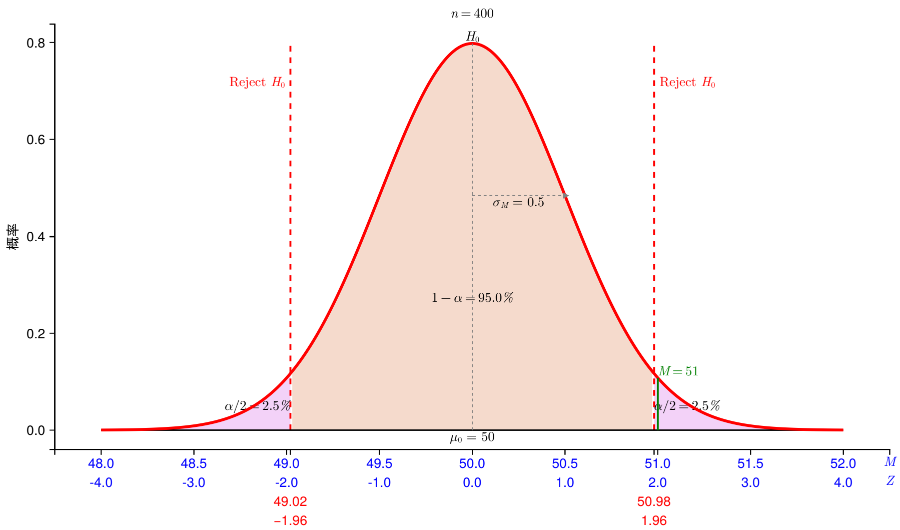
```
PlotZ(50, 10, n = 1, PVal = 0.05, showZ = true, XVal = 51)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



```
PlotZ(50, 10, n = 400, PVal = 0.05, showZ = true, XVal = 51)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.6.2 测量效应大小

如前所述，假设检验的一个问题是，假设检验实际上没有评估处理效应的绝对大小。为了解决这个问题，建议报告统计显著效应时，还应该报告**效应大小** (Effect Size)。因此，当我们介绍不同的假设检验时，我们还将介绍不同的测量和报告效应大小的选项。目标是以不受样本数量影响的方式来测量和描述处理效应的绝对大小。

测量效应大小的一种最简单和最直接的方法是 *Cohen's d*。Cohen（1988）建议，效应大小可以通过使用标准差来度量均值差异来进行标准化。结果的效应大小测量计算如下：

$$\begin{aligned}\text{Cohen's } d &= \frac{\text{平均值差异}}{\text{标准差}} \\ &= \frac{\mu_{\text{有操作}} - \mu_{\text{无操作}}}{\sigma}\end{aligned}$$

对于 z 分数假设检验，均值差异是由处理前的总体均值与处理后的总体均值之间的差异确定的。但是，处理后的总体均值是未知的。因此，我们必须使用受试样本的均值来代替。请记住，样本均值预期代表总体均值并提供了处理效应的最佳度量。因此，实际的计算实际上是估计 Cohen's d 的值，如下所示：

$$
\begin{aligned}
\text{估计的 Cohen's } d &= \frac{\text{均值差异}}{\text{标准差}} \\
&= \frac{M_{\text{有操作}} - \mu_{\text{无操作}}}{\sigma}
\end{aligned}
$$

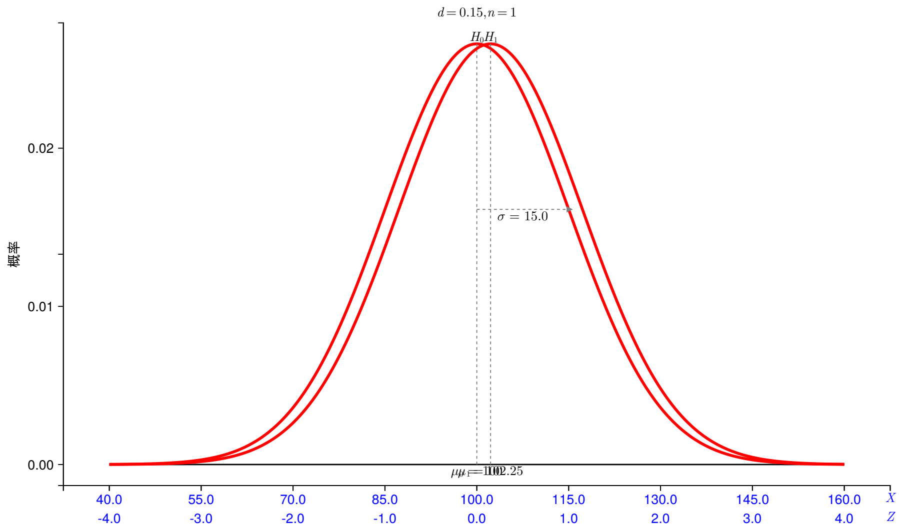
重叠分布：Cohen's d 度量了两个分布之间的分离程度，而一个标准差的分离度（$d = 1.00$）代表了一个大的差异。

8.6.3 用 Cohen's d 评估效应大小

8.6.3.1 效应大小和均值差异

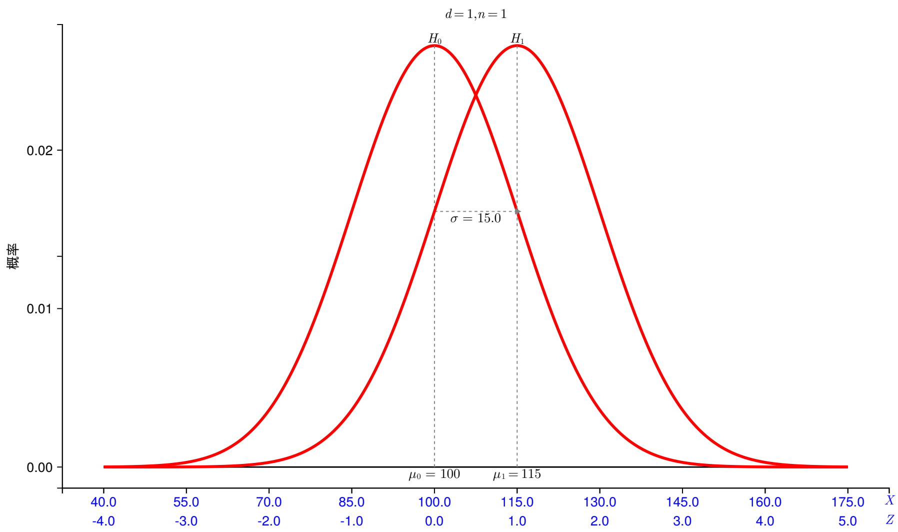
```
PlotZ(100, 15, n = 1, d = 0.15, showAlternative = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(100, 15, n = 1, d = 1, showAlternative = true)
```

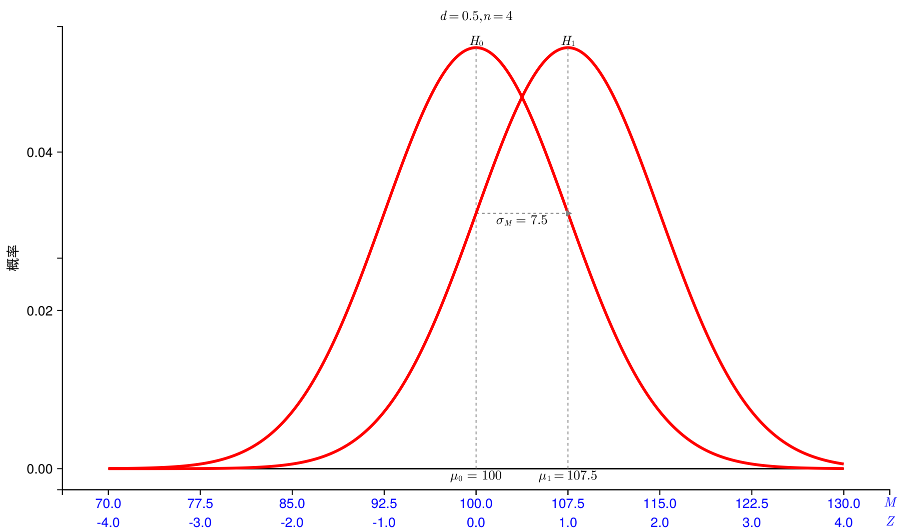
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.6.3.2 效应大小和样本大小

```
PlotZ(100, 15, n =  4, d = 0.5, showAlternative = true, XExtrema = (70, 130))
```

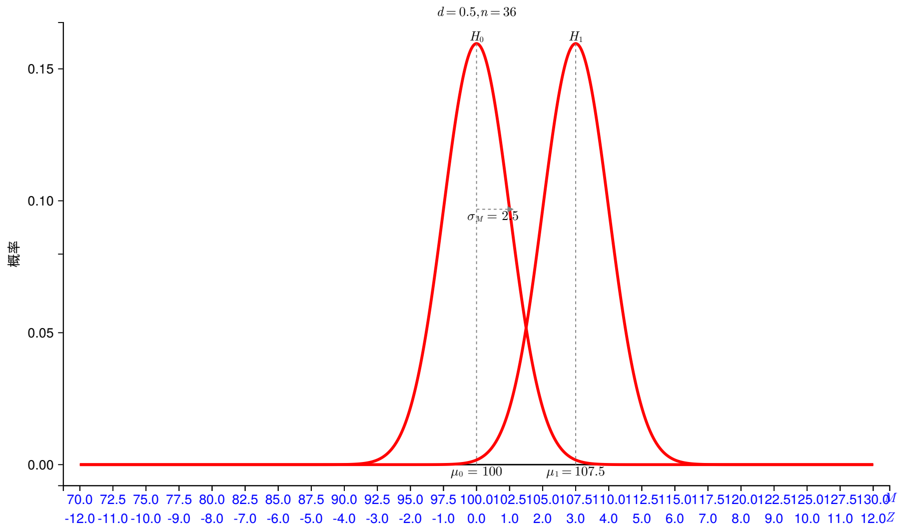
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(100, 15, n = 36, d = 0.5, showAlternative = true, XExtrema = (70, 130))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
```

```
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.7 统计效力

8.7.1 统计效力

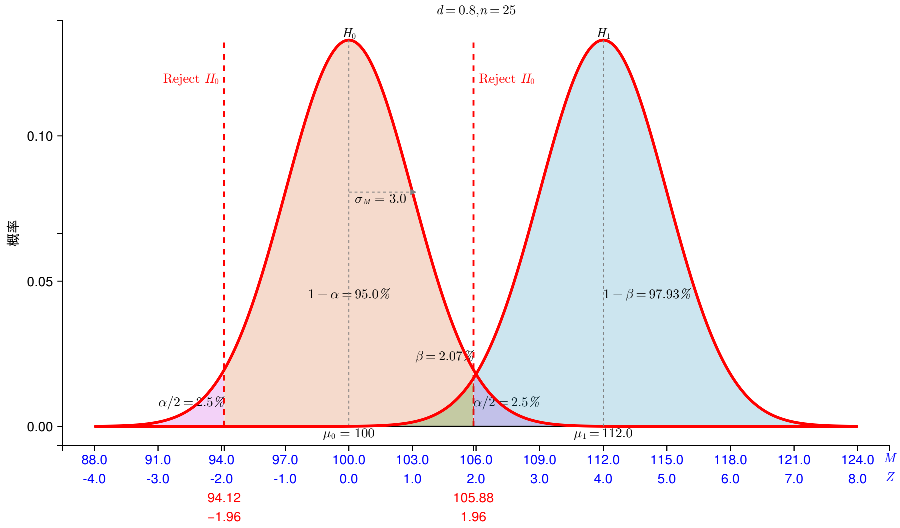
与直接测量效应大小不同，确定处理效应的大小或强度的另一种方法是测量统计检验的统计效力。**统计效力**（Statistical Power）是统计检验将正确拒绝虚无假设的概率。也就是说，统计效力是检验将能够识别出一个真正存在的处理效应的概率。

每当处理产生效应时，假设检验只有两种可能的结果：要么不能拒绝 H_0，要么拒绝 H_0。由于只有两种可能的结果，第一种结果，当存在真实效应时不能拒绝 H_0，在先前被定义为第二类错误，其概率被标识为 $p = \beta$。因此，第二种结果必须具有概率 $1 - \beta$。然而，第二种结果，当存在真实效应时拒绝 H_0，就是检验的统计效力。因此，假设检验的统计效力等于 $1 - \beta$。

8.7.2 零分布和备择分布

```julia
PlotZ(100, 15, n = 25, d = 0.8, PVal = 0.05, showAlternative = true)
```

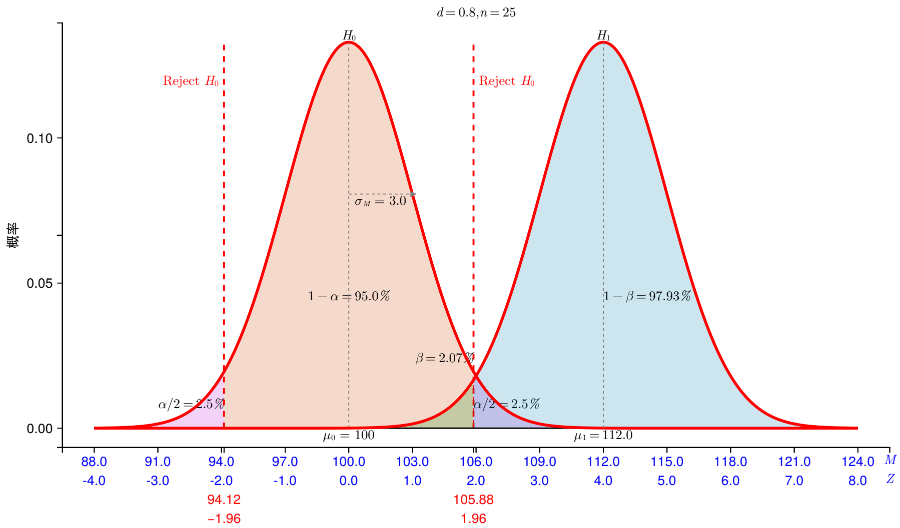
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.7.3 统计效力和效应大小

```
PlotZ(100, 15, n = 25, d = 0.8, PVal = 0.05, showAlternative = true)
```

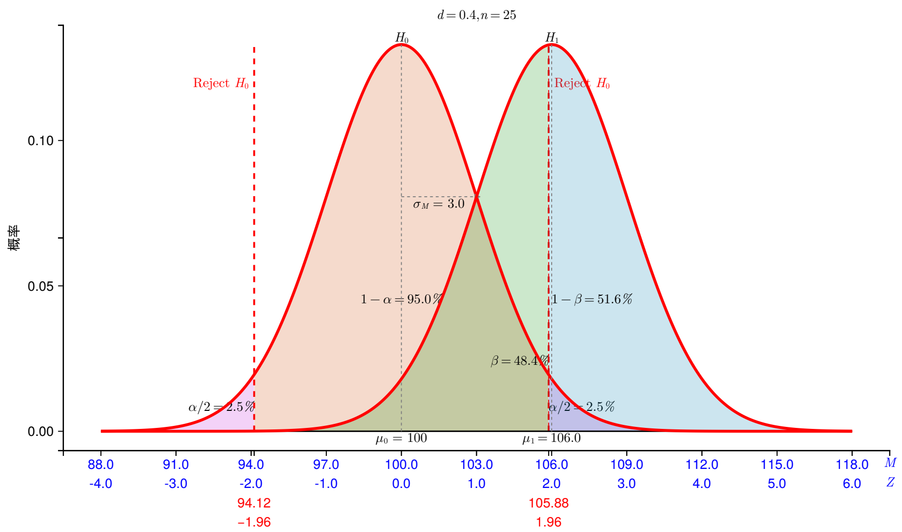
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(100, 15, n = 25, d = 0.4, PVal = 0.05, showAlternative = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
```

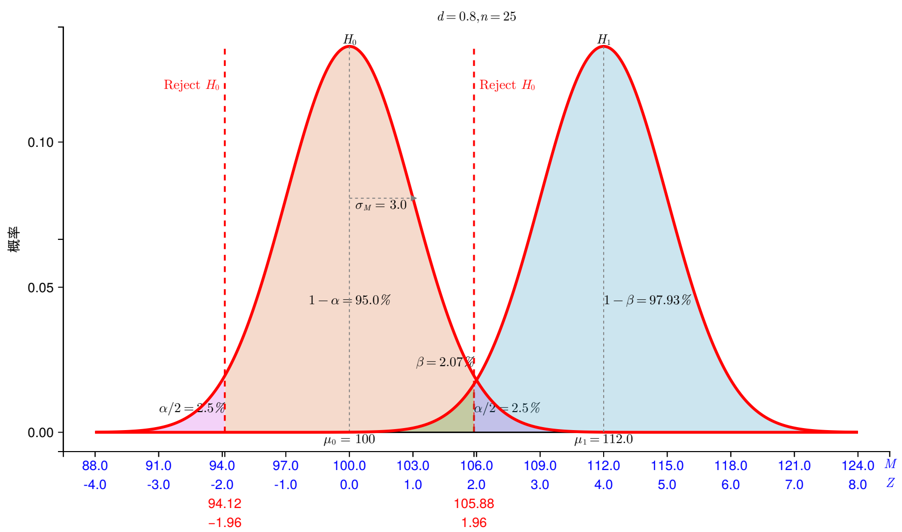
```
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.7.4 统计效力和样本大小

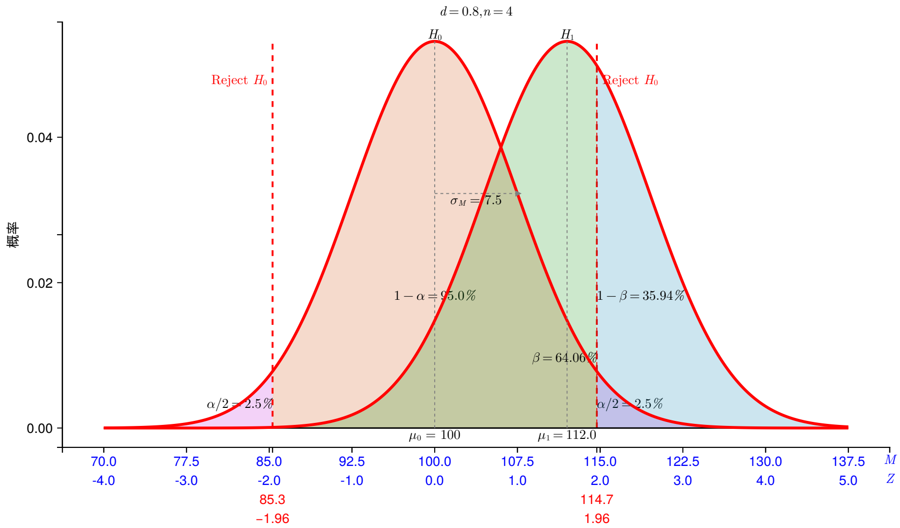
```
PlotZ(100, 15, n = 25, d = 0.8, PVal = 0.05, showAlternative = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



```
PlotZ(100, 15, n =  4, d = 0.8, PVal = 0.05, showAlternative = true)
```

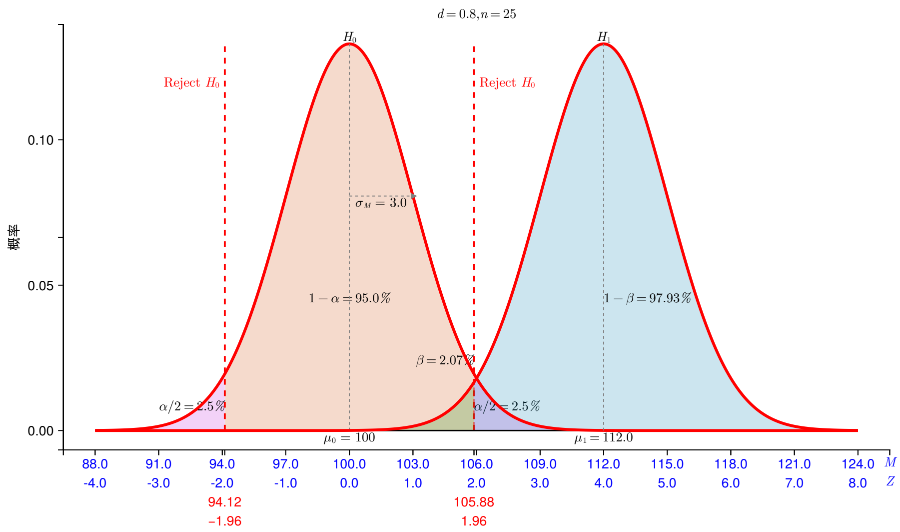
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.7.5 统计效力和显著性水平

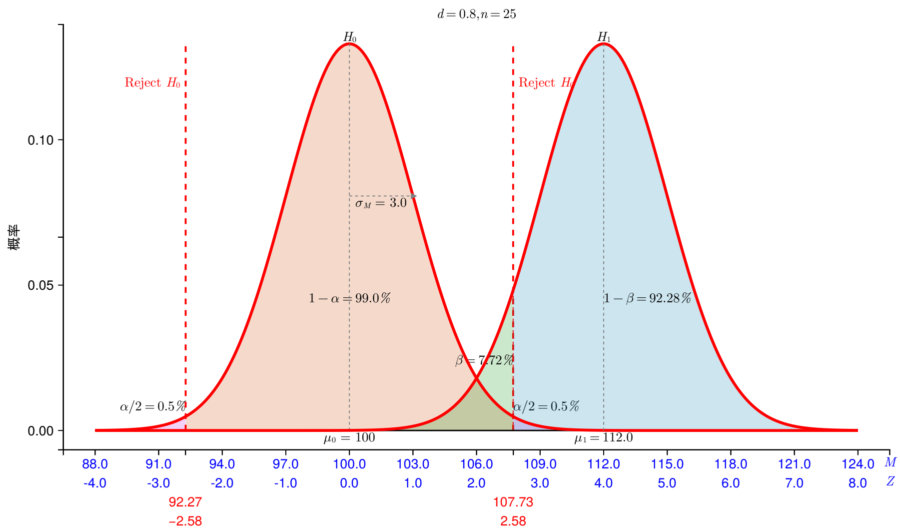
```
PlotZ(100, 15, n = 25, d = 0.8, PVal = 0.05, showAlternative = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(100, 15, n = 25, d = 0.8, PVal = 0.01, showAlternative = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```

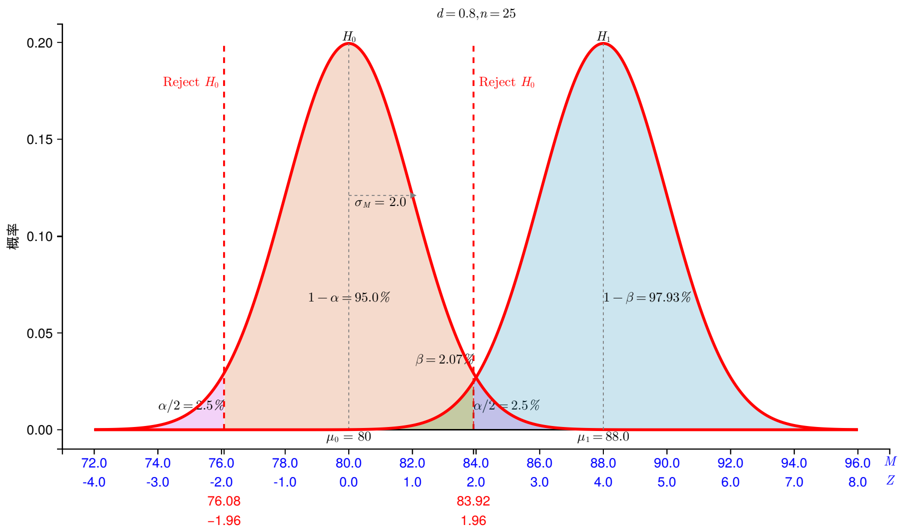


8.7.6 统计效力和单尾 vs. 双尾

```
PlotZ(80, 10, n = 25, d = 0.8, PVal = 0.05, showAlternative = true)
```

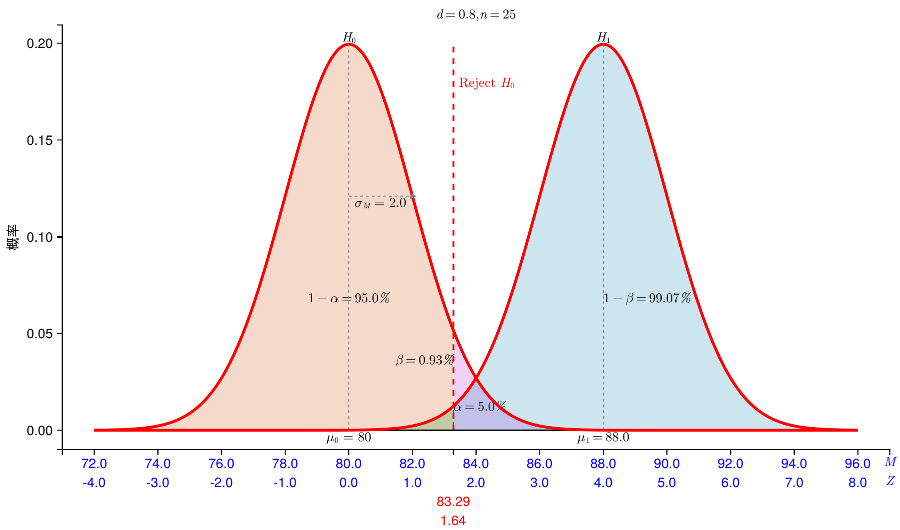
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
```

```
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
PlotZ(80, 10, n = 25, d = 0.8, PVal = 0.05, showAlternative = true, Tail = "Right")
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



8.8 正确理解 P 值

8.8.1 错误和正确理解

The p-value is the probability of observing a value for our test statistic that is as extreme (or more extreme) than the value we have calculated from our sample data, given that the null hypothesis is true.

P-值是指在零假设为真的前提下，检验统计量观测到当下值或比当下值更极端的概率。

– May I Show You My Collection of p-Values?

- *If* $P = .05$*, the null hypothesis has only a 5% chance of being true.*
  如果 $P = .05$，零假设只有 5% 的概率是真实的。
- *A nonsignificant difference (eg,* $P > .05$*) means there is no difference between groups.*
  一个不显著的差异（例如，$P > .05$）意味着两组之间没有差异。
- *A statistically significant finding is clinically important.*
  一个在统计上显著的发现在临床上具有重要意义。
- *Studies with P values on opposite sides of .05 are conflicting.*
  具有 P 值在显著性水平为 $.05$ 的邻近区域两端的研究结果相互矛盾。
- *Studies with the same P value provide the same evidence against the null hypothesis.*
  具有相同 P 值的研究提供相同的反对零假设的证据。
- $P = .05$ *means that we have observed data that would occur only 5% of the time under the null hypothesis.*
  $P = .05$ 意味着我们观察到的数据只会在零假设下发生 5% 的时间。
- $P = .05$ *means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
  $P = .05$ 意味着如果你拒绝零假设，发生类型 I 错误的概率只有 5%。
- $P = .05$ *and* $P \leq .05$ *mean the same thing.*
  $P = .05$ 与 $P \leq .05$ 的含义相同。
- *P values are properly written as inequalities (eg,* $P < .02$ *when* $P = .015$*).*
  P 值应该以不等式形式书写（例如，当 $P = .015$ 时写作 "$P < .02$"）。
- *With a* $P = .05$ *threshold for significance, the chance of a type I error will be 5%.*
  以 $P = .05$ 作为显著性阈值，发生类型 I 错误的概率将为 5%。
- *You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
  当你不关心某个方向的结果，或者在那个方向上的差异是不可能的时，应该使用单侧 P 值。
- *A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*
  科学结论或操作策略应该基于 P 值是否显著来制定。

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematollogy, 45*(3), 135-140

8.8.2 显著性检验：争论

- Nuzzo, B. (2014). Statistical errors. *Nature, 506*, 150-152.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133.
- Benjamin, D. J., et al.,. (2017). Redefine Statistical Significance. *Nature Human Behavior, 2*, 6-10.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician, 73*(sup1).
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567*, 305-307.

第三部分

学生氏 t 检验

第九章　t 统计量简介

9.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using Distributions: TDist, Normal
using Distributions: ccdf, cdf, pdf, quantile, cquantile
using Statistics: mean, var, std
using StatsAPI: pvalue, confint
using HypothesisTests: OneSampleTTest
using CairoMakie
using StatsReIntro
using DataFrames
using GLM
```

9.2　t 统计量：z 统计量的替代

9.2.1　z 统计量检验的逻辑假设

在前一章中，我们介绍了允许研究人员使用样本均值来检验关于未知总体均值的假设的统计程序。这些统计程序基于一些基本概念，我们总结如下。

1. 预期样本均值（M）将接近其总体均值（μ）。
2. 标准误差提供了样本均值接近总体均值的程度的度量。

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

3. 为了量化我们关于总体的推断，我们通过计算 z-*分数检验统计量*将获得的样本均值（M）与假定的总体均值（μ）进行比较。

$$z = \frac{M-\mu}{\sigma_M} = \frac{\text{数据与假设之间的获得差异}}{\text{M 与 }\mu\text{ 之间的标准距离}}$$

假设检验的目标是确定数据与假设之间的获得差异是否显著大于随机情况下预期的。当 z 分数形成正态分布时，我们能够使用单位正态表来找到假设检验的临界区域。

9.2.2 z 分数存在的问题

使用 z 分数进行假设检验的不足之处在于，z 公式需要的信息通常比可用的信息更多。具体来说，z 分数需要我们了解总体标准差（或方差）值，这是计算标准误差所需的。然而，在大多数情况下，总体的标准差是未知的。实际上，进行假设检验的整个原因是获得关于未知总体的信息。这种情况似乎制造了一个悖论：你想要使用 z 分数来了解未知的总体，但在计算 z 分数之前，你必须了解关于总体的信息。幸运的是，这个问题有一个相对简单的解决方案。当总体的方差（或标准差）未知时，我们使用相应的样本值代替它。

9.2.3 介绍 t 统计量

9.2.3.1 什么是 t 统计量

估计的标准误差（s_M）是在 σ 的值未知时，用作实际标准误差 σ_M 的估计。它是从样本方差或样本标准差计算出来的，并提供了样本均值 M 与总体均值 μ 之间的标准距离的估计。当 σ 已知时，标准误差的公式为：

$$\text{标准误差} = \sigma_M = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

当 σ 未知时，估计的标准误差为：

$$\text{估计的标准误差} = s_M = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

样本方差提供了相应总体方差的无偏估计：

$$\text{样本方差} = s^2 = \frac{SS}{n-1} = \frac{SS}{df}$$

$$\text{样本标准差} = s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

现在，我们可以将估计的标准误差代入 z 分数公式的分母。结果是一个称为 t *统计量*（t statistic）的新的检验统计量：

$$t = \frac{M-\mu}{s_M} = \frac{M-\mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

t 统计量（t statistic）用于测试关于未知总体均值 μ 的假设，当 σ 的值未知时。t 统计

量的公式结构与 z 分数的公式相同，只是 t 统计量在分母中使用了估计的标准误差。

$$z = \frac{M-\mu}{\sigma_M} = \frac{M-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$
$$t = \frac{M-\mu}{s_M} = \frac{M-\mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

9.2.3.2 自由度和 *t* 统计量

简要回顾，你必须在计算样本方差之前知道样本均值。这对样本变异性施加了限制，以便只有样本中的 *n*-1 个分数是独立且自由变化的。值 *n*-1 称为样本方差的**自由度** (degrees of freedom 或 df)。

$$\text{自由度} = df = n-1$$

对于样本的自由度 df 越大，样本方差 s^2 越好地代表了总体方差 σ^2，t 统计量越好地逼近 z 分数。这是因为样本（n）越大，样本越好地代表其总体。因此，与 s^2 关联的自由度还描述了 t 如何逼近 z。

9.2.4 *t* 分布

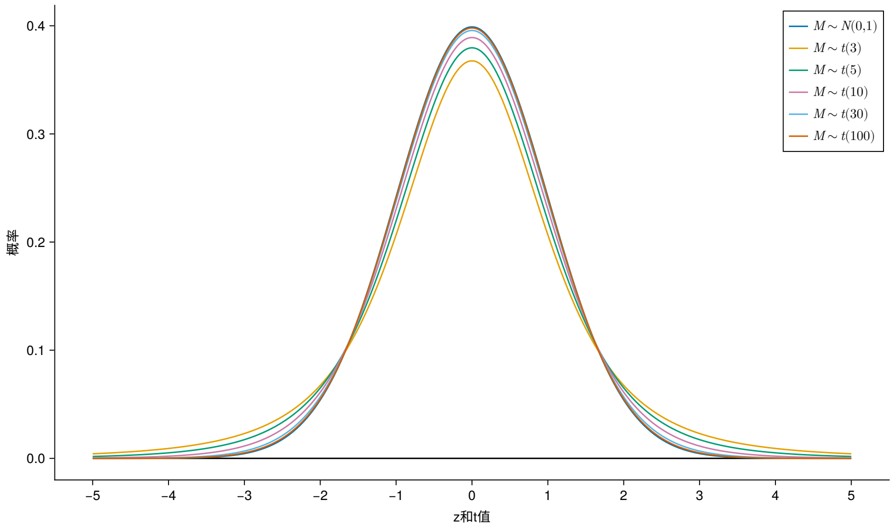
9.2.4.1 *t* 分布

一个 **t 分布**（t 分布）是为特定样本大小（*n*）或特定自由度（df）计算的每个可能随机样本的 t 值的完整集合。t 分布近似于正态分布的形状。

9.2.4.2 不同自由度下 *t* 统计量的分布

```
PlotT([3, 5, 10, 30, 100], showBoundary = false, showNormal = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



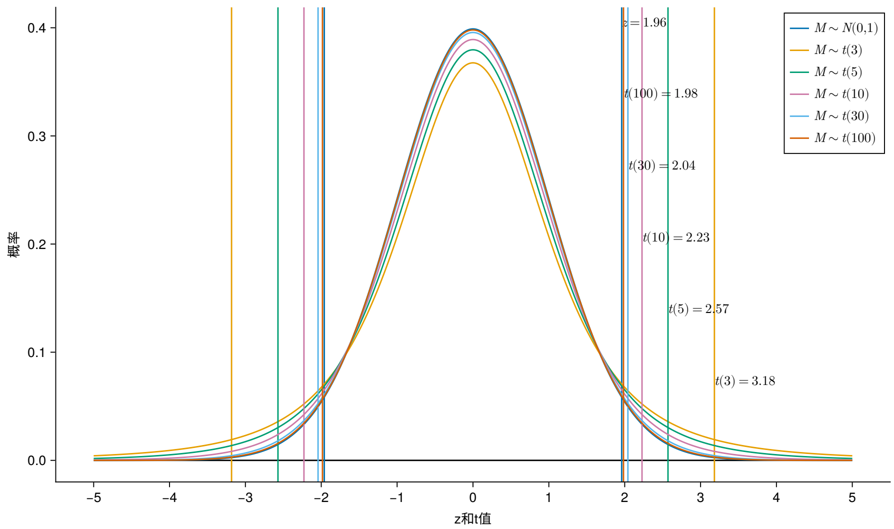
t 分布的确切形状随着自由度的变化而变化。事实上，统计学家谈论的是一个“t 分布家族”。对于每个可能的自由度数量，都有一个不同的 t 的抽样分布（所有可能样本 t 值的分布）。随着自由度（df）变得非常大，t 分布在形状上越来越接近单元正态 z 分布。

t 的分布呈钟形，对称，并且均值为零。但是，t 分布比正态 z 分布具有更多的变异性，特别是在自由度值较小的情况下。t 分布往往较平坦且分散，而正态 z 分布具有更多的中央峰值。

