

# R 语言在心理学研究中的应用：从原始数据到可重复的论文手稿

胡传鹏

2023 年 1 月 5 日

# 目录

<b>1</b>	<b>教学内容与课时</b>	<b>2</b>
1.1	目录	2
<b>2</b>	<b>第一讲：为什么要学习 R</b>	<b>5</b>
2.1	R 在心理科学及社会科学中的运用	5
2.2	R 语言使用的示例展示	18
2.3	课程安排	20
2.4	如何学好这门课	24
<b>3</b>	<b>第二讲：如何开始使用 R</b>	<b>26</b>
3.1	要解决的数据分析问题简介	26
3.2	如何安装	28
3.3	如何方便使用？Rstudio 的安装与界面介绍	29
<b>4</b>	<b>第三讲：如何导入数据</b>	<b>31</b>
4.1	路径与工作目录	31
4.2	读取数据	32
4.3	了解 R 里的数据（R 语言中的对象）	34

# Chapter 1

## 教学内容与课时

### 1.1 目录

第一讲：为什么要学习 R（3 学时）

1.1 R 在心理科学及社会科学中的运用

1.2 R 语言使用的示例展示

1.3 课程安排

1.4 如何学好这门课

第二讲：如何开始使用 R:（3 学时）

2.1 要解决的数据分析问题简介 [介绍我们的数据和拟解决的问题，对比 R 和传统 flow]

2.1 如何安装？

2.2 如何方便使用？Rstudio 的安装与界面介绍

第三章：如何导入数据（3 学时）

3.1 路径与工作目录

3.2 读取数据

3.3 了解 R 里的数据（R 语言中的对象）

第四章：如何清理数据一 R 语言编程基础（3 学时）

4.1 R 对象的操控

4.2 逻辑运算

4.3 函数

第五章：如何清理数据二数据的预处理（3 学时）

5.1 数据预处理准备

5.2 数据预处理的基本操作

5.3 数据预处理的进阶操作

第六章：如何探索数据：描述性统计与数据可视化基础（3 学时）

6.1 描述性统计

6.2 ggplot2 的基本使用

6.3 ggplot2 的元素控制

第七章：如何进行基本的数据分析：t-test 和 anova（3 学时）

7.1 语法实现

7.2 分析的流程

第八章：如何进行基本的数据分析：相关与回归（3 学时）

8.1 语法实现

8.2 分析的流程

第九章：如何进行基本的数据分析：中介分析（3 学时）

9.1 语法实现

9.2 分析的流程

第十章：结果稳健吗？使用 Multiverse 比较方法选择对结果的影响（3 学时）

10.1. 多种分析方法的实现

10.2 代码整合与规范化

第十一章：如何得到可发表的图像：数据可视化进阶（3 学时）

11.1 ggplot2 的图层与面板控制

11.2 ggplot2 与其他工具的结合

第十二章：从分析到手稿（3 学时）

12.1 Rmarkdown

12.2 Latex 语法基本介绍

12.3 papaja 工具包的介绍

第十三章：多人协作版本控制:Git? (3 学时)

13.1 版本控制与 git

13.2 多人协作与 git

第十四章：如何帮助我们计划下一个研究? (3 学时)

14.1 计算效应量：Meta-analysis

14.2 计划样本量：Power analysis (模拟)

14.3 计划分析方法：假数据与分析代码 (模拟)

14.4 并行处理

第十五章：如何让导师/合作者完全重复我的分析? (3 学时)

15.1 软件版本记录

15.2 容器技术与 docker 的使用

# Chapter 2

## 第一讲：为什么要学习 R

### 序

第一节课没有太多实质性的内容，但是可以帮助大家为之后的课程做好心理的准备，这一过程是很重要的。从第一节课大家能知道接下来要上课的内容、想要上好这门课需要做什么准备、最后能从这门课获得什么。使用 RStudio 软件完成编辑和转换功能。在 RStudio 中，安装 bookdown 等必要的扩展包。

## 2.1 R 在心理科学及社会科学中的运用

### 2.1.1 数据科学

#### 数据科学是什么？

首先我们讲一下这门课的大背景。虽然我们作为心理学人在心理学院学习这门课，我们会说本课是 R 语言在心理学研究当中的应用。但实际上，R 语言会在一个更广的领域中应用，叫做数据科学，data science。那么什么是 data science 呢？在科学研究中有人认为，科学的革命是经过了几次范式转换的。早期的是“实验”的科学，我们通过做实验，一个一个地去验证假设；随着计算机越来越发达，我们进入了“计算”的范式，通过用各种计算模型模拟的方法，帮助我们去理解世界。但是现在，随着数据越来越多，我们实际上是通过数据驱动的方式进行探索。最近这些年，很多在科技领域尤其是在计算机领域取得的重大突破和进展都是依赖于大量数据的，也就是通过对数据进行提炼从而得到新的发现。比方说最近非常火的 ChatGPT。作为现在全球最火的科技界产品之一，它背后的模型叫做 LM，就 Language Model。这里说的 Language 实际上就是一个大的语言模型，它依靠的就是大量语言材料的训练。