9.2.4.3 确定 t 分布的概率

```
PlotT([3, 5, 10, 30, 100], showBoundary = true, showNormal = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



9.2.4.4 确定 *t* 分布的概率

	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails Combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

图 9.1: t 分布表

本书中的 t 分布表已经被删减，不包括每个可能的 df 值的条目。在这些情况下，你应该查找列出的周围 df 值的临界 t，然后使用较大的 t 值。例如，加入研究的自由度为 35，显著性水平为.05，那么临界值应该为?

9.3 用 *t* 统计量进行假设检验

9.3.1 *t* 统计量检验的基本逻辑

9.3.1.1 使用 *t* 统计量进行假设检验

在假设检验的情境中，我们开始于一个具有未知均值和未知方差的总体，通常是一个接受了某种处理的总体。目标是使用来自经过处理的总体的样本（经过处理的样本）来确定处理是否有任何效果。

与以往一样，零假设表明处理没有效果；具体来说，H_0 表明总体均值没有改变。因此，零假设为未知总体均值提供了一个具体的值。样本数据提供了样本均值的值。最后，方差和估计的标准误差是从样本数据中计算出来的。

当这些值用于 t 公式时，结果如下：

$$t = \frac{\text{样本均值 (来自数据) - 总体均值 (从 } H_0 \text{ 假设中推断)}}{\text{估计的标准误差 (从样本数据中计算)}}$$

与 z 分数公式一样，t 统计量形成一个比率。分子测量样本数据（M）和总体假设（μ）之间的实际差异。分母中的估计标准误差测量在样本均值和总体均值之间可以合理期望多大差异。

当数据与假设之间的实际差异（分子）远大于期望差异（分母）时，我们得到了一个较大的 t 值（可以是正值或负值）。在这种情况下，我们得出结论，数据与假设不一致，我们的决策是"拒绝 H_0"。

另一方面，当数据与假设之间的差异相对于标准误差很小时，我们得到了一个接近零的 t 统计量，我们的决策是"无法拒绝 H_0"。

9.3.2 未知的总体

9.3.2.1 *t* 统计量假设检验的基本研究情境。

假设在处理前已知总体参数 μ。研究的目的是确定处理是否有影响。请注意，处理后的总体均值和方差的值未知。我们将使用一个样本来测试关于总体均值的假设。

9.3.3 假设检验示例

9.3.3.1 假设检验的四个步骤

1. 提出假设并选择 alpha 水平。
2. 确定临界区域。
3. 计算检验统计量。
4. 对 H_0 做出决定。

9.3.3.2 示例：对漂亮面孔的偏好

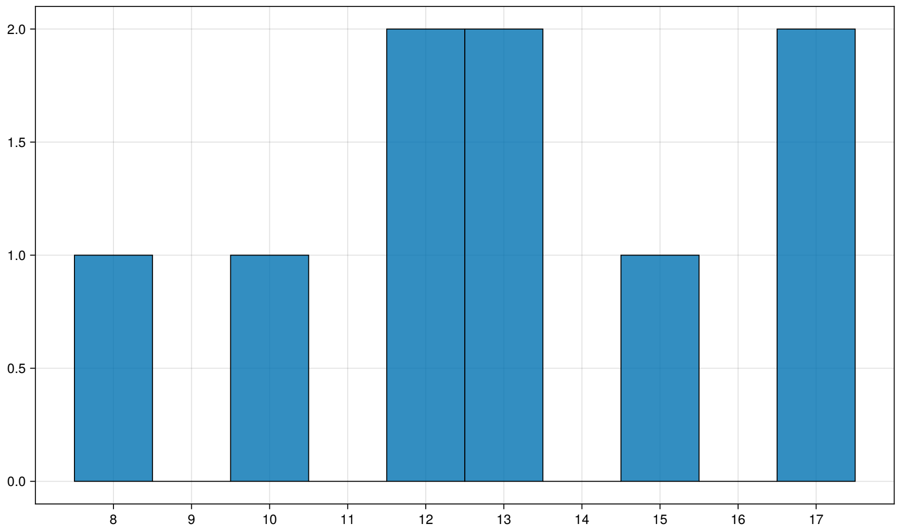
婴儿，甚至新生儿，更喜欢看漂亮的面孔而不是不太漂亮的面孔。在这项研究中，1-6 天大的婴儿被展示了两张女性面孔的照片。事先，一组成年人已经评价了其中一张面孔比另一张更漂亮。婴儿被放置在一个屏幕前，屏幕上呈现照片。这两张图片始终保持在屏幕上，直到婴儿注视两张图片的总时间达到 20 秒。记录了每个婴儿观看漂亮面孔的秒数。假设该研究使用了一个包含 $n = 9$ 名婴儿的样本，并获得了以下数据。

```
sample = [8, 10, 12, 12, 13, 13, 15, 17, 17]
```

请注意，所有可用的信息都来自样本。具体来说，我们不知道总体均值或总体标准差。

```
hist(sample, bins = 7.5:1:17.5,
  strokewidth = 1, axis = (; xticks = 8:17))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



9.3.3.3 步骤 1. 提出假设并选择 alpha 水平

- 零假设：婴儿对两张面孔没有偏好。

$$H_0 : \mu_{\text{漂亮}} = 10\text{秒}$$

- 备择假设：存在偏好，一张面孔优先于另一张面孔。

$$H_1 : \mu_{\text{漂亮}} \neq 10\text{秒}$$

- 我们将显著性水平设置为 $\alpha = 0.05$，双尾检验。

9.3.3.4 步骤 2. 确定临界区域

- 自由度是

```
n = length(sample)
df = n - 1
```

8

- 临界区域是

```
quantile(TDist(df),  0.05/2),
cquantile(TDist(df), 0.05/2)
```

(-2.3060041352041662, 2.3060041352041662)

9.3.3.5 步骤 3. 计算检验统计量

- 计算离差平方和

```
using LinearAlgebra
O = ones(n)
P =  I - O * inv(O'O) * O'
SX = norm(P * sample)^2
SY = sum((sample .- mean(sample)).^2)
SS = sum(sample .^ 2) - sum(sample)^2 / n
SX, SY, SS
```

(71.99999999999999, 72.0, 72.0)

- 计算方差

```
VS = var(sample)
VV = SS / df
VS, VV
```

(9.0, 9.0)

- 计算估计的标准误差

```
SE = sqrt(VV / n)
```

1.0

- 计算样本数据的 t 统计量

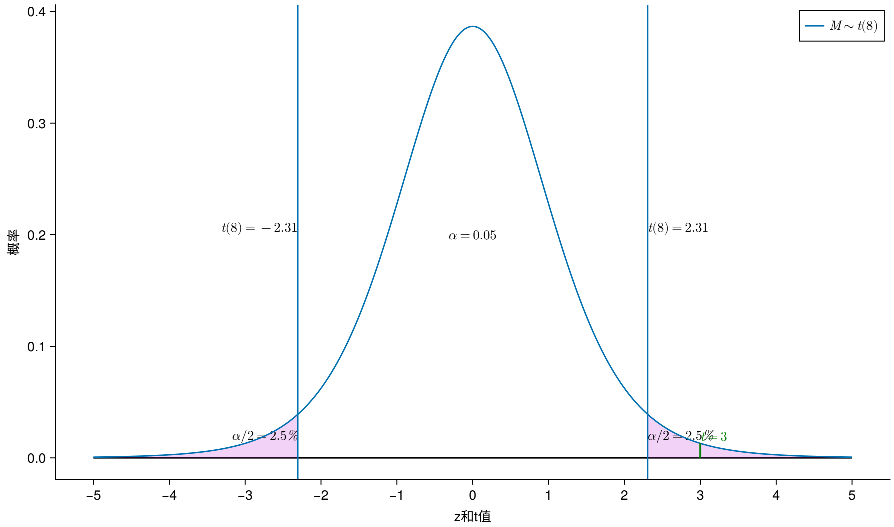
```
μ, M = 10, mean(sample)
t = (M - μ) / SE
```

3.0

9.3.3.6 步骤 4. 对 H_0 做出决定

```
PlotT(df, TVal = 3, showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



得到的 t 统计量为 3.00，位于 t 分布右侧的临界区域内。我们的统计决策是拒绝 H_0，并得出结论，婴儿在面临选择时确实显示出对漂亮面孔的偏好。

9.3.3.7 用 Julia 语言计算精确 p 值

直接计算累计概率密度

```
ccdf(TDist(df), t) * 2
```

```
0.01707168123378264
```

用 HypothesisTests 包算法

```
TT = OneSampleTTest(sample, 10)
pvalue(TT)
```

```
0.01707168123378264
```

9.3.4 t 检验的假设

- 样本中的值必须是独立观测的。
- 所抽样的总体必须服从正态分布。

9.3.5 样本大小和样本方差的影响

t 公式的结构:

$$t = \frac{M - \mu}{s_M}$$

由于估计的标准误差 s_M 出现在公式的分母中，所以较大的 s_M 值会产生较小（接近零）的 t 值。因此，任何影响标准误差的因素也会影响拒绝零假设并发现显著处理效应的可能性。

$$s_M = \sqrt{\frac{s^2}{n}}$$

9.4 t 统计量的效应大小

9.4.1 估计的 Cohen's d

原始的 Cohen's d:

$$\text{Cohen's } d = \frac{\text{均值差异}}{\text{标准差}} = \frac{\mu_{\text{处理}} - \mu_{\text{无处理}}}{\sigma}$$

估计的 Cohen's d:

$$\text{估计的 } d = \frac{M - \mu}{s}$$

我们的示例

```
μ, M, s = 10, mean(sample), std(sample)
(M - μ) / s
```

```
1.0
```

9.4.2 解释的方差百分比，r^2

9.4.2.1 方差的分解

$$\begin{aligned}
\sum_1^n (X_i-\mu)^2 &= \sum_1^n [(X_i-M)+(M-\mu)]^2 \\
&= \sum_1^n (X_i-M)^2 + \sum_1^n (M-\mu)^2 + \sum_1^n [2\times(X_i-M)(M-\mu)] \\
&= \sum_1^n (X_i-M)^2 + \sum_1^n (M-\mu)^2
\end{aligned}$$

所以方差可分解为:

$$\begin{array}{ccccc}
\sum_1^n (X_i-\mu)^2 & = & \sum_1^n (X_i-M)^2 & + & \sum_1^n (M-\mu)^2 \\
\Downarrow & & \Downarrow & & \Downarrow \\
SST & = & SSE & + & SSR \\
 & & \Uparrow & & \Uparrow \\
 & & \text{离差平方和}(SS) & & n\times(M-\mu)^2
\end{array}$$

处理占方差的百分比被称为 r^2:

$$r^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \frac{SST-SSE}{SST} = 1-\frac{SSE}{SST}$$

被解释方差占总方差的比率是由方差占比来决定的。方差是离差平方和，是差别的二阶数据；而 t 值是则是差别的一阶数据。所以可以先把 t 值平方来推导二者的关系。如下公式所示。

$$\begin{aligned}
t^2 &= \frac{(M-\mu)^2}{SE^2} = \frac{SSR/n}{SSE/(n\times df)} = \frac{SSR}{SSE}\times df \\
&\Updownarrow \\
\frac{t^2}{df} &= \frac{SSR}{SSE}
\end{aligned}$$

有了上述结论，可以如下过程推到 r^2 和 t 值之间的关系:

$$
\begin{aligned}
r^2 &= \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{SSR/SSE}{SSR/SSE + 1} \\
&= \frac{t^2/df}{t^2/df + 1} = \frac{t^2}{t^2 + df}
\end{aligned}
$$

我们的例子

```
dt = DataFrame(数据 = sample, 样本均值 = mean(sample), 整体均值 = 10.0)
```

数据	样本均值	整体均值
8	13.0	10.0
10	13.0	10.0
12	13.0	10.0
12	13.0	10.0
13	13.0	10.0
13	13.0	10.0
15	13.0	10.0
17	13.0	10.0
17	13.0	10.0

```
X, M, μ = dt.数据, dt.样本均值, dt.整体均值
SST = sum((X - μ) .^ 2)
SSE = sum((X - M) .^ 2)
SSR = sum((M - μ) .^ 2)
SST, SSE, SSR
```

```
(153.0, 72.0, 81.0)
```

```
SSR / SST
```

```
0.5294117647058824
```

公式二

```
df, t = TT.df, TT.t
t^2/(t^2+df)
```

```
0.5294117647058824
```

9.4.2.2 测量解释方差的百分比，r^2

$r^2 = 0.01$	小效应
$r^2 = 0.09$	中等效应

$r^2 = 0.25$	大效应

9.4.2.3 t 检验作为线性模型的特殊形式

```
fm = lm(@formula((数据 - 10) ~ 1), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	3.0	1.0	3.0	0.0170717	0.693996	5.306

9.4.3 估计 μ 的置信区间

9.4.3.1 置信区间

描述处理效应大小的另一种方法是计算处理后总体均值的估计值。估计未知总体均值要构建一个置信区间。置信区间基于这样一个观察：样本均值往往能够相对准确地估计总体均值。

样本均值趋向于接近总体均值的事实意味着总体均值应该接近样本均值。因此，置信区间包括围绕样本均值的一系列值的区间，我们可以相对有信心地认为未知的总体均值位于该区间内。**置信区间**（confidence interval）是围绕样本统计量的一个区间或数值范围。置信区间背后的逻辑是，样本统计量，如样本均值，应该与相应的总体参数相对接近。因此，我们可以有信心地估计参数的值应该位于该区间内。

9.4.3.2 构建置信区间

构建置信区间的过程始于这个观察：每个样本均值都有一个相应的 t 值，由方程定义：

$$t = \frac{M - \mu}{s_M}$$

尽管 M 和 s_M 的值可以从样本数据中获得，但我们不知道 t 或 μ 的值。但是，我们可以估算 t 值。注意，t 值在 $t = 0$ 附近堆积，因此我们可以估算出样本的 t 值应该在 0 附近。此外，t 分布表列出了与 t 分布的特定部分对应的不同 t 值。

为了构建 μ 的置信区间，我们将估算 t 值代入 t 方程，然后可以计算 μ 的值。因为目标是计算 μ 值，所以我们使用简单的代数来解出方程中的 μ。$\pm t$ 可以纳入方程中以得出：

$$\mu = M \pm t \times s_M$$

这是置信区间的基本方程。请注意，该方程产生了一个围绕样本均值的值范围。该区间的一端位于 $M + t_{s_M}$，另一端位于 $M - t_{s_M}$。

在我们的示例中，80% 的 t 值位于：

```
pval = (1-0.8) / 2
n , df = length(sample), length(sample) - 1
pst, ngt = quantile(TDist(df), pval), cquantile(TDist(df), pval)
```

```
(-1.396815309743865, 1.396815309743865)
```

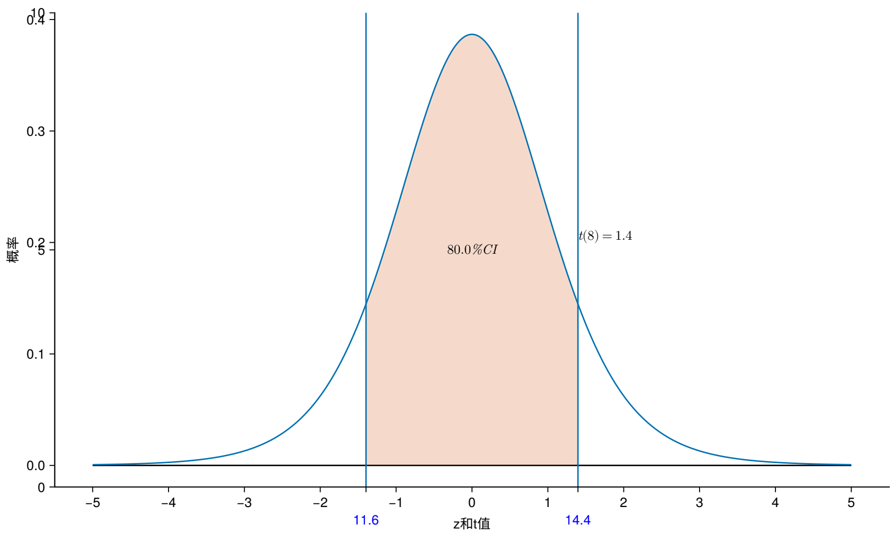
因为所有可能的 t 值的 80% 位于 ± 1.397 之间，所以我们可以有 80% 的信心，我们的样本均值对应于此区间内的一个 t 值。使用样本数据和估算的 t 值范围，我们得到了 μ 的 80% 置信区间：

```
M, sₘ = mean(sample), std(sample) / sqrt(n)
M + pst * sₘ, M + ngt * sₘ
```

```
(11.603184690256136, 14.396815309743864)
```

```
PlotT(8, PVal = 1-0.8, showCI = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



9.4.3.3 使用 julia 语言自动计算

```
TT = OneSampleTTest(sample, 10)
confint(TT; level = 0.8)
```

```
(11.603184690256136, 14.396815309743864)
```

9.4.3.4 置信区间宽度的影响因素

当您改变置信水平（百分之多少的置信度）时，请注意区间宽度会发生什么变化。为了对您的估计更有信心，您必须增加区间的宽度。相反，如果要获得更小、更精确的区间，则必须放弃信心。

注意，如果您改变样本大小，会发生什么区间宽度的变化。样本越大（n），区间就越小。

9.4.4 文献中的内容

9.4.4.1 报告 *t* 检验的结果

首先，请记住，科学报告通常使用术语“显著”来表示拒绝零假设，使用术语“不显著”来表示未能拒绝 H_0。此外，关于报告 t 检验的计算值、自由度和阿尔法水平，有一种规定的格式。

在科学报告中，这些信息以简明的陈述方式传达，如下所示：

The infants spent an average of $M = 13$ out of 20 seconds looking at the attractive face, with $SD = 3.00$. Statistical analysis indicates that the time spent looking at the attractive face was significantly greater than would be expected if there were no preference, $t(8) = 3.00$, $p < .05$, $r^2 = 0.5294$.

婴儿平均花费 $M = 13$ 秒钟观看吸引人的面孔，标准差 $SD = 3.00$。统计分析表明，观看吸引人的面孔的时间显著大于如果没有偏好时所期望的，$t(8) = 3.00$，$p < .05$，$r^2 = 0.5294$。

第一句报告描述性统计数据，均值（$M = 13$）和标准差（$SD = 3$），如前所述。接下来的陈述提供了推论统计分析的结果。请注意，自由度在 t 符号之后用括号报告。获得的 t 统计量的值跟在后面（3.00），接下来是发生第一类错误的概率（小于 5%）。最后，报告效应大小，$r^2 = 52.94$。

如果报告中包含了 80% 置信区间作为效应大小的描述，那么它将在假设检验结果之后添加如下所示：

$t(8) = 3.00$，$p < .05$，$80\%\ CI\ [11.603, 14.397]$。

除了计算均值、标准差和数据的 t 统计量之外，计算机通常还会计算并报告与 t 值相关的确切概率（或 α 水平）。

$t(8) = 3.00$，$p = .017$，$r^2 = 0.5294$。

最后要注意的一点是，偶尔 t 值可能极端到计算机报告 $p = 0.000$。零值不表示概率实际上是零；相反，它意味着计算机将概率值四舍五入到三位小数并获得结果 0.000。在这种情况下，您不知道确切的概率值，但可以报告 $p < .001$。

9.5 方向性假设和单尾检验

9.5.0.1 方向性假设和单尾检验

与单尾（方向性）替代方案相比，双尾（非方向性）检验更常用。在某些研究情况下，如探索性研究、试点研究或有先验理由（例如理论或先前的研究）时，可以使用单尾检验。

9.5.0.2 步骤 1. 提出假设并选择临界水平

- 零假设：不超过 20 秒的一半时间看吸引人的脸，即

$$H_0 : \mu_{吸引人} \leq 10秒$$

- 备择假设：超过 20 秒的一半时间看吸引人的脸，即

$$H_1 : \mu_{吸引人} > 10秒$$

我们将显著性水平设定为 $\alpha = .01$。

9.5.0.3 步骤 2. 确定临界区域

- 自由度为

```
n, df = length(sample), length(sample) - 1
```

```
(9, 8)
```

- 临界区域为

```
tval = cquantile(TDist(df), 0.01)
```

```
2.896459447709622
```

可以将此步骤分为两个阶段，消除了确定临界区域应包含哪个尾部（右侧或左侧）的需要。这个过程的第一阶段只是确定样本均值是否与原始研究问题预测的方向一致。过程的第二阶段是确定效应是否足够显著。

9.5.0.4 步骤 3. 计算检验统计量

- 计算样本方差和标准误

```
SS = sum(sample .^ 2) - sum(sample)^2 / n
vv = SS / df
SE = sqrt(vv / n)
```

```
1.0
```

```
SE = std(sample) / sqrt(n)
```

```
1.0
```

- 计算检验统计量

```
M, μ = mean(sample), 10
t = (M - μ) / SE
```

```
3.0
```

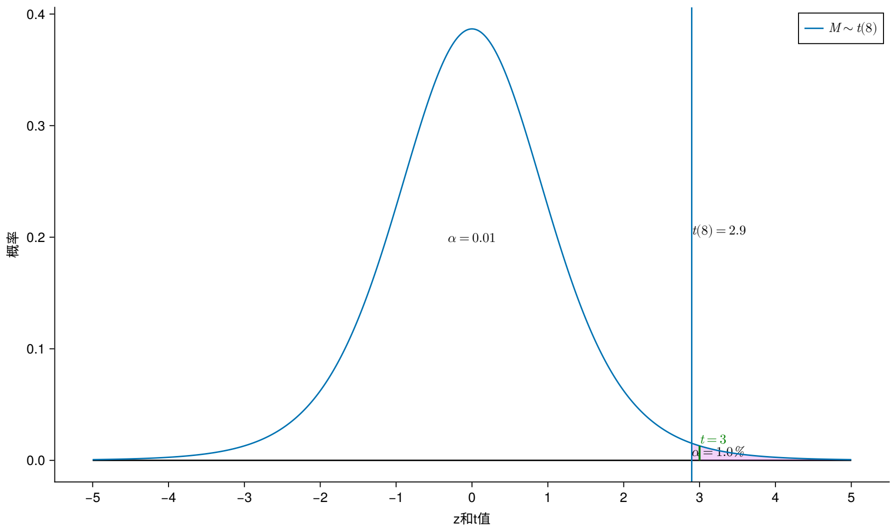
```
TT = OneSampleTTest(sample, 10)
TT.t
```

```
3.0
```

9.5.0.5 步骤 4. 做出决策

```
PlotT(8, TVal = 3, PVal = 0.01, Tail = "Right", showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



检验统计量位于临界区域内，因此我们拒绝了 H_0。在研究报告中，结果将如下呈现：

The time spent looking at the attractive face was significantly greater than would be expected if there were no preference, $t(8) = 3.00$, $p < .01$, one tailed.

观看吸引人的脸的时间明显大于如果没有偏好时所期望的，$t(8) = 3.00$, $p < .01$，单尾。

请注意，报告明确承认使用了单尾检验。

9.5.0.6 Julia 语言的自动计算

```julia
pvalue(TT, tail = :right)
```

```
0.00853584061689132
```

```julia
confint(TT, tail = :right)
```

```
(11.140451962469102, Inf)
```

第十章　独立样本 t 检验

10.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using DataFrames
using CairoMakie, StatsReIntro
using Distributions
using HypothesisTests: EqualVarianceTTest, UnequalVarianceTTest, VarianceFTest, pvalue, confi
using GLM
```

10.2　独立测量设计

10.2.0.1　单样本设计

直到目前为止，我们考虑的所有推断统计都涉及使用一个样本作为推断一个总体的基础。尽管**单样本**技术偶尔在实际研究中使用，但大多数研究需要比较*两个（或更多）组*的数据。

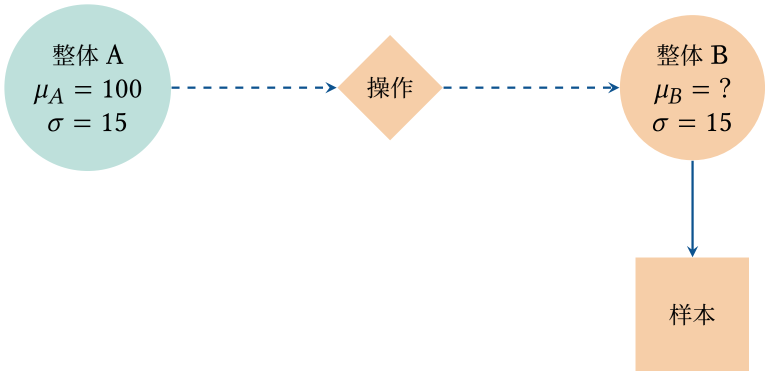
10.2.0.2　被试间设计和被试内设计

用于获取两组数据的研究设计可分为两个一般类别：两组数据可以来自完全不同的参与者组；两组数据可以来自同一组参与者。

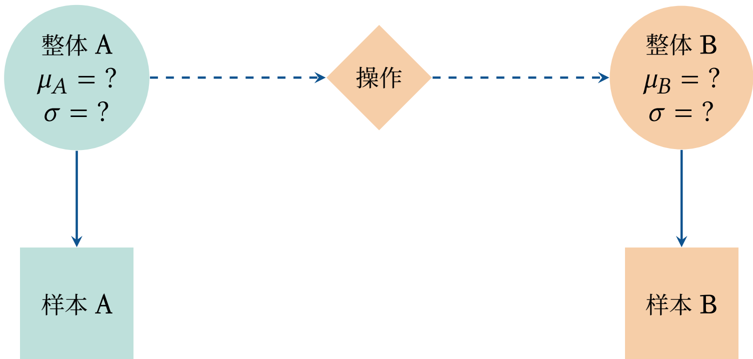
每个处理条件（或每个总体）使用单独被试组的研究设计被称为**独立测量**研究设计（Independent-Measures Design）或**被试间设计**（Between-Subjects Design）。

重复测量设计（Repeated-Measures Design）或**被试内设计**（Within-Subjects Design）是一种依赖变量被多次测量的研究设计，对于每个样本中的每个个体，都会进行多次测量。所有处理条件都使用相同的参与者组。

10.2.0.3 单样本设计研究的结构



10.2.0.4 独立测量设计研究的结构



10.3 零假设和独立测量 t 统计量

10.3.1 独立测量检验的假设

10.3.1.1 独立测量 t 的一些符号

由于独立测量研究涉及两个不同的样本，因此我们需要一些特殊的符号来帮助指定哪些数据与哪个样本有关。这个符号使用下标，下标是写在样本统计量旁边的小数字。例如：n_1，n_2；M_1，M_2；SS_1，SS_2。

10.3.1.2 独立测量检验的假设

独立测量研究的目标是评估两个总体（或两个处理条件）之间的均值差异。使用下标来区分两个总体，第一个总体的均值是 μ_1，第二个总体的均值是 μ_2。均值之间的差异为 $\mu_1 - \mu_2$。

与往常一样，零假设表示没有变化、没有效应，或者在这种情况下没有差异。用符号表示独立测量检验的零假设

$$H_0 : \mu_1 - \mu_2 = 0$$

（两个总体均值没有差异）

备择假设则表示两个总体之间存在均值差异

$$H_1 : \mu_1 - \mu_2 \neq 0$$

（存在均值差异）

10.3.2 独立测量检验的公式

10.3.2.1 *t* 统计量基本结构

独立测量和单样本的假设检验中，t 统计量的基本结构是相同的，即

$$t = \frac{\text{样本统计量} - \text{假设的总体参数}}{\text{估计标准误}}$$

独立测量的 t 基本上是单样本 t 的所有元素都加倍的两样本 t 公式。

10.3.2.2 总体 *t* 公式

- 单样本 t 公式

$$t = \frac{\text{样本均值} - \text{总体均值}}{\text{估计标准误}} = \frac{M - \mu}{s_M}$$

- 独立测量 t 公式

$$t = \frac{\text{样本均值差异} - \text{总体均值差异}}{\text{估计标准误}} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

在假设检验中，零假设设定了总体均值差异为零，因此独立测量 t 的公式可以进一步简化为

$$t = \frac{\text{样本均值差异}}{\text{估计标准误}}$$

10.3.2.3 估计标准误

在每个 t-分数公式中，分母中的标准误度量样本统计量代表总体参数的准确度。在单样本 t 公式中，标准误度量样本均值的误差量，表示为 s_M。对于独立测量 t 公式，标准误度量当您使用样本均值差异 $(M_1 - M_2)$ 来代表总体均值差异 $(\mu_1 - \mu_2)$ 时，预期会出现多少误差。样本均值差异的标准误由符号 $s_{M_1 - M_2}$ 表示。

通常情况下，标准误度量样本统计量代表总体参数的准确程度。标准误的符号采用 $s_{\text{统计量}}$ 的形式。当统计量是样本均值 M 时，标准误的符号为 s_M。对于独立测量检验，统计量是样本均值差异 $(M_1 - M_2)$，标准误的符号为 $s_{M_1 - M_2}$。在每种情况下，标准误表示样本统计量与相应总体参数之间合理的差异有多大。

有两种方法可以解释估计均值差异标准误 $(M_1 - M_2)$：标准误被定义为样本统计量 $(M_1 - M_2)$ 与相应总体参数 $(\mu_1 - \mu_2)$ 之间的标准或平均距离的度量。它测量了如果零假设为真，则 $(M_1 - M_2)$ 的标准或平均大小。也就是说，它测量了可以合理预期两个样本均值之间的差异有多大。然而，当零假设为真时，总体均值差异为零。在这种情况下，标准误度量了样本均值差异 $(M_1 - M_2)$ 平均比零值偏多远。但是，测量它距离零有多远等同于测量它有多大。因此，当零假设为真时，标准误测量了样本均值差异 $(M_1 - M_2)$ 平均有多大。

10.3.3 计算估计标准误

10.3.3.1 理解估计标准误

样本一的标准误

$$s_{M_1} = \sqrt{\frac{s_1^2}{n_1}}$$

样本二的标准误

$$s_{M_2} = \sqrt{\frac{s_2^2}{n_2}}$$

两个独立样本的标准误 不是：

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

10.3.3.2 混合方差

两个独立样本的标准误为

$$s_{(M_1-M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p\sqrt{\frac{n_1+n_2}{n_1 n_2}} = s_p/\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

两个样本的 **混合方差**（Pooled variance）s_p^2 可以通过两个样本方差的加权平均来计算，每个单独样本方差的权重由该样本的**自由度**（贝塞尔矫正）而非样本大小决定。

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{SS_1 + SS_2}{df_1 + df_2}$$

当两个样本完全相同时，即 $n_1 = n_2 = n$，$df_1 = df_2 = n - 1$ 时，

$$s_p^2 = \frac{SS_1 + SS_2}{2(n-1)}$$

估计标准误的公式可以简化为

$$s_{(M_1-M_2)} = \sqrt{\frac{2s_p^2}{n}} = \sqrt{\frac{SS_1 + SS_2}{(n-1)\times n}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

10.3.4 最终公式和自由度

独立测量 t 统计量的完整公式为

$$t = \frac{\text{样本均值差异} - \text{总体均值差异}}{\text{估计标准误}} = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1-M_2)}}$$

自由度

$$\begin{aligned} df_{(M_1-M_2)} &= df_1 + df_2 \\ &= (n_1 - 1) + (n_2 - 1) \\ &= n_1 + n_2 - 2 \end{aligned}$$

10.3.4.1 t 统计量的基本元素

	样本统计量	总体参数	估计标准误	样本方差
单样本 t 检验	M	μ	$\sqrt{\frac{s^2}{n}}$	$s^2=\frac{SS}{df}$
独立样本 t 检验	(M_1-M_2)	$(\mu_1-\mu_2)$	$\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}$	$s_p^2=\frac{SS_1+SS_2}{df_1+df_2}$

10.3.4.2 基本理论总结

如果 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，则 $X-Y\sim N(\mu_1-\mu_2,\sigma_1^2+\sigma_2^2)$。若从两个分布中选出容量分别为 n_1 和 n_2 的样本，则样本均值的分布为 $M_1-M_2\sim N(\mu_1-\mu_2,\sigma_1^2/n_1+\sigma_2^2/n_2)$。假定两个整体的整体方差相同，即 $\sigma_1=\sigma_2=\sigma$，则 $M_1-M_2\sim N(\mu_1-\mu_2,\sigma^2/n_1+\sigma^2/n_2)$。零假设情况下（$\mu_1-\mu_2$），整体方差未知时，可用联合样本方差代替整体方差，此时正态分布分布变成了自由度为 n_1+n_2-2 的 t 分布，即 $(M_1-M_2)/(s_p^2/n_1+s_p^2/n_2)\sim t(n_1+n_2-2)$。联合样本方差的计算方式与单样本的样本方差计算方式类似，需要进行贝塞尔矫正。

参考资料:

- Wagaman, A. S., & Dobrow, R. P. (2021). Probability with applications and R. Wiley. 285 页，基本理论。
- Demidenko, E. (2020). Advanced Statistics With Applications In R. Wiley-Blackwell. 第 284 页，提供了证明过程。

10.4 使用独立测量 t 统计量的假设检验

10.4.1 例子：光线与诚实

研究表明，人们在黑暗环境中更有可能表现出不诚实和自私的行为，而在良好照明的环境中则不太可能。在一个实验中，参与者在 5 分钟内解决了一组 20 道难题，每解决一道难题获得 0.50 美元。然而，参与者报告自己的表现，没有明显的方法来检查他们的诚实性。因此，这项任务提供了一个明确的机会来作弊和获得不应得的钱。

一组参与者在灯光昏暗的房间中接受测试，另一组在明亮的房间中接受测试。记录每个个体报告的解决难题数量。

以下数据代表与该研究类似的结果。

```
亮房间 = [11,  9,  4,  5, 6,  7, 12, 10]
暗房间 = [ 7, 13, 14, 16, 9, 11, 15, 11]
```

数据框：宽格式

```
dt = DataFrame(亮房间 = 亮房间, 暗房间 = 暗房间)
```

亮房间	暗房间
11	7
9	13
4	14
5	16
6	9
7	11
12	15
10	11

数据整洁（tidy）的三个相互关联的规则

- 每个变量必须有自己的列。
- 每个观测必须有自己的行。
- 每个数值必须有自己的单元格。

数据框：长格式

```
dt = stack(dt, [:亮房间, :暗房间],
  variable_name = "房间类型", value_name = "分数")
```

房间类型	分数
亮房间	11
亮房间	9
亮房间	4
亮房间	5
亮房间	6
亮房间	7
亮房间	12
亮房间	10
暗房间	7
暗房间	13
暗房间	14
暗房间	16
暗房间	9
暗房间	11
暗房间	15
暗房间	11

10.4.1.1 假设检验的四个步骤

- 陈述假设并选择 alpha 水平。
- 设定决策标准。
- 获取数据并计算 t 统计量。

- 做出决定。

10.4.2 标准假设和双尾检验

10.4.2.1 陈述假设并选择 alpha 水平

两个假设

$H_0 : \mu_1 - \mu_2 = 0$ (无差异)

$H_1 : \mu_1 - \mu_2 \neq 0$ (有差异)

将设置 $\alpha = .05$

10.4.2.2 定位临界区域

这是一个独立测量设计。这些数据的 t 统计量的自由度由以下确定

```
dtn = combine(groupby(dt, :房间类型), nrow => :人数)
n₁, n₂ = dtn.人数
df₁,df₂ = n₁-1, n₂-1
df = df₁ + df₂
```

14

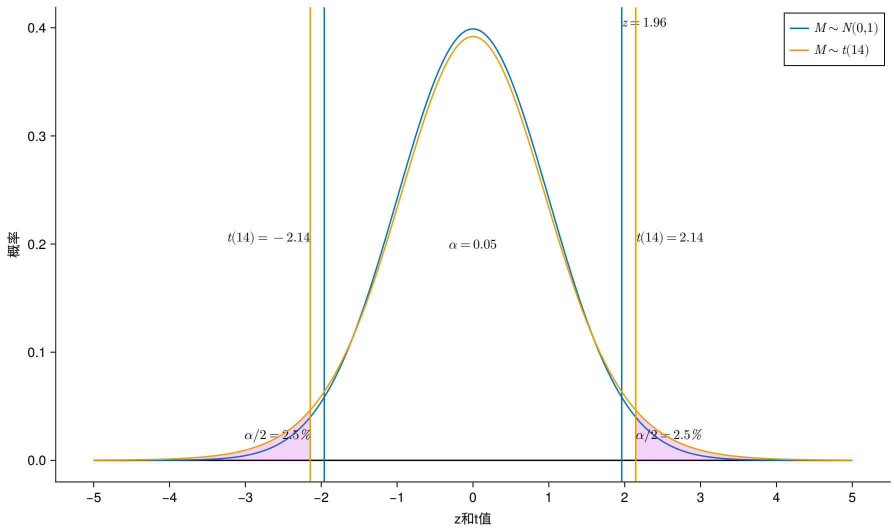
- 对于 $\alpha = .05$，临界区域包括分布中的极端 5%，边界为

```
quantile(TDist(df), 0.05/2),
cquantile(TDist(df), 0.05/2)
```

(-2.1447866879178035, 2.1447866879178035)

```
PlotT(14, showNormal = true, showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



10.4.2.3 获取数据并计算 t 统计量

- 找到两个样本的混合方差

方法一

```
transform!(groupby(dt, :房间类型), :分数 => mean => :Mj)
transform!(dt, :分数 => mean => :Mij)
```

房间类型	分数	Mj	Mij
亮房间	11	8.0	10.0
亮房间	9	8.0	10.0
亮房间	4	8.0	10.0
亮房间	5	8.0	10.0
亮房间	6	8.0	10.0
亮房间	7	8.0	10.0
亮房间	12	8.0	10.0
亮房间	10	8.0	10.0
暗房间	7	12.0	10.0
暗房间	13	12.0	10.0
暗房间	14	12.0	10.0
暗房间	16	12.0	10.0
暗房间	9	12.0	10.0
暗房间	11	12.0	10.0
暗房间	15	12.0	10.0
暗房间	11	12.0	10.0

```
X, Mj = dt.分数, dt.Mj
SSE  = sum((X - Mj).^2)
```

```
sp2 = SSE / df
```

```
9.0
```

方法二

```
SS₁ = sum(亮房间 .^ 2) - (sum(亮房间))^2 / n₁
SS₂ = sum(暗房间 .^ 2) - (sum(暗房间))^2 / n₂
sp2 = (SS₁ + SS₂) / (df₁ + df₂)
```

```
9.0
```

- 使用混合方差计算估计标准误

```
SE = sqrt(sp2 / n₁ + sp2 / n₂)
```

```
1.5
```

- 计算 t 统计量

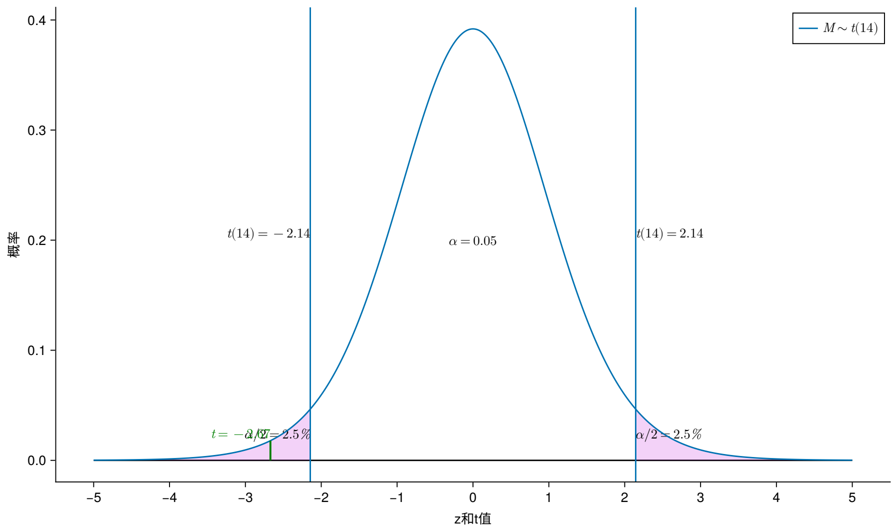
```
M₁, M₂ = mean(亮房间),  mean(暗房间)
t = (M₁ - M₂) / SE
```

```
-2.6666666666666665
```

10.4.2.4 做出决策

```
PlotT(14, TVal = round(t, digits = 2), showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



获得的值（$t = -2.67$）位于临界区域内。因此，我们拒绝 H_0，并得出在昏暗房间报告的分数与明亮房间的分数之间存在显著差异的结论。具体来说，在昏暗的房间里，学生报告的分数明显高于明亮房间里的分数。

如果进一步假设光线的明亮程度不影响被试在这些题目中的实际表现。该结果说明个体在昏暗环境中确实更容易撒谎。

10.4.2.5 用 Julia 语言进行统计

```
TT = EqualVarianceTTest(亮房间, 暗房间)
TT.t, pvalue(TT), confint(TT)
```

```
(-2.6666666666666665, 0.01841953514292608, (-7.217180031876705, -0.7828199681232948))
```

```
f  = @formula(分数 ~ 1 + 房间类型)
s  = schema(f, dt, Dict(:房间 => DummyCoding(base = "暗房间")))
fs = apply_schema(f, s)
resp, pred = modelcols(fs, dt)
pred
```

```
16×2 Matrix{Float64}:
 1.0  0.0
 1.0  0.0
 1.0  0.0
 1.0  0.0
 1.0  0.0
 1.0  0.0
 1.0  0.0
```

```
1.0  0.0
1.0  1.0
1.0  1.0
1.0  1.0
1.0  1.0
1.0  1.0
1.0  1.0
1.0  1.0
1.0  1.0
```

```
fm1 = fit(LinearModel, pred, resp)
DataFrame(coeftable(fm1))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
x1	8.0	1.06066	7.54247	2.69949e-6	5.72511	10.2749
x2	4.0	1.5	2.66667	0.0184195	0.78282	7.21718

```
fm2 = fit(LinearModel, f, dt)
DataFrame(coeftable(fm2))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	8.0	1.06066	7.54247	2.69949e-6	5.72511	10.2749
房间类型: 暗房间	4.0	1.5	2.66667	0.0184195	0.78282	7.21718

```
fm3 = lm(f, dt)
DataFrame(coeftable(fm3))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	8.0	1.06066	7.54247	2.69949e-6	5.72511	10.2749
房间类型: 暗房间	4.0	1.5	2.66667	0.0184195	0.78282	7.21718

```
ft = ftest(fm3.model)
sqrt(ft.fstat)
```

```
2.6666666666666665
```

10.4.3 定向假设和单侧检验

10.4.3.1 提出假设并选择 α 水平

- 零假设:

$$H_0 : \mu_{\text{昏暗环境}} \leq \mu_{\text{充分照明环境}}$$

（报告的分数在昏暗环境中不更高）

- 备择假设：

$$H_1 : \mu_{昏暗环境} > \mu_{充分照明环境}$$

（报告的分数在昏暗环境中更高）

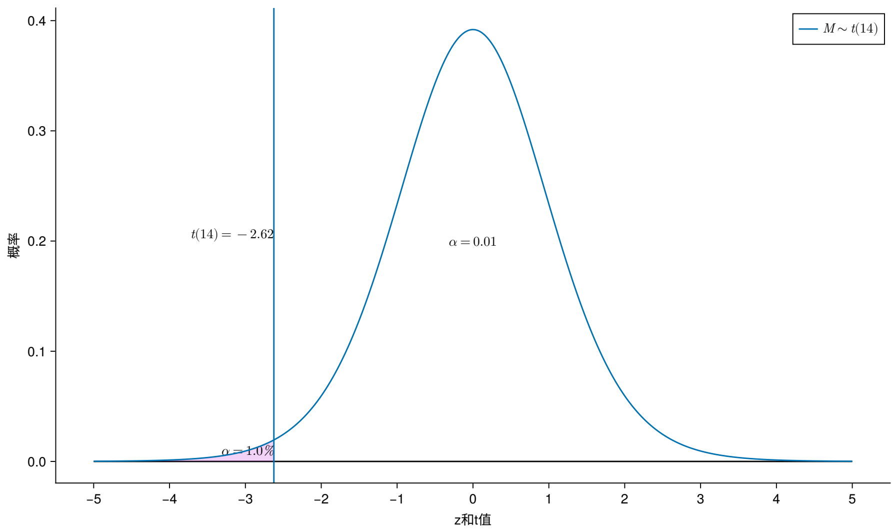
- 设定 $\alpha = .01$

10.4.3.2 定位临界区域

查看数据，确定样本均值差异是否朝着预测的方向。如果差异符合预测的方向，那么第二步就是确定差异是否足够显著。

```
PlotT(14, Tail = "Left", PVal = 0.01, showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



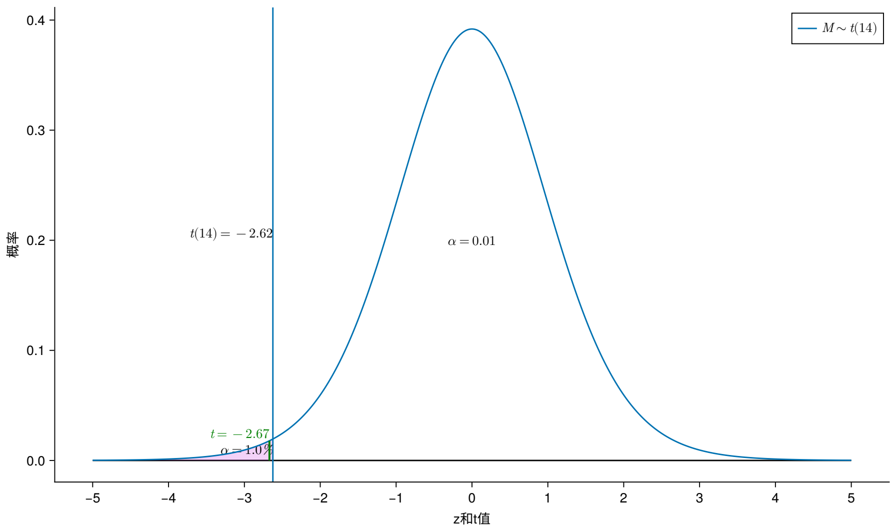
10.4.3.3 剩下的两个步骤

- 收集数据并计算检验统计量
- 做出决策

10.4.3.4 定位临界区域

```
PlotT(14, TVal = round(t, digits = 2), PVal = 0.01, Tail = "Left", showFalseAlarm =
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



10.4.3.5 报告结果

在研究报告中，单尾检验将明确标注为：

Reported scores were significantly higher for students in the dimly lit room, t(14) = -2.67, p < .01, one tailed.

在昏暗环境中，学生的报告分数显著更高， $t(14) = -2.67, p < .01$，*单尾。*

10.4.3.6 在 Julia 计算

```
pvalue(TT, tail = :left)
```

```
0.00920976757146304
```

10.4.4 独立样本 t 公式的基础假设

10.4.4.1 基本假设

- 每个样本内的观察值必须相互独立。
- 从样本中选取的两个总体必须服从正态分布。
- 从样本中选取的两个总体的方差必须相等。**方差齐性**（homogeneity of variance）

10.4.4.2 方差齐性

t 统计公式中的汇总方差是将两个样本方差平均在一起得到的。只有当这两个值都在估计相同的总体方差时，才有将它们平均在一起的意义，也就是方差齐性的假设成立。如果这两个样本方差在估计不同的总体方差，那么平均值就没有意义。

当样本大小之间存在较大差异时，方差齐性假设变得最为重要。在样本大小相等（或几乎相等）的情况下，这个假设就不太关键了，但仍然重要。违反方差齐性假设可能会使独立样本实验数据失去任何有意义的解释。

10.4.4.3 Hartley 的 F-Max 检验

如何知道方差齐性假设是否成立？一个简单的检验涉及查看两个样本方差。逻辑上，如果两个总体方差相等，那么两个样本方差应该非常相似。当两个样本方差相差不大时，可以相对有信心地认为方差齐性假设成立，并继续进行检验。但是，如果一个样本方差比另一个大三四倍以上，那就有理由担忧了。

一种更客观的方法涉及使用统计检验来评估方差齐性假设。尽管有许多不同的统计方法可用于确定方差齐性假设是否成立，但 Hartley 的 F-Max 检验是计算和理解最简单的方法之一。

另一个优点是，这个检验也可以用于检验超过两个独立样本的方差齐性。F-max 检验基于一个原则，即样本方差提供总体方差的无偏估计。此检验的零假设表明总体方差相等，因此样本方差应该非常相似。使用 F-max 检验的步骤如下：

- 计算每个单独样本的样本方差，$s^2 = \frac{SS}{df}$。
- 选择最大和最小的样本方差，计算

$$F\text{-max} = \frac{s^2(\text{最大值})}{s^2(\text{最小值})}$$

- 用样本数据计算的 F-max 值与临界值进行比较。

TABLE B.3 Critical Values for the *F*-Max Statistic*

*The critical values for α = .05 are in lightface type, and for α = .01, they are in boldface type.

	k = Number of Samples										
n − 1	2	3	4	5	6	7	8	9	10	11	12
4	9.60 **23.2**	15.5 **37.**	20.6 **49.**	25.2 **59.**	29.5 **69.**	33.6 **79.**	37.5 **89.**	41.4 **97.**	44.6 **106.**	48.0 **113.**	51.4 **120.**
5	7.15 **14.9**	10.8 **22.**	13.7 **28.**	16.3 **33.**	18.7 **38.**	20.8 **42.**	22.9 **46.**	24.7 **50.**	26.5 **54.**	28.2 **57.**	29.9 **60.**
6	5.82 **11.1**	8.38 **15.5**	10.4 **19.1**	12.1 **22.**	13.7 **25.**	15.0 **27.**	16.3 **30.**	17.5 **32.**	18.6 **34.**	19.7 **36.**	20.7 **37.**
7	4.99 **8.89**	6.94 **12.1**	8.44 **14.5**	9.70 **16.5**	10.8 **18.4**	11.8 **20.**	12.7 **22.**	13.5 **23.**	14.3 **24.**	15.1 **26.**	15.8 **27.**
8	4.43 **7.50**	6.00 **9.9**	7.18 **11.7**	8.12 **13.2**	9.03 **14.5**	9.78 **15.8**	10.5 **16.9**	11.1 **17.9**	11.7 **18.9**	12.2 **19.8**	12.7 **21.**
9	4.03 **6.54**	5.34 **8.5**	6.31 **9.9**	7.11 **11.1**	7.80 **12.1**	8.41 **13.1**	8.95 **13.9**	9.45 **14.7**	9.91 **15.3**	10.3 **16.0**	10.7 **16.6**
10	3.72 **5.85**	4.85 **7.4**	5.67 **8.6**	6.34 **9.6**	6.92 **10.4**	7.42 **11.1**	7.87 **11.8**	8.28 **12.4**	8.66 **12.9**	9.01 **13.4**	9.34 **13.9**
12	3.28 **4.91**	4.16 **6.1**	4.79 **6.9**	5.30 **7.6**	5.72 **8.2**	6.09 **8.7**	6.42 **9.1**	6.72 **9.5**	7.00 **9.9**	7.25 **10.2**	7.48 **10.6**
15	2.86 **4.07**	3.54 **4.9**	4.01 **5.5**	4.37 **6.0**	4.68 **6.4**	4.95 **6.7**	5.19 **7.1**	5.40 **7.3**	5.59 **7.5**	5.77 **7.8**	5.93 **8.0**
20	2.46 **3.32**	2.95 **3.8**	3.29 **4.3**	3.54 **4.6**	3.76 **4.9**	3.94 **5.1**	4.10 **5.3**	4.24 **5.5**	4.37 **5.6**	4.49 **5.8**	4.59 **5.9**
30	2.07 **2.63**	2.40 **3.0**	2.61 **3.3**	2.78 **3.5**	2.91 **3.6**	3.02 **3.7**	3.12 **3.8**	3.21 **3.9**	3.29 **4.0**	3.36 **4.1**	3.39 **4.2**
60	1.67 **1.96**	1.85 **2.2**	1.96 **2.3**	2.04 **2.4**	2.11 **2.4**	2.17 **2.5**	2.22 **2.5**	2.26 **2.6**	2.30 **2.6**	2.33 **2.7**	2.36 **2.7**

Table 31 of E. Pearson and H.O. Hartley, Biometrika Tables for Statisticians, 2nd ed. New York: Cambridge University Press, 1958. Adapted and reprinted with permission of the Biometrika trustees.

图 10.1: F-Max 表

要查找表中的临界值，您需要知道：

- $k =$ 单独样本的数量。（对于独立样本 t 检验，$k = 2$。）
- $df = n - 1$ 每个样本方差的自由度。Hartley 检验假定所有样本的大小相同。
- α 水平。表格提供了 $\alpha = .05$ 和 $\alpha = .01$ 的临界值。通常，方差齐性检验会使用更大的 α 水平。

10.4.4.4 用 Julia 进行计算

```
var₁, var₂ = SS₁ / df₁, SS₂/df₂
FMax = max(var₁, var₂) / min(var₁, var₂)
1 - 2 * abs(cdf(FDist(df₁, df₂), FMax) - 0.5)
```

```
0.903174397811275
```

```
pvalue(VarianceFTest(亮房间, 暗房间))
```

```
0.9031743978112748
```

10.4.4.5 汇总方差的替代方法

使用汇总方差计算独立样本 t 统计量要求数据满足方差齐性假设。具体来说，从样本中选取的两个分布必须具有相等的方差。为了避免这一假设，许多统计学家建议使用一种不需要汇总方差或方差齐性假设的替代计算独立样本 t 统计量的公式。

替代程序包括以下两个步骤：

- 第一步是使用两个单独样本方差来计算标准误差，如方程

$$s_{(M_1-M_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- t 统计量的自由度值使用以下公式进行调整：

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1-1} + \frac{V_2^2}{n_2-1}}, \text{ 其中} V_1 = \frac{s_1^2}{n_1} \text{ 且} V_2 = \frac{s_2^2}{n_2}$$

自由度的调整会降低 df 的值，从而将临界区域的边界推得更远。使用上述方程手动计算自由度

```
V₁, V₂ = var₁ / n₁, var₂ / n₂
df_Adj = (V₁ + V₂)^2 / (V₁^2/df₁ + V₂^2/df₂)
```

13.968325791855204

使用 Julia 计算

```
UnequalVarianceTTest(亮房间, 暗房间).df
```

13.968325791855204

10.5 效应大小和置信区间

10.5.1 效应大小

尽管方差和样本大小都会影响假设检验，但只有方差会对 Cohen's d 和 r^2 等效应大小指标产生较大影响；较大的方差会导致较小的效应大小指标。而样本大小对 Cohen's d 的值没有影响，对 r^2 的值只有很小的影响。

10.5.1.1 效应大小：Cohen's *d*

在一般情况下，Cohen's d 被定义为

$$d = \frac{\text{均值差异}}{\text{标准差}} = \frac{\mu_1 - \mu_2}{\sigma}$$

在独立样本研究中，估算 Cohen's d 的公式变为

$$\text{估算}d = \frac{\text{估算的均值差异}}{\text{估算的标准差}} = \frac{M_1 - M_2}{\sqrt{s_p^2}}$$

具体计算过程如下

```
sp2 = (SS₁ + SS₂) / (df₁ + df₂)
d = (M₂ - M₁) / sqrt(sp2)
```

1.3333333333333333

10.5.1.2 效应大小：方差百分比

计算独立样本 t 的 r^2 与单一样本 t 的计算方式完全相同

$$r^2 = \frac{SSR}{SST} = \frac{t^2}{t^2 + df}$$

1. 基于 t 和 df 的计算

```
r_sqr = t ^ 2 / (t^2 + df)
```

0.3368421052631579

2. 基于 SSR 和 SSE 的计算

方法一、

```
总数据 = [亮房间; 暗房间]
M₀, M₁, M₂ = mean(总数据), mean(亮房间), mean(暗房间)
SST = sum((总数据 .- M₀).^2)
SSE = sum((亮房间 .- M₁).^2) + sum((暗房间 .- M₂).^2)
SSR = n₁ * (M₁ - M₀)^2 + n₂ * (M₂ - M₀)^2
SST, SSR, SSE
```

(190.0, 64.0, 126.0)

方法二、

```
SST = sum((dt.分数 - dt.Mij).^2)
SSR = sum((dt.Mj  - dt.Mij).^2)
SSE = sum((dt.分数 -  dt.Mj).^2)
SST, SSR, SSE
```

(190.0, 64.0, 126.0)

```
r_sqr = SSR / SST
```

```
0.3368421052631579
```

3. julia 的自动计算

```
r2(fm3)
```

```
0.33684210526315794
```

10.5.2 估计 $\mu_1 - \mu_2$ 的置信区间

10.5.2.1 置信区间

如前所述，置信区间可作为衡量和描述处理效应大小的替代方法。对于单一样本 t，我们使用单个样本均值 M 来估计单个总体均值。对于独立样本 t，我们使用样本均值差异 $M_1 - M_2$ 来估计总体均值差异$\mu_1 - \mu_2$。此时，置信区间实际上估计了两个总体或处理条件之间的总体均值差异的大小。

与单一样本 t 一样，第一步是解 t 方程求未知总体参数。对于独立样本 t 统计量可得：

$$
t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}
$$

$$
\Downarrow
$$

$$
\mu_1 - \mu_2 = (M_1 - M_2) \pm t \times s_{(M_1 - M_2)}
$$

在方程中，值 $M_1 - M_2$ 和 $s_{M_1 - M_2}$ 都是从样本数据中获得的。虽然 t 统计量的值未知，但可用 t 统计量的自由度和 t 分布表来估计 t 值。使用估算的 t 值和样本的已知值，可算出 $\mu_1 - \mu_2$ 值。

手动计算

```
qt = quantile(TDist(df), 0.05/2)
(M₁ - M₂) + qt * SE,
(M₁ - M₂) - qt * SE
```

```
(-7.217180031876705, -0.7828199681232948)
```

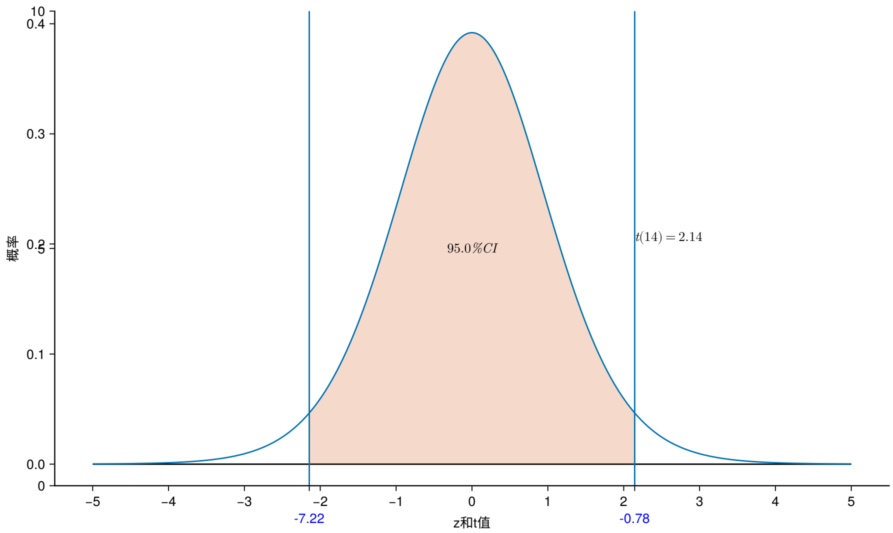
用 Julia 计算

```
confint(TT)
```

```
(-7.217180031876705, -0.7828199681232948)
```

```
PlotT(df, PVal = 0.05, M = M₁ - M₂, sM = SE, showCI = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



10.5.3 文献中的应用

10.5.3.1 报告结果

现在，我们使用 APA 格式来报告前面的例子，一个独立样本 t 检验的结果：

The students who were tested in a dimly lit room reported higher performance scores (M = 12, SD = 2.93) than the students who were tested in the well-lit room (M = 8, SD = 3.07). The mean difference was significant, t(14) = 2.67, p < .05, d = 1.33.

在昏暗的房间中受测的学生报告的表现分数（$M = 12, SD = 2.93$）比在充分照明的房间中受测的学生（M = 8, SD = 3.07）的分数更高。均值差异显著，$t(14) = 2.67, p < .05, d = 1.33$。

您应该注意，标准差不是独立样本 t 检验计算中的一步，但在为每个处理组提供描述性统计信息时很有用。在进行 t 检验时容易计算，因为您需要两个组的 SS 和 df 来确定汇总方差。请注意，报告 t 的格式与前面描述的格式完全相同，效应大小的度量是在假设检验结果之后立即报告的。此外，如果从计算机分析中获得了精确的概率，应该报告。

The difference was significant, t(14) = 2.67, p = .018, d = 1.33.

差异显著，$t(14) = 2.67, p = .018, d = 1.33$。

最后，如果报告置信区间以描述效应大小，它会立即出现在假设检验结果之后。

The difference was significant, t(14) = 2.67, p = .018, 95% CI [0.782, 7.218].

差异显著，$t(14) = 2.67, p = .018, 95$。

10.6 样本方差和样本大小的作用

10.6.0.1 独立样本 t 统计量的公式

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} = \frac{SS_1 + SS_2}{df_1 + df_2}$$

10.6.0.2 样本方差的作用

尽管方差和样本大小都会影响假设检验，但只有方差会对 Cohen's d 和 r^2 等效应大小指标产生较大影响；较大的方差会导致较小的效应大小指标。而样本大小对 Cohen's d 的值没有影响，对r^2 的值只有很小的影响。

以下示例提供了一个可视化演示，说明大样本方差如何可能掩盖样本之间的均值差异，并降低了在独立测量研究中拒绝 H_0 的可能性。

- 实验一

```
tr11 = [6, 7, 7, 8, 8, 8, 9, 9, 10]
tr12 = [11, 12, 12, 13, 13, 13, 14, 14, 15]
mean(tr11), mean(tr12)
```

```
(8.0, 13.0)
```

```
pvalue(EqualVarianceTTest(tr11, tr12))
```

```
1.9529832025352932e-7
```

```
n11, n12 = length(tr11), length(tr12)
df11, df12 = n11 - 1, n12 - 1
var11, var12 = var(tr11), var(tr12)
sp12 = (var11 * df11 + var12 * df12) / (df11 + df12)
```

```
SE1 = sqrt(sp12 / n11 + sp12 / n12)
var11, var12, SE1
```

```
(1.5, 1.5, 0.5773502691896257)
```

- 实验二

```
tr21 = [0, 1, 2, 4, 8, 11, 12, 16, 18]
tr22 = [3, 5, 9, 10, 13, 17, 19, 20, 21]
mean(tr21), mean(tr22)
```

```
(8.0, 13.0)
```

```
n21, n22 = length(tr21), length(tr22)
df21, df22, var21, var22 = n21 - 1, n22 - 1, var(tr21), var(tr22)
sp22 = (var21 * df21 + var22 * df22) / (df21 + df22)
SE2 = sqrt(sp22 / n21 + sp22 / n22)
var21, var22, SE2
```

```
(44.25, 44.25, 3.13581462037113)
```

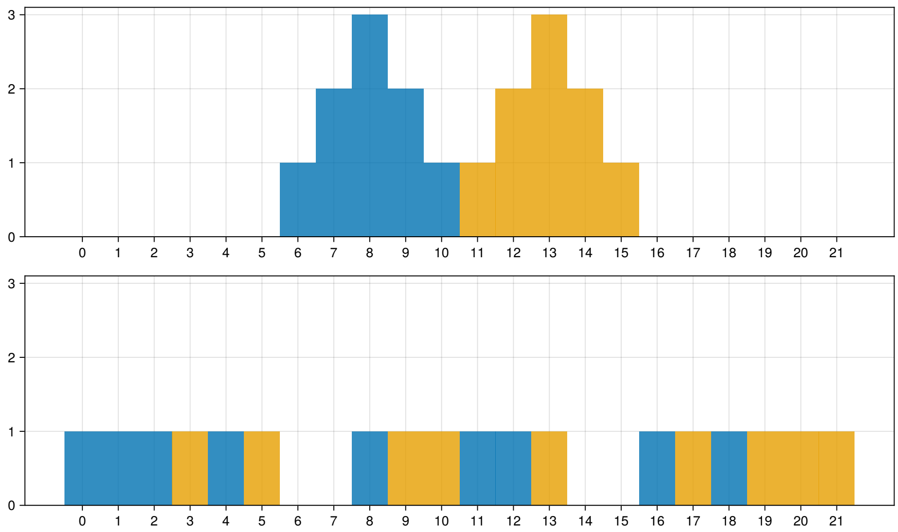
```
pvalue(EqualVarianceTTest(tr21, tr22))
```

```
0.13038853998413932
```

- 作图

```
fig = Figure(size = (1000, 600))
ax1, ax2 = [Axis(fig[i, 1],
  limits = (nothing, (0, 3.1)), xticks = 0:21) for i in 1:2]
[hist!(ax1, X, bins = -0.5:1:21.5) for X in [tr11, tr12]]
[hist!(ax2, X, bins = -0.5:1:21.5) for X in [tr21, tr22]]
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



10.6.0.3 Mann-Whitney 检验

与高方差相关的问题通常可以通过将原始分数转化为排名，然后进行一种称为曼-惠特尼检验的备选统计分析来最小化，该检验专门设计用于序数数据。`Mann-Whitney` 检验在附录 E 中介绍，该附录还讨论了将数值分数转换为排名的一般目的和过程。如果数据违反了独立样本 t 检验的假设之一，也可以使用 `Mann-Whitney` 检验。

第十一章　两组相关样本的 t 检验

11.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using DataFrames
using Distributions: TDist
using Distributions: quantile, cquantile
using Statistics: mean, std, var
using HypothesisTests: pvalue, OneSampleTTest, EqualVarianceTTest
using GLM, MixedModels
using JuliaLearn
using CairoMakie, StatsReIntro
```

11.2　重复测量设计简介

11.2.0.1　重复测量设计

重复测量设计，或者被试内设计，是一种在单个样本中为每个个体测量两次或更多次因变量的设计。重复测量研究的主要优势在于它在所有处理条件中都使用完全相同的个体。

11.2.0.2　匹配组设计

在匹配组研究中，一个样本中的每个个体与另一个样本中的个体进行匹配。匹配是这样进行的，以使这两个个体在研究人员想要控制的特定变量方面是等价的（或者几乎等价）。

11.3 重复测量设计的 t 统计量

11.3.0.1 差异分数

重复测量设计的 t 统计量在结构上与我们已经研究过的其他 t 统计量非常相似。它基本上与单样本 t 统计量相同。相关样本 t 的主要区别在于它是基于差异分数而不是原始分数（X 值）的。通常，差异分数（*difference scores*）或 D 值是通过将每个个体的第一个分数（操作前）减去第二个分数（操作后）来获得的：

$$\text{差异分数} = D = X_2 - X_1$$

差异分数的样本（D 值）作为假设检验的样本数据，所有计算都使用 D 分数进行。

11.3.0.2 相关样本检验的假设

- 零假设：

$$H_0 : \mu_D = 0$$

总体的平均差异为零。

- 备择假设：

$$H_1 : \mu_D \neq 0$$

存在一种操作效应，使得一个操作条件中的分数比另一条件中的分数明显更高（或更低）。

11.3.0.3 相关样本的 t 统计量

- t 统计量的一般形式：

$$t = \frac{\text{样本统计量} - \text{总体参数}}{\text{估计标准误差}}$$

- 单样本 t 统计量的公式：

$$t = \frac{M - \mu}{s_M}$$

- 重复测量设计的 t 统计量的公式：

$$t = \frac{M_D - \mu_D}{s_{M_D}}$$

11.3.0.4 相关样本的 t 统计量

- D 分数样本的方差（或标准差）：

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df}$$
$$s = \sqrt{\frac{SS}{df}}$$

- 单样本 t 统计量的估计标准误差：

$$s_{M_D} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

11.4 重复测量设计的假设检验

11.4.0.1 假设检验的四个步骤

- 提出假设并选择显著性水平。
- 找到临界区域。
- 计算 t 统计量。
- 做出决策。

11.4.0.2 例子：诅咒与疼痛

诅咒是对疼痛的一种常见、几乎是一种本能反应。无论是不小心撞到咖啡桌的桌角还是用锤子砸到拇指，我们大多数人都会用一串粗口来应对疼痛。然而，有一个问题，那就是诅咒是否会让人更加关注疼痛，从而增加疼痛的强度，还是作为一种减轻疼痛的分心方法。

在这项研究中，参与者被要求把一只手放入冰冷的水中，尽量忍受疼痛。一半的参与者被告知在手放入水中的时间内反复说出他们最喜欢的脏话。另一半则重复一个中性的词语。参与者评价了疼痛的强度，在短暂的休息后，两组交换了词语并重复了放入冰水的过程。因此，所有参与者都经历了两种条件（诅咒和中性），其中一半在第一次放入冰水时诅咒，另一半在第二次放入冰水时诅咒。

研究收集到的数据为，在一个从 1 到 10 的疼痛水平评分标尺上，对那些发誓或重复中性词的参与者进行评分。

```
被试号 = 'A':'I'
中性词 = [9, 8, 7, 7, 8, 9, 7, 7, 8]
诅咒词 = [7, 7, 3, 8, 6, 4, 6, 7, 4]
```

11.4.0.3 提出假设并选择显著性水平

- 零假设：

$$H_0 : \mu_D = 0$$

两种条件之间没有差异。

- 备择假设：

$$H_1 : \mu_D \neq 0$$

存在差异。

- 对于这个检验，我们使用 $\alpha = 0.05$。

11.4.0.4 找到临界区域

- 自由度

```
n = length(被试号)
df = n - 1
```

8

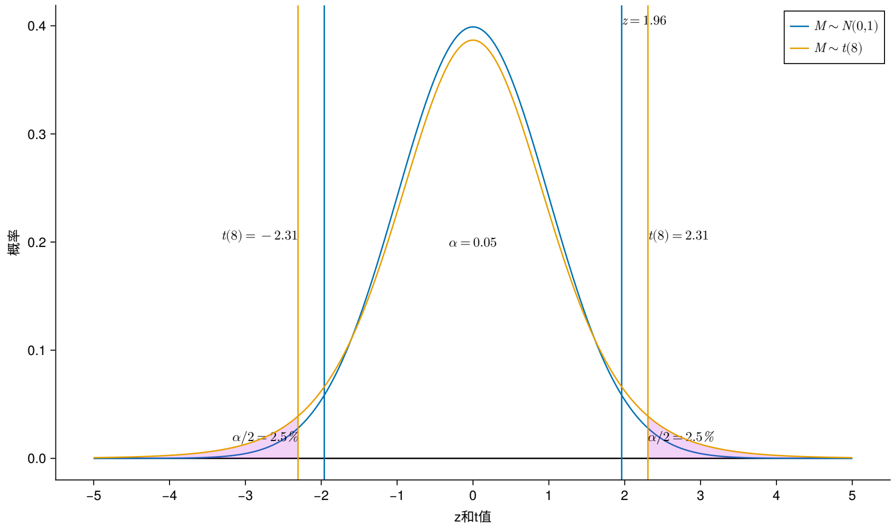
- 对于 $\alpha = 0.05$，t 分布中的临界值是

```
quantile(TDist(df), .05/2),
cquantile(TDist(df), .05/2)
```

(-2.3060041352041662, 2.3060041352041662)

```
PlotT(df, showNormal = true, showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



11.4.0.5 计算 t 统计量

- 计算差异分数 D

```
D = 诅咒词 - 中性词
```

```
9-element Vector{Int64}:
 -2
 -1
 -4
  1
 -2
 -5
 -1
  0
 -4
```

- 计算平方和

```
SS = sum((D .- mean(D)).^2)
SS = sum(D.^2) - sum(D)^2 / n
```

```
32.0
```

- 计算样本标准差

```
Var = var(D)
Var = SS / df
```

```
SD = sqrt(Var)
SD = std(D)
```

```
2.0
```

- 计算估计的标准误差

```
SE = SD / sqrt(n)
```

```
0.6666666666666666
```

- 计算 t 统计量的值

```
μ, MD = 0, mean(D)
t = (MD - μ) / SE
```

```
-3.0
```

11.4.0.6 用 Julia 计算

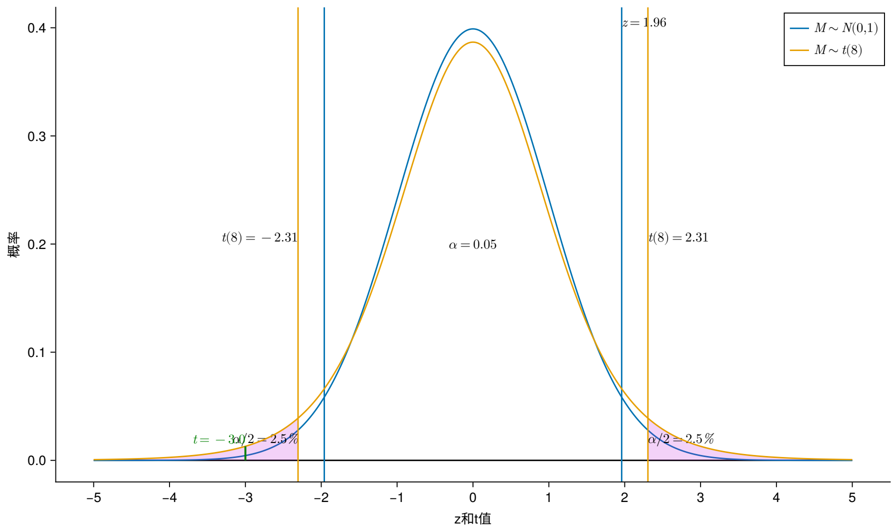
```
TT = OneSampleTTest(D)
TT.t, pvalue(TT)
```

```
(-3.0, 0.01707168123378264)
```

11.4.0.7 找到临界区域

```
PlotT(df, TVal = t, showNormal = true, showFalseAlarm = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



11.4.0.8 做出决策

我们得到的 t 值位于临界区域内。拒绝零假设，得出结论，即与重复中性词相比，诅咒对疼痛感知有显著影响。

11.4.0.9 方向性假设和单尾检验

- 零假设：

$$H_0 \ : \ \mu_D \geq 0(\text{诅咒不会降低疼痛})$$

- 备择假设：

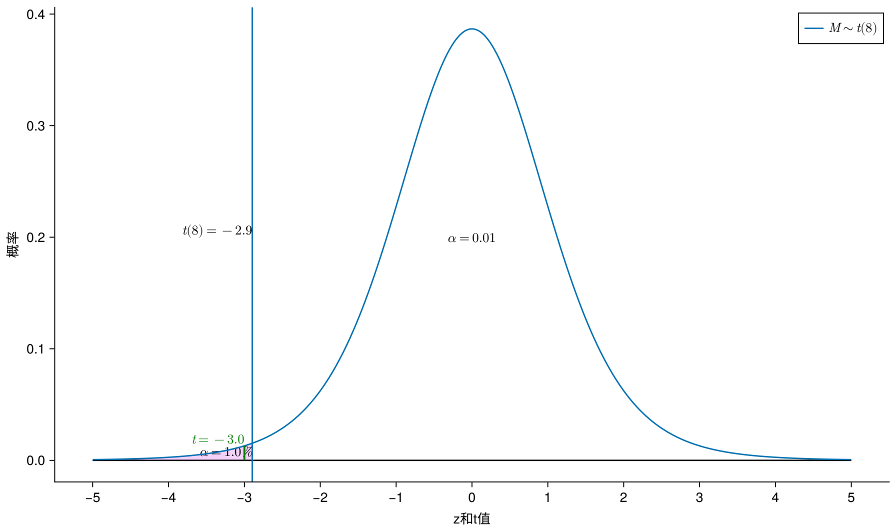
$$H_1 \ : \ \mu_D < 0(\text{评分下降})$$

- 对于这个检验，我们使用 $\alpha = 0.01$。

11.4.0.10 找到临界区域

```
PlotT(df, PVal = 0.01, TVal = t, showFalseAlarm = true, Tail = "Left")
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



11.4.0.11 相关样本 t 检验的假设

每个操作条件中的观测必须是独立的。差异分数（D 值）的总体分布必须是正态的。

如果有理由怀疑重复测量 t 检验的任何一个假设被违反，那么附录 E 中将介绍一种称为威尔科克森检验（Wilcoxon test）的备用分析方法。威尔科克森检验要求在评估两种操作条件之间的差异之前，将原始分数转换为秩次。

11.5 效应大小和置信区间

11.5.0.1 Cohen's d

- 对于重复测量研究，Cohen's d 定义为

$$d = \frac{\text{总体均值差异}}{\text{标准差}} = \frac{\mu_D}{\sigma_D}$$

- 估计的 Cohen's d 是

$$\text{估计的}d = \frac{\text{样本均值差异}}{\text{样本均值偏差}} = \frac{M_D}{s}$$

11.5.0.2 诅咒和疼痛：估计的 Cohen's d

```
d =  MD / SD
```

```
-1.0
```

11.5.0.3 方差占比

- 使用 SSR 和 SST 计算 r^2

```
SST = sum((D .- μ).^2)
SSE = sum((D .- MD).^2)
SSR = n * (MD - μ)^2
SST, SSE, SSR
```

```
(68, 32.0, 36.0)
```

```
SSR / SST
```

```
0.5294117647058824
```

- 使用 t 统计量计算 r^2

```
t^2 / (t^2 + df)
```

```
0.5294117647058824
```

11.5.0.4 估计 μ_D 的置信区间

- 对于重复测量实验，t 统计量定义为

$$t = \frac{M_D - \mu_D}{s_{M_D}}$$

- 解出总体 μ_D，我们得到

$$\mu_D = M_D \pm t \cdot S_{M_D}$$

- 对于给定的示例，我们得到

```
TL = quantile(TDist(df), 0.05/2)
TH = cquantile(TDist(df), 0.05/2)
```

```
MD + TL * SE,
MD + TH * SE
```

```
(-3.537336090136111, -0.46266390986388917)
```

11.5.0.5 做为 GLM 或 GLMM 的特殊形式

- 宽格式转换成长格式

```
dt = DataFrame(被试号 = 被试号, 中性词 = 中性词, 诅咒词 = 诅咒词)
dt = stack(dt, [:中性词, :诅咒词], variable_name = "类型", value_name = "分数")
sort!(dt, order(:类型, rev = true))
```

被试号	类型	分数
A	诅咒词	7
B	诅咒词	7
C	诅咒词	3
D	诅咒词	8
E	诅咒词	6
F	诅咒词	4
G	诅咒词	6
H	诅咒词	7
I	诅咒词	4
A	中性词	9
B	中性词	8
C	中性词	7
D	中性词	7
E	中性词	8
F	中性词	9
G	中性词	7
H	中性词	7
I	中性词	8

- 线性模型

```
form  = @formula(分数 ~ 1 + 类型 + 被试号)
contr = Dict(
    :类型   => DummyCoding(base = "诅咒词"),
    :被试号 => DummyCoding(base = 'A'))
mm = fit(LinearModel, form, dt; contrasts = contr)
DataFrame(coeftable(mm))
```

```
HM = HypothesisMatrix(c = 9, Interaction = false)
```

```
convert(Array{Float64}, Matrix(HM[:, 2:end]) * dt.分数)
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	7.0	1.05409	6.64078	0.000162352	4.56926	9.43074
类型: 中性词	2.0	0.666667	3.0	0.0170717	0.462664	3.53734
被试号: B	-0.5	1.41421	-0.353553	0.73281	-3.76118	2.76118
被试号: C	-3.0	1.41421	-2.12132	0.066688	-6.26118	0.261182
被试号: D	-0.5	1.41421	-0.353553	0.73281	-3.76118	2.76118
被试号: E	-1.0	1.41421	-0.707107	0.499576	-4.26118	2.26118
被试号: F	-1.5	1.41421	-1.06066	0.319813	-4.76118	1.76118
被试号: G	-1.5	1.41421	-1.06066	0.319813	-4.76118	1.76118
被试号: H	-1.0	1.41421	-0.707107	0.499576	-4.26118	2.26118
被试号: I	-2.0	1.41421	-1.41421	0.195016	-5.26118	1.26118

Coef	R1C1	R1C2	R1C3	R1C4	R1C5	R1C6	R1C7	
Intercept	5//9	1//18	1//18	1//18	1//18	1//18	1//18	...
R2	-1//9	-1//9	-1//9	-1//9	-1//9	-1//9	-1//9	...
C2	-1//2	1//2	0//1	0//1	0//1	0//1	0//1	...
C3	-1//2	0//1	1//2	0//1	0//1	0//1	0//1	...
C4	-1//2	0//1	0//1	1//2	0//1	0//1	0//1	...
C5	-1//2	0//1	0//1	0//1	1//2	0//1	0//1	...
C6	-1//2	0//1	0//1	0//1	0//1	1//2	0//1	...
C7	-1//2	0//1	0//1	0//1	0//1	0//1	1//2	...
C8	-1//2	0//1	0//1	0//1	0//1	0//1	0//1	...
C9	-1//2	0//1	0//1	0//1	0//1	0//1	0//1	...