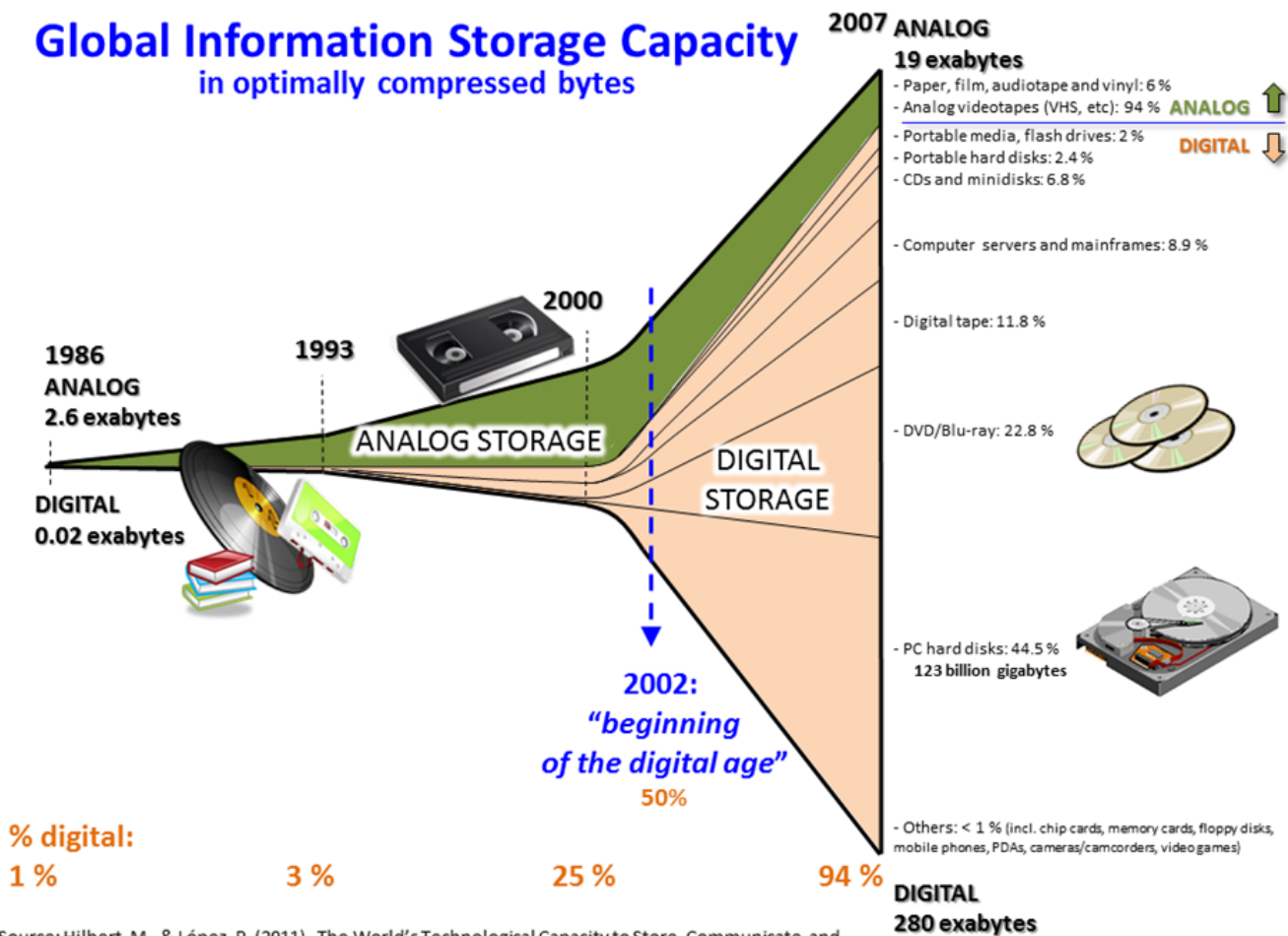
#### 数据科学的内容

大概 10 多年前，数据科学在我读研究生的时候实际上就出现了。最近这两年大家应该对数据科学已经不再陌生了，可以看到在数据科学里面有传统的计算科学，也有数学和统计。它也有具体的应用领域，比方说应用到商业，或者是我们科研领域。但是不论是哪个领域，它都是需要 domain specific language 的，就是说要有这个领域专属的特

殊性知识的。这意味着什么？意味着如果你仅仅懂计算机，那你其实是不能说自己懂 data science 的，如果你仅仅是懂数学和统计，那也不意味这你能解决一个 data science 的问题，必须要将这三方面进行结合。这实际上也是对我们每一个人提出了一个新的要求。

### 2.1.2 数据科学的诞生——数字化时代

为什么会有 data science？其实大家应该能感受到，随着我们电脑的普及，互联网越来越发达，我们产生的数据实际上产生了爆炸式的增长。这里有一个可视化的例子。我们可以看到，在计算机出现之前人类产生的数据是非常少的，而计算机出现之后产生的数据越来越多。



我们也有了越来越先进的仪器，它们所观察到的、产生的数据也是非常大的。去年的这个图片相信很多人在朋友圈都被刷屏过。这是人类所能观察到的一个划时代的新的图像，尽管我们可能不知道它具体的内涵是什么，但是都知道它很酷。