```
10-element Vector{Float64}:
  7.0
  2.0
 -0.5
 -3.0
 -0.5
 -1.0
 -1.5
 -1.5
 -1.0
 -2.0
```

- 线性混合模型

```
form  = @formula(分数 ~ 类型 + (1 + 类型|被试号))
contr = Dict(:类型   => DummyCoding(base = "诅咒词"))
mm = fit(MixedModel, form, dt; contrasts = contr)
DataFrame(coeftable(mm))
```

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	5.77778	0.539271	10.7141	8.74396e-27
类型: 中性词	2.0	0.628547	3.18194	0.00146291

11.5.0.6 样本方差和样本大小

尽管方差和样本大小都会影响假设检验，但只有方差对效应大小的度量，如 Cohen's d 和 r^2 有较大的影响；较大的方差会产生较小的效应大小。另一方面，样本大小对 Cohen's d 的值没有影响，只对 r^2 有较小的影响。

11.5.0.7 报告重复测量 t 检验的结果

这项研究的发表报告可以总结如下：

Changing from a neutral word to a swear word reduced the perceived level of pain by an average of $M = 2.00$ *points with* $SD = 2.00$*. The treatment effect was statistically significant,* $t(8) = -3.00$*,* $p < .05$*,* $r^2 = 0.529$*.*

从中性词转变为脏话平均降低了感知疼痛水平 $M = 2.00$ *分，标准差* $SD = 2.00$。*操作效应在统计上是显著的，*$t(8) = -3.00$，$p < 0.05$，$r^2 = 0.529$。

当使用计算机程序进行假设检验时，打印输出通常包括显著性水平的精确概率。

⋯ *statistically significant,* $t(8) = -3.00$*,* $p = .017$*,* $r^2 = 0.529$

⋯ *统计上显著，*$t(8) = -3.00$，$p = 0.017$，$r^2 = 0.529$。

如果置信区间与假设检验的结果一起作为效应大小的描述报告，它将如下所示：

Changing from a neutral word to a swear word significantly reduced the perceived level of pain, $t(8) = -3.00$*,* $p < .05$*, 95% CI* $[-0.462, -3.538]$*.*

从中性词转变为脏话显著降低了感知疼痛水平，$t(8) = -3.00$，$p < 0.05$，95% *置信区间* $[-0.462, -3.538]$。

11.6 比较重复测量设计和独立测量设计

11.6.0.1 重复测量设计的优势

- 受试者数量
- 随时间的变化
- 个体差异。个体差异是指年龄、智商、性别和个性等特征，这些特征因个体而异。

11.6.0.2 重复测量和独立测量设计

- **顺序效应** (order effects)。由于参与早期操作的评分变化被称为顺序效应，可能会扭曲重复测量研究中找到的均值差异。

- **平衡**（Counterbalancing）。平衡的目的是控制顺序效应。

11.6.0.3 重复测量和独立测量设计

给定下面的数据

```
tr1 = [18, 27, 33]
tr2 = [15, 20, 28]
```

11.6.0.4 独立测量设计

```
BT = DataFrame(
    被试 = 'A':'F',
    条件 = 'C' .* string.(repeat(1:2, inner = 3)),
    数据 = [tr1; tr2]
)
```

被试	条件	数据
A	C1	18
B	C1	27
C	C1	33
D	C2	15
E	C2	20
F	C2	28

- t 检验

```
TT2 = EqualVarianceTTest(tr2, tr1)
TT2.t, pvalue(TT2)
```

```
(-0.8660254037844386, 0.435330942514376)
```

- 线性模型

```
fm = lm(@formula(数据 ~ 条件), BT)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	26.0	4.08248	6.36867	0.00311704	14.6652	37.3348
条件: C2	-5.0	5.7735	-0.866025	0.435331	-21.0298	11.0298

- 线性混合模型

```
form = @formula(数据 ~ 条件 + (1 |被试))
fm = fit(MixedModel, form, BT)
```

```
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	26.0	3.33333	7.8	6.19072e-15
条件: C2	-5.0	4.71405	-1.06066	0.288844

11.6.0.5 重复测量设计

```
WT = DataFrame(
    被试 = repeat('A':'C', outer = 2),
    条件 = 'C' .* string.(repeat(1:2, inner = 3)),
    数据 = [tr1; tr2]
)
```

被试	条件	数据
A	C1	18
B	C1	27
C	C1	33
A	C2	15
B	C2	20
C	C2	28

- t 检验

```
D2 = tr2 - tr1
TT3 = OneSampleTTest(D2)
TT3.t, pvalue(TT3)
```

```
(-4.330127018922193, 0.04941362421328312)
```

- 线性模型

```
form = @formula(数据 ~ 条件)
fm = fit(LinearModel, form, WT)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	26.0	4.08248	6.36867	0.00311704	14.6652	37.3348
条件: C2	-5.0	5.7735	-0.866025	0.435331	-21.0298	11.0298

```
form = @formula(数据 ~ 条件 + 被试)
fm = fit(LinearModel, form, WT)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	19.0	1.1547	16.4545	0.00367311	14.0317	23.9683
条件: C2	-5.0	1.1547	-4.33013	0.0494136	-9.96828	-0.0317246
被试: B	7.0	1.41421	4.94975	0.0384761	0.91513	13.0849
被试: C	14.0	1.41421	9.89949	0.0100505	7.91513	20.0849

- 线性混合模型

```
form = @formula(数据 ~ 条件 + (1 |被试))
fm = fit(MixedModel, form, WT)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	z	Pr(>|z|)
(Intercept)	26.0	3.33333	7.8	6.19072e-15
条件: C2	-5.0	0.942809	-5.3033	1.13727e-7

第四部分

方差分析

第十二章　方差分析介绍

12.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using StatsReIntro
using DataFrames
using CairoMakie
using Statistics: mean, std
using Distributions: FDist, pdf, cquantile, cdf, ccdf
using HypothesisTests: OneWayANOVATest, pvalue, teststatistic
using GLM
using SimpleANOVA
```

12.2　方差分析概述

12.2.1　芯片制造的良品率

如果某个芯片制造涉及 100 道工序。其中每道工序的良品率均为 95%。那么该芯片制造的总良品率是多少?

12.2.2　类型 I 错误和多重比较

如果我们进行一次假设检验，犯错的概率是多少?

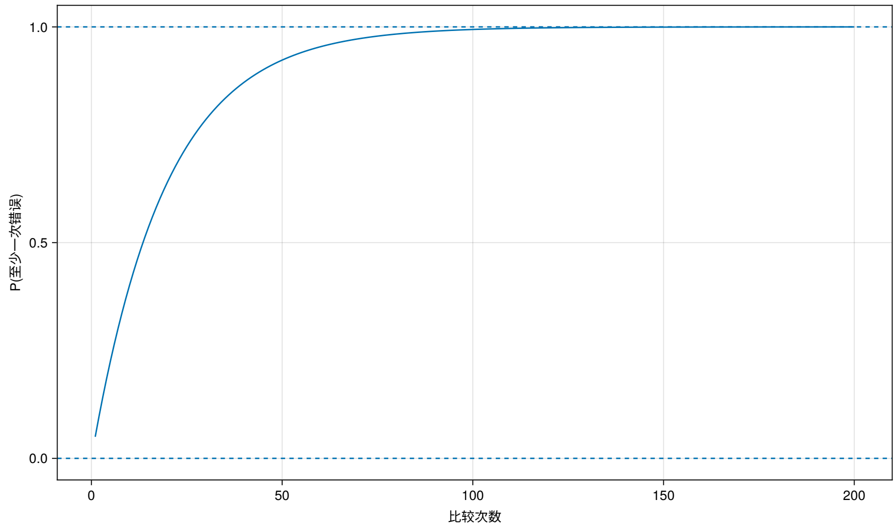
- P（犯错）$= \alpha$
- P（不犯错）$= 1 - \alpha$

如果我们进行 m 次假设检验，至少出现一次犯错的概率是多少?

- P（m 次检验不犯错）$= (1-\alpha)^m$
- P（至少有一次犯错在 m 次检验中）$= 1 - (1-\alpha)^m$

```
fwe(m; a = .05) = 1 - (1-a)^m
lines(1:200, fwe,
    axis = (; xlabel = "比较次数", ylabel = "P(至少一次错误)"))
hlines!([0, 1], linestyle = :dash)
current_figure()
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



12.2.3 类型 I 错误和多重假设检验

单个检验的 α 水平（testwise alpha level）是针对单个假设检验的类型 I 错误风险，或者 α 水平。

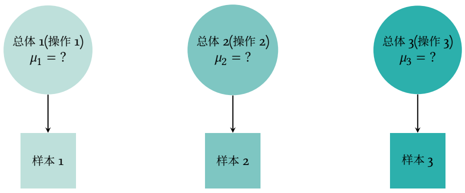
当实验涉及多个不同的假设检验时，**实验级别的 α 水平**（experimentwise alpha level）或**整体 α 水平**（familywise alpha level）是从实验中的所有单个检验积累的类型 I 错误的总概率。

通常情况下，实验级 α 水平远大于用于任何一个单独检验的 α 值。

12.2.4 方差分析

方差分析（**ANOVA**）是一种用于评估*两个或多个处理*（或总体）均值差异的假设检验程序。

12.2.4.1 单因素设计（Single-factor designs）



12.2.4.2 两因素设计或因子设计（Two-factor design or a factorial design）

		时间		
		治疗前	治疗后	6 个月后
治疗	疗法 1	样本 11	样本 12	样本 13
	疗法 2	样本 21	样本 22	样本 23

12.2.4.3 方差分析的术语

在方差分析中，指定要比较的组的变量（独立或准独立）称为**因子**（factor）。

构成因子的个别条件或值称为因子的**水平**（levels）。

12.2.4.4 单因子 ANOVA 的统计假设

- *零假设*（H_0）：不同处理的均值都相同。

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots$$

- *备择假设*（H_1）：总体之间至少存在一个均值差异。

12.3 方差分析的逻辑

12.3.0.1 方差分析的检验统计量：F 比率

- t 检验统计量

$$t = \frac{\text{获得的两个样本均值之间的差异}}{\text{标准误差（没有处理效应时预期的差异）}}$$

- 方差分析的检验统计量：F 比率

$$F = \frac{\text{样本均值之间的方差（差异）}}{\text{没有处理效应时预期的方差（差异）}}$$

- 方差分析的检验统计量：F 比率

$$\begin{aligned} F &= \frac{\text{组间方差}}{\text{组内方差}} \\ &= \frac{\text{包括任何处理效应的差异}}{\text{没有处理效应的差异}} \\ &= \frac{\text{系统性处理效应} + \text{随机非系统性差异}}{\text{随机非系统性差异}} \end{aligned}$$

12.3.1 F 比率：方差分析的检验统计量

对于方差分析，F 比率的分母被称为**误差项**（error term）。误差项提供了由随机非系统性差异引起的方差的度量。

当处理效应为零（H_0 成立）时，误差项度量的是 F 比率的分子中与之相同的方差来源，因此 F 比率的值预计会接近 1.00。

12.4 方差分析的符号和公式

12.4.1 假设的单因素方差分析

	操作 1	⋯	操作 j	⋯	操作 k	
	X_{11}	⋯	X_{j1}	⋯	X_{ki}	
	⋮	⋱	⋮	⋱	⋮	
	X_{1i}	⋯	X_{ji}	⋯	X_{ki}	
	⋮	⋱	⋮	⋱	⋮	
	X_{1n_1}	⋯	X_{jn_j}	⋯	X_{kn_k}	
总和	T_1	⋯	T_j	⋯	T_k	G
个数	n_1	⋯	n_j	⋯	n_k	N
均值	$\overline{X}_{1\cdot}$	⋯	$\overline{X}_{j\cdot}$	⋯	$\overline{X}_{k\cdot}$	$\overline{X}_{\cdot\cdot}$

12.4.2 方差的分解

$$
\begin{aligned}
SS_{\text{总体}} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2 - \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij})^2}{N} \\
&= \sum X^2 - \frac{G^2}{N} \\
SS_{\text{组间}} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_{j.}-\overline{X}_{..})^2 = \sum_{j=1}^{k} n_j(\overline{X}_{j.}-\overline{X}_{..})^2 \\
&= \sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{G^2}{N} \\
SS_{\text{组内}} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_{j.})^2 = \sum_{j=1}^{k} SSE_j
\end{aligned}
$$

$$
\begin{array}{ccccc}
\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_{..})^2 & = & \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_{j.}-\overline{X}_{..})^2 & + & \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji}-\overline{X}_{j.})^2 \\
\Downarrow & & \Downarrow & & \Downarrow \\
\sum X^2 - \frac{G^2}{N} & = & \sum_{j=1}^{k} n_j(\overline{X}_{j.}-\overline{X}_{..})^2 & + & \sum_{j=1}^{k} SSE_j \\
\Downarrow & & \Downarrow & & \Downarrow \\
SS_{\text{总体}} & = & SS_{\text{组间}} & + & SS_{\text{组内}}
\end{array}
$$

12.4.3 自由度的分解

$$
\begin{array}{ccccc}
\sum_{j=1}^{k} n_j - 1 & = & k-1 & + & \sum_{j=1}^{k}(n_j-1) \\
\Downarrow & & \Downarrow & & \Downarrow \\
N-1 & = & k-1 & + & N-k \\
\Downarrow & & \Downarrow & & \Downarrow \\
df_{\text{总体}} & = & df_{\text{组间}} & + & df_{\text{组内}}
\end{array}
$$

12.4.4 均方差（MS）和 F 比率

- 均方差（MS，Mean sum of squares）

– 组间方差

$$MS_{组间} = \frac{SS_{组间}}{df_{组间}}$$

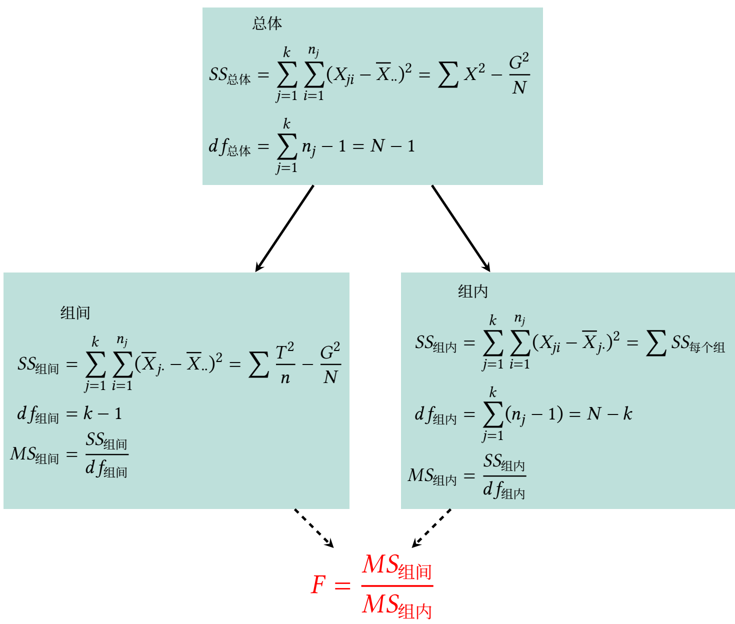
– 组内方差

$$MS_{组内} = \frac{SS_{组内}}{df_{组内}}$$

- F 比率是比较这两个方差

$$F = \frac{MS_{组间}}{MS_{组内}}$$

12.4.5 单因素设计



12.4.6 方差分析汇总表

来源	SS	df	MS	
组间	$SS_{组间}$	$df_{组间}$	$MS_{组间}$	$F = \frac{MS_{组间}}{MS_{组内}}$
组内	$SS_{组内}$	$df_{组内}$	$MS_{组内}$	
总体	$SS_{总体}$	$df_{总体}$		

12.4.7 一个示例：驾驶表现和手机使用

```
不打电话 = [4, 3, 6, 3, 4]
耳机电话 = [0, 1, 3, 1, 0]
手持电话 = [1, 2, 2, 0, 0]
```

```
dt = DataFrame(不打电话 = 不打电话, 耳机电话 = 耳机电话, 手持电话 = 手持电话)
dt = stack(dt, [:不打电话, :耳机电话, :手持电话],
    variable_name = "条件", value_name = "分数")
```

```
SSf(vec) = sum((vec).^2) - sum(vec)^2/length(vec)
nms(vec) = (n = length(vec), M = mean(vec), SS = SSf(vec))
combine(groupby(dt, :条件), :分数 => (nms) => AsTable)
```

条件	n	M	SS
不打电话	5	4.0	6.0
耳机电话	5	1.0	6.0
手持电话	5	1.0	4.0

12.4.7.1 驾驶表现和手机使用：方差分解

```
transform!(groupby(dt, :条件), :分数 => (mean) => :Mj)
transform!(dt, :分数 => (mean) => :Mij)
```

```
X, Mj, Mij = dt.分数, dt.Mj, dt.Mij
SST = sum((X  - Mij).^2)
SSB = sum((Mj - Mij).^2)
SSW = sum((X  - Mj).^2)
SST, SSB, SSW
```

```
(46.0, 30.0, 16.0)
```

条件	分数	Mj	Mij
不打电话	4	4.0	2.0
不打电话	3	4.0	2.0
不打电话	6	4.0	2.0
不打电话	3	4.0	2.0
不打电话	4	4.0	2.0
耳机电话	0	1.0	2.0
耳机电话	1	1.0	2.0
耳机电话	3	1.0	2.0
耳机电话	1	1.0	2.0
耳机电话	0	1.0	2.0
手持电话	1	1.0	2.0
手持电话	2	1.0	2.0
手持电话	2	1.0	2.0
手持电话	0	1.0	2.0
手持电话	0	1.0	2.0

12.4.7.2 驾驶表现和手机使用：自由度分解

```
n = combine(groupby(dt, :条件), nrow).nrow
k = length(levels(dt.条件))
N = nrow(dt)
dft = N - 1
dfb = k - 1
dfw = sum(n .- 1)
dft, dfb, dfw
```

```
(14, 2, 12)
```

12.4.7.3 驾驶表现和手机使用：F 比率

```
MSB = SSB / dfb
MSW = SSW / dfw
F = MSB / MSW
```

```
11.25
```

12.4.7.4 驾驶表现和手机使用：方差分析汇总表

来源	SS	df	MS	
组间	30.0	2	15.0	$F = 11.25$
组内	16.0	12	1.33	
汇总	46.0	14		

12.5 方差分析的假设检验和效应大小

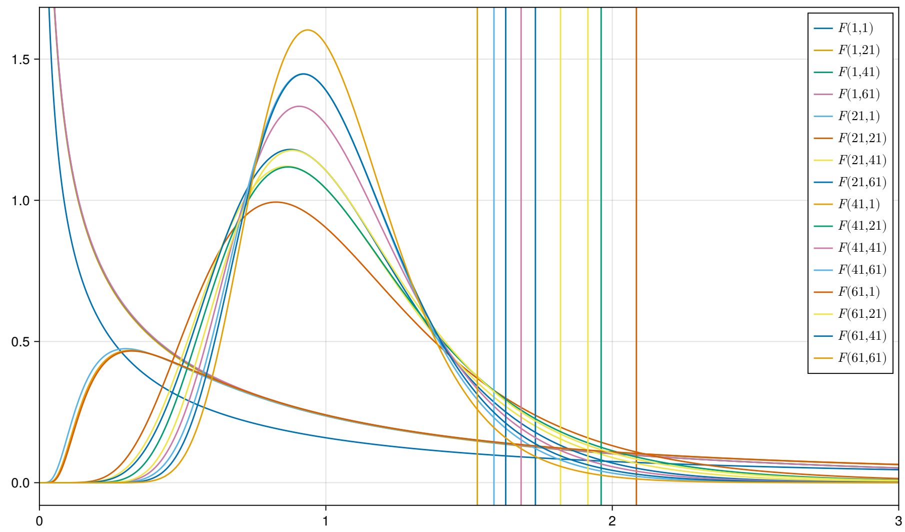
12.5.1 F 比率：方差分析的检验统计量

因为 F 比率是由两个方差（比率的分子和分母）计算得出的，所以 F 值始终是正数。请记住，方差总是*正*数。

当 H_0 成立时，F 比率的分子和分母测量相同的方差。在这种情况下，两个样本方差应该是差不多的大小，所以 F 比率应该*接近 1.00*。换句话说，F 比率的分布应该在 1.00 附近堆积。

```
fig = Figure()
ax = Axis(fig[1, 1], limits = ((0, 3), nothing))
for df1 in 1:20:61, df2 in 1:20:61
    dist = FDist(df1, df2)
    fval = quantile(dist, 0.05)
    lines!(ax, 0:.001:10, x -> pdf(dist, x),
        label = L"F(%$df1, %$df2)")
    vlines!(ax, [fval])
end
axislegend()
fig
```

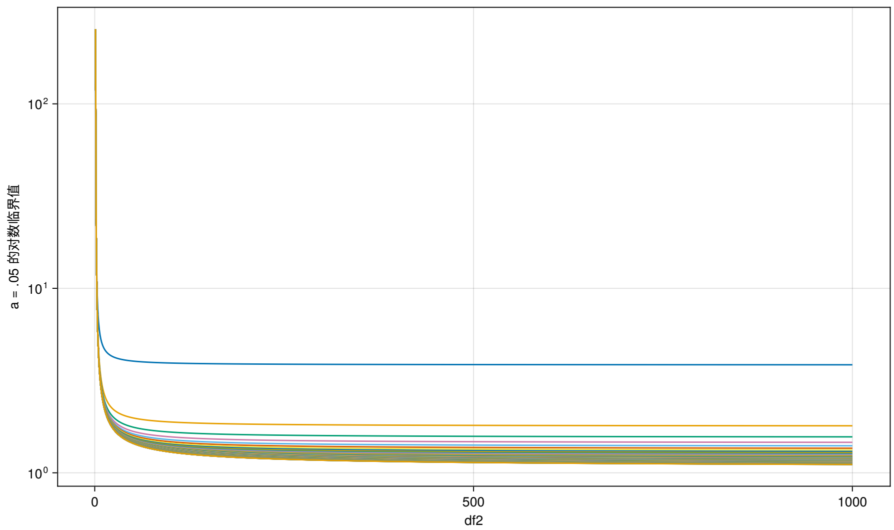
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



$\alpha = 0.05$ 情况下不同自由度 F 比率临界值的变化趋势图:

```
using Distributions
fig = Figure()
ax = Axis(fig[1, 1], yscale = log10,
    xlabel = "df2",
    ylabel = "a = .05 的对数临界值"
    )
for df1 in 1:10:1000
    fs = Float64[]
    for df2 in 1:1000
        fv = cquantile(FDist(df1, df2), 0.05)
        push!(fs, fv)
    end
    lines!(ax, fs)
end
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



12.5.1.1 方差分析表

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator					
	1	2	3	4	5	6
10	4.96	4.10	3.71	3.48	3.33	3.22
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**
11	4.84	3.98	3.59	3.36	3.20	3.09
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**
12	4.75	3.88	3.49	3.26	3.11	3.00
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**
13	4.67	3.80	3.41	3.18	3.02	2.92
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**
14	4.60	3.74	3.34	3.11	2.96	2.85
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**

图 12.1: 方差分析表

12.5.2 假设检验的一个示例

多年来，学生和老师们制定了各种策略来帮助准备即将到来的考试。但你怎么知道哪个最好？在这项研究中，学生阅读一篇材料，知道他们将在该材料上进行测试。在一种条件（RRR）下，参与者只是重新阅读即将测试的材料。在第二种条件（RPQ）下，学生回答与材料相关的理解问题，和在第三种条件（RCQ）下，学生提出并回答自己的问题。

```
RRR = [2, 3,  8,  6,  5,  6]
RPQ = [5, 9, 10, 13,  8,  9]
RCQ = [8, 6, 12, 11, 11, 12]
```

```
dt2 = DataFrame(RRR = RRR, RPQ = RPQ, RCQ = RCQ)
dt2 = stack(dt2, :,
    variable_name = "条件", value_name = "分数")
transform!(groupby(dt2, :条件), :分数 => (mean) => :Mj)
```

```
transform!(dt2, :分数 => (mean) => :Mij)
```

条件	分数	Mj	Mij
RRR	2	5.0	8.0
RRR	3	5.0	8.0
RRR	8	5.0	8.0
RRR	6	5.0	8.0
RRR	5	5.0	8.0
RRR	6	5.0	8.0
RPQ	5	9.0	8.0
RPQ	9	9.0	8.0
RPQ	10	9.0	8.0
RPQ	13	9.0	8.0
RPQ	8	9.0	8.0
RPQ	9	9.0	8.0
RCQ	8	10.0	8.0
RCQ	6	10.0	8.0
RCQ	12	10.0	8.0
RCQ	11	10.0	8.0
RCQ	11	10.0	8.0
RCQ	12	10.0	8.0

12.5.2.1 第 1 步：陈述假设并选择 α 水平

H_0 ：$\mu_1 = \mu_2 = \mu_3$（没有处理效应）。

H_1：至少有一个处理均值不同。

我们将使用 $\alpha = .05$。

12.5.2.2 第 2 步：定位临界区域

- 自由度

```
n = combine(groupby(dt2, :条件), nrow).nrow
k = length(levels(dt2.条件))
N = nrow(dt2)
dft = N - 1
dfb = k - 1
dfw = sum(n .- 1)
return dft, dfb, dfw
```

```
(17, 2, 15)
```

- 临界区域

```
cquantile(FDist(dfb, dfw), 0.05)
```

3.682320343673239

12.5.2.3 第 3 步：计算 F 比率

- 计算平方和

```
X, Mj, Mij = dt2.分数, dt2.Mj, dt2.Mij
SST = sum((X  - Mij).^2)
SSB = sum((Mj - Mij).^2)
SSW = sum((X  - Mj).^2)
SST, SSB, SSW
```

(172.0, 84.0, 88.0)

- 均方的计算

```
MSB = SSB / dfb
MSW = SSW / dfw
```

5.866666666666666

- 计算 F 统计量

```
F = MSB / MSW
```

7.159090909090909

12.5.2.4 第 4 步：作出决策

我们得到的 F 值，$F = 7.16$，位于临界区域。如果 H_0 成立，我们很不可能获得这么大的值（$p < .05$）。因此，我们拒绝 H_0，并得出结论：存在显著的处理效应。

12.5.2.5 分析汇总表

来源	SS	df	MS	
组间	84.0	2	42.0	$F = 7.16$
组内	88.0	15	5.87	
汇总	172.0	17		

12.5.2.6 方差分析的效应大小

- 方差解释的百分比

$$\eta^2 = \frac{SS_{\text{组间}}}{SS_{\text{总体}}}$$

- 我们示例的效应大小

```
η_sqr = SSB / SST
```

```
0.4883720930232558
```

12.5.2.7 用 Julia 进行方差分析

```
FF = OneWayANOVATest(RRR, RPQ, RCQ)
teststatistic(FF),
pvalue(FF)
```

```
(7.159090909090909, 0.0065638203998640765)
```

12.5.2.8 用 Julia 进行方差分析

```
anova(dt2, :分数, [:条件])
```

```
Analysis of Variance Results

Effect      SS  DF        MS        F           p        ω²
-----------------------------------------------------------
 Total   172.0  17
  条件    84.0   2  42.0      7.15909  0.00656382  0.406297
 Error    88.0  15   5.86667
```

12.5.2.9 用 Julia 进行方差分析

```
fmnull = lm(@formula(分数 ~ 1), dt2)
fmfull = lm(@formula(分数 ~ 1 + 条件), dt2)
GLM.ftest(fmnull.model, fmfull.model)
```

```
F-test: 2 models fitted on 18 observations
```

	DOF	ΔDOF	SSR	ΔSSR	R²	ΔR²	F*	p(>F)
[1]	2		172.0000		0.0000			
[2]	4	2	88.0000	-84.0000	0.4884	0.4884	7.1591	0.0066

```
chval = 2 * abs(
    loglikelihood(fmfull.model) -
    loglikelihood(fmnull.model))
ccdf(Chisq(2), chval)
```

0.0024020830969630562

```
fmnull = lm(@formula(分数 ~ 1), dt2)
fmfull = lm(@formula(分数 ~ 1 + 条件), dt2)
lrtest(fmnull.model, fmfull.model)
```

Likelihood-ratio test: 2 models fitted on 18 observations

	DOF	ΔDOF	LogLik	Deviance	Chisq	p(>Chisq)
[1]	2		-45.8550	172.0000		
[2]	4	2	-39.8236	88.0000	12.0628	0.0024

12.5.3 在文献中：报告结果

The means and standard deviations are presented in Table 1. The analysis of variance indicates that there are significant differences among the three strategies for studying, $F(2, 15) = 7.16$, $p < .05$, $\eta^2 = 0.488$

表 1 中提供了均值和标准差。方差分析表明，三种学习策略之间存在显著差异， $F(2, 15) = 7.16$，$p < .05$，$\eta^2 = 0.488$。

表 1. 使用三种不同学习策略后的小测验分数。

```
msd(vec) = (M = mean(vec), SD = std(vec))
MM = combine(groupby(dt2, :条件), :分数 => (msd) => AsTable)
```

条件	M	SD
RRR	5.0	2.19089
RPQ	9.0	2.60768
RCQ	10.0	2.44949

12.6 后续检验

单因素方差分析的统计假设：（1）*零假设*（H_0）：不同处理的均值都相同。（2）*备择假设*（H_1）：各总体之间至少有一个均值不同。

当（1）拒绝 H_0 且（2）有三个或更多处理时（$k \geq 3$）。**后续检验**（或**事后检验**）是在方差分析之后进行的额外假设检验，用于确定哪些均值差异是显著的，哪些不是。

12.6.1 Tukey 的可靠显著差异（HSD）检验

Tukey 的测试允许你计算一个确定显著性所需的处理均值最小差异的*单个值*。这个值被称为*可靠显著差异*，或 *HSD*（Honestly Significant Difference），然后用于比较任意两个处理条件。如果均值差异超过了 Tukey 的 HSD，你就可以得出处理之间存在显著差异的结论。否则，你无法得出处理显著不同的结论。

Tukey 的 HSD 的公式是

$$HSD = q \cdot \sqrt{\frac{MS_{组内}}{n}}$$

其中 q 的值可以在 Table B.5（附录 B，第 656 页）中找到，$MS_{组内}$ 是从方差分析中得出的组内方差，n 是每个处理中分数的个数。Tukey 的测试要求所有处理的样本大小 n 相同。

要找到 q 的合适值，你必须知道整个实验中的处理数 k，$MS_{组内}$（F-比值中的误差项）的*自由度*，并且必须选择一个 *alpha 水平*（通常与用于方差分析的 α 相同）。

```
dfw, k
```

```
(15, 3)
```

```
q, n = 3.67, 6
HSD = q * sqrt(MSW / n)
```

```
3.628993126352144
```

```
diff(MM.M)
```

```
2-element Vector{Float64}:
 4.0
 1.0
```

12.6.2 Scheffé 检验

由于它使用了一种极端谨慎的方法来降低第 I 型错误的风险，Scheffé 检验具有所有可能的后续检验中风险最小的特点。**Scheffé 检验**（雪费检验）使用 F-比值来评估任意两个处理条件之间的差异的显著性。F-比值的分子只使用了你要比较的两个处理的 MS 组间，分母是用于整个方差分析的 MS 组内。Scheffé 检验的" 安全因素" 来自以下两个考虑因素：

- 尽管你只比较两种处理，但 Scheffé 检验使用原始实验中的 k 值来计算 MS 组间的自由度。因此，F-比值的分子的*自由度是*$k-1$。
- Scheffé 的 F-比值的临界值与用于评估来自整个方差分析的 F-比值的临界值相同。因此，Scheffé 要求每个后续检验都满足用于完整方差分析的相同标准。

12.6.2.1 方差分析汇总表

来源	SS	df	MS	
组间	$SS_{组间}$	$df_{组间}$	$MS_{组间}$	$F=\dfrac{MS_{组间}}{MS_{组内}}$
组内	$SS_{组内}$	$df_{组内}$	$MS_{组内}$	
总体	$SS_{总体}$	$df_{总体}$		

12.6.2.2 Scheffe 检验：条件 RPQ 和 RCQ

- 计算 $SS_{组间}$ RPQ 和 RCQ

```
dt2n = subset(dt2, :条件 => ByRow(!=("RRR")))
dt2n.Mij .= mean(dt2n.分数)
SSBn = sum((dt2n.Mj - dt2n.Mij).^2)
```

3.0

- 计算 $MS_{组间}$ RPQ 和 RCQ，和 F-比值

```
MSBn = SSBn / dfb
```

1.5

```
F = MSBn / MSW
```

0.2556818181818182

- 计算相应的 p 值

```
ccdf(FDist(dfb, dfw), F)
```

```
0.7776955227873097
```

12.6.3 Julia 和线性模型

```
contr = Dict(:条件 => DummyCoding(
        base = "RRR", levels = ["RRR", "RPQ", "RCQ"]))
fmfull = lm(@formula(分数 ~ 1 + 条件), dt2; contrasts = contr)
DataFrame(coeftable(fmfull))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	5.0	0.988826	5.0565	0.000141974	2.89237	7.10763
条件: RPQ	4.0	1.39841	2.86039	0.0119134	1.01936	6.98064
条件: RCQ	5.0	1.39841	3.57548	0.00276105	2.01936	7.98064

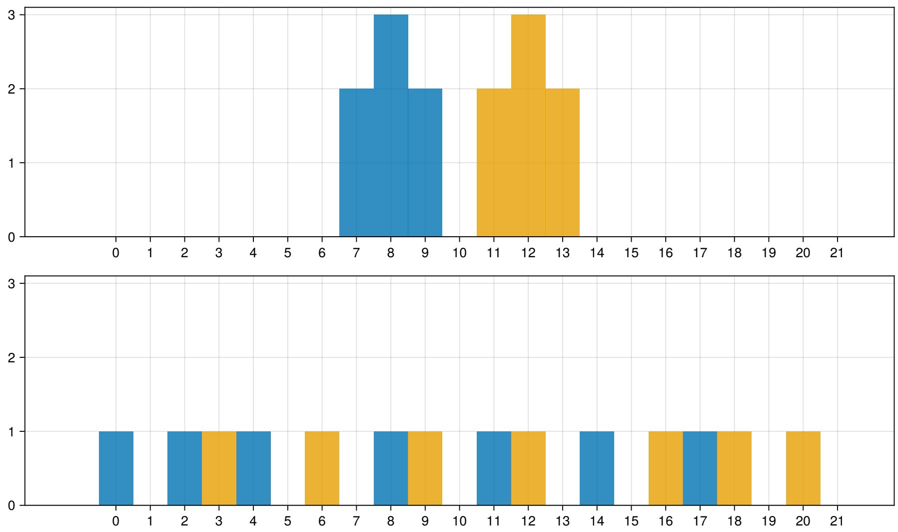
12.7 更多关于方差分析的内容

12.7.1 方差分析的概念视图

```
dt3 = DataFrame(
    TR = repeat(["T1", "T2"], inner = 7),
    CA = [8,  8, 7,  9, 8, 9,  7, 12, 13, 12, 11, 13, 12, 11],
    CB = [4, 11, 2, 17, 0, 8, 14, 12,  9, 20,  6, 16, 18,  3]
)
```

```
fig = Figure(size = (1000, 600))
ax1 = Axis(fig[1, 1], limits = (nothing, (0, 3.1)), xticks = 0:21)
ax2 = Axis(fig[2, 1], limits = (nothing, (0, 3.1)), xticks = 0:21)
[hist!(ax1, dt3[dt3.TR .== cl, "CA"], bins = -1.5:1:21.5) for cl in ["T1", "T2"]]
[hist!(ax2, dt3[dt3.TR .== cl, "CB"], bins = -1.5:1:21.5) for cl in ["T1", "T2"]]
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
OneWayANOVATest(dt3[dt3.TR .== "T1", "CA"], dt3[dt3.TR .== "T2", "CA"])
```

```
One-way analysis of variance (ANOVA) test
-----------------------------------------
Population details:
    parameter of interest:   Means
    value under h_0:         "all equal"
    point estimate:          NaN

Test summary:
    outcome with 95% confidence: reject h_0
    p-value:                     <1e-06

Details:
    number of observations: [7, 7]
    F statistic:            84.0
    degrees of freedom:     (1, 12)
```

```
anova(dt3, :CA, [:TR])
```

```
Analysis of Variance Results

Effect     SS  DF        MS     F            p       ω²
-------------------------------------------------------
 Total   64.0  13
    TR   56.0   1   56.0       84.0   9.10813e-7   0.85567
 Error    8.0  12    0.666667
```

```
OneWayANOVATest(dt3[dt3.TR .== "T1", "CB"], dt3[dt3.TR .== "T2", "CB"])
```

```
One-way analysis of variance (ANOVA) test
-----------------------------------------
Population details:
    parameter of interest:   Means
    value under h_0:         "all equal"
    point estimate:          NaN

Test summary:
    outcome with 95% confidence: fail to reject h_0
    p-value:                     0.2615

Details:
    number of observations: [7, 7]
    F statistic:            1.38843
    degrees of freedom:     (1, 12)
```

12.7.2 方差分析和 t 检验之间的关系

t 统计量比较的是*距离*：两个样本均值之间的距离（分子）以及基于标准误差计算出来的距离（分母）。而 *F-比*率比较的是 方差。方差度量的是距离的平方。因此可以预期二者有以下关系：$F = t^2$。换句话说，t 统计量是 F 比率的平方根。

当只有两个处理条件时，t 检验和方差分析检验的假设是相同的：

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 \\
H_1 &: \mu_1 \neq \mu_2
\end{aligned}
$$

t 统计量的自由度和 F-比值分母的自由度相同。如果考虑到 $F = t^2$ 的关系，t 统计量的分布和 F-比值的分布完全吻合。

12.7.2.1 证明 $F = t^2$

假定两个独立样本的样本容量和样本平均值分别为：n_1、M_1 和 n_2 和 M_2。两样本合并后的整体平均值 M_g 为两个样本平均值的加权平均值，即 $M_g = \frac{n_1 \times M_1 + n_2 \times M_2}{n_1 + n_2}$。

当只有两个组别（$k = 2$）时，F 统计量的组间方差和组间自由度可作如下运算：

$$
\begin{aligned}
SS_{组间} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{X}_{j.} - \overline{X}_{..})^2 \\
&= n_1 \times (M_1 - M_g)^2 + n_2 \times (M_2 - M_g)^2 \\
&= n_1 \times \left(M_1 - \frac{n_1 \times M_1 + n_2 \times M_2}{n_1 + n_2}\right)^2 + n_2 \times \left(M_2 - \frac{n_1 \times M_1 + n_2 \times M_2}{n_1 + n_2}\right)^2 \\
&= (M_1 - M_2)^2 \times \frac{n_1 \times n_2}{n_1 + n_2} \\
df_{组间} &= k - 1 = 1 \\
MS_{组间} &= \frac{SS_{组间}}{df_{组间}} = (M_1 - M_2)^2 \times \frac{n_1 \times n_2}{n_1 + n_2}
\end{aligned}
$$

F 统计量的组内方差和组内自由度可作如下运算:

$$
\begin{aligned}
SS_{组内} &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ji} - \overline{X}_{j.})^2 = \sum_{j=1}^{k} SSE_j = SS_1 + SS_2 \\
df_{组内} &= \sum_{j=1}^{k} df_j = \sum_{j=1}^{k}(n_j - 1) = df_1 + df_2 = (n_1 - 1) + (n_2 - 1) \\
s^2_{(M_1 - M_2)} &= s^2_p \times (\frac{1}{n_1} + \frac{1}{n_2}) = s^2_p \times \frac{n_1 + n_2}{n_1 \times n_2} \\
MS_{组内} &= \frac{SS_{组内}}{df_{组内}} = \frac{SS_1 + SS_2}{df_1 + df_2} = sp^2 = S^2_{M_1 - M_2} \times \frac{n_1 \times n_2}{n_1 + n_2}
\end{aligned}
$$

此时 F 统计量可做如下计算:

$$
\begin{aligned}
F &= \frac{MS_{组间}}{MS_{组内}} \\
&= \frac{(M_1 - M_2)^2 \times \frac{n_1 \times n_2}{n_1 + n_2}}{S^2_{M_1 - M_2} \times \frac{n_1 \times n_2}{n_1 + n_2}} = (\frac{M_1 - M_2}{S_{M_1 - M_2}})^2 \\
&= t^2
\end{aligned}
$$

第十三章　重复测量方差分析

13.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using DataFrames, StatsReIntro
using Statistics: mean, std
using Distributions: FDist, quantile, cquantile, cdf, ccdf
using CairoMakie
using SimpleANOVA
using GLM
using MixedModels
using HypothesisTests
```

13.2　重复测量 ANOVA 概述

13.2.0.1　单因素独立组设计：示例

给定下面三个组别的数据

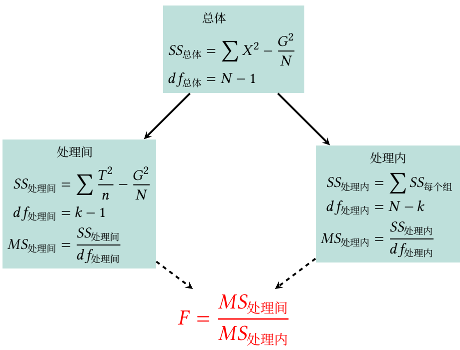
```
RRR = [2, 3,  8,  6,  5,  6]
APQ = [5, 9, 10, 13,  8,  9]
CAQ = [8, 6, 12, 11, 11, 12]
```

被试间设计的数据可以有如下形式:

```
数据 = [RRR; APQ; CAQ]
被试 = collect('A':'Z')[1:(length(数据))]
处理 = repeat(["RRR", "APQ", "CAQ"], inner = 6)
DataFrame(被试 = 被试, 处理 = 处理, 数据 = 数据)
```

其方差和自由度可做如下分解:

被试	处理	数据
A	RRR	2
B	RRR	3
C	RRR	8
D	RRR	6
E	RRR	5
F	RRR	6
G	APQ	5
H	APQ	9
I	APQ	10
J	APQ	13
K	APQ	8
L	APQ	9
M	CAQ	8
N	CAQ	6
O	CAQ	12
P	CAQ	11
Q	CAQ	11
R	CAQ	12



单因素独立组设计的 F 比率：

$$F = \frac{\text{处理间方差差异}}{\text{无处理效应时的方差}} = \frac{\text{处理间方差差异 (包括个体差异)}}{\text{无处理效应时的方差 (包括个体差异)}}$$

13.2.0.2 重复测量 ANOVA 的假设

- *零假设*表明对于总体来说，正在比较的处理条件之间没有均值差异。

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots$$

- *备择假设*表明处理条件之间存在均值差异。

$$H_1 \ : \text{至少有一个处理均值}\ (\mu)\ \text{与其他不同}$$

13.2.0.3 重复测量 *ANOVA* 示例

前述数据来自被试内设计时，具有以下形式

```
数据 = [RRR; APQ; CAQ]
被试 = repeat(collect('A':'Z')[1:(length(RRR))], 3)
处理 = repeat(["RRR", "APQ", "CAQ"], inner = 6)
dt = DataFrame(被试 = 被试, 处理 = 处理, 数据 = 数据)
```

被试	处理	数据
A	RRR	2
B	RRR	3
C	RRR	8
D	RRR	6
E	RRR	5
F	RRR	6
A	APQ	5
B	APQ	9
C	APQ	10
D	APQ	13
E	APQ	8
F	APQ	9
A	CAQ	8
B	CAQ	6
C	CAQ	12
D	CAQ	11
E	CAQ	11
F	CAQ	12

13.2.0.4 重复测量 ANOVA 的 F 比率

$$F = \frac{\text{处理间差异的方差}}{\text{没有处理效应的方差}} = \frac{\text{处理间差异的方差 (不包括个体差异)}}{\text{没有处理效应的方差 (排除个体差异)}}$$

重复测量 ANOVA 中，F-比率的分母称为**残差方差**（residual variance）或误差方差，它度量了如果没有系统性处理效应和个体差异对得分的变异性贡献，那么可以期望的方差有多少。

13.3 重复测量 ANOVA 的逻辑

13.3.0.1 重复 ANOVA 的方差分解

被试	处理 1	⋯	处理 j	⋯	处理 k	个数	求和	均值	SS
S_1	X_{11}	⋯	X_{j1}	⋯	X_{k1}	k	$P_{\cdot 1}$	$\overline{X}_{\cdot 1}$	$SS_{\cdot 1}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮	⋮
S_i	X_{1i}	⋯	X_{ji}	⋯	X_{ki}	k	$P_{\cdot i}$	$\overline{X}_{\cdot i}$	$SS_{\cdot i}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋮	⋮	⋮
S_n	X_{1n}	⋯	X_{jn}	⋯	X_{kn}	k	$P_{\cdot n}$	$\overline{X}_{\cdot n}$	$SS_{\cdot n}$
个数	n	⋯	n	⋯	n	N			
求和	$T_{1\cdot}$	⋯	$T_{j\cdot}$	⋯	$T_{k\cdot}$		G		
均值	$\overline{X}_{1\cdot}$	⋯	$\overline{X}_{j\cdot}$	⋯	$\overline{X}_{k\cdot}$			$\overline{X}_{\cdot\cdot}$	
SS	$SS_{1\cdot}$	⋯	$SS_{j\cdot}$	⋯	$SS_{k\cdot}$				$SS_{\cdot\cdot}$

13.3.0.2 重复 ANOVA 的方差分解

$$SS_{\text{总体}} = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ji}-\overline{X}_{\cdot\cdot})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}X_{ij}^2 - \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n}X_{ij})^2}{N} = \sum X^2 - \frac{G^2}{N}$$

$$SS_{\text{处理间}} = \sum_{j=1}^{k}\sum_{i=1}^{n}(\overline{X}_{j\cdot}-\overline{X}_{\cdot\cdot})^2 = \sum_{j=1}^{k} n(\overline{X}_{j\cdot}-\overline{X}_{\cdot\cdot})^2 = \sum_{j=1}^{k}\frac{T_j^2}{n} - \frac{G^2}{N}$$

$$SS_{\text{处理内}} = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ji}-\overline{X}_{j\cdot})^2 = \sum_{j=1}^{k} SS_{j\cdot}$$

13.3.0.3 重复 ANOVA 的方差分解

$$
\begin{aligned}
SS_{\text{总体}} &= \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ji}-\overline{X}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}X_{ij}^2 - \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n}X_{ij})^2}{N} = \sum X^2 - \frac{G^2}{N} \\
SS_{\text{被试间}} &= \sum_{j=1}^{k}\sum_{i=1}^{n}(\overline{X}_{\cdot i}-\overline{X}_{..})^2 = k\sum_{i=1}^{n}(\overline{X}_{\cdot i}-\overline{X}_{..})^2 = \sum_{i=1}^{n}\frac{P_i^2}{k} - \frac{G^2}{N} \\
SS_{\text{被试内}} &= \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ji}-\overline{X}_{\cdot i})^2 = \sum_{i=1}^{n}SS_{\cdot i}
\end{aligned}
$$

13.3.0.4 误差方差的计算

$$
\begin{aligned}
SS_{\text{误差}} &= SS_{\text{处理内}} - SS_{\text{被试间}} \\
&= SS_{\text{被试内}} - SS_{\text{处理间}}
\end{aligned}
$$

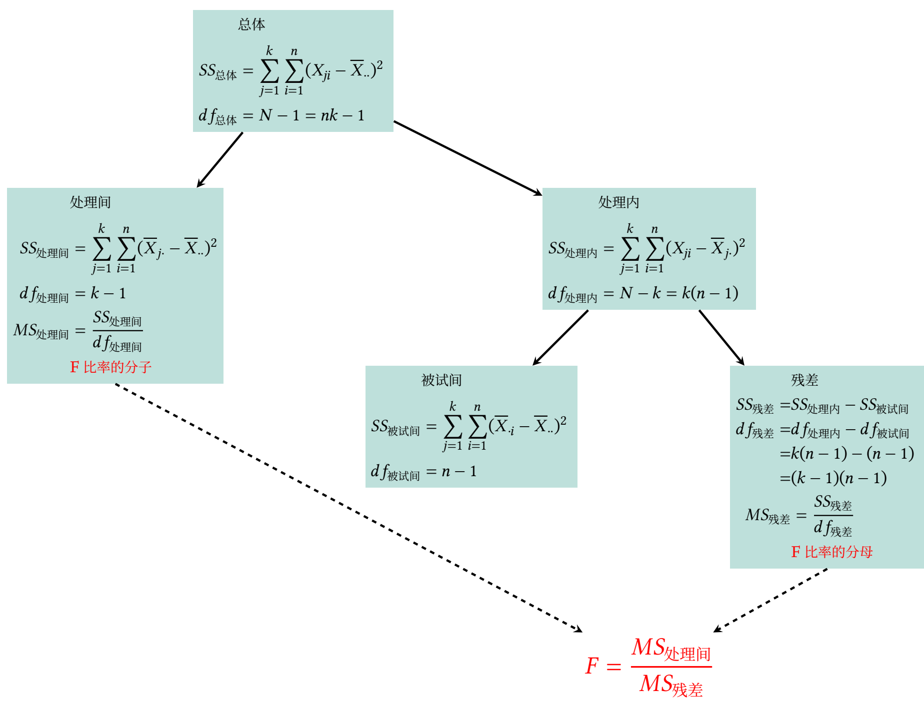
13.3.0.5 自由度的分配

$$
\begin{aligned}
df_{\text{总体}} &= N-1 = nk-1 \\
df_{\text{处理间}} &= k-1 \\
df_{\text{处理内}} &= k(n-1) = N-k \\
df_{\text{被试间}} &= n-1 \\
df_{\text{被试内}} &= n(k-1) = N-n \\
df_{\text{误差}} &= df_{\text{处理内}} - df_{\text{被试间}} = k(n-1)-(n-1) = (k-1)(n-1) \\
df_{\text{误差}} &= df_{\text{被试内}} - df_{\text{处理间}} = n(k-1)-(k-1) = (k-1)(n-1)
\end{aligned}
$$

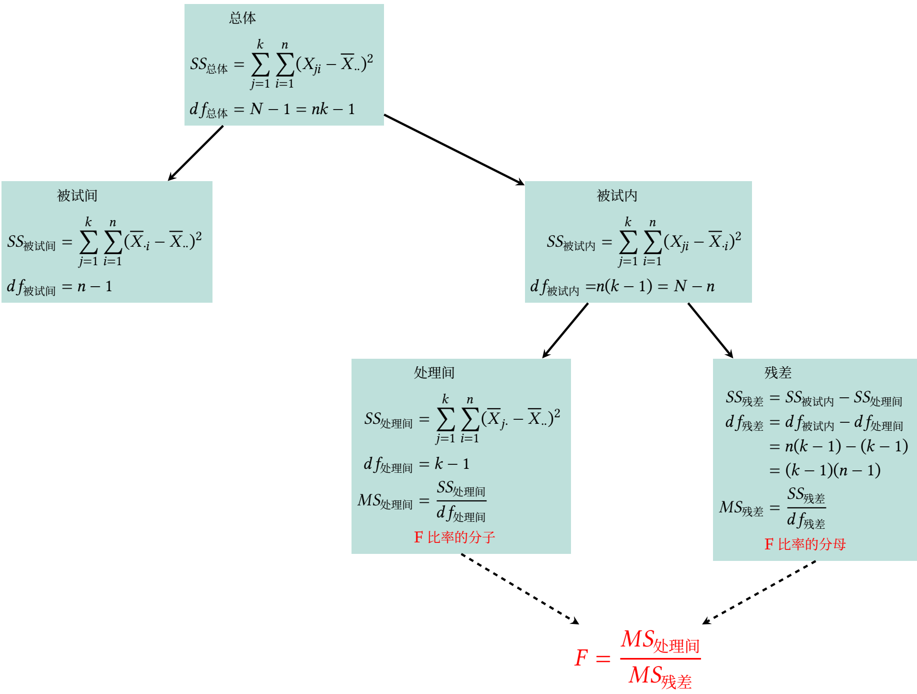
13.3.0.6 方差的计算（MS 值）和 F-比率

$$
\begin{aligned}
MS_{\text{处理间}} &= \frac{SS_{\text{处理间}}}{df_{\text{处理间}}}\\
MS_{\text{误差}} &= \frac{SS_{\text{误差}}}{df_{\text{误差}}}\\
F &= \frac{MS_{\text{处理间}}}{MS_{\text{误差}}}
\end{aligned}
$$

13.3.0.7 重复测量 ANOVA 的方差分割



13.3.0.8 重复测量 ANOVA 的方差分割



13.3.0.9 重复测量 ANOVA 汇总表

来源	SS	df	MS	F
处理间	$SS_{组间}$	$df_{组间}$	$MS_{组间}$	$F = \frac{MS_{组间}}{MS_{残差}}$
处理内	$SS_{组内}$	$df_{组内}$		
被试间	$SS_{被试间}$	$df_{被试间}$		
残差	$SS_{残差}$	$df_{残差}$	$MS_{残差}$	
总	$SS_{总体}$	$df_{总体}$		

13.3.0.10 重复测量 ANOVA 的效应大小测量

$$\eta^2 = \frac{SS_{处理间}}{SS_{总体} - SS_{被试间}} = \frac{SS_{处理间}}{SS_{处理间} + SS_{误差}}$$

13.3.0.11 重复测量 ANOVA：示例

```
transform!(groupby(dt, :被试), :数据 => (mean) => :Mi)
transform!(groupby(dt, :处理), :数据 => (mean) => :Mj)
transform!(dt, :数据 => (mean) => :M)
```

被试	处理	数据	Mi	Mj	M
A	RRR	2	5.0	5.0	8.0
B	RRR	3	6.0	5.0	8.0
C	RRR	8	10.0	5.0	8.0
D	RRR	6	10.0	5.0	8.0
E	RRR	5	8.0	5.0	8.0
F	RRR	6	9.0	5.0	8.0
A	APQ	5	5.0	9.0	8.0
B	APQ	9	6.0	9.0	8.0
C	APQ	10	10.0	9.0	8.0
D	APQ	13	10.0	9.0	8.0
E	APQ	8	8.0	9.0	8.0
F	APQ	9	9.0	9.0	8.0
A	CAQ	8	5.0	10.0	8.0
B	CAQ	6	6.0	10.0	8.0
C	CAQ	12	10.0	10.0	8.0
D	CAQ	11	10.0	10.0	8.0
E	CAQ	11	8.0	10.0	8.0
F	CAQ	12	9.0	10.0	8.0

13.3.0.12 重复测量 ANOVA：平方和

```
X, Mi, Mj, M = dt.数据, dt.Mi, dt.Mj, dt.M
SST  = sum((X  - M ).^2)
SSbt = sum((Mj - M ).^2)
SSwt = sum((X  - Mj).^2)
SSbp = sum((Mi - M ).^2)
SSwp = sum((X  - Mi).^2)
SSE = SSwt - SSbp
SSEp = SSwp - SSbt
SST, SSbt, SSwt, SSbp, SSwp, SSE, SSEp
```

```
(172.0, 84.0, 88.0, 66.0, 106.0, 22.0, 22.0)
```

13.3.0.13 重复测量 ANOVA：自由度

```
N, k, n = nrow(dt), length(levels(dt.处理)), length(levels(dt.被试))
dft  = N - 1
dfbt = k - 1
dfwt = k * (n - 1)
dfbp = n - 1
dfwp = n * (k-1)
dfe  = (k-1)*(n-1)
dft, dfbt, dfwt, dfbp, dfwp, dfe
```

```
(17, 2, 15, 5, 12, 10)
```

13.3.0.14 重复测量 ANOVA：均方

```
MSbt = SSbt / dfbt
MSE  = SSE / dfe
Fval   = round(MSbt / MSE, digits = 2)
MSbt, MSE, Fval
```

```
(42.0, 2.2, 19.09)
```

13.3.0.15 重复测量 ANOVA：汇总表

来源	SS	df	MS	F
处理间	84.0	2	42.0	$F(2, 10) = 19.09$
处理内	88.0	15		
被试间	66.0	5		
残差	22.0	10	2.2	
总	172.0	17		

13.3.0.16 重复测量 ANOVA：临界值

- 观察到的 F 比率为 $F = 19.09$
- 临界 F 比率

```
cquantile(FDist(dfbt, dfe), 0.05)
```

```
4.1028210151304005
```

```
ccdf(FDist(dfbt, dfe), Fval)
```

```
0.00038518203356503044
```

13.3.0.17 重复测量 ANOVA：效应值

```
eta_sqr = SSbt / (SST - SSbp)
```

```
0.7924528301886793
```

13.3.0.18 用 Julia 重复测量 ANOVA

```
anova(dt, :数据, [:处理, :被试], [fixed, subject])
```

```
Analysis of Variance Results

   Effect      SS   DF    MS         F            p         ω²
-----------------------------------------------------------
    Total   172.0   17
     处理    84.0    2   42.0   19.0909   0.000385109   0.429806
     被试    66.0    5   13.2
Remainder    22.0   10    2.2
```

13.3.0.19 用 Julia 重复测量 ANOVA

```
fm1 = lm(@formula(数据 ~ 1 + 被试 + 处理), dt)
fm2 = lm(@formula(数据 ~ 1 + 被试), dt)
GLM.ftest(fm2.model, fm1.model)
```

F-test: 2 models fitted on 18 observations

	DOF	ΔDOF	SSR	ΔSSR	R^2	ΔR^2	F*	p(>F)
[1]	7		106.0000		0.3837			
[2]	9	2	22.0000	-84.0000	0.8721	0.4884	19.0909	0.0004

13.3.0.20 用 Julia 重复测量 ANOVA

```
fm3 = fit(MixedModel, @formula(数据 ~ 1 + 处理 + (1|被试)), dt)
fm4 = fit(MixedModel, @formula(数据 ~ 1 + (1|被试)), dt)
MixedModels.likelihoodratiotest(fm3, fm4)
```

	model-dof	deviance	2	2-dof	P(> 2)
数据 ~ 1 + (1 | 被试)	3	92			
数据 ~ 1 + 处理 + (1 | 被试)	5	73	19	2	<1e-04

13.3.0.21 文献中的结果报告

方差分析的报告应该包含描述统计的总结（至少包括处理均值和标准差，以及需要的表格或图表）和对 ANOVA 结果的简明陈述。

A repeated-measures analysis of variance indicated significant mean differences in the participants' test scores for the three study strategies, $F(2,10) = 19.09$, $p < .01$, $\eta^2 = 0.79$.

三种策略的均值和方差如表 1 所示。重复测量方差分析表明，三种学习策略对参与者的测试分数存在显著差异， $F(2,10) = 19.09$，$p < .01$，$\eta^2 = 0.79$。

```
combine(groupby(dt, :处理), :数据 => (
    x -> (均值 = mean(x), 标准差 = round(std(x), digits = 2))
    ) => AsTable)
```

表 13.2: 使用三种不同学习策略的测试分

处理	均值	标准差
RRR	5.0	2.19
APQ	9.0	2.61
CAQ	10.0	2.45

13.3.0.22 重复测量后事实检验

对于重复测量 ANOVA，可以像独立测量 ANOVA 一样使用 Tukey's HSD 和 Scheffè 测试，只要在公式中用 $MS_{误差}$ 替代 $MS_{处理内}$，在查找统计表中的临界值时使用 $df_{误差}$ 替代 $df_{处理内}$。

13.3.0.23 重复测量 ANOVA 的假设

重复测量 ANOVA 的假设包括:

1. 每个处理条件内的观察必须相互独立。
2. 每个处理内的总体分布必须是正态的。
3. 每个处理的总体分布方差应当相等。
4. 协方差同质性（homogeneity of covariance）：如果处理的效应对所有参与者不一致，或者某些参与者存在次序效应而其他参与者不存在，则该假设被违反。

如果有理由怀疑重复测量 ANOVA 的某一假设被违反，可以使用称为弗里德曼检验（Friedman test）的替代分析，该检验在附录 E 中介绍。弗里德曼检验要求在评估处理条件之间的差异之前，将原始分数转换为“等级”。

13.4 更多关于重复测量设计的信息

13.4.0.1 影响结果的因素

- 处理效应的大小
- 样本的大小
- 分数的方差

13.4.0.2 重复测量设计的优点

独立测量设计和重复测量设计的 F 比率。

$$F_{\text{整体结构}} = \frac{\text{处理效应} + \text{随机的非系统差异}}{\text{随机的非系统差异}}$$
$$F_{\text{独立测量}} = \frac{\text{处理效应} + \text{个体差异和其他误差}}{\text{个体差异和其他误差}}$$
$$F_{\text{重复测量}} = \frac{\text{处理效应} + \text{误差（不包括个体差异）}}{\text{误差（不包括个体差异）}}$$

当个体差异较大时，重复测量实验可能对处理效应提供更敏感的检验。从统计学角度看，重复测量测试比独立测量测试具有更高的“统计效力”，即更有可能检测到真实的处理效应。

13.4.0.3 处理效应的一致性

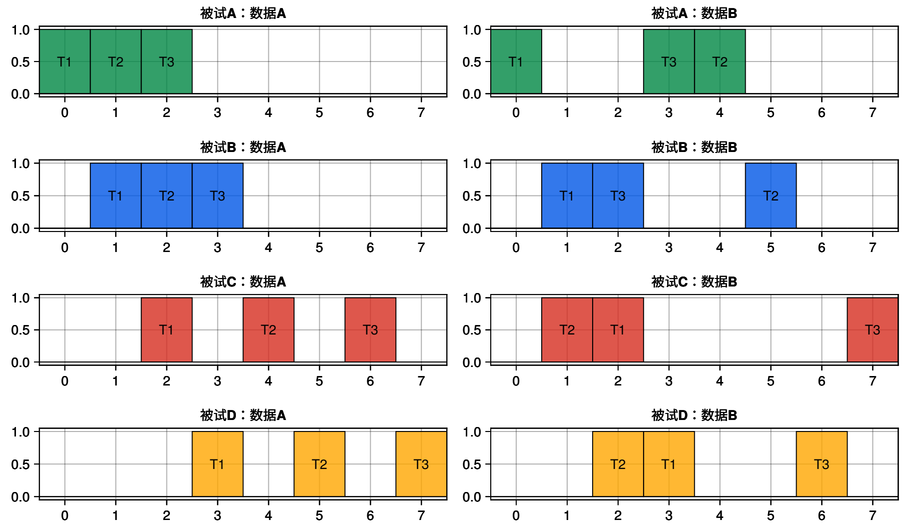
当处理效应在一个个体和另一个个体之间保持一致时，个体差异也往往保持一致且相对较大。大的个体差异会从 F 比率的分母中减去，产生更大的 F 值，并增加 F 比率处于临界区域的可能性。

```
dt2 = DataFrame(
    被试   = repeat('A':'D', outer = 3),
    处理   = repeat('T' .* string.(1:3), inner = 4),
    数据A = [0, 1, 2, 3, 1, 2, 4, 5, 2, 3, 6, 7],
    数据B = [0, 1, 2, 3, 4, 5, 1, 2, 3, 2, 7, 6])
sort!(dt2, :被试)
```

被试	处理	数据 A	数据 B
A	T1	0	0
A	T2	1	4
A	T3	2	3
B	T1	1	1
B	T2	2	5
B	T3	3	2
C	T1	2	2
C	T2	4	1
C	T3	6	7
D	T1	3	3
D	T2	5	2
D	T3	7	6

```
fig = Figure(size = (1000, 600))
for (sid,  sbj) in enumerate(unique(dt2.被试)),
    (cid,  cnd) in enumerate(unique(dt2.处理)),
    (tid, test) in enumerate(names(dt2)[3:4])
    col = ["#008744", "#0057e7", "#d62d20", "#ffa700"][sid]
    dtp = dt2[(dt2.被试 .== sbj) .&& (dt2.处理 .== cnd), test]
    ax = Axis(fig[sid, tid],
        xticks = 0:7,  limits = ((-0.5, 7.5), nothing),
        title = string("被试" * sbj * ": " * test))
    hist!(ax, dtp, bins = -0.5:1:7.5, color = (col, 0.8),
        strokewidth = 1, strokecolor = :black)
    text!(ax, dtp[], 0.5; text = cnd, align = (:center, :center))
end
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```julia
anova(dt2, :数据A, [:处理, :被试], [fixed, subject])
```

```
Analysis of Variance Results

   Effect     SS  DF          MS      F       p         ω²
-----------------------------------------------------
    Total   50.0  11
     处理   18.0   2    9.0        27.0  0.001  0.288889
     被试   30.0   3   10.0
Remainder    2.0   6    0.333333
```

```julia
anova(dt2, :数据B, [:处理, :被试], [fixed, subject])
```

```
Analysis of Variance Results

   Effect         SS  DF         MS         F          p         ω²
-------------------------------------------------------------
    Total   50.0       11
     处理   18.0        2   9.0       1.88372  0.231799  0.165217
     被试    3.33333    3   1.11111
Remainder   28.6667     6   4.77778
```

```julia
fm1 = lm(@formula(数据A ~ 1 + 被试 + 处理), dt2)
fm2 = lm(@formula(数据A ~ 1 + 被试), dt2)
GLM.ftest(fm2.model, fm1.model)
```

```
F-test: 2 models fitted on 12 observations
──────────────────────────────────────────────────────────────────
     DOF  ΔDOF        SSR        ΔSSR         R²        ΔR²          F*   p(>F)
```

```
[1]     5        20.0000            0.6000
[2]     7     2   2.0000  -18.0000  0.9600  0.3600  27.0000  0.0010
```

```
fm1 = lm(@formula(数据B ~ 1 + 被试 + 处理), dt2)
fm2 = lm(@formula(数据B ~ 1 + 被试), dt2)
GLM.ftest(fm2.model, fm1.model)
```

```
F-test: 2 models fitted on 12 observations
```

	DOF	ΔDOF	SSR	ΔSSR	R^2	ΔR^2	F*	p(>F)
[1]	5		46.6667		0.0667			
[2]	7	2	28.6667	-18.0000	0.4267	0.3600	1.8837	0.2318

13.4.0.4 重复测量 ANOVA 和重复测量 t 检验

这两个测试总是对零假设达成相同的结论。两个测试统计量之间的基本关系是 $F = t^2$。t 统计量的自由度值与 F 比率的分母的自由度值是相同的。如果你将双侧 t 检验的临界值平方，你将得到 F 比率的临界值。再次强调，基本关系是 $F = t^2$。

```
dt3 = DataFrame(
    被试 = repeat('A':'D', outer = 2),
    处理 = repeat('T' .* string.(1:2), inner = 4),
    数据 = [3,  4, 5, 4, 5, 14, 7, 6])
```

被试	处理	数据
A	T1	3
B	T1	4
C	T1	5
D	T1	4
A	T2	5
B	T2	14
C	T2	7
D	T2	6

```
DD = combine(groupby(dt3, :被试), :数据 => (diff) => :D)
TT = OneSampleTTest(DD.D)
TT.t, pvalue(TT)
```

```
(2.0, 0.13932596855884313)
```

```
anova(dt3, :数据, [:处理, :被试], [fixed, subject])
```

```
Analysis of Variance Results

   Effect    SS  DF        MS    F         p        ω²
------------------------------------------------------
    Total  84.0   7
     处理  32.0   1  32.0      4.0  0.139326  0.257143
     被试  28.0   3   9.33333
Remainder  24.0   3   8.0
```

```
fm1 = lm(@formula(数据 ~ 1 + 被试 + 处理), dt3)
fm2 = lm(@formula(数据 ~ 1 + 被试), dt3)
GLM.ftest(fm2.model, fm1.model)
```

F-test: 2 models fitted on 8 observations

	DOF	ΔDOF	SSR	ΔSSR	R^2	ΔR^2	F*	p(>F)
[1]	5		56.0000		0.3333			
[2]	6	1	24.0000	-32.0000	0.7143	0.3810	4.0000	0.1393

第十四章　两因素方差分析

14.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie
using DataFrames, StatsReIntro
using Statistics: mean, std
using Distributions: FDist, cquantile, ccdf
using GLM
using SimpleANOVA
using StatsModels
```

14.2　两因素独立测量方差分析概述

14.2.0.1　暴力视频游戏和玩家行为?