此外，我们国家是互联网普及最高的国家之一，我们现在有百分之七十四（可能现在又更多了）的人都已经开始接入到互联网。这么多的人接入到互联网，所产生的数据可想而知，一定是海量的。所以现在我国很多电商，像淘宝这样的各类购物平台，它们在中国做的是非常好的，这也得益于海量的数据。包括最近像拼多多，听说在美国也是势如破竹，态势很猛。还有像 TikTok，就是字节跳动，前一段时间在美国甚至被封杀了，为什么它被封杀了，因为大家都很喜欢他。就比如中国的这个字节跳动，它有一个特点就是它很好地利用了我国大量的网民产生的海量数据，通过网民不断地使用它们的产品，不断地进行迭代。所以当它能出海的时候，去给海外的用户提供服务的时候，它的迭代已经非常成熟了。当然迭代的过程中需要大量数据的产生和调试，产品才能越来越成熟。





### 数字化对心理学研究的影响

其实很早就有人关注数字化我们心理学的影响了，国内的研究者也经常会提到心理学和大数据。我们这里做一个不完全的概括，主要是这三个方面的显著变化。

#### Big n (sample size)

首先就是样本量很大。现在的数字化平台生产的数据非常大。我们传统的行为学实验可能就是几十、几百的数据，上千已经是很不错的了，上万就比较费劲了。但是如果说我们能从互联网上抓取数据的话，那动辄就是上万甚至是百万级别的。

#### Big v (variables)

不仅仅是样本的体量更大，也意味着收集到的变量更多了。比如我们使用手机，那其实我们产生了非常多的数据。你一天使用多久、点击多少次、点击了什么、在哪个地方、用的是什么 APP，甚至包括你所处的地址，你在地球上的经纬度、当地的气温、湿度这些数据信息。如果大家还记得的话，有一段时间有一个非常火的一个图片。显示的是美国陆军还是海军运动的一个轨迹。因为他们使用了一个运动的 APP。大家就可以看到他们在哪跑步在哪个健身房健身，然后去哪活动等等非常的详细的地理的资料。当然这是属于一个看起来很好玩的东西。那么能不能够用于心理学的研究呢？几年前，北大的王连老师和我们心理学院的韦文琦老师在 human behavior 上面发了一篇文章就是讲不同的地区跟人格之间的关系。所以实际上就是我们能够产生很多的数据，这些数据能进行大量的挖掘，这些挖掘可能是以前你没法想象的。

#### Big t (time)

还有就是时间的跨度比较长。现在很多的 APP 一旦用户开始使用之后就会长期使用，如果能用于收集心理学的数据，就可以在很长的一段时间里记录很多的数据。这对于了解人类的心理和行为的规律来说其实是非常好的。对于发展心理学家来说，有这样的一个例子：有一个做语言发展的一个研究者他从自己的孩子出生开始就一直用视频记录孩子的成长。儿童产生语言目前对于人类来说还是一个不能够完全理解的过程，从语言学上看这是一个很大的跳跃。这位研究者完全记录了孩子生长的过程，积累了大量的数据，并通过对视频数据的全方位分析，得到了以前心理语言学比较少能够得到的东西。


数字化时代的心理学研究


这里有几个例子。比如这个是利用手机里的数据预测人格，我们这里看到的纵轴就是人格的类型，不同的颜色表示使用不同的方法进行预测，横轴是相关系数。可以发现说手机里的一些数据和人格的倾向是密切相关的。






另外，像我们心理学的顶刊 Psychological Science，上面也不定期的有研究是探索我们在数字的视觉中留下的痕迹跟我们的行为之间的关系。比如这篇文章发现我们个体在使用手机的行为上是非常的一致。研究者通过大量的数据是得到了一个比较强的一个结论，尽管这个结论比较简单，但是基于数据量比较大，说服力是比较强的。

Psychological Science



 Open access | 
  | Research article | First published online February 17, 2022

Behavioral Consistency in the Digital Age

[Heather Shaw](#)  , [Paul J. Taylor](#) , [...], and [Stacey M. Conchie](#)   [View all authors and affiliations](#)

[Volume 33, Issue 3](#) | <https://doi.org/10.1177/09567976211040491>

我们也可以利用互联网的平台来收集行为的数据。传统上我们是把被试带到实验室做实验，但现在由于互联网的发达，我们可以进行线上实验。在疫情期间，不少同学可能尝试过 JsPsych 或者巴普洛夫这样的平台，把实验编写好放在网上，并通过链接给被试发放被试费。这样我们在短时间内可以收到比传统实验室更大量的数据。同时，因为我们使用的其实还是传统的实验任务，只要事先验证过在线的实验和实验室实验的可比性，就可以利用互联网在线去收集更大量的数据来研究我们感兴趣的问题。

比如这里列举的实验，去年的 R 课上我们以此作为示例。它就是通过在线的平台收集的数据，收集到了很多变量。它的研究一个目的是为了探测 self regulation，自我控制或者硕士自我调节，不同的测量方法是不是一致的，哪一个能更好地预测生活中的一些行为，比方说这里就是看哪一些与饮食控制有更强的相关。



### 数字化时代心理学研究方式的变化

数字化时代不仅给心理学提供了新的数据，实际上也变革了我们做研究的方式。比如说合作。在疫情以前，研究者建立国际合作主要是通过导师的联系、邀请讲座或者是开会时建立联系。但在疫情期间，因为大家都困在家里同时又都使用很发达的互联网，所以就有研究者直接在社交媒体上发起合作的倡议。比如有的人说我有想法需要在不

同国家收数据，想要有人和我一起收集数据，最后可以一起发表文章。在疫情期间很多研究就是这么展开的，并且因为这样的研究往往样本量比较大，也能发到很好的期刊上去。同时，除了个人的发起，研究者们也成立了更多的学术组织，比方说 Psychological Science accelerator 心理科学加速器，这个组织就是专门组织在全球范围通过互联网进行合作研究。