Bartholow 和 Anderson（2002）随机分配了男性和女性本科生来玩一个暴力视频游戏或一个非暴力游戏。游戏结束后，每个参与者被要求参加一个与另一个实际上是研究团队一员（同谋者）的学生一起进行的竞争反应时间游戏。两名学生都被要求尽快对刺激音进行反应，每次失败者都会通过耳机传送的白噪声爆炸来惩罚。说明中的一部分允许参与者设置惩罚噪声的强度级别，并所选级别被用作该参与者攻击性行为的度量，较高级别表示更具攻击性的行为。研究结果显示，视频游戏中的暴力程度基本上对女性参与者的行为没有影响，但与非暴力游戏相比，男性在玩暴力游戏后更具攻击性。

14.2.0.2　两因素独立测量等 n 设计

当研究涉及*多个因素*时，它被称为*因子设计*（factorial design）。我们将研究适用于具有*两个*因素的研究。我们限制我们的讨论仅适用于使用各个处理条件的单独样本的研究设计，即*独立测量*设计（independent-measures）。最后，我们只考虑样本大小（n）对所有处理条件都是*相同*的研究设计。

14.2.0.3 以矩阵形式呈现两因素实验

	因素 B				
因素 A	B_1	$\cdots$	B_c	$\cdots$	B_k
	X_{111}		X_{1c1}		X_{1k1}
A_1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{11n}		X_{1cn}		X_{1kn}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	X_{r11}		X_{rc1}		X_{rk1}
A_r	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{r1n}		X_{rcn}		X_{rkn}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	X_{m11}		X_{mc1}		X_{mk1}
A_m	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	X_{m1n}		X_{mcn}		X_{mkn}

该矩阵中可以得出一下一些平均值

	因素 B					
因素 A	B_1	$\cdots$	B_c	$\cdots$	B_k	均值$_A$
A_1	$M_{11\cdot}$	$\cdots$	$M_{1c\cdot}$	$\cdots$	$M_{1k\cdot}$	$M_{1\cdot\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	$M_{r1\cdot}$	$\cdots$	$M_{rc\cdot}$	$\cdots$	$M_{rk\cdot}$	$M_{r\cdot\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_m	$M_{m1\cdot}$	$\cdots$	$M_{mc\cdot}$	$\cdots$	$M_{mk\cdot}$	$M_{k\cdot\cdot}$
均值$_B$	$M_{\cdot 1\cdot}$	$\cdots$	$M_{\cdot c\cdot}$	$\cdots$	$M_{\cdot k\cdot}$	$M_{\cdot\cdot\cdot}$

14.2.0.4 主效应

一个因素的水平之间的平均差异被称为该因素的**主效应**（main effect）。当研究设计以一个因素决定行和第二因素决定列的*矩阵*表示时，*行*之间的平均差异描述了一个因素的主效应，以及 *列*之间的平均差异描述了第二因素的主效应。

14.2.0.5 交互作用

当两个因素之间的平均差异，或单元格之间的平均差异与从因素的整体主效应预测的不同时，就发生了两个因素之间的**交互作用**（interaction）。当一个因素的影响*取决*于第二因素的不同水平时，就会发生因素之间的交互作用。当两因素研究的结果以图形方式呈现时，*存在非平行线*（交叉或汇合的线）表明两个因素之间存在交互作用。

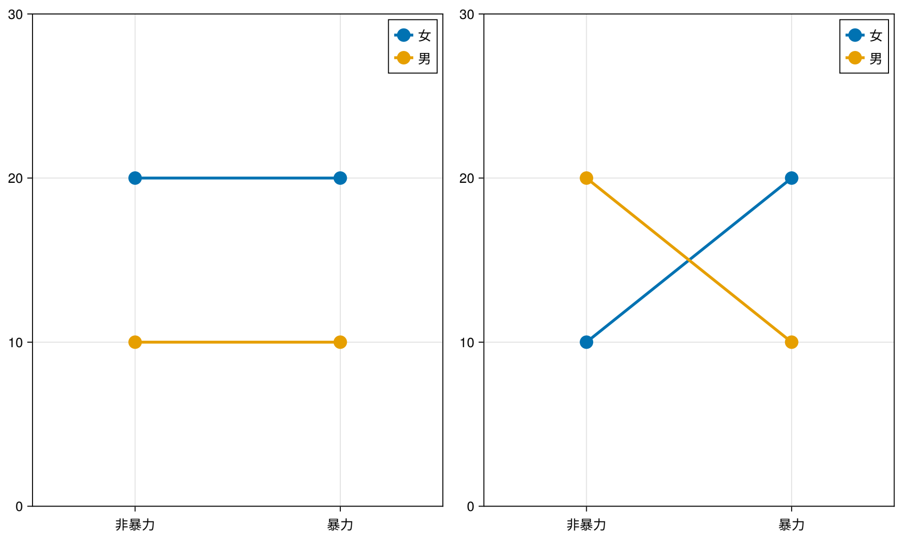
14.2.0.6 主效应和交互作用

```julia
function MainInter(dts)
    fig = Figure(size = (1000, 500))
    axs = [Axis(fig[1, id],
                limits = ((0.5, 2.5), (0, 30)),
                xticks = ([1, 2], ["非暴力", "暴力"]))
        for id in 1:length(dts)]
    for (id, dt) in enumerate(dts)
        for ky in keys(dt)
            scatter!(axs[id], 1:2, dt[ky], markersize = 20, label = string(ky))
            lines!(axs[id], 1:2, dt[ky], linewidth = 3, label = string(ky))
        end
        axislegend(axs[id], merge = true)
    end
    fig
end
```

```
MainInter (generic function with 1 method)
```

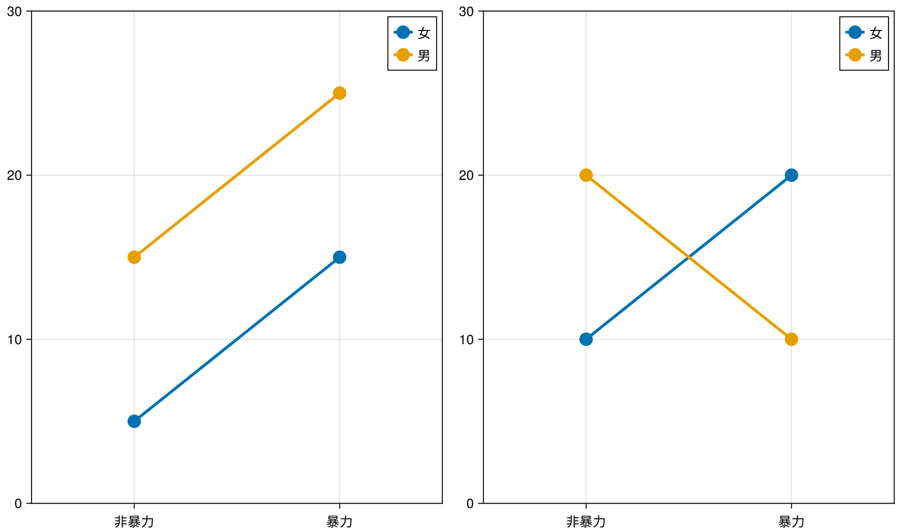
```julia
dts1 = [(女 = [20, 20], 男 = [10, 10]),
        (女 = [10, 20], 男 = [20, 10])]
MainInter(dts1)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



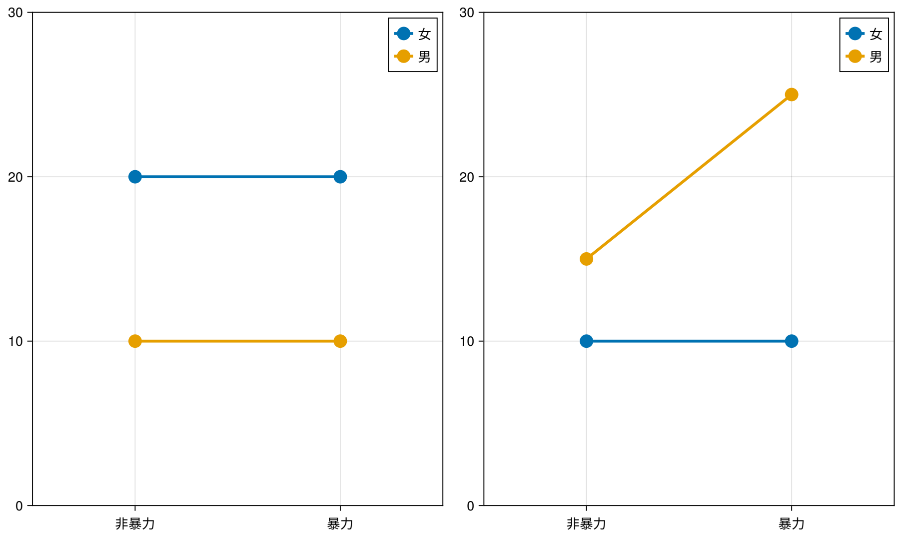
```
dts2 = [(女 = [5, 15], 男 = [15, 25]),
        (女 = [10, 20], 男 = [20, 10])]
MainInter(dts2)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
dts3 = [(女 = [20, 20], 男 = [10, 10]),
        (女 = [10, 10], 男 = [15, 25])]
MainInter(dts3)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



14.3 两因素 ANOVA 和效应大小示例

14.3.0.1 F-比率

$$F = \frac{\text{处理间方差（平均差异）}}{\text{如果没有处理效应则预期的方差（平均差异）}}$$

14.3.0.2 两因素 ANOVA 的矩阵结构

	因素 B				
因素 A	B_1	$\cdots$	B_c	$\cdots$	B_k
A_1	X_{111} $\cdots$ X_{11n}	$\cdots$	X_{1c1} $\cdots$ X_{1cn}	$\cdots$	X_{1k1} $\cdots$ X_{1kn}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
A_r	X_{r11} $\cdots$ X_{r1n}	$\cdots$	X_{rc1} $\cdots$ X_{rcn}	$\cdots$	X_{rk1} $\cdots$ X_{rkn}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
A_m	X_{m11} $\cdots$ X_{m1n}	$\cdots$	X_{mc1} $\cdots$ X_{mcn}	$\cdots$	X_{mk1} $\cdots$ X_{mkn}

与该矩阵相关的平均值和求和为

	因素 B						
因素 A	B_1	$\cdots$	B_c	$\cdots$	B_k	Total$_A$	均值$_A$
A_1	n $M_{11\cdot}$ $T_{11\cdot}$ $SS_{11\cdot}$	$\cdots$	n $M_{1c\cdot}$ $T_{1c\cdot}$ $SS_{1c\cdot}$	$\cdots$	n $M_{1k\cdot}$ $T_{1k\cdot}$ $SS_{1k\cdot}$	$T_{1\cdot\cdot}$	$M_{1\cdot\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
A_r	n $M_{r1\cdot}$ $T_{r1\cdot}$ $SS_{r1\cdot}$	$\cdots$	n $M_{rc\cdot}$ $T_{rc\cdot}$ $SS_{rc\cdot}$	$\cdots$	n $M_{rk\cdot}$ $T_{rk\cdot}$ $SS_{rk\cdot}$	$T_{r\cdot\cdot}$	$M_{r\cdot\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
A_m	n $M_{m1\cdot}$ $T_{m1\cdot}$ $SS_{m1\cdot}$	$\cdots$	n $M_{mc\cdot}$ $T_{mc\cdot}$ $SS_{mc\cdot}$	$\cdots$	n $M_{mk\cdot}$ $T_{mk\cdot}$ $SS_{mk\cdot}$	$T_{m\cdot\cdot}$	$M_{m\cdot\cdot}$
T$_B$	$T_{\cdot 1\cdot}$	$\cdots$	$T_{\cdot c\cdot}$	$\cdots$	$T_{\cdot k\cdot}$	$T_{\cdots}(G)$	
均值$_B$	$M_{\cdot 1\cdot}$	$\cdots$	$M_{\cdot c\cdot}$	$\cdots$	$M_{\cdot k\cdot}$		$M_{\cdots}$

14.3.0.3　两因素 ANOVA 的方差分解

$$
\begin{aligned}
SST &= \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(X_{rci}-M_{\cdots})^2 = \sum X_{rci}^2 - \frac{G^2}{N} \\
SS_{\text{处理间}} &= \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(M_{rc\cdot}-M_{\cdots})^2 = \sum T_{rc\cdot}^2 - \frac{G^2}{N} \\
SS_{\text{处理内}} &= \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(X_{rci}-M_{rc\cdot})^2 = \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(SS_{rc\cdot})^2 \\
SST &= SS_{\text{处理间}} + SS_{\text{处理内}}
\end{aligned}
$$

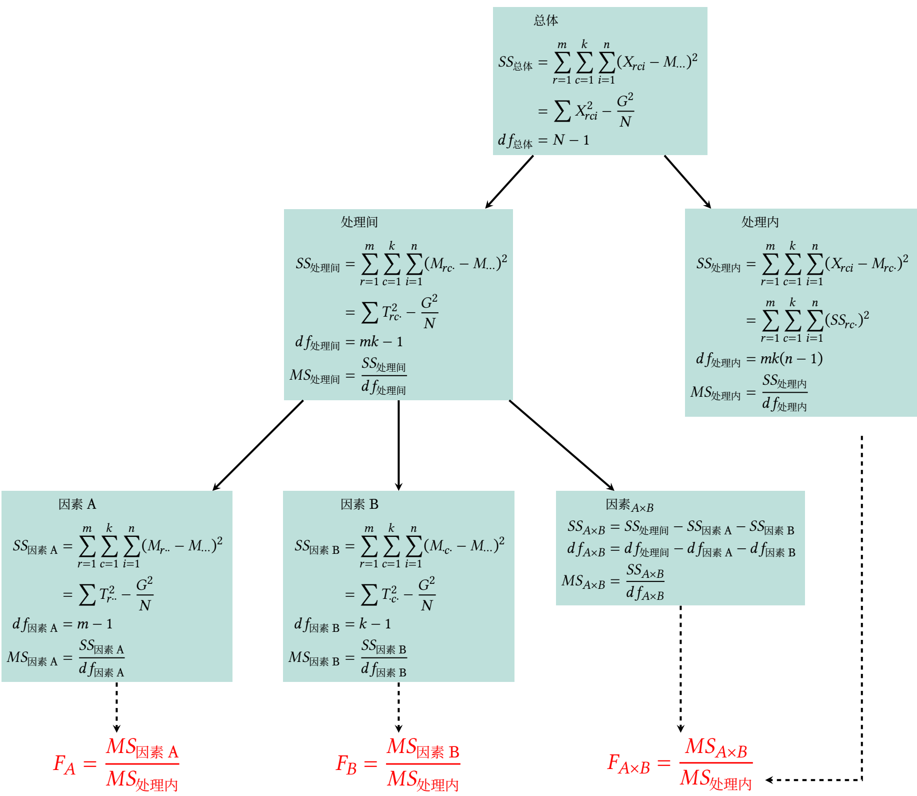
14.3.0.4 两因素 ANOVA 的方差分解

$$SS_{\text{因素 A}} = \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(M_{r\cdot\cdot} - M_{\cdot\cdot\cdot})^2 = \sum T_{r\cdot\cdot}^2 - \frac{G^2}{N}$$

$$SS_{\text{因素 B}} = \sum_{r=1}^{m}\sum_{c=1}^{k}\sum_{i=1}^{n}(M_{\cdot c\cdot} - M_{\cdot\cdot\cdot})^2 = \sum T_{\cdot c\cdot}^2 - \frac{G^2}{N}$$

$$SS_{A\times B} = SS_{\text{处理间}} - SS_{\text{因素 A}} - SS_{\text{因素 B}}$$

14.3.0.5 两因素 ANOVA 的方差分解



14.3.0.6 两因素 ANOVA：示例

```
dt = DataFrame(
    方式 = repeat(["自定步速", "固定步速"], inner = 5, outer = 2),
    媒介 = repeat(["纸质", "电子"], inner = 10),
    分数 = [11, 8,  9, 10, 7, 10, 7, 10,  6, 7, 4, 4,  8,  5, 4, 10, 6, 10, 10, 9]
)
```

方式	媒介	分数
自定步速	纸质	11
自定步速	纸质	8
自定步速	纸质	9
自定步速	纸质	10
自定步速	纸质	7
固定步速	纸质	10
固定步速	纸质	7
固定步速	纸质	10
固定步速	纸质	6
固定步速	纸质	7
自定步速	电子	4
自定步速	电子	4
自定步速	电子	8
自定步速	电子	5
自定步速	电子	4
固定步速	电子	10
固定步速	电子	6
固定步速	电子	10
固定步速	电子	10
固定步速	电子	9

```
transform!(groupby(dt, [:方式, :媒介]), :分数 => (mean) => :Mab)
transform!(groupby(dt, :方式), :分数 => (mean) => :Ma)
transform!(groupby(dt, :媒介), :分数 => (mean) => :Mb)
transform!(dt, :分数 => (mean) => :Mt)
```

```
X, Mab, Ma, Mb, Mt = dt.分数, dt.Mab, dt.Ma, dt.Mb, dt.Mt
m, k = length(levels(dt.方式)), length(levels(dt.媒介))
N, n = nrow(dt), combine(groupby(dt, [:方式, :媒介]), nrow).nrow
```

14.3.0.7 两因素分析的第一阶段

- 总变异和总自由度

```
SST = sum((X - Mt).^2)
dft = N - 1
```

方式	媒介	分数	Mab	Ma	Mb	Mt
自定步速	纸质	11	9.0	7.0	8.5	7.75
自定步速	纸质	8	9.0	7.0	8.5	7.75
自定步速	纸质	9	9.0	7.0	8.5	7.75
自定步速	纸质	10	9.0	7.0	8.5	7.75
自定步速	纸质	7	9.0	7.0	8.5	7.75
固定步速	纸质	10	8.0	8.5	8.5	7.75
固定步速	纸质	7	8.0	8.5	8.5	7.75
固定步速	纸质	10	8.0	8.5	8.5	7.75
固定步速	纸质	6	8.0	8.5	8.5	7.75
固定步速	纸质	7	8.0	8.5	8.5	7.75
自定步速	电子	4	5.0	7.0	7.0	7.75
自定步速	电子	4	5.0	7.0	7.0	7.75
自定步速	电子	8	5.0	7.0	7.0	7.75
自定步速	电子	5	5.0	7.0	7.0	7.75
自定步速	电子	4	5.0	7.0	7.0	7.75
固定步速	电子	10	9.0	8.5	7.0	7.75
固定步速	电子	6	9.0	8.5	7.0	7.75
固定步速	电子	10	9.0	8.5	7.0	7.75
固定步速	电子	10	9.0	8.5	7.0	7.75
固定步速	电子	9	9.0	8.5	7.0	7.75

```
SST, dft
```

```
(101.75, 19)
```

- 处理间变异

```
SSbt = sum((Mab - Mt).^2)
dfbt = m * k - 1
SSbt, dfbt
```

```
(53.75, 3)
```

- 处理内变异

```
SSwt = sum((X   - Mab).^2)
dfwt = sum(n .- 1)
MSwt = SSwt / dfwt
SSwt, dfwt, MSwt
```

```
(48.0, 16, 3.0)
```

14.3.0.8 两因素分析的第二阶段

- A 因素

```
SSa = sum((Ma - Mt).^2)
dfa = m - 1
MSa = SSa / dfa
SSa, dfa, MSa
```

```
(11.25, 1, 11.25)
```

- B 因素

```
SSb = sum((Mb - Mt).^2)
dfb = k - 1
MSb = SSb / dfb
SSa, dfb, MSb
```

```
(11.25, 1, 11.25)
```

- $A \times B$

```
SSab = SSbt - SSa - SSb
dfab = dfbt - dfa - dfb
MSab = SSab / dfab
SSab, dfab, MSab
```

```
(31.25, 1, 31.25)
```

14.3.0.9 两因素 ANOVA 的 F 比率

- A 因素的 F 比率

```
Fa = MSa / MSwt
```

```
3.75
```

- B 因素的 F 比率

```
Fb = MSb / MSwt
```

```
3.75
```

- $A \times B$ 交互作用的 F 比率

```
Fab = MSab / MSwt
```

```
10.416666666666666
```

14.3.0.10 两因素 ANOVA 的临界值

- 我们将 α 值设定为 $\alpha = 0.05$
- A 因素的临界值

```
cquantile(FDist(dfa, dfwt), 0.05)
```

```
4.493998477666358
```

- B 因素的临界值

```
cquantile(FDist(dfb, dfwt), 0.05)
```

```
4.493998477666358
```

- A x B 交互作用的临界值

```
cquantile(FDist(dfab, dfwt), 0.05)
```

```
4.493998477666358
```

14.3.0.11 两因素 ANOVA 的汇总表

来源	SS	df	MS	F
处理间	53.75	3		
因素 A	11.25	1	11.25	F(1, 16) = 3.75
因素 B	11.25	1	11.25	F(1, 16) = 3.75
A × B	31.25	1	31.25	F(1, 16) = 10.42
处理内	48.0	16	3.0	
总	101.75	19		

14.3.0.12 两因素 ANOVA 的效应大小测量

- A 因素

$$\eta^2 = \frac{SS_A}{SS_{总体} - SS_B - SS_{A\times B}} = \frac{SS_A}{SS_A + SS_{处理间}}$$

- B 因素

$$\eta^2 = \frac{SS_B}{SS_{总体} - SS_A - SS_{A\times B}} = \frac{SS_B}{SS_B + SS_{处理间}}$$

- $A \times B$ 交互作用

$$\eta^2 = \frac{SS_{A\times B}}{SS_{总体} - SS_A - SS_B} = \frac{SS_A}{SS_{A\times B} + SS_{处理间}}$$

14.3.0.13 两因素 ANOVA 的效应大小测量

- A 因素的 η^2

```
SSa / (SSa + SSwt)
```

```
0.189873417721519
```

- B 因素的 η^2

```
SSb / (SSb + SSwt)
```

```
0.189873417721519
```

- $A \times B$ 交互作用的 η^2

```
SSab / (SSab + SSwt)
```

```
0.3943217665615142
```

14.3.0.14 文献中的结果报告

首先，报告平均值和标准差。因为两因素设计通常涉及多个处理条件，这些描述性统计通常以表格或图形的形式呈现。接下来，报告所有三个假设检验（F-比率）的结果。

表 1 显示了所有处理条件的均值和标准差。两因素方差分析显示，时间控制没有显著的主效应，$F(1,16) = 3.75$，$p > .05$，$\eta^2 = 0.190$，*呈现模式也没有显著的主效应，*$F(1,16) = 3.75$，$p > .05$，$\eta^2 = 0.190$。*然而，两个因素之间的交互作用是显著的，*$F(1,16) = 10.41$，$p < .01$，$\eta^2 = 0.394$。

The means and standard deviations for all treatment conditions are shown in Table 1. The two-factor analysis of variance showed no significant main effect for time control, $F(1,16) = 3.75$, $p > .05$, $\eta^2 = 0.190$ *or for presentation mode,* $F(1,16) = 3.75$, $p > .05$, $\eta^2 = 0.190$. *However, the interaction between factors was significant,* $F(1,16) = 10.41$, $p < .01$, $\eta^2 = 0.394$.

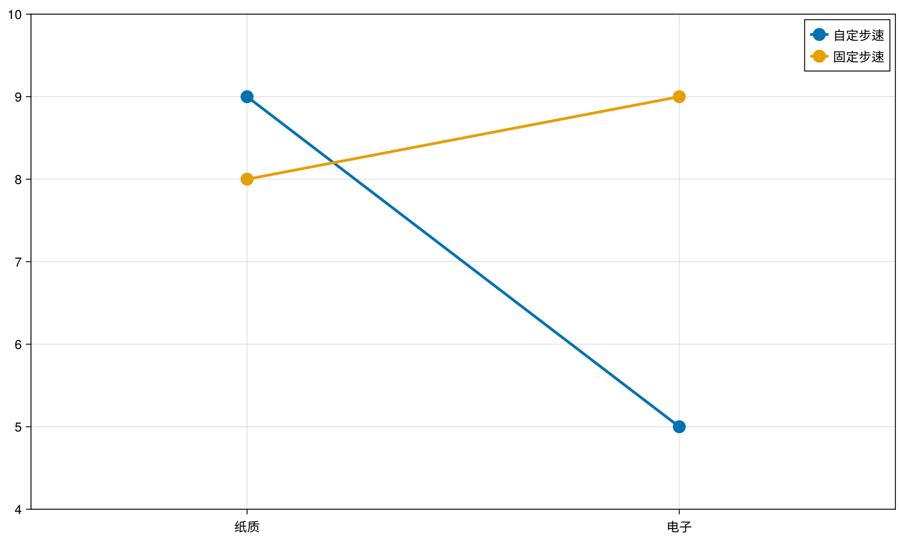
```
dtc = combine(groupby(dt, [:方式, :媒介]),
    :分数 => (x -> (均值 = mean(x), 方差 = std(x))) => AsTable)
```

方式	媒介	均值	方差
自定步速	纸质	9.0	1.58114
固定步速	纸质	8.0	1.87083
自定步速	电子	5.0	1.73205
固定步速	电子	9.0	1.73205

14.3.0.15 理解两因素 ANOVA 的结果

```
gdtc = groupby(dtc, :方式)
fig = Figure(resolution = (1000, 600))
ax1 = Axis(fig[1, 1],
    limits = ((0.5, 2.5), (4, 10)),
    xticks = (1:2, ["纸质", "电子"]))
for dt in gdtc
    lb = unique(dt.方式)
    scatter!(ax1, 1:2, dt.均值, markersize = 20, label = lb)
    lines!(ax1, 1:2, dt.均值, linewidth = 3, label = lb)
end
axislegend(ax1, merge = true)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



14.3.0.16　用 Julia 语言分析：传统方式

```
anova(dt, :分数, [:方式, :媒介])
```

```
Analysis of Variance Results

 Effect       SS  DF     MS         F           p        ω²
----------------------------------------------------------
  Total  101.75  19
   媒介    11.25   1  11.25   3.75    0.0706743   0.120879
   方式    11.25   1  11.25   3.75    0.0706743   0.120879
媒介 × 方式   31.25   1  31.25  10.4167  0.00526454  0.320113
  Error   48.0   16   3.0
```

14.3.0.17　用 Julia 语言分析：线性模型

```
dtn = insertcols(dt, "方式媒介" => dt.方式 .* dt.媒介)
contrasts = Dict(
    :方式媒介 => HypothesisCoding(
      [
        -1 -1 +1 +1          # 方式
        -1 +1 -1 +1          # 媒介
        +1 -1 -1 +1          # 方式 x 媒介
      ];
      levels = reverse(StatsModels.levels(dtn."方式媒介")),
      labels = ["方式", "媒介", "方式 x 媒介"]
    ),
  )
```

```
md = lm(@formula(分数 ~ 方式媒介), dtn; contrasts)
DataFrame(coeftable(md))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	7.75	0.387298	20.0104	9.49253e-13	6.92896	8.57104
方式媒介: 方式	3.0	1.54919	1.93649	0.0706743	-0.284143	6.28414
方式媒介: 媒介	-3.0	1.54919	-1.93649	0.0706743	-6.28414	0.284143
方式媒介: 方式 x 媒介	5.0	1.54919	3.22749	0.00526454	1.71586	8.28414

14.4 更多关于两因素 ANOVA 的内容

14.4.0.1 测试简单主效应

显著的交互作用表明一个因素的效应（均值差异）*取决于*第二因素的水平。当数据以显示处理均值的矩阵形式呈现时，显著的交互作用表明在一个*列*（或行）内的均值差异显示与在*另一个列*（或行）内的均值差异不同的模式。

在这种情况下，研究人员可能希望对每个单独的列（或行）进行单独的分析。实际上，研究人员正在将两因素实验分为一系列单独的单因素实验。在两因素设计的一个列（或一行）内测试均值差异的显著性的过程被称为测试**简单主效应**。

14.4.0.2 使用第二因素减少方差

$$F = \frac{\text{处理间方差（平均差异）}}{\text{如果没有处理效应则预期的方差（平均差异）}}$$

- 以自定步速条件为例

```
dts = subset(dt, :方式 => ByRow(==("自定步速")))
transform!(groupby(dts, [:方式, :媒介]), :分数 => (mean) => :Mab)
transform!(groupby(dts, :方式), :分数 => (mean) => :Ma)
transform!(groupby(dts, :媒介), :分数 => (mean) => :Mb)
transform!(dts, :分数 => (mean) => :Mt)
```

方式	媒介	分数	Mab	Ma	Mb	Mt
自定步速	纸质	11	9.0	7.0	9.0	7.0
自定步速	纸质	8	9.0	7.0	9.0	7.0
自定步速	纸质	9	9.0	7.0	9.0	7.0
自定步速	纸质	10	9.0	7.0	9.0	7.0
自定步速	纸质	7	9.0	7.0	9.0	7.0
自定步速	电子	4	5.0	7.0	5.0	7.0
自定步速	电子	4	5.0	7.0	5.0	7.0
自定步速	电子	8	5.0	7.0	5.0	7.0
自定步速	电子	5	5.0	7.0	5.0	7.0
自定步速	电子	4	5.0	7.0	5.0	7.0

```
SSb = sum((dts.Mb - dts.Mt).^2)
dfb = 1
MSb = SSb / dfb
```

```
40.0
```

```
Fratio = MSb / MSwt
```

```
13.333333333333334
```

```
ccdf(FDist(dfb, dfwt), Fratio)
```

```
0.0021519618126000523
```

```
fm = @formula(分数 ~ 媒介 * 方式)
contr = Dict(
    :方式 => DummyCoding(base = "自定步速")
  )
DataFrame(coeftable(lm(fm, dt; contrasts = contr)))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	
(Intercept)	5.0	0.774597	6.45497	7.94995e-6	3.35793	…
媒介: 纸质	4.0	1.09545	3.65148	0.00215196	1.67776	…
方式: 固定步速	4.0	1.09545	3.65148	0.00215196	1.67776	…
媒介: 纸质 & 方式: 固定步速	-5.0	1.54919	-3.22749	0.00526454	-8.28414	…

14.4.0.3 两因素 ANOVA 的假设

- 每个样本内的观测必须是独立的。
- 从样本中选择的总体必须是正态的。
- 从样本中选择的总体必须具有相等的方差（方差的均匀性）。

第五部分

相关和回归

第十五章　相关性

15.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using StatsReIntro, DataFrames
using CairoMakie
using GLM
using Distributions: MvNormal, TDist, ccdf
using Statistics: mean, cor, cov, var
using HypothesisTests: CorrelationTest, EqualVarianceTTest
```

15.2　介绍

15.2.0.1　什么是相关性

相关性是一种用于衡量和描述*两个变量*之间关系的统计技术。相关性需要*每个个体的两个*分数（来自两个变量中的每一个的一个分数）。

15.2.0.2　一个例子

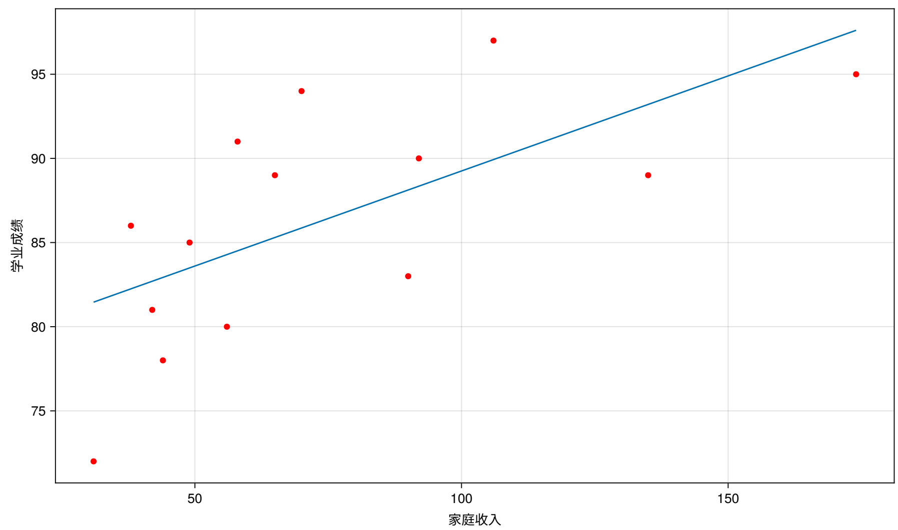
```
dt = DataFrame(
    学号     = ('A':'Z')[1:14],
    家庭收入 = [31, 38, 42, 44, 49, 56, 58, 65, 70, 90, 92, 106, 135, 174],
    学业成绩 = [72, 86, 81, 78, 85, 80, 91, 89, 94, 83, 90, 97, 89, 95]
)
```

```
pred = predict(lm(@formula(学业成绩 ~ 家庭收入), dt))
scatter(dt.家庭收入, dt.学业成绩, color = :red, axis = (xlabel = "家庭收入", ylabel = "学业成
lines!(dt.家庭收入, pred)
```

学号	家庭收入	学业成绩
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95

```
current_figure()
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.2.0.3 关系的特征

1. 关系的形式
2. 关系的方向
3. 关系的强度或一致性

15.2.0.4 关系的方向

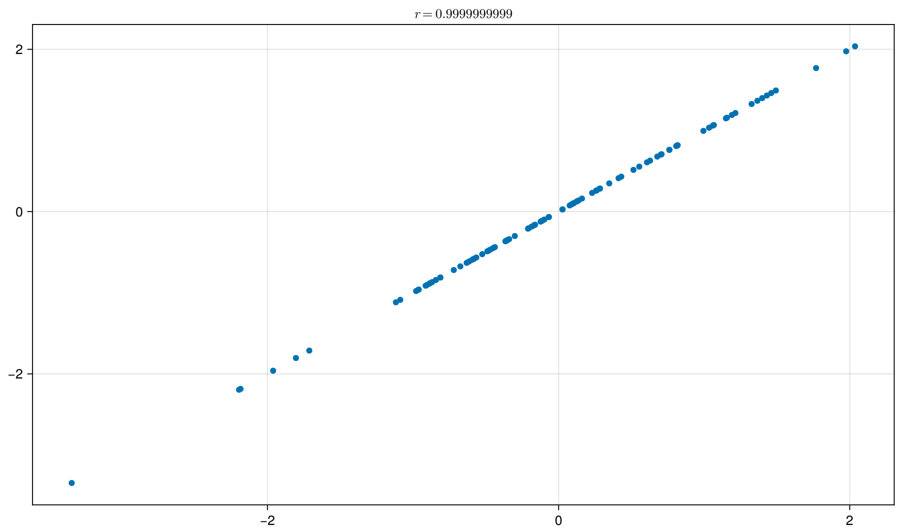
- 在**正相关**中，两个变量倾向于以相同的方向变化：随着 X 变量的值从一个个体增加到另一个个体，Y 变量也倾向于增加；当 X 变量减小时，Y 变量也减小。

- 在**负相关**中，两个变量倾向于朝相反方向变化。随着 X 变量增加，Y 变量减小。也就是说，这是一种反向关系。

15.2.0.5 相关性的方向和强度

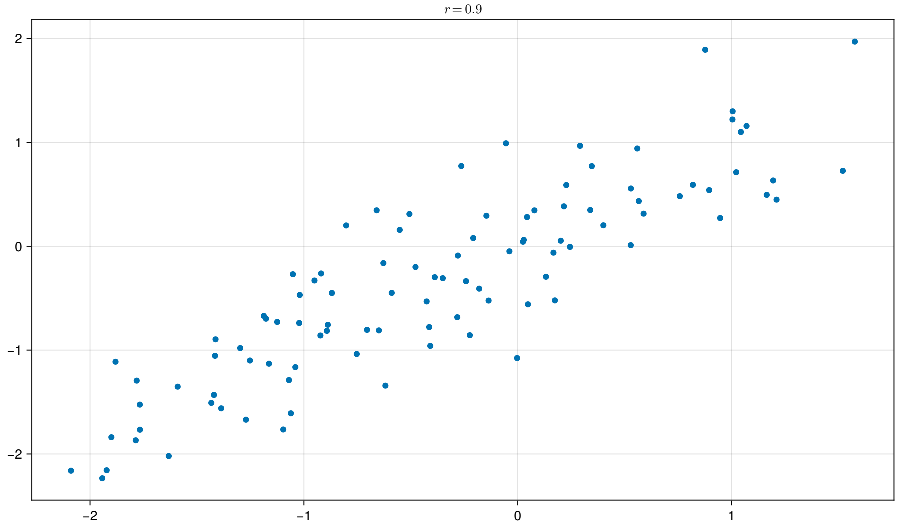
```
r = 1-1e-10
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



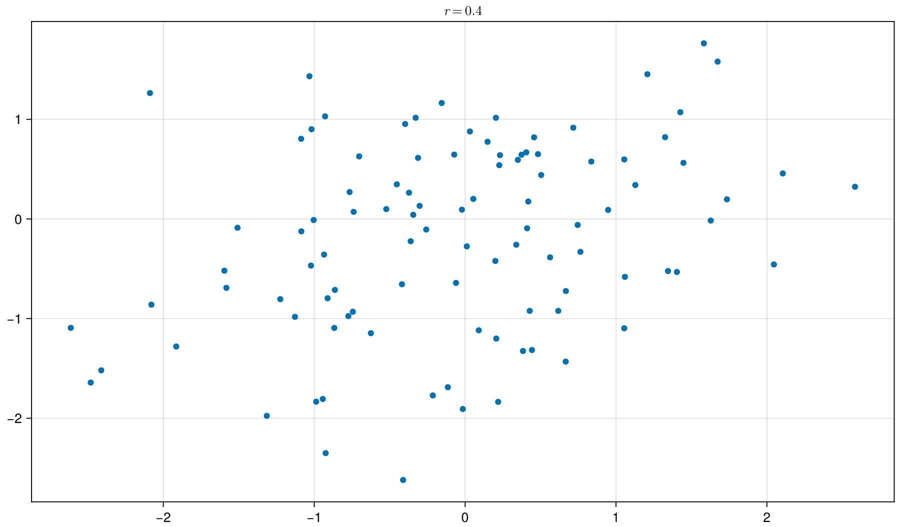

```
r = 0.9
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
r = 0.4
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

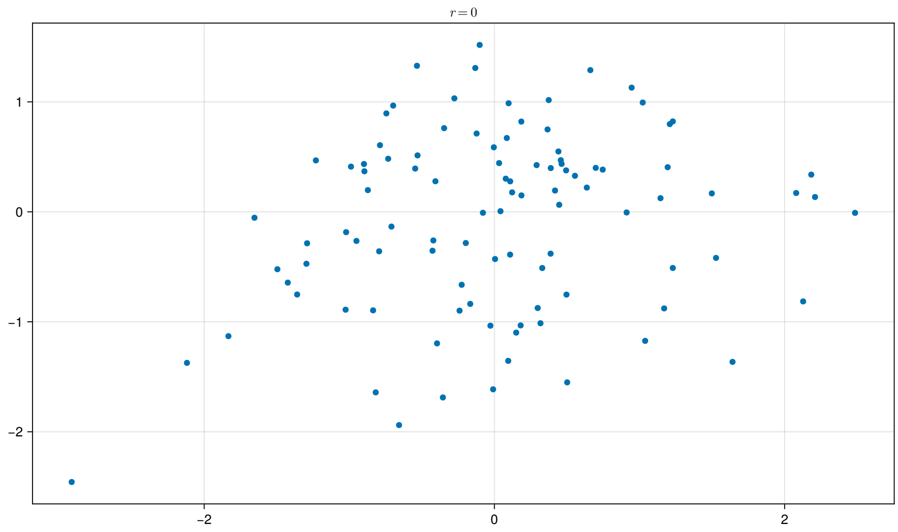
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
r = 0
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

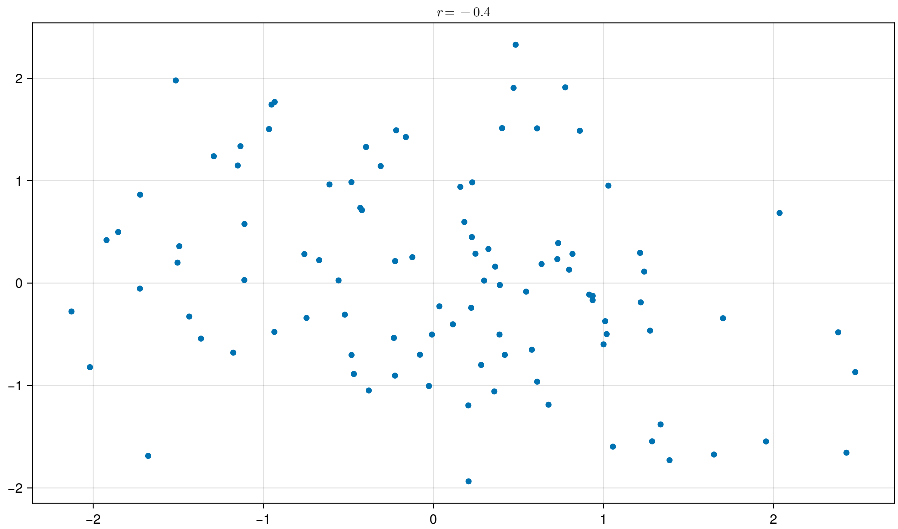
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
```

```
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
r = -0.4
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

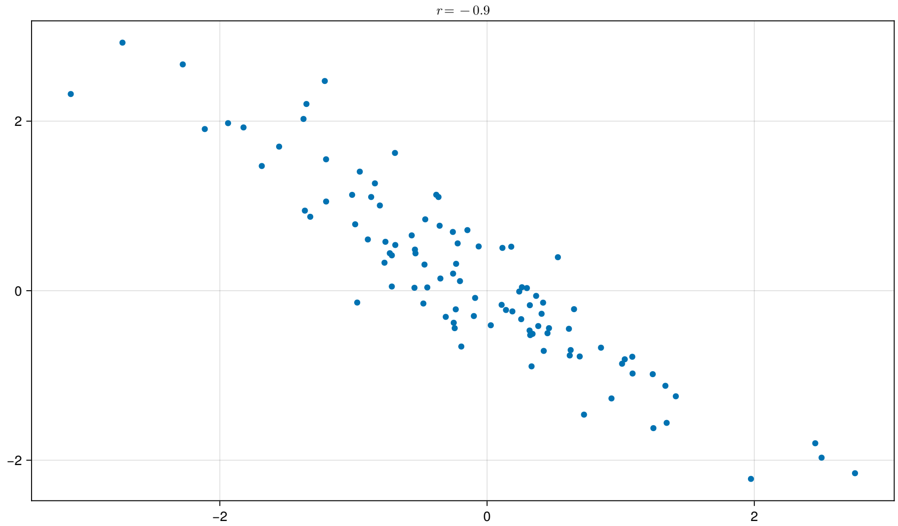
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
r = -0.9
dt = rand(MvNormal([1 r; r 1]), 100)
```