此外，以前研究者需要去线下开会或者是参加工作坊，现在即便是疫情过去了，大家还是越来越习惯和更多地采用在线的方式进行学术讲座或是工作坊。所以呢互联网实际上是改变了我们心理学的方方面面。

### 2.1.3 为什么要学习 R 语言？{1-why-learn-R}

我个人总结了一些我们学习 R 语言或者说学习 R 而不是 matlab 或者 python 的理由。

首先是一个大的背景。R 语言是一个开源的一个软件，他跟 python、junior 一样，基本绝大部分的基于 R 语言的工具都是开源的免费的，也说你基本上都能够（只要你的互联网是畅通的话）免费得到所有的内容。

第二，它是一个高级的语言，不需要和计算机的硬件直接进行交流，和我们日常的语言差不多。

第三，它有一个强大的 community。因为现在的所有的开源的语言，他依赖于有多少人在使用它、有多少人在不断的进行开发、尤其是谁在开发这些新的东西。对于我们心理学或者是社会科学而言，绝大部分是使用 R 做数据分析。简单来讲，我们一开始作为新手肯定不会去开发什么工具的，就必须要把别人开发的工具拿过来。那谁为我们开发呢？肯定是这个 community 里的人，这就需要我们有一个比较成熟和强大的社区。而 R 语言本身就是由统计学家所开发的，所以它就是为了做数据分析而生的一门语言。同时，在这么多年的发展当中有大量的研究者，尤其是社会科学的研究者不断加入这个 community，从初学者变成使用者最后变成开发者。

从做研究的角度来说，R 可以在这三个方面提供强大的支持。

#### 科学性

它有助于增强计算的可重复性，帮助我们找到更加适合的新的统计的方法。

#### 美观

它可以帮我们得到非常好的图片，有非常精细的图片的细节控制。

#### 实用性

它可以适用于我们做研究的各个阶段，也能够适应这个数字化时代的需求，比如大数据的需求。

#### 2.1.3.1 心理学的可重复性危机

在我们心理学领域从 2011 年开始出现了一个比较大的问题，就是可重复性的问题。大量发表的研究的结果无法被其他的研究者独立重复。那这个问题到底有多严重呢？最有代表性的应该就是这篇文章。在 2015 年的时候，一篇 Science 的文章专门报道了整个心理学领域的可重复性的问题。Science 是跨领域的多学科的一个综合的期刊，能够发表到 Science 这个期刊的文章都是能够引起广泛的兴趣的，也是对整个科学界来说都很重要的。在这个文章中当中 100 个团队重复了 2008 年发表在心理学顶刊上的 100 个研究。他们的分析的结果发现大概只有 36% 的结果是能

够被重复出来的。2015 年这个结果是引起了非常大的震撼也被 nature 评价为 2015 年的年度的十大重要的论文之一。因为这个问题出现，研究者就做了很多的反思。当然我也是被这个问题所深深的震撼，现在还是在一直在寻求能够去做到更加严谨、可重复的透明的这种研究。这个是我在 2016 年的时候跟我们课题组的同学一起写的一篇对可重复性问题的一个介绍和思考，大家有兴趣的话可以看一下。

我这里没有去把这些心理学的计算上可重复的研究拿过来，有人对心理科学在 science 这个期刊上面有公开数据的文章进行了计算的可重复性的检验，也就是说按照研究者描述的方法去做一遍分析，看能不能得到跟研究者一模一样的结果。大家猜一下这个比例大概有多高。做一个区间吧 30% 以下、30% 到 50%、50% 到 80%、还是 80% 到百分之百。约为 30% 以下的举手，30%-50% 呢，50% 到 80% 呢。大家都很乐观啊，80% 我就不问了吧。如果说我们考虑完全能够重复的话，他们在 14 篇文章里面只有 1 篇能够重复，是 1 篇还是 2 篇能够完全重复。然后有的是在作者的协助之下都得不到原来的结果，所以这个问题并没有那么简单。

### 2.1.3.2 利用 R 语言增强计算的可重复性

既然他是这么大的一个问题，那么为什么说 r 语言可以帮助我们解决计算的可重复性呢？首先是说可重复性它是多个层面的。大家可以可以想一下，如果说你的这个结果是可以重复的，那么最简单的一个可重复是什么？就是计算的可输入性对吧。computational reproducibility，这个 computation reproducibility 说的是什么？假如你有一个数据，然后你做了一套分析，你把它报告出来了，我拿到你这个数据，我按照你描述的方法，我能不能得到跟你一模一样的结果。假如说你的计算的过程当中，没有一些随机的生成的过程，全部都是说我们用的这种可以求到解析解的这种这种算法的话，那就意味着我不仅仅要跟你的结论是一致的，而且是在数值上应该是一模一样的。你原来得到比方说  $t$  等于 2.1，那我应该也得到就是  $t$  等于 2.1，或者你得到的是  $f$  等于 10，我应该也得到  $f$  值等于 10，不然就说明什么计算上他是他是不可重复的。

**2.1.3.2.1 记录数据分析的全过程** 你们现在可能绝大部分是研一的同学，那么你们自己有做过数据分析吗？有的举手做过数据分析吗？就是你们做毕业论文应该做过数据分析吧，都不举手，没有做过毕业论文？做个数据分析大家都是用的什么软件，SPSS？那你们都要通过手点击吧，你们现在还记得怎么点击的吗，一步一步的怎么点击过来的。

所以我们实际上按照传统的做数据分析的方法，我们都是用手动的去点击对吧，尤其是前面那一部分。我们不是说你把数据录入到 SPSS 是以后的那部分，（说的是）比方说你用你在问卷上面，用问卷星收一批问卷，那里面可能有一些不太认真的吧，你要把它给删除删除掉对吧？有可能你会删除一些你认为是比较极端的也确实可能是极端的数据对吧？有可能你就是 100 个人里面或者是 300 个人里面你把某一两个（极端数据）你当时看到你就删除掉了，然后你最后认为你得到了一个干净的数据，你把它存起来以 final 或者是以最终数据作为后缀对吧，然后你就会基于那个数据把它打到 SPS 里面对吧。但是如果说你要重复的话从前面到你那个最终数据你能够（重复出来吗？）可能一个月之后你就不一定记得为什么你删除某个数据了。那么这是很普遍的一个（原因）导致我们最后结果无法得到（重复结果）。如果我们用 R 语言编程语言来记录数据分析流程的话，就可以把我们整个数据分析的过程全部记录下来，也就说任何一个步骤出错了我们都可以找到，因为我们代码全部在那里。

我们这门课最后就是从实验软件导出的数据出发，这是最原始的数据，以那个数据开始怎么预处理，一直到我们后面得到可分析的数据，一直到后面做统计分析。我们会展示如何把每一个步骤都用代码记录下来，这样一个好处

就是即便过了一两年之后，即便我们已完全忘记了当时是怎么处理的，但是代码还是可以告诉我们当时怎么做的，这一点就可以帮助我们去保证计算的可重复性。

**2.1.3.2.2 跨机器的一致结果** 另外一点，可以帮助我们达到跨系统或者是跨机器的结果。这个其实是在我们心理学的数据当中比如行为学的数据当中是不是很大的问题。为什么呢？因为我们行为数据的处理涉及到的步骤很少，即便里面包括一些随机化的过程，他的错误不会累积和放大。但是如果你们要去处理一些分析流程更长的一些数据，比方说像 fMRI 的数据，那么你在不同的机器之间的随机性或者浮点数据导致的这个差异，他就会随着你研究的步骤慢慢积累起来，也就是说即便你的这个系统刚开始的时候输了原始数据。经过了不同的系统不同的机器有不同的随机的非常微小的差异，经过一段时间之后也会累积成为很大的一个差异。我们后面会讲如何控制这种随机性导致的这个结果，如果我们使用比较好的使用包括像 pandas 或其他的一些软件，我们实际上是能够达到某种程度上跨机器的一致性的。当然到了一个精度非常高的程度的话，其实就不是我们心理学家能够解决的问题，因为他涉及到一些计算机内部如何去控制浮点的精确度等一些技术细节的问题。但是我们可以怎么样呢？当我们学习了这些编程语言之后，我们能够去把计算机科学家在这方面做的改进纳入到我们的分析当中从而去改进我们自己的分析的结果。

那么在这个记录分析的全过程中我自己有一个例子。就是我们在 2020 年有一篇文章当中公开了数据。但是呢最后其实我们的数据跟统计的结果稍微有一点不一致，有一个读者他读了我们文章之后给我发邮件，我们后来是追溯到内部。之所以出问题就是我们用 Excel 来操作的，然后我们用 Excel 操作的时候不小心删除了几行。所以如果我们使用 R 做全部的数据分析的话，应该就不会出现这个问题，这是说 R 可以做全过程的一个记录。

那么另外一个例子就是我自己 2020 年另外一篇文章当中，因为他的数据和代码全部是公开的，所以有一个叫做 reproducible 的团队，他们去对已经发表的这个文章的结果的可重复性进行评估。他们就对我的这个数据和代码尝试进行了一次重复。那么他们就是说行为的这个结果绝大部分是能够重复出来的，但是有一部分是没有重复出来，原因是因为他安不安装不了那个软件。这不是一个很小的问题啊，我后来花了两年的时间去专门把那个软件打了一个 docker 的包，我们后面会学到的这个 docker 就是为了去让我们能够保证跨机器的一致性。这个看起来是个很简单的问题，你发了一个文章，然后你做了某一个分析，结果别人连你的软件都装不上。那么我们如何去提高自己的研究结果的可重复性，那当然就是我们把软件让它变得更好安装。那个重复的它都是用 python 写的，不是用 R 语言写的。他重复这个 R 语言部分呢，他的 comments 都是非常好的，就是说即便他没有在他自己的工作中没有使用 R，但他也能够很好的读到我的（注释），他也能够知道我在做什么，非常详细。大家以后能够比较好的比较规范的写自己的 R 代码的话，那么同行的反应也是类似的，他会发现你的这个数据，第一个结构非常清晰，第二个你的代码非常清晰易懂，他很很快就知道你在干什么，他也能够很快的重复你的一个结果。至少这样的话，你就保证了自己的结果是非常的稳定的也非常靠谱的。

### 2.1.3.3 R 有更合适的新方法

**2.1.3.3.1 强大的使用群体和不断更新的方法库** 然后另外一个就是说我们在用 R 因为他有一个比较强大的社区使用群体，并且不断有人在开发新的方法，这就意味着我们可以不断的使用更加新的方法，而不是拘泥于我们在课本上学到的比方说常规的这些统计 test、anova、线性回归。我们可以使用一些更加适合我们研究问题的方法。