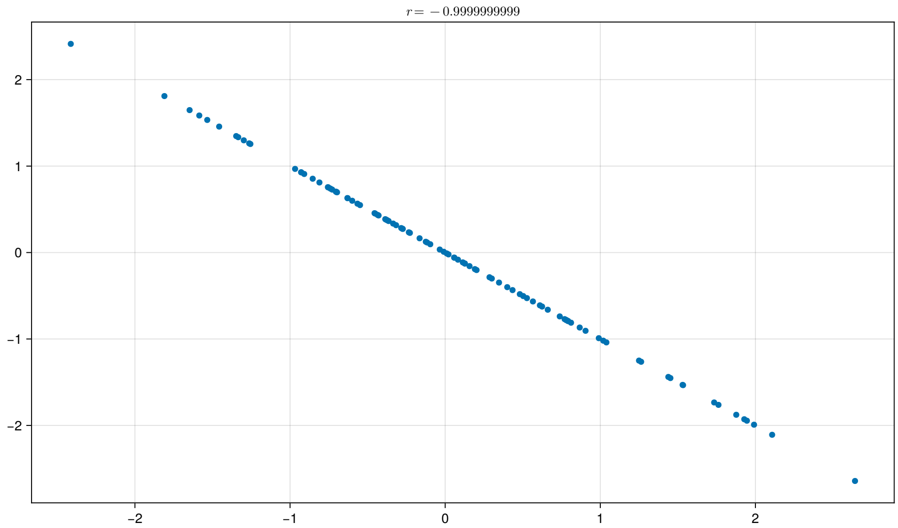
```
scatter(dt, axis = (title = L"r = %$(r)", ))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
r = -1 + 1e-10
dt = rand(MvNormal([1 r; r 1]), 100)
scatter(dt, axis = (title = L"r = %$(r)", ))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.3 皮尔逊相关

15.3.0.1 关系的方向

- **皮尔逊相关**（Pearson Correlation）度量了两个变量之间的*线性关系*的*程度*和*方向*。

- 样本的皮尔逊相关性由字母 r 表示。

$$
\begin{aligned}
r &= \frac{\text{X 和 Y 一起变化的程度}}{\text{X 和 Y 分开变化的程度}} \\
&= \frac{\text{X 和 Y 的协变性}}{\text{X 和 Y 的分散度}}
\end{aligned}
$$

15.3.0.2 离差平方和与离差积和

- **离差平方和** (SS)

$$
\begin{aligned}
SS &= \sum_{i=1}^{n}(X_i - \overline{X.})^2 \\
&= \sum_{i=1}^{n}(X_i - \overline{X.})(X_i - \overline{X.}) \\
&= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \\
&= \sum_{i=1}^{n} X_i X_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i}{n}
\end{aligned}
$$

- **离差积和**（sum of products of deviations, SP）

$$
\begin{aligned}
SP &= \sum_{i=1}^{n}(X_i - \overline{X.})(Y_i - \overline{Y.}) \\
&= \sum_{i=1}^{n}(X_iY_i - \overline{X.}Y_i - X_i\overline{Y.} + \overline{X.Y.}) \\
&= \sum_{i=1}^{n}(X_iY_i) - \sum_{i=1}^{n}(\overline{X.}Y_i) - \sum_{i=1}^{n}(X_i\overline{Y.}) + \sum_{i=1}^{n}(\overline{X.Y.}) \\
&= \sum_{i=1}^{n}(X_iY_i) - \overline{X.}\sum_{i=1}^{n}(Y_i) - \overline{Y.}\sum_{i=1}^{n}(X_i) + \overline{X.Y.}\sum_{i=1}^{n}(1) \\
&= \sum_{i=1}^{n}(X_iY_i) - \frac{\sum_{i=1}^{n}X_i}{n}\sum_{i=1}^{n}Y_i - \frac{\sum_{i=1}^{n}Y_i}{n}\sum_{i=1}^{n}X_i + \frac{\sum_{i=1}^{n}X_i}{n}\frac{\sum_{i=1}^{n}Y_i}{n}\sum_{i=1}^{n}(1) \\
&= \sum_{i=1}^{n}X_iY_i - \frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}
\end{aligned}
$$

15.3.0.3 协方差和方差

- *方差*（s^2）

$$
s^2 = \frac{SS}{df} = \frac{SS}{n-1}
$$

- **协方差**（covariance）

$$
cov(X_i, Y_i) = \frac{SP}{df} = \frac{SP}{n-1}
$$

15.3.0.4 皮尔逊相关的计算

- *皮尔逊相关的公式*

$$
\begin{aligned}
r &= \frac{SP}{\sqrt{SS_xSS_y}} \\
&= \frac{SP/df}{\sqrt{(SS_xSS_y)/df}} \\
&= \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}
\end{aligned}
$$

- 每个 X 值和每个 Y 值可以转换为 z 分数，分别是 z_X 和 z_Y。

$$
\begin{aligned}
r &= \frac{cov(z_X, Z_Y)}{\sqrt{var(z_X)var(z_Y)}} \\
&= cov(z_X, z_Y) \\
&= \frac{SP}{df} \\
&= \frac{\sum_{i=1}^{n}(z_{X_i} - \overline{z_X})(z_{Y_i} - \overline{z_Y})}{df} \\
&= \frac{\sum_{i=1}^{n} z_{X_i} z_{Y_i}}{df} \\
&= \frac{\sum z_X z_Y}{n-1} \\
\rho &= \frac{\sum z_X z_Y}{N}
\end{aligned}
$$

15.3.0.5 手动计算皮尔逊相关

```
X = [0, 10, 4, 8, 8]
Y = [2,  6, 2, 4, 6]
dt = DataFrame(X = X, Y = Y)
transform!(dt, :X => (mean) => :MX)
transform!(dt, :Y => (mean) => :MY)
```

X	Y	MX	MY
0	2	6.0	4.0
10	6	6.0	4.0
4	2	6.0	4.0
8	4	6.0	4.0
8	6	6.0	4.0

```
SSX = sum((dt.X - dt.MX).^2)
SSY = sum((dt.Y - dt.MY).^2)
SP  = sum((dt.X - dt.MX) .* (dt.Y - dt.MY))
SSX, SSY, SP
```

```
(64.0, 16.0, 28.0)
```

```
r = SP / sqrt(SSX * SSY)
```

```
0.875
```

15.3.0.6 用 julia 计算皮尔逊相关

```
cov(X, Y) / sqrt(var(X)var(Y)),
cor(X, Y)
```

```
(0.875, 0.875)
```

15.3.0.7 相关性和数据点的模式

```
@show X1 = X .+  20;
@show X2 = X .*  20;
@show X3 = X .* -20;
```

```
X1 = X .+ 20 = [20, 30, 24, 28, 28]
X2 = X .* 20 = [0, 200, 80, 160, 160]
X3 = X .* -20 = [0, -200, -80, -160, -160]
```

```
[cor(X, Y) for X in [X1, X2, X3]]
```

```
3-element Vector{Float64}:
  0.875
  0.8749999999999999
 -0.8749999999999999
```

15.4 使用和解释皮尔逊相关

15.4.0.1 何时和为何使用相关性

- **预测** (Prediction)
- **效度** (Validity)
- **信度** (Reliability)
- **理论验证** (Theory verification)

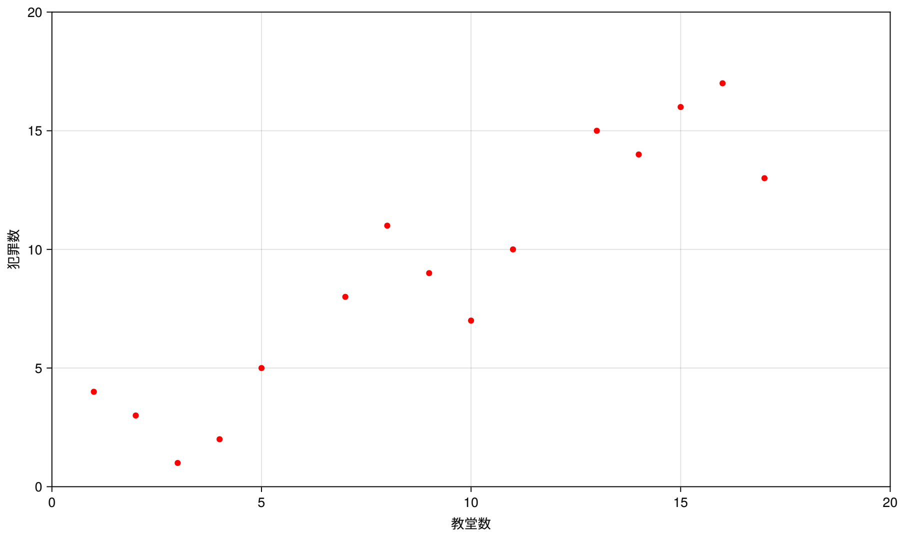
15.4.0.2 解释相关性

相关性只是描述两个变量之间的关系。相关性的值可以受数据中所代表的分数范围的影响很大。一个或两个极端数据点，通常称为*极端值*（outliers），对相关性的值有巨大的影响。在判断一个关系是否“好”时，很容易专注于相关性的*数值*。

15.4.0.3 相关性与因果关系

```
教堂数 = [1:5; 7:11; 13:17]
犯罪数 = [4,  3,  1,  2,  5,
         8, 11,  9,  7, 10,
         15, 14, 16, 17, 13]
scatter(教堂数, 犯罪数, color = :red, axis = (limits = ((0, 20), (0, 20)), xlabel = "教堂数"
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```

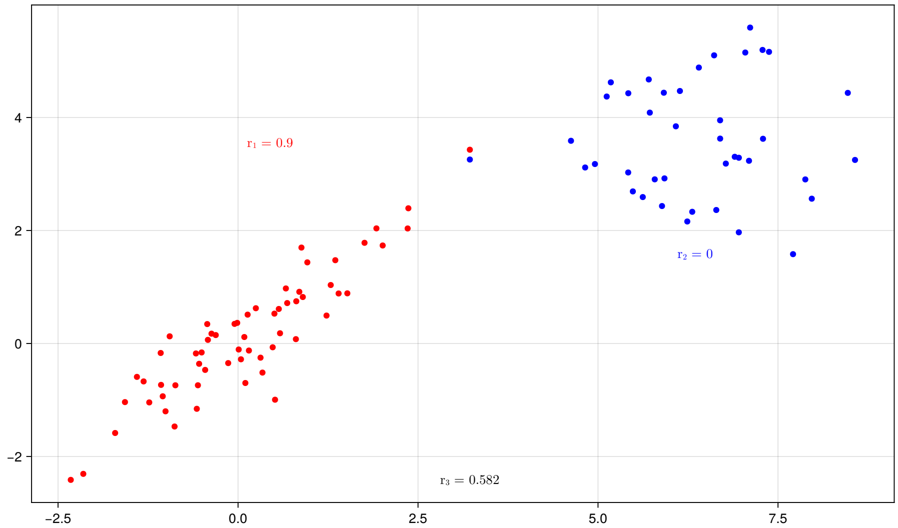


15.4.0.4 相关性和数据范围

```
rr1, rr2 = 0.9, 0
dt1 = rand(MvNormal([1 rr1; rr1 1]), 60)
dt2 = rand(MvNormal([1 rr2; rr2 1]), 40)
rr3 = cor(hcat(dt1, dt2), dims = 2)
dfx = minimum(dt2[1, :]) - maximum(dt1[1, :])
dt2[1, :] = dt2[1, :] .- dfx
dt2[2, :] = dt2[2, :] .+ maximum(dt1[2, :])
fig = Figure()
ax = Axis(fig[1, 1])
scatter!(ax, dt1, color = :red)
text!(ax, mean(dt1[1, :]), maximum(dt1[2, :]);
    text = L"r$_1$ = %$(rr1)", color = :red)
scatter!(ax, dt2, color = :blue)
```

```
text!(ax, mean(dt2[1, :]), minimum(dt2[2, :]);
    text = L"r$_2$ = %$(rr2)", color = :blue, align = (:bottom, :center))
text!(ax,
    maximum(dt1, dims = 2)[1], minimum(dt1[2, :]);
    text = L"r$_3$ = %$(round(rr3[3], digits = 3))",
    align = (:bottom, :center))
fig
```

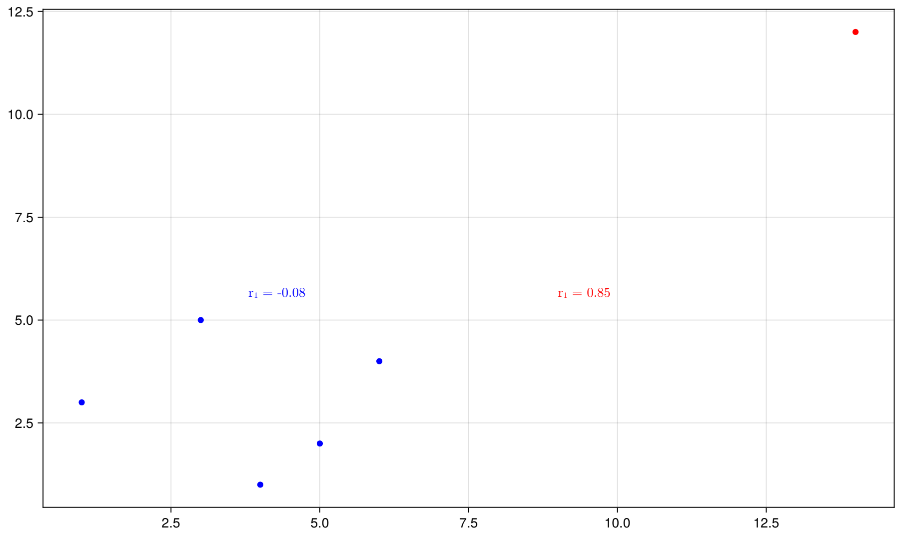
```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.4.0.5 极端值

```
X = [1, 3, 6, 4, 5, 14]
Y = [3, 5, 4, 1, 2, 12]
scatter(X[1:end-1], Y[1:end-1], color = :blue)
text!(mean(X[1:end-1]), maximum(Y[1:end-1]) + 0.5;
    text = L"r$_1$ = %$(round(cor(X[1:end-1], Y[1:end-1]), digits = 2))", color = :
scatter!(X[end], Y[end], color = :red)
text!(X[end] - 5, maximum(Y[1:end-1]) + 0.5;
    text = L"r$_1$ = %$(round(cor(X, Y), digits = 2))", color = :red)
current_figure()
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.4.0.6 相关性和关系的强度

值 r^2 被称为 **决定系数**，因为它衡量了一个变量中可以从与另一个变量的关系中确定的可变性的比例。例如，相关性 $r = 0.80$（或 -0.80）意味着 $r^2 = 0.64$（或 64%）的 Y 得分中的可变性可以从与 X 的关系中预测。

15.4.0.7 向均值回归

- 当两个变量之间的关系不完美相关时，一个变量的极端分数（高或低）倾向于与第二个变量的均值更接近的分数配对。这一事实被称为**回归向均值**（regression toward the mean.）。

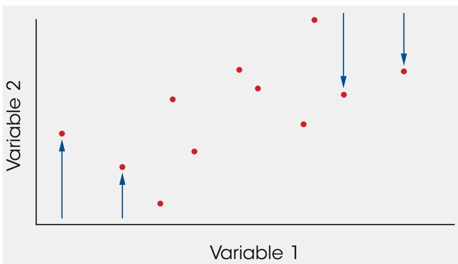


图 15.1: 向均值回归

15.4.0.8 部分相关

有时研究人员可能怀疑两个变量之间的关系受到第三个变量的影响而被扭曲。**部分相关**（Partial Correlations）测量了两个变量之间的关系，同时通过将第三个变量保持不变来

控制第三个变量的影响。

$$r_{XY\cdot Z} = \frac{r_{XY} - (r_{XZ}r_{YZ})}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

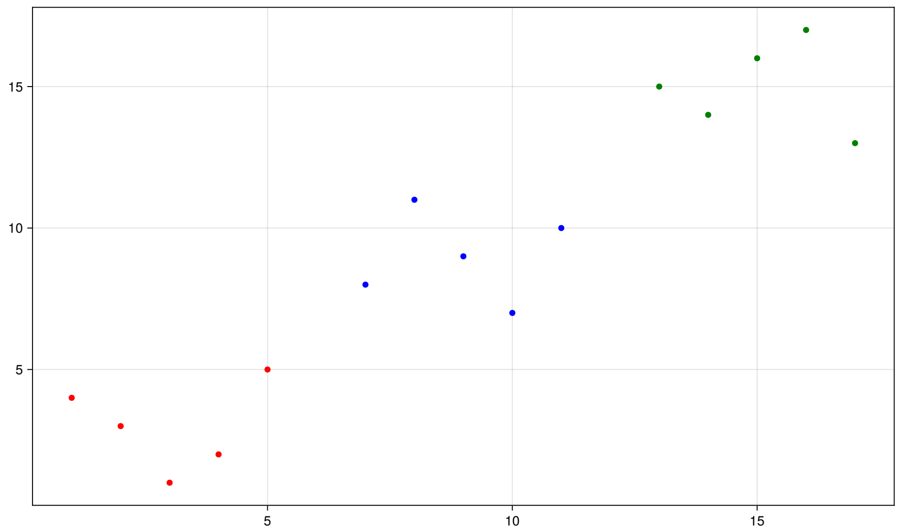
15.4.0.9 部分相关

```
dt = DataFrame(
    城市大小 = repeat(1:3, inner = 5),
    教堂数 = 教堂数, 犯罪数 = 犯罪数
)
```

城市大小	教堂数	犯罪数
1	1	4
1	2	3
1	3	1
1	4	2
1	5	5
2	7	8
2	8	11
2	9	9
2	10	7
2	11	10
3	13	15
3	14	14
3	15	16
3	16	17
3	17	13

```
fig = Figure()
ax = Axis(fig[1, 1])
for (dt, col) in zip(groupby(dt, :城市大小), [:red, :blue, :green])
    scatter!(ax, dt.教堂数, dt.犯罪数, color = col)
end
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.4.0.10 部分相关

```
rxy = cor(dt.教堂数, dt.犯罪数)
rxz = cor(dt.教堂数, dt.城市大小)
ryz = cor(dt.犯罪数, dt.城市大小)
rxy, rxz, ryz
```

```
(0.9230769230769231, 0.9607689228305228, 0.9607689228305228)
```

```
rxyz = (rxy - rxz * ryz) / sqrt((1 - rxz^2) * (1 - ryz^2))
```

```
0.0
```

```
X, Y, Z = dt.教堂数, dt.犯罪数, dt.城市大小
CorrelationTest(X, Y, Z)
```

```
Test for nonzero partial correlation
------------------------------------
Population details:
    parameter of interest:   Partial correlation
    value under h_0:         0.0
    point estimate:          0.0
    95% confidence interval: (-0.5306, 0.5306)

Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           1.0000
```

```
Details:
    number of observations:          15
    number of conditional variables: 1
    t-statistic:                     0.0
    degrees of freedom:              12
```

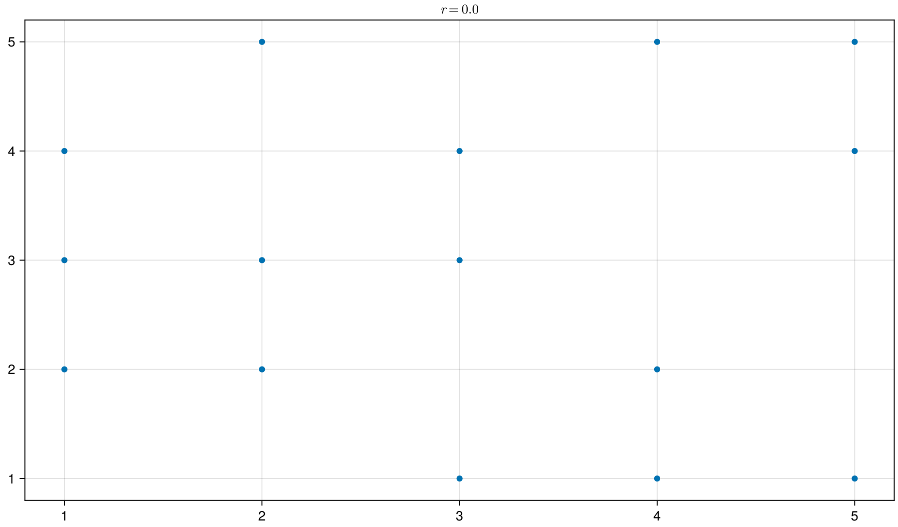
15.4.0.11 部分相关

```
transform!(groupby(dt, :城市大小),
    [:教堂数, :犯罪数] .=> (x -> x .- mean(x)) .=> [:X, :Y])
transform!(dt,
    [:X, :Y] .=> (x -> x .- minimum(x) .+ 1) .=> [:X, :Y])
```

城市大小	教堂数	犯罪数	X	Y
1	1	4	1.0	4.0
1	2	3	2.0	3.0
1	3	1	3.0	1.0
1	4	2	4.0	2.0
1	5	5	5.0	5.0
2	7	8	1.0	2.0
2	8	11	2.0	5.0
2	9	9	3.0	3.0
2	10	7	4.0	1.0
2	11	10	5.0	4.0
3	13	15	1.0	3.0
3	14	14	2.0	2.0
3	15	16	3.0	4.0
3	16	17	4.0	5.0
3	17	13	5.0	1.0

```
scatter(dt.X, dt.Y,
    axis = (title = L"r = %$(cor(dt.X, dt.Y))", ))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```

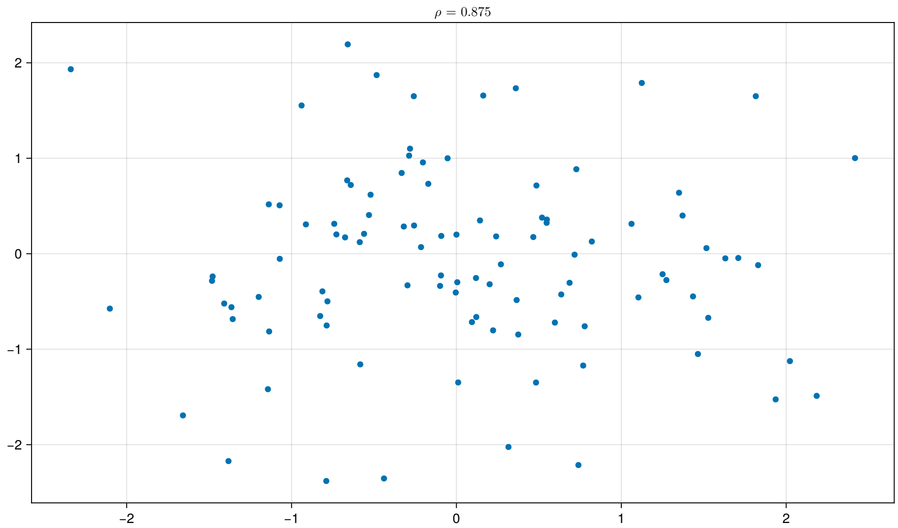


15.5 皮尔逊相关的假设检验

15.5.0.1 样本和总体的相关性

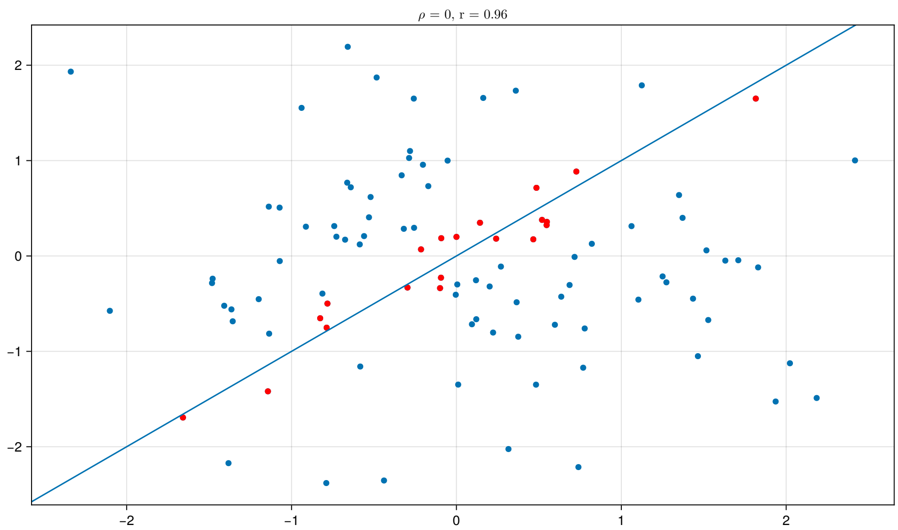
```
ρ = 0
dt = rand(MvNormal([1 ρ; ρ 1]), 100)
scatter(dt, axis = (title = L"$\rho$ = %$(r)",))
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
sp = sortperm(abs.(diff(dt, dims = 1)), dims=2)[1:20]
dts = dt[:, sp]
r = round(cor(dts, dims = 2)[2], digits = 2)
scatter(dt, axis = (title = L"$\rho$ = %$(ρ), r = %$(r)",))
scatter!(dts, color = :red)
ablines!(0, 1)
current_figure()
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



15.5.0.2 假设

- 双尾检验

$$
\begin{aligned}
H_0 &: \rho = 0(\text{总体没有相关性。}) \\
H_0 &: \rho = 0(\text{总体没有相关性。})
\end{aligned}
$$

- 单尾检验

$$
\begin{aligned}
H_0 &: \rho \leq 0(\text{总体相关性不是正的。}) \\
H_0 &: \rho > 0(\text{总体相关性是正的。})
\end{aligned}
$$

15.5.0.3 假设检验

- t 统计量的一般结构

$$t = \frac{\text{样本统计量} - \text{总体参数}}{\text{标准误}}$$

- r 的标准误是

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- 完整的 t 统计量是

$$t = \frac{r - \rho}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

- t 统计量的自由度

$$df = n - 2$$

15.5.0.4 假设检验：一个例子

```
X = [0, 10, 4, 8, 8]
Y = [2,  6, 2, 4, 6]
n = length(X)
r = cor(X, Y)
r, n-2
```

```
(0.875, 3)
```

```
t = r / sqrt((1 - r^2)/(n - 2))
p = 2 * ccdf(TDist(n - 2), t)
t, p
```

```
(3.1304951684997055, 0.052045548743758 74)
```

15.5.0.5 假设检验：一个例子

```
CorrelationTest(X, Y)
```

```
Test for nonzero correlation
----------------------------
Population details:
    parameter of interest:   Correlation
    value under h_0:         0.0
    point estimate:          0.875
    95% confidence interval: (-0.03187, 0.9917)

Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           0.0520

Details:
    number of observations:           5
    number of conditional variables: 0
    t-statistic:                      3.1305
    degrees of freedom:               3
```

```
dt = DataFrame(X = X, Y = Y)
fm = lm(@formula(Y ~ X), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	1.375	0.976281	1.40841	0.253758	-1.73196	4.48196
X	0.4375	0.139754	3.1305	0.0520455	-0.00726039	0.88226

```
pred = [ones(n) X]
fm = fit(LinearModel, pred, Y)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
x1	1.375	0.976281	1.40841	0.253758	-1.73196	4.48196
x2	0.4375	0.139754	3.1305	0.0520455	-0.00726039	0.88226

15.5.0.6 文献中的报告相关性

报告相关性有标准的 APA 格式。报告应包括*样本大小*，*相关性*的计算值，它是否是*统计显著的关系*，*概率水平*，以及所使用的*检验类型*（单尾或双尾）。例如，相关性可以如下报告：

A correlation for the data revealed a significant relationship between amount of education and annual income, r = +.65, n = 30, p < .01, two tails.

数据的相关性显示了教育水平与年收入之间的显著关系， $r = +.65$，$n = 30$，$p < .01$，*双尾。*

有时一项研究可能涉及多个变量，计算了所有可能的变量对之间的相关性。多个相关性的结果最容易在称为*相关性矩阵*（correlation matrix）的表中报告，使用脚注指出哪些相关性是显著的。对于四个变量，存在六种可能的配对，导致六种不同的相关性。比如，一项研究可能测量人们的年收入、教育水平、年龄和智商。报告可以如下陈述：

The analysis examined the relationships among income, amount of education, age, and intelligence for n = 30 participants. The correlations between pairs of variables are reported in Table 1. Significant correlations are noted in the table.

该分析研究了 30 名参与者的收入、教育水平、年龄和智商之间的关系。变量之间的相关性在表 1 中报告。表中标明了显著的相关性。

Table 1. Correlation matrix for income, amount of education, age, and intelligence

表 1. 收入、教育水平、年龄和智商的相关性矩阵

	教育	年龄	智商
收入	+.65*	+.41**	+.27
教育		+.11	+.38**
年龄			-.02

备注：n = 30; * $p < .05$, 双尾; ** $p < .01$, 双尾

15.6 皮尔逊相关的替代方法

15.6.0.1 皮尔逊相关的替代方法

当数据（X 和 Y 值）由一个*区间*或*比率*测量尺度的数值分数组成时，皮尔逊相关性度量了两个变量之间的*线性关系程度*。但是，已经开发了用于*非线性关系*和*其他类型*数据的其他相关性。在本节中，我们将研究三种额外的相关性：*斯皮尔曼相关性*（Spearman correlation），*点二系列相关性*（point-biserial correlation）和 *Phi 相关系数*（phi-coefficient）。正如您将看到的，这三种相关性都可以被视为皮尔逊相关性的特殊应用。

15.6.0.2 斯皮尔曼相关性

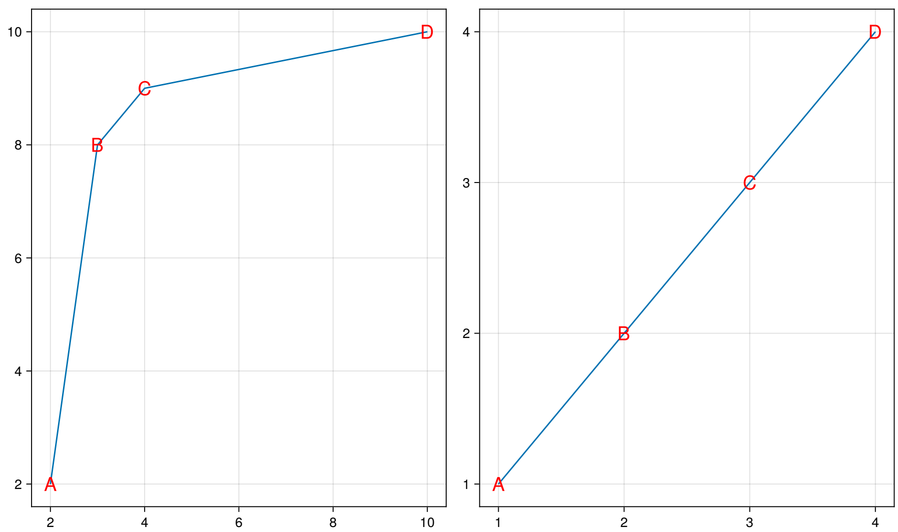
斯皮尔曼相关性度量了两个变量之间的关系，当两者都在有序尺度（排名）上测量时。有两种一般情况下使用斯皮尔曼相关性：当原始数据是*有序的*时，也就是说，X 和 Y 的值是排名。在这种情况下，只需将排名集应用于一组排名上的*皮尔逊相关性*公式。当两个变量一致（*单调*）相关时，它们的*排名是线性相关*的。

```
X = [2, 3, 4, 10]
Y = [2, 8, 9, 10]
X, Y
```

```
([2, 3, 4, 10], [2, 8, 9, 10])
```

```
fig = Figure(size = (1000, 500))
ax1 = Axis(fig[1, 1])
ax2 = Axis(fig[1, 2])
scatterlines!(ax1, X, Y,
    marker = 'A':'D', markersize = 20, markercolor = :red)
scatterlines!(ax2, sortperm(X), sortperm(Y),
    marker = 'A':'D', markersize = 20, markercolor = :red)
fig
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



当研究人员想要测量 X 和 Y 之间的关系的*一致性*，独立于关系的*具体形式*时，使用斯皮尔曼。

15.6.0.3 斯皮尔曼相关性的特殊公式

X_i 和 Y_i 的排名实际上只是一组整数：$1, 2, \cdots, n$。这些整数的均值是该系列的 *中点*，即

$$M = \frac{(n+1)}{2}$$

同样，这些整数的 SS 是

$$SS = \frac{n(n^2-1)}{12}$$

斯皮尔曼相关性的简化公式是

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

其中 D_i 是 X_i 的排名和 Y_i 的排名之间的差异。

15.6.0.4 斯皮尔曼相关性：一个例子

```
X = [ 3,  4, 10, 11, 12]
Y = [12, 10, 11,  9,  2]
RX = sortperm(X)
RY = sortperm(Y)
@show X RX Y RY;
```

```
X = [3, 4, 10, 11, 12]
RX = [1, 2, 3, 4, 5]
Y = [12, 10, 11, 9, 2]
RY = [5, 4, 2, 3, 1]
```

- 使用皮尔逊公式

```
SP  = sum((RX .- mean(RX)) .* (RY .- mean(RY)))
SSX = sum((RX .- mean(RX)).^2)
SSY = sum((RY .- mean(RY)).^2)
r = SP / sqrt(SSX * SSY)
```

```
-0.9
```

- 使用简化公式

```
D = RX - RY
n = length(D)
rs = 1 - 6 * sum(D.^2)/(n*(n^2-1))
```

```
-0.8999999999999999
```

15.6.0.5 用 julia 计算斯皮尔曼相关

- 相关分数

```julia
cor(RX, RY)
```

```
-0.9
```

- 假设检验

```julia
CorrelationTest(RX, RY)
```

```
Test for nonzero correlation
----------------------------
Population details:
    parameter of interest:   Correlation
    value under h_0:         0.0
    point estimate:          -0.9
    95% confidence interval: (-0.9934, -0.0861)

Test summary:
    outcome with 95% confidence: reject h_0
    two-sided p-value:           0.0374

Details:
    number of observations:           5
    number of conditional variables: 0
    t-statistic:                      -3.57624
    degrees of freedom:               3
```

15.6.0.6 排名并列分数

```julia
TS = [3, 3, 5, 6, 6, 6, 12]
sortperm(TS)
```

```
7-element Vector{Int64}:
 1
 2
 3
 4
 5
 6
```

7

15.6.0.7 点二系列相关性和效应大小

点二系列相关性用于衡量两个变量之间的关系，情况是一个变量由常规数值分数组成，而*第二个变量只有两个值*。只有两个值的变量称为*二分变量*或*二值变量*。

15.6.0.8 点二系列相关性和效应大小

- 数据

```
亮环境 = [11,  9,   4, 5, 6,  7, 12, 10]
暗环境 = [ 7, 13, 14, 16, 9, 11, 15, 11]
```

- 宽格式

```
dt = DataFrame(亮环境 = 亮环境, 暗环境 = 暗环境)
```

亮环境	暗环境
11	7
9	13
4	14
5	16
6	9
7	11
12	15
10	11

- 长格式

```
dt = stack(dt, [:亮环境, :暗环境],
    variable_name = "环境类型", value_name = "分数"
)
insertcols!(dt, :条件 => repeat([0, 1], inner = 8))
```

15.6.0.9 点二系列相关性和效应大小

- 相关系数

```
cor(dt.分数, dt.条件)
```

```
0.5803810000880094
```

环境类型	分数	条件
亮环境	11	0
亮环境	9	0
亮环境	4	0
亮环境	5	0
亮环境	6	0
亮环境	7	0
亮环境	12	0
亮环境	10	0
暗环境	7	1
暗环境	13	1
暗环境	14	1
暗环境	16	1
暗环境	9	1
暗环境	11	1
暗环境	15	1
暗环境	11	1

- 相关系数的显著性检验

```
CorrelationTest(dt.分数, dt.条件)
```

```
Test for nonzero correlation
----------------------------
Population details:
    parameter of interest:   Correlation
    value under h_0:         0.0
    point estimate:          0.580381
    95% confidence interval: (0.1189, 0.8357)

Test summary:
    outcome with 95% confidence: reject h_0
    two-sided p-value:           0.0184

Details:
    number of observations:          16
    number of conditional variables: 0
    t-statistic:                     2.66667
    degrees of freedom:              14
```

- 独立样本 t 检验

```
tt = EqualVarianceTTest(暗环境, 亮环境)
```

```
Two sample t-test (equal variance)
----------------------------------
Population details:
    parameter of interest:   Mean difference
```

```
    value under h_0:         0
    point estimate:          4.0
    95% confidence interval: (0.7828, 7.217)

Test summary:
    outcome with 95% confidence: reject h_0
    two-sided p-value:           0.0184

Details:
    number of observations:   [8,8]
    t-statistic:              2.6666666666666665
    degrees of freedom:       14
    empirical standard error: 1.5
```

- 线性回归方程

```
fm = lm(@formula(分数 ~ 环境类型), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	8.0	1.06066	7.54247	2.69949e-6	5.72511	10.2749
环境类型: 暗环境	4.0	1.5	2.66667	0.0184195	0.78282	7.21718

```
fm = lm(@formula(分数 ~ 条件), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	8.0	1.06066	7.54247	2.69949e-6	5.72511	10.2749
条件	4.0	1.5	2.66667	0.0184195	0.78282	7.21718