**2.1.3.3.2 可使用最新的方法** 那么我在这里列举的一个方法，我们后面也可以在课堂中进行实现的，那就是 IJzerman2018 年的 Collabra 这篇文章。那么这篇文章我也是合作者之一，当时也通过互联网我们来合作收集的数据。在这个文章当中，他就使用了机器学习的方法，叫做（条件）随机森林，叫做 conditional random forest。他实际上是在机器学习里面非常常见的一个方法

他的特点就是说即便你只有比较小的数据，你也能够得到比较稳健的一个结果。当然这个小的数据是相对于机器学习里面的小的数据，因为机器学习里面可能动则就是上十万百万的数据。相比而言，我们的数据其实每一个都是很小的，就几百人上千人。所以当拿到这 1,000 多人的数据之后，他想去探索这么多变量之间到底哪些变量之间有一个比较稳定的关系，他就采用了随机森林的方法，最后也发现他感兴趣的那个变量，就是身体的温度和这个社交网络的复杂程度是有关系的。

**2.1.3.3.3 可使用更合适的方法** 另外一个就是比方说我们实验室实验当中非常常用的反应时间，它基本上都是偏态的一个分布，对于这种偏态分布的数据我们到底应该采用什么样的一个模型，到底是用传统的线性模型还是应该用广义的线性模型。如果说我们是使用 r，那我们可以很灵活的使用 r 里面比较新的一些回归模型的包。在这包里面我们可以使用最适合这个模型的，比方说 GLM。我们甚至可以通过模型比较的方式找到哪一个模型是最适合的。也就是说正是因为 r 里面有一个很强大的 community，然后这里面有众多可以选择的 r 的工具包。这样我们就能够不仅仅是使用新的方法，它也可以帮助我们不断的去选出更加适合的方法。

# Reaction Times and other Skewed Distributions: Problems with the Mean and the Median

Guillaume A. Rousselet

Institute of Neuroscience and Psychology, College of Medical, Veterinary and Life Sciences,  
University of Glasgow, Glasgow, UK

Rand R. Wilcox

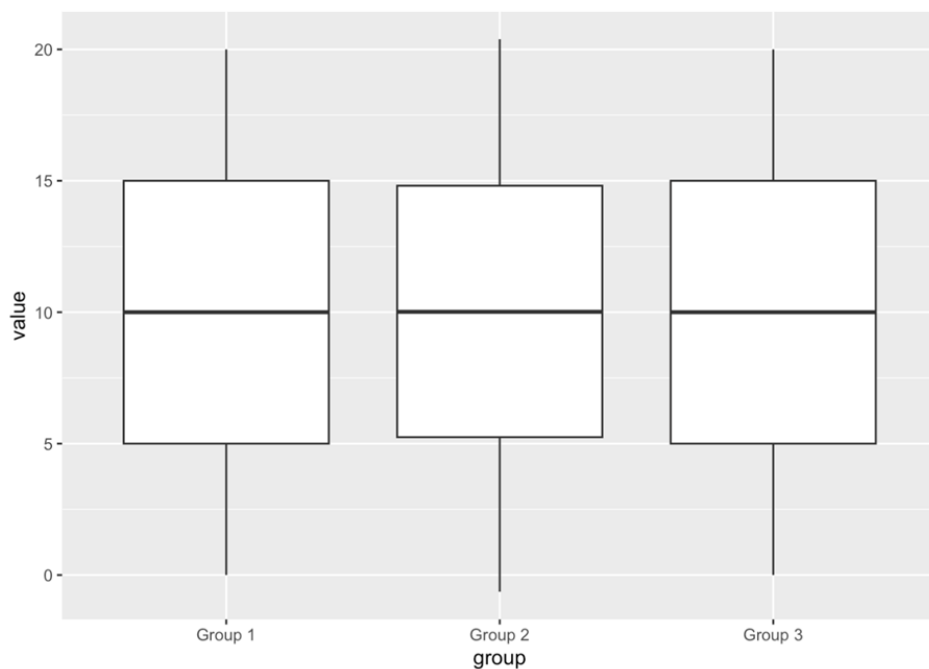
Dept. of Psychology, University of Southern California, Los Angeles, CA, USA

## 2.1.3.3.4 R 更合适可视化 R 语言绘图可调节每个细节

然后从美观的角度来讲，R 画图可以精确的调节每一个细节。这个我们后面再讲可视化的进阶，那一章的时候我们会把那个 ggplot 这个最常见的画图软件里面的每一个细节都掰开讲，这里我们只是稍微展示一下。

然后这个最简单的图，box plot 就很一般了。我们可以把原始数据和 group level 数据结合到一起，然后再把每个被试的数据，把它的分布画出来。

```
library(ggplot2)
ggplot(data, aes(x = group, y = value)) +
  geom_boxplot()
```



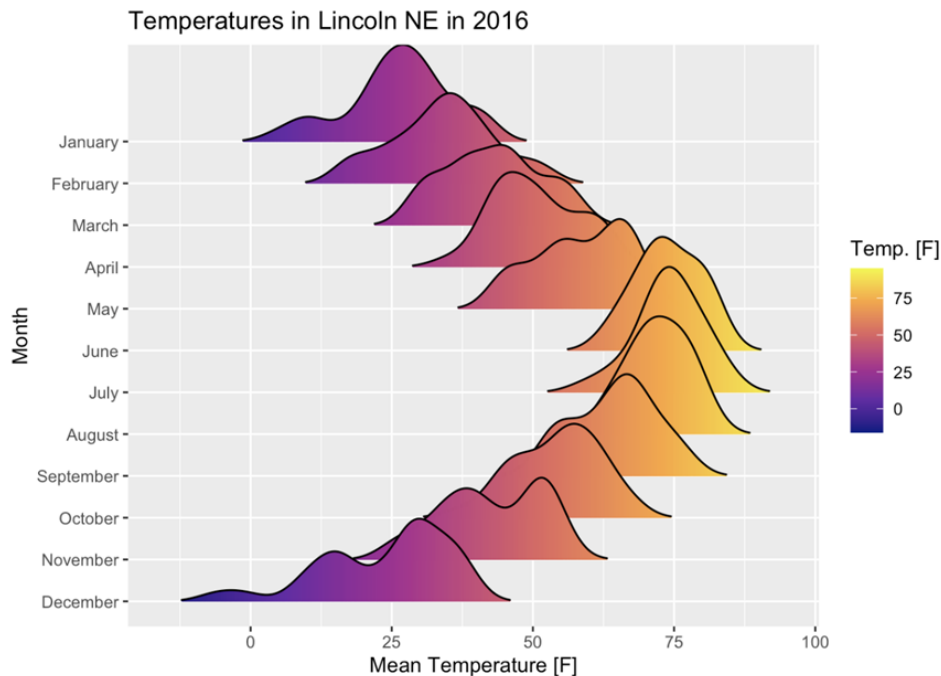
最近几年非常流行的雨云图。当然我们还可以把多个（图）进行叠加，像这种被试类的实验设计我们可以把每个点都连到一起，可以看到在不同之间的一个变化，它是不是完全具有一致性的。然后我们把把这个 box plot 也加上，这样的话我们能够看到极端点。我们同时还把这个分布加上，当然这个分布目前的 值比较高，我们还可以把它调的低一点，就是说让他透明度再低一点，让我们看到这个分布之间的一个叠加。这样我们就可以在一个图上看到非常丰富的信息。

当然还有一个叫做 ggrudges 的一个图，这个上面我们不仅仅看到可视化的效果，还可以直接把值标到上面。这样一个图给我们的信息量就非常大，当然在画图的时候我们不是单纯的追求这个信息量很大，我们要美观。要有足够的信息量同时也能够让大家不会一看到之后就不想看了，而是说看到之后能够立刻 get 到你想要传达一个什么样的想法，这个是很重要的。所以可视化这一点上面说实话我们即便在这个课上有两次课，但是我们只能教大家一些方法，大家最后画图的实际效果要依赖自己的 taste，就是自己的一个口味和不断提升的感觉。



```
# install.packages("ggribes")
library(ggribes)
ggplot(lincoln_weather, aes(x = `Mean Temperature [F]`, y = Month, fill = after_stat(x))) +
  ggribes::geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_fill_viridis_c(name = "Temp. [F]", option = "C") +
  labs(title = "Temperatures in Lincoln NE in 2016")
```

```
## Picking joint bandwidth of 3.37
```



我们也可以把很多 distribution 把它进行叠加，叫做 GG Region，我们也可以画地图。我们可以把一些相关的数据在地图上进行映射，随着我们越来越多的能够得到不同地区的数据，把这些数据映射到地图上的时候，我们就会发现很有价值的信息。这个图是我最近画的一个图，就是我们在分析大团队科学中的样本被试到底是不是真的具有代表性。因为很多做这种就是跨国的研究的研究者总是会 claim 我们的这个研究从几十个国家来的，那么这个数据是能够推广到全人类的。是不是真的如此呢？我们看一下被试在我们这个图当中（的位置），我们就这边是中国的人口的一个（分布）那边就是他的那个被试的群体在不同省份的一个分布图，我们可以看到其实他选取的样本主要就集中在这两个地方，一个是广西一个是上海，其他地方的话其实数据量非常少。