15.6.0.10 点二系列相关性和效应大小

相关性测量了两个变量之间的关系强度。t 检验评估了关系的*显著性*。t 和 r 之间的关系

$$r^2 = \frac{t^2}{t^2 + df}$$

$$t^2 = \frac{r^2}{(1 - r^2)/df}$$

```
t, df = tt.t, tt.df
r_sqr = t^2 / (t^2 + df)
```

```
0.3368421052631579
```

15.6.0.11 Phi 相关系数

当为每个个体测量的两个变量（X 和 Y）都是二分的时，两个变量之间的相关性被称为 Phi 相关系数。为计算 Phi（Φ），您遵循一个两步过程。(1) 通过为每个变量分配一个 0 和另一个类别的 1，将每个二分变量转换为数值值。(2) 使用转换后的分数使用常规的皮尔逊公式。

15.6.0.12 出生顺序和个性

```
出生顺序 = ["第一", "第三", "独生", "第二", "第四", "第二", "独生", "第三"]
人格特征 = ["内向", "外向", "外向", "外向", "外向", "内向", "内向", "外向"]
dt = DataFrame(出生顺序 = 出生顺序, 人格特征 = 人格特征)
```

出生顺序	人格特征
第一	内向
第三	外向
独生	外向
第二	外向
第四	外向
第二	内向
独生	内向
第三	外向

```
transform!(dt,
    :出生顺序 => ByRow(x -> x ∈ ["第一", "独生"] ? 0 : 1) => :X,
    :人格特征 => ByRow(x -> x == "内向" ? 0 : 1)         => :Y
)
```

出生顺序	人格特征	X	Y
第一	内向	0	0
第三	外向	1	1
独生	外向	0	1
第二	外向	1	1
第四	外向	1	1
第二	内向	1	0
独生	内向	0	0
第三	外向	1	1

```
cor(dt.X, dt.Y)
```

0.4666666666666667

第十六章　回归简介

16.1　本章 Julia 包

```
using Pkg; Pkg.activate(@__DIR__)
using CairoMakie
using Distributions
using StatsReIntro, DataFrames
using GLM
```

16.2　线性方程和回归

16.2.0.1　线性方程

两个变量 X 和 Y 之间的线性关系可以用以下方程表示:

$$Y = bX + a$$

b 称为 *斜率*（slope)。a 称为 Y *轴-截距* (Y-intersept)。

16.2.0.2　线性方程

```
function mvn(; r=1-1e-10, n = 20, intercept = 2)
    r = r == 1 ? 1-1e-10 : r
    dt   = rand(MvNormal([1 r; r 1]), n)
    dt[2, :] =  dt[2, :] .+ intercept
    pred = [ones(size(dt, 2)) dt[1, :]]
    res  = dt[2, :]
    fm = fit(LinearModel, pred, res)
    a, b, ftd = coef(fm)..., predict(fm)
```

```
    r    = cor(pred[:, 2], res)
    a, b, r = [round(x, digits = 2) for x in [a, b, r]]
    return dt, a, b, r, ftd
end
```

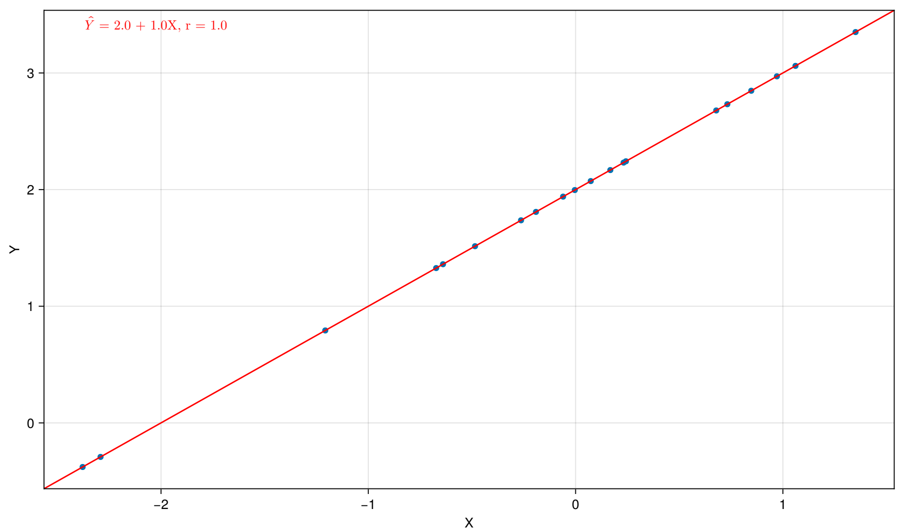
mvn (generic function with 1 method)

```
function mvnp(r; stem = false)
    dt, a, b, r, ftd = mvn(r = r)
    fig = Figure()
    ax = Axis(fig[1,1], xlabel = "X", ylabel = "Y")
    stem && stem!(ax, dt[1, :], dt[2, :], offset = ftd, color = :gray)
    scatter!(ax, dt)
    ablines!(ax, a, b, color  = :red)
    text!(ax, minimum(dt[1, :]), maximum(dt[2, :]); color = :red,
        text = L"$\hat{Y}$ = %$(a) + %$(b)X, r = %$(r)")

    fig
end
```

mvnp (generic function with 1 method)

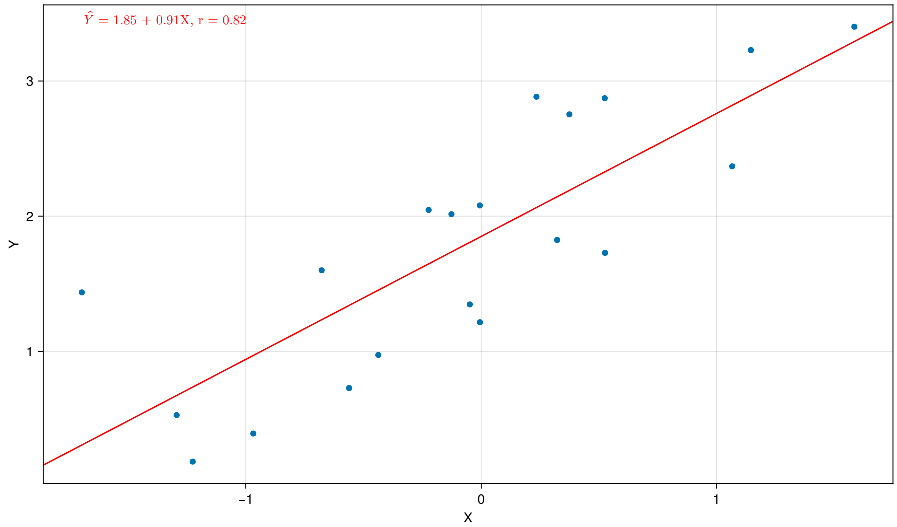
```
mvnp(1)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```




```
mvnp(0.8)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keyword for `Sc
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



这条线标识了关系的*中心*或中心趋势，就像均值描述了一组分数的中心趋势一样。这条线使 X 和 Y 之间的关系更容易看到。这条线可以用于*预测*。

16.2.0.3 回归

用于找到一组数据的最佳拟合直线的统计技术称为**回归**（regression），结果的直线称为**回归线**（regression line）。

16.2.0.4 最小二乘法解

对于数据中的每个 X 值，线性方程确定了线上的 Y 值。这个值是*预测的* Y，称为 $\widehat{Y}$（“Y hat”）。这个预测值与数据中的实际 Y 值之间的距离由以下公式确定：

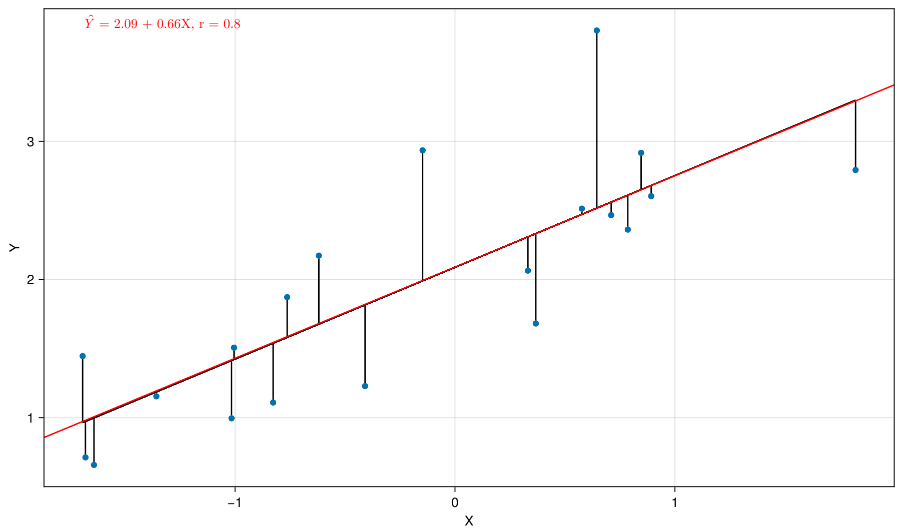
$$\text{距离} = Y_i - \widehat{Y}_i$$

线和数据之间的总平方误差为

$$\text{总平方误差} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

```
mvnp(0.8; stem = true)
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



最小二乘误差解（least-squared-error solution）：我们将最佳拟合线定义为具有最小总平方误差的线。**Y 的回归方程**是线性方程。

$$\hat{Y} = bX + a$$

当常数 b 和 a 如下确定时，该方程将使数据点与线之间的平方误差最小。

- 常数 b 由以下公式确定：

$$b = \frac{SP}{SS_x} = r\frac{s_Y}{s_X}$$

其中，SP 是乘积之和，SS_X 是 X 分数的平方和；s_Y 是 Y 分数的标准差，s_X 是 X 分数的标准差，r 是 X 和 Y 的 Pearson 相关系数。

- 常数 a 由以下公式确定：

$$a = M_Y - bM_X$$

其中，M_Y 和 M_X 分别是所有 Y 分数和 X 分数的均值。

16.2.0.5 最小二乘法解：一个示例

```
X = [ 5, 1, 4,  7,  6, 4, 3, 2]
Y = [10, 4, 5, 11, 15, 6, 5, 0]
```

```
8-element Vector{Int64}:
 10
  4
  5
 11
 15
  6
  5
  0
```

- 手动计算两个系数

```
MX, MY = mean(X), mean(Y)
DevX, DevY = X .- MX, Y .- MY
SSX = sum(DevX .^ 2)
SP  = sum(DevX .* DevY)
b   = SP / SSX
a   = MY - b * MX
b, a
```

```
(2.0, -1.0)
```

- 结果方程是

$$\widehat{Y} = 2X - 1$$

16.2.0.6 最小二乘法解：Julia 计算

```
dt = DataFrame(X = X, Y = Y)
```

X	Y
5	10
1	4
4	5
7	11
6	15
4	6
3	5
2	0

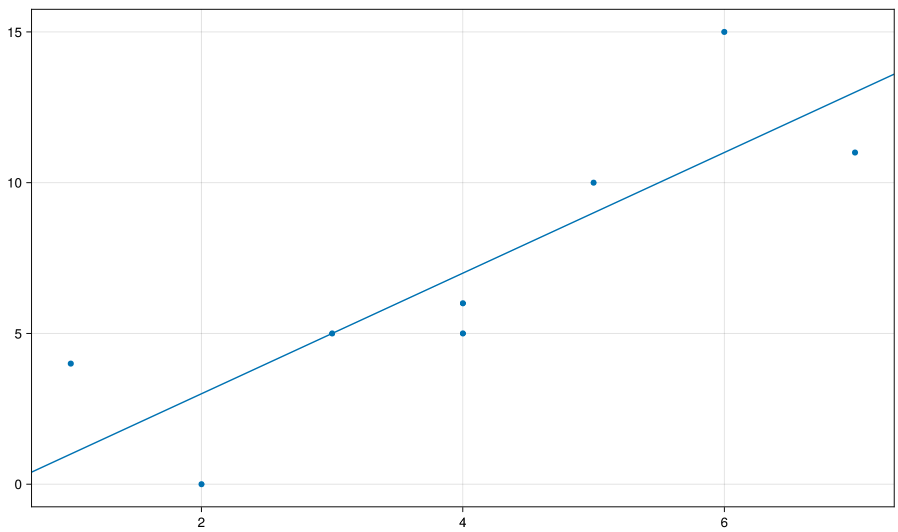
```
fm = lm(@formula(Y ~ X), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	-1.0	2.2599	-0.442498	0.673634	-6.52977	4.52977
X	2.0	0.511766	3.90803	0.00791031	0.747753	3.25225

16.2.0.7 利用回归方程进行预测

```
scatter(X, Y)
ablines!(a, b)
current_figure()
```

```
┌ Warning: Found `resolution` in the theme when creating a `Scene`. The `resolution` keywo
└ @ Makie ~/.julia/packages/Makie/6NLuU/src/scenes.jl:220
```



- 预测值并不完美。
- 不应该使用回归方程来预测超出原始数据范围的 X 值。

16.2.0.8 回归方程的标准化形式

- 回归方程

$$\widehat{Y} = bX + a$$

- 回归方程的标准化形式是

$$\hat{z}_Y = \beta \cdot z_X$$

- 公式中的常数 b 为

$$b = r\frac{s_{z_Y}}{s_{z_X}} = r$$

- 回归方程的标准化形式也可以写为

$$\hat{z}_Y = rz_X$$

16.3 回归方程的显著性

16.3.0.1 方差分析

- 方差的分区

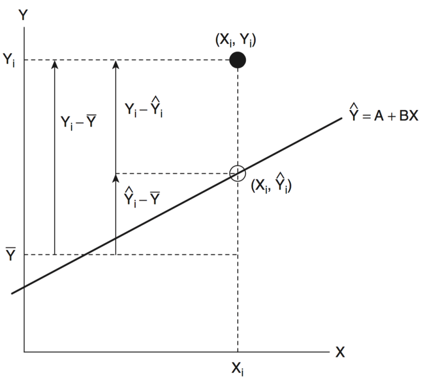


图 16.1: 方差分区

$$\begin{array}{ccccc} \sum_{i=1}^{n}(Y_i - \overline{Y}.)^2 & = & \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 & + & \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}.)^2 \\ \Downarrow & & \Downarrow & & \Downarrow \\ SST & = & SSE & + & SSR \end{array}$$

- 标准误差与相关系数之间的关系

$$
\begin{aligned}
SSR &= r^2 \cdot SST \\
SSE &= (1 - r^2) \cdot SST
\end{aligned}
$$

16.3.0.2 自由度的划分

- 总自由度

$$df_{总} = n - 1$$

- 回归自由度

$$df_{回归} = 1$$

- 残差自由度

$$df_{残差} = n - 2$$

16.3.0.3 回归分析

总体

$$SS_{总体} = \sum_{i=1}^{n} (Y_i - \overline{Y}.)^2$$

$$df_{总体} = n - 1$$

回归

$$SS_{回归} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}.)^2 = r^2 \cdot SST$$

$$df_{回归} = 1$$

$$MS_{回归} = \frac{SS_{回归}}{df_{回归}}$$

残差

$$SS_{残差} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = (1 - r^2) \cdot SST$$

$$df_{残差} = df_{总体} - df_{回归} = n - 2$$

$$MS_{残差} = \frac{SS_{残差}}{df_{残差}}$$

$$F = \frac{MS_{回归}}{MS_{残差}}$$

16.3.0.4 回归与相关性

- 回归方程的零假设

$$H_0 \text{ ：回归方程的斜率 } (b \text{ 或 } \beta) \text{ 为零}$$

- 斜率 b 和相关系数 r 之间的关系

$$b = r \cdot \frac{s_Y}{s_X}$$

- 它等于两个变量在总体中没有关系的零假设。

$$H_0 \text{ ：总体相关性为} \rho = 0$$

- F 比率为

$$MSR = \frac{SSR}{dfR} = \frac{SSR}{1}$$
$$MSE = \frac{SSE}{dfE} = \frac{SSE}{n-2}$$
$$F = \frac{MSR}{MSE}$$

- 用于测试相关性显著性的 t 统计量

$$t = \frac{r-\rho}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

- 从 t 方程中去除总体相关性 ρ，并对 t 统计量进行平方运算以得到相应的 F 比率

$$t^2 = F = \frac{r}{\frac{(1-r^2)}{(n-2)}}$$

- 通过 SS_Y 乘以分子和分母得到

$$t^2 = F = \frac{r^2 \cdot SS_Y}{\frac{(1-r^2)\cdot SS_Y}{(n-2)}} = \frac{MSR}{MSE}$$

16.3.0.5 估计标准误差

- **估计标准误差**提供了回归线上的预测 Y 值与数据中的实际 Y 值之间的标准距离的度量。
- 最终方程为

$$\text{估计标准误差} = \sqrt{MSE} = \sqrt{\frac{SSE}{dfE}}$$

16.3.0.6 回归分析：一个示例

- 计算 $\widehat{Y}$

```
Y_hat = @. b * X + a
```

```
8-element Vector{Float64}:
  9.0
  1.0
  7.0
 13.0
 11.0
  7.0
  5.0
  3.0
```

- 计算平方和

```
SST = sum((Y     .- mean(Y)).^2)
SSR = sum((Y_hat .- mean(Y)).^2)
SSE = sum((Y     .-   Y_hat).^2)
SST, SSR, SSE
```

```
(156.0, 112.0, 44.0)
```

- 相关性和回归

```
r = cor(X, Y)
SSR = r^2 * SST
SSE = (1 - r^2) * SST
SST, SSR, SSE
```

```
(156.0, 112.0, 44.0)
```

- 计算 F 比率

```
dft, dfr, dfe = length(Y) - 1, 1, length(Y) - 2
MSR, MSE = SSR / dfr, SSE / dfe
F_ratio  = MSR / MSE
p_value  = ccdf(FDist(dfr, dfe), F_ratio)
MSE, F_ratio = round(MSE, digits = 2), round(F_ratio, digits = 2)
MSR, MSE, F_ratio, p_value
```

```
(112.0, 7.33, 15.27, 0.007910312203698278)
```

- 汇总表

来源	SS	df	MS	F
回归	112.0	1	112.0	15.27
残差	44.0	6	7.33	
汇总	156.0	7		

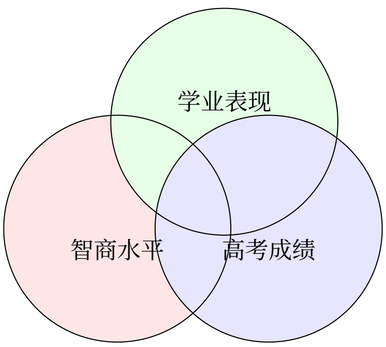
16.3.0.7 回归分析：用 julia 处理

```
fm = lm(@formula(Y ~ X), dt)
ftest(fm.model)
```

```
F-test against the null model:
F-statistic: 15.27 on 8 observations and 1 degrees of freedom, p-value: 0.0079
```

16.4 两个预测变量的多元回归

16.4.0.1 学业表现、智商和高考成绩



16.4.0.2 两个预测变量的多元回归

- 带有两个预测变量的多元回归方程的一般形式是

$$\widehat{Y} = b_1 \cdot X_1 + b_2 \cdot X_2 + a$$

- 标准化形式为

$$\widehat{z_Y} = \beta_1 \cdot z_{X_1} + \beta_2 \cdot z_{X_2} + a$$

16.4.0.3 三个系数的最小二乘法解

$$\begin{aligned}
b_1 &= \frac{SP_{X_1Y} \cdot SS_{X_2} - SP_{X_1X_2} \cdot SP_{X_2Y}}{SS_{X_1} \cdot SS_{X_2} - (SP_{X_1X_2})^2} \\
b_2 &= \frac{SP_{X_2Y} \cdot SS_{X_1} - SP_{X_1X_2} \cdot SP_{X_1Y}}{SS_{X_1} \cdot SS_{X_2} - (SP_{X_1X_2})^2} \\
a &= M_Y - b_1 M_{X_1} - b_2 M_{X_2}
\end{aligned}$$

16.4.0.4 带两个预测变量的回归方程：一个示例

```
X1  = [ 4, 5, 3, 2, 1,  7,  8, 2,  7, 1]
X2  = [10, 6, 7, 4, 3,  5,  8, 4, 10, 3]
Y   = [11, 5, 7, 3, 4, 12, 10, 4,  8, 6]
dt = DataFrame(被试 = ('A':'Z')[1:10],
    Y   = Y, X1  = X1, X2  = X2
)
```

被试	Y	X1	X2
A	11	4	10
B	5	5	6
C	7	3	7
D	3	2	4
E	4	1	3
F	12	7	5
G	10	8	8
H	4	2	4
I	8	7	10
J	6	1	3

- 手动计算

```
X = [ones(nrow(dt)) dt.X1 dt.X2]
```

```
10×3 Matrix{Float64}:
 1.0  4.0  10.0
 1.0  5.0   6.0
 1.0  3.0   7.0
 1.0  2.0   4.0
 1.0  1.0   3.0
 1.0  7.0   5.0
 1.0  8.0   8.0
 1.0  2.0   4.0
 1.0  7.0  10.0
 1.0  1.0   3.0
```

```
bs = inv(X'X) * (X'Y)
```

```
3-element Vector{Float64}:
 2.5517241379310316
 0.6724137931034474
 0.29310344827586177
```

```
bs = X \ Y
```

```
3-element Vector{Float64}:
 2.551724137931034
 0.6724137931034483
 0.2931034482758619
```

- 使用 Julia 计算

```
fm = lm(@formula(Y ~ X1 + X2), dt)
DataFrame(coeftable(fm))
```

Name	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	2.55172	1.94376	1.31278	0.230657	-2.04454	7.14799
X1	0.672414	0.406908	1.65249	0.142414	-0.289772	1.6346
X2	0.293103	0.4005	0.731844	0.488036	-0.653929	1.24014

16.4.0.5 R^2 和残差方差

- 一般公式

$$R^2 = \frac{SSR}{SST}$$

- 对于带有两个预测变量的回归，R^2 可以直接从回归方程中计算如下

$$R^2 = \frac{b_1 SP_{X_1 Y} + b_2 SP_{X_2 Y}}{SST}$$

- 使用第一个公式

```
# Y_hat = predict(fm)
Y_hat = X * bs
SST = sum((Y     .- mean(Y)).^2)
SSR = sum((Y_hat .- mean(Y)).^2)
SSE = sum((Y      -   Y_hat).^2)
SST, SSR, SSE
```

```
(90.0, 50.08620689655171, 39.91379310344828)
```

```
r_sqr = SSR / SST
```

```
0.5565134099616856
```

- 使用第二个公式

```
SPx1y = sum((X1 .- mean(X1)) .* (Y .- mean(Y)))
SPx2y = sum((X2 .- mean(X2)) .* (Y .- mean(Y)))
SST   = sum((Y  .- mean(Y)).^2)
r_sqr = (bs[2] * SPx1y + bs[3] * SPx2y) / SST
```

```
0.5565134099616857
```

- Julia 中的 GLM 包自动计算

```
r_sqr = r2(fm)
```

```
0.5565134099616857
```

16.4.0.6 回归的方差分析表

- 手动计算

```
dft, dfr = length(Y) - 1, 2
dfe = dft - dfr
MSR, MSE = SSR / dfr, SSE / dfe
F_ratio = MSR / MSE
p_value = ccdf(FDist(dfr, dfe), F_ratio)
SSR, MSR, F_ratio, SSE, MSE =
    [round(x, digits = 2) for x in [SSR, MSR, F_ratio, SSE, MSE]]
MSR, MSE, F_ratio, p_value
```

```
(25.04, 5.7, 4.39, 0.05808737129705226)
```

- 汇总表

来源	和方	自由度	均方	F
回归	50.09	2	25.04	4.39
残差	39.91	7	5.7	
汇总	90.0	9		

16.4.0.7 julia 的 GLM 包

- 用 ftest 函数计算

```julia
ftest(fm.model)
```

```
F-test against the null model:
F-statistic: 4.39 on 10 observations and 2 degrees of freedom, p-value: 0.0581
```

- 模型比较

```julia
fmfull = lm(@formula(Y ~ 1 + X1 + X2), dt)
fmnull = lm(@formula(Y ~ 1), dt)
ftest(fmfull.model, fmnull.model)
```

```
F-test: 2 models fitted on 10 observations
```

	DOF	ΔDOF	SSR	ΔSSR	R²	ΔR²	F*	p(>F)
[1]	4		39.9138		0.5565			
[2]	2	-2	90.0000	50.0862	0.0000	-0.5565	4.3920	0.0581

16.4.0.8 每个预测变量的贡献

- 模型解释的方差比例

```julia
r_sqr
```

```
0.5565134099616857
```

- 仅由 X_1 解释的方差比例

```
rx1y = cor(X1, Y)
rx1y_sqr = rx1y ^ 2
```

```
0.5225806451612902
```

- 添加 X_2 作为预测变量后的额外变异性

```
SSx2 = SST * (r_sqr - rx1y_sqr)
```

```
3.053948832035598
```

16.4.0.9 每个预测变量的贡献

- 均方和

```
MSx2 = SSx2 / 1
```

```
3.053948832035598
```

- 评估 X_2 贡献的 F 比率

```
F_ratio  = MSx2 / MSE
```

```
0.5357804968483505
```

- F 比率的显著性

```
ccdf(FDist(1, dfe), F_ratio)
```

```
0.48796326015976643
```

16.4.0.10 用模型比较的方式计算：julia

```
fm0 = lm(@formula(Y ~ 1), dt)
fm1 = lm(@formula(Y ~ 1 + X1), dt)
fm2 = lm(@formula(Y ~ 1 + X1 + X2), dt)
ftest(fm0.model, fm1.model, fm2.model)
```

F-test: 3 models fitted on 10 observations

	DOF	ΔDOF	SSR	ΔSSR	R^2	ΔR^2	F*	p(>F)
[1]	2		90.0000		0.0000			
[2]	3	1	42.9677	-47.0323	0.5226	0.5226	8.7568	0.0182

```
[3]    4     1  39.9138   -3.0539  0.5565  0.0339  0.5356  0.4880
```

16.4.0.11 多元回归和偏相关

- 回归分析评估了考虑了其他预测变量的影响后，每个预测变量的贡献。
- 因此，您可以确定每个预测变量是否单独对关系有贡献，还是仅仅重复了其他变量已经作出的贡献。

第六部分

补充材料

第十七章 投射矩阵与均值和方差

17.1 投射矩阵

一个方型矩阵 $\mathbf{P}$ 被称为投影矩阵 (projection matrix), 若它等于它的平方, 即 $\mathbf{P}^2 = \mathbf{P}$。一个方型矩阵 $\mathbf{P}$ 被称为正交投影矩阵（orthogonal projection matrix），如果对于实矩阵满足 $\mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$，对于复矩阵满足 $\mathbf{P}^2 = \mathbf{P} = \mathbf{P}^*$，其中 $\mathbf{P}^T$ 表示 $\mathbf{P}$ 的转置, 而 $\mathbf{P}^*$ 表示 $\mathbf{P}$ 的伴随或共轭转置。一个不是正交投影矩阵的投影矩阵被称为斜投影矩阵 (oblique projection matrix)。

设 λ 是矩阵 $\mathbf{P}$ 的特征值, 对应的特征向量为 v。有 $\lambda^2 v = \mathbf{P}^2 v = \mathbf{P}v = \lambda v$。因为 $v \neq 0$, 所以必须有 $\lambda^2 = \lambda$。这个方程的解为 $\lambda_1 = 0$ 和 $\lambda_2 = 1$。所以投影矩阵的特征值必须为 0 或 1。

令 $\mathscr{M}$ 为 $\mathfrak{R}^n$ 的 r-维子空间（subspace），$\mathbf{M}_{n\times r}$ 和 $\mathbf{N}_{n\times n-r}$ 分别为 $\mathscr{M}$ 和 $\mathscr{N}^\perp$ 的基 (bases)，则 $\mathscr{M}$ 和 $\mathscr{M}^\perp$ 的正交投射矩阵为 (Meyer, 2000, p. 430)

$$\begin{aligned}\mathbf{P}_{\mathscr{M}} &= \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \\ \mathbf{P}_{\mathscr{M}^\perp} &= \mathbf{N}(\mathbf{N}^T\mathbf{N})^{-1}\mathbf{N}^T\end{aligned}$$

若 $r = 1$，即 $\mathscr{M}$ 是一个线性子空间。则其基是一个 $n \times 1$ 维向量。设其基为 n 个 1 组成的向量，即 $\mathbf{M}_{n\times 1} = \mathbf{O} = (1, 1, \ldots)$，有 $\mathbf{O}^T\mathbf{O} = n$。此时 $\mathscr{M}$ 上的投射矩阵可写为

$$\begin{aligned}\mathbf{P}_{\mathscr{M}} &= \mathbf{O}(\mathbf{O}^T\mathbf{O})^{-1}\mathbf{O}^T \\ &= \frac{\mathbf{O}\mathbf{O}^T}{\mathbf{O}^T\mathbf{O}}\end{aligned}$$

17.2 平均值和方差

设长度为 n 的向量 $X = (x_1, \cdots, x_i, \cdots, x_n)$，其中 x_i 为向量的第 i 个元素。X 的平均值可写为

$$\mu = \frac{\mathbf{O}^T X}{\mathbf{O}^T\mathbf{O}}$$

向量 X 在线性子空间 $\mathscr{M}$ 的正交投射为

$$
\begin{aligned}
\mathbf{P}_{\mathscr{M}}X &= \frac{\mathbf{O}\mathbf{O}^T}{\mathbf{O}^T\mathbf{O}}X \\
&= \mathbf{O}\frac{\mathbf{O}^T X}{\mathbf{O}^T\mathbf{O}} \\
&= \mathbf{O}\mu \\
&= \mu\mathbf{O}
\end{aligned}
$$

即平均值 μ 为向量 X 投射到线性子空间 $\mathscr{M}$ 的投射 $\mathbf{P}_{\mathscr{M}}X$ 的系数。

向量 X 的离差（Deviance）可写成

$$
\begin{aligned}
\mathrm{Dev}(X) &= X - \mathbf{O}\mu \\
&= X - \mathbf{P}_{\mathscr{M}}X \\
&= (\mathbf{I} - \mathbf{P}_{\mathscr{M}})X \\
&= \mathbf{P}_{\mathscr{M}^\perp}X
\end{aligned}
$$

即 X 的离差为向量 X 在向量 O 的垂直子空间 $\mathbf{P}_{\mathscr{M}^\perp}$ 上的投射。

样本方差是投射长度的平方除以自由度 $n-1$，即

$$
\mathrm{Var}(x) = \frac{\left\|\mathbf{P}_{\mathscr{M}^\perp}X\right\|^2}{m-1}
$$

这个 $n-1$ 被称为Bessel 校正。事实上，$n-1$ 分母与这个正交投影的列空间只有 $n-1$ 个维度（$\mathbf{O}$ 张成的线空间的补集）的事实密切相关：在减去均值后，只剩下 $n-1$ 个自由度。

17.3 贝塞尔矫正

17.3.1 正交矩阵的特征

在线性代数中，**正交矩阵**（orthogonal matrix），或称**正交归一矩阵**（orthonormal matrix），是一个实数方阵，其列和行都是正交向量（orthonormal vectors）。一种表达方式是

$$
\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{I}
$$

其中 $\mathbf{Q}^T$ 是 $\mathbf{Q}$ 的转置，$\mathbf{I}$ 是单位矩阵。这导致了等价的特征：如果一个矩阵 $\mathbf{Q}$ 的转置等于其逆矩阵，则矩阵 $\mathbf{Q}$ 是正交的：

$$
\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}
$$

其中 $\mathbf{Q}^{-1}$ 是 $\mathbf{Q}$ 的逆矩阵。正交矩阵 $\mathbf{Q}$ 必然是可逆的（其逆矩阵为 $\mathbf{Q}^{-1}=\mathbf{Q}^T$），酉(unitary）的（$\mathbf{Q}^{-1}=\mathbf{Q}^*$），其中 $\mathbf{Q}^*$ 是 $\mathbf{Q}$ 的共轭转置（Hermitian adjoint），因此在实数范围内是正规的（$\mathbf{Q}^*\mathbf{Q}=\mathbf{Q}\mathbf{Q}^*$）。任何正交矩阵的行列式要么是 $+1$，要么是 -1。作为线性变换，正交矩阵保持向量的内积，因此作为欧几里德空间的等距变换，如旋转、反射或旋转反射。换句话说，它是一个酉变换。

$$\|\mathbf{Q}X\|^2=X^T\mathbf{Q}^T\mathbf{Q}X=X^TX=\|X\|^2$$

$n\times n$ 正交矩阵的集合，在乘法下形成群 $O(n)$，称为正交群。由行列式为 $+1$ 的正交矩阵组成的子群 $SO(n)$ 被称为特殊正交群，它的每个元素都是**特殊正交矩阵**（special orthogonal matrix）。作为线性变换，每个特殊正交矩阵都可以看作是一个旋转。

参考资料：*正交矩阵*

17.3.2 正交矩阵与向量的乘积

令 $\mathbf{X}\sim\mathbf{N}(0,\mathbf{I}_n)$ 为 $\mathbb{R}^n$ 中的一个多元正态向量，$\mathbf{A}\in\mathbb{R}^{n\times n}$ 为一个固定矩阵（假设为可逆/满秩矩阵），则 $\mathbf{AX}\sim\mathbf{N}(0,\mathbf{AA}^\mathbf{T})$。若 $\mathbf{A}$ 为正交矩阵，即 $\mathbf{AA}^\mathbf{T}=\mathbf{I}$ 且 $\det(\mathbf{AA}^\mathbf{T})=1$ 时，则 $\mathbf{AX}$ 的密度与标准正态密度相同，即 $\mathbf{AX}\sim\mathbf{N}(0,\mathbf{I}_n)$。参考资料：*网页链接*

17.3.3 矩阵的特征值的分解

设 $\mathbf{A}$ 是一个具有 n 个线性无关特征向量 q_i（其中 $i=1,\ldots,n$）的 $n\times n$ 方阵。那么 $\mathbf{A}$ 可以分解为：

$$\mathbf{A}=\mathbf{Q}\Lambda\mathbf{Q}^{-1}$$

其中 $\mathbf{Q}$ 是一个 $n\times n$ 方阵，其第 i 列是 $\mathbf{A}$ 的特征向量 $\mathbf{q}_i$，Λ 是对角矩阵，其对角元素是相应的特征值，$\Lambda_{ii}=\lambda_i$。注意，只有可对角化的矩阵才能以这种方式分解。例如，不可对角化的矩阵 $\begin{bmatrix}1&1\\0&1\end{bmatrix}$（这是一个剪切矩阵）不能对角化。

这 n 个特征向量 q_i 通常被归一化，但不一定需要归一化。一组未归一化的 n 个特征向量 $\mathbf{v}_i$ 也可以用作 $\mathbf{Q}$ 的列。这可以通过注意到，在分解中，特征向量在 $\mathbf{Q}^{-1}$ 的存在下会被消除来理解。如果其中一个特征值 λ_i 具有多于一个线性无关的特征向量（即 λ_i 的几何重数大于 1），那么可以选择这些特征值 λ_i 的特征向量为相互正交的；但如果两个特征向量属于不同的特征值，它们可能无法相互正交。一个特殊情况是，如果 $\mathbf{A}$ 是一个正规矩阵(normal matrix)，那么根据**谱定理**（spectral theorem），在正交基 $\mathbf{q}_i$ 中总是可以对角化 $\mathbf{A}$。这种分解可以从特征向量的基本性质中导出：

$$\begin{aligned}\mathbf{Av}&=\Lambda\mathbf{v}\\\mathbf{AQ}&=\mathbf{Q}\Lambda\\\mathbf{A}&=\mathbf{Q}\Lambda\mathbf{Q}^{-1}\end{aligned}$$

具有非零特征值的线性无关特征向量 $\mathbf{q}_i$ 形成了所有可能的 $\mathbf{AX}$ 的基础（不一定是正交的），其中 $\mathbf{X}\in\mathbf{C}^n$，这与相应矩阵变换的像（或范围）相同，也是矩阵 $\mathbf{A}$ 的列空间。具有非零特征值的线性无关特征向量 $\mathbf{q}_i$ 的数量等于矩阵 $\mathbf{A}$ 的秩，也等于相应矩阵变换的像（或范

围）的维度，以及其列空间的维度。带有特征值为零的线性无关特征向量 $\mathbf{q}_i$ 形成了矩阵变换 $\mathbf{A}$ 的零空间（也称为核空间）的基础，这些特征向量可以选择为正交的。

参考资料: *矩阵分解*

17.3.4 投射矩阵的特征分解

根据上述分析，可以把投射矩阵 $\mathbf{P}$ 特征分解写为 $\mathbf{P} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$，其中 $\mathbf{\Lambda}$ 有一个对角元素为 0（比如第一个元素），其余为 1。因此，如果 $\mathbf{X}$ 是一个单元正态随机变量，那么 $\mathbf{QX}$ 也是一个单元正态随机变量，且数学上无法区分 $\mathbf{X}$ 和 $\mathbf{QX}$。

$$\|\mathbf{PQX}\|^2 = \|\mathbf{Q\Lambda X}\|^2 = \|\mathbf{\Lambda X}\|^2 = x_2^2 + \ldots + x_n^2$$

```
using LinearAlgebra, Statistics
function PQ(n)
    X = randn(n)

    O = ones(n)
    P =  I - O * inv(O'O) * O'
    # rank(P)
    λ, Q = eigen(P)
    Λ    = diagm(λ)

    VAR  = var(X) * (n-1)
    PX   = norm(P * X)^2
    PQX  = norm(Λ * X)^2

    return [VAR, PX, PQX]
end
```

```
PQ (generic function with 1 method)
```

```
mean(hcat([PQ(9) for _ in 1:10^5]...), dims = 2)
```

```
3×1 Matrix{Float64}:
 8.006295257816436
 8.006295257816436
 8.003653149922263
```

参考文献

Meyer, C. (2000). *Matrix analysis and applied linear algebra* [Book]. SIAM.