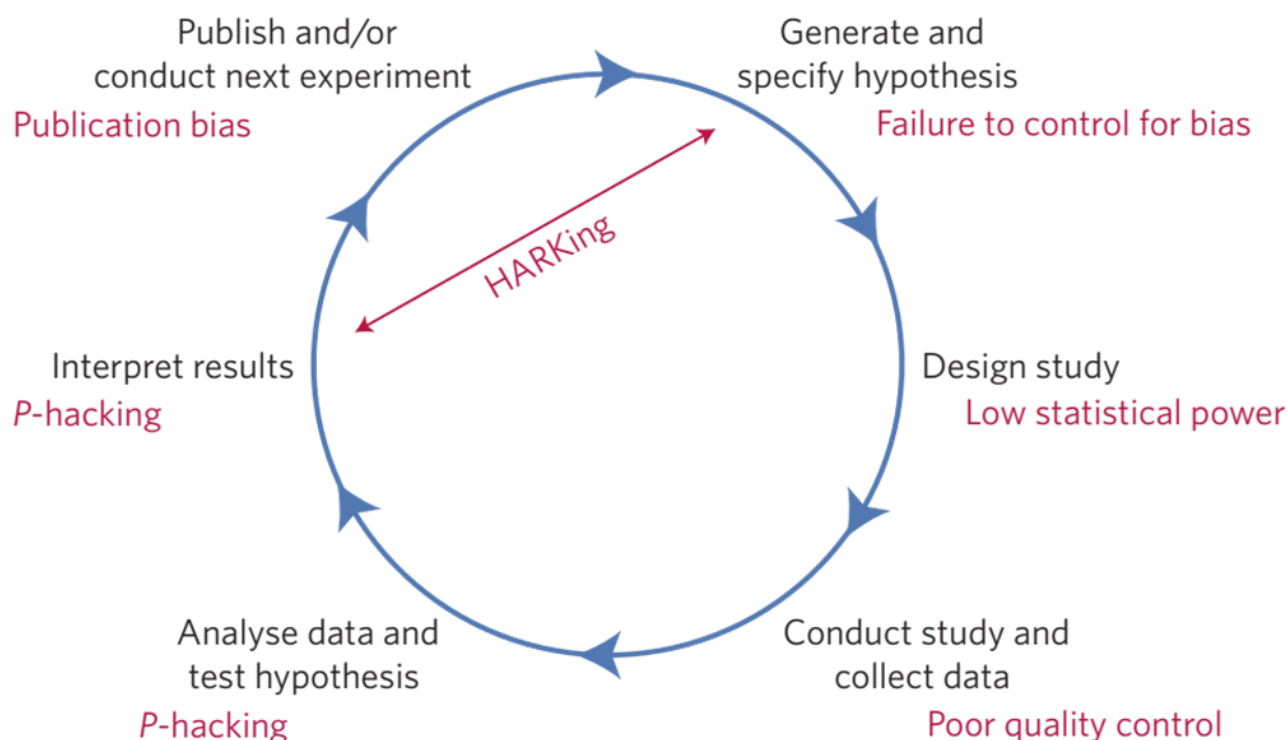
当然我们还可以从其他的维度对他的样本代表性分析，这里主要是展示我们可以把数据映射到地图上面，这样一眼就看到他的数据到底行不行。现在把数据映射到地图上是越来越多的使用在心理学的理念当中。

### 2.1.3.4 R 的实用性

**2.1.3.4.1 R 的实用性之一：适用于数据分析的各个阶段** 几乎科研每个阶段中涉及到的数据处理，均有对应的 R 包。

我们刚刚说的是他能够帮助我们增强计算可重复性、能够帮助我们使用更好的方法、能够帮助我们得到比较漂亮的适合我们目的的可视化。那么还有一个就是 `r` 是非常实用的。目前基本上我们数据分析的各个阶段都可以使用，比如说我们如果使用 `r` 就是整个数据流程一条龙就下来了。我们到课程的后期，会介绍如何把我们的数据分析和我们的代码以及我们的论文的描述文字直接把它整合到一个文档里面，然后生成一个 PDF。这个 PDF 就是可以拿去投的 PDF。

我们从原始的数据，然后再读取数据，然后得到一个比较整洁的数据，然后这个整洁数据我们就可以进行可视化，然后也可以进行统计分析，最后我们就可以到 word 文档，甚至可以得到 PPT，就可以直接给大家展示。当然要掌握这里的每一个流程的话其实是要花很多时间的，但是呢我们可以找到一个对心理学研究者来说最快的（方式）。比方说我们这个课上就会把整个过程中所有心理学的数据整个演示一遍。大家就可以照着这个流程去走，就不需要再去重新探索，这是我们这门课的意义。



那么其实每一个涉及到数据处理的科研阶段，我们前面讲的是数据分析，实际上对应就是我们这个阶段得到的数据之后我们分析数据检验我们的假设。但是在我们的研究中其实我们在实验设计的时候，我们也需要数据分析。我们在设计实验的时候我们需要知道我们大概需要什么样的一个实验设计。比方说我们是做问卷还是说我们要做一个实验，那么做实验被试间设计还是被试内设计，然后被试间设计或者被试内设计之后我们需要多少被试。这里就会有一个大家通常会碰到的一个问题就是 power，叫 statistic power。我们如何去确认我们得到的 power 它是符合我们的目的的，实际上也是可以通过 R 语言的一些相关的工具包来解决。还有比方说我们在 quality control 里面其实也涉及到我们如何去处理这个缺失值。比方说你收了一批数据，80% 的数据都是没有用的，那他到底算不算一个好的研究？或者说我们有少量的缺失值的时候我们应该如何处理缺失值。实际上他既是属于质量控制也是属于数

据分析中的一部分，然后另外一个实用性是可以适用我们现在这个数字化时代的需求。因为我们现在越来越多的大的数据，所以我们要使用一些更加 fashion 的一些方法，机器学习、深度学习什么的。那么 r 语言现在已经有很多这种框架了，如果我们能够掌握 r 的知识以后我们后面去拓展到这些部分相对来说是容易的。因为里面已经有一些比较成熟的框架那么就像我们能够调用 Tidyverse，我们同样也能够去调用这些机器学习的包，只不过我们要真正的合理的使用还是需要去了解了解背后的一些知识，不能盲目的使用 R 语言。

### 2.1.3.4.2 R 的实用性之二：适应数字化时代的需求 广泛使用

他（R）现在也是一个非常广泛使用的语言，大家可以看到 Python 可能是非常火的一个语言，还有像其他的 Javascript 还有各种各样的。像在这个表里面能够做数据分析的大概就是 python 然后 r、matlab，其他的可能都是用于其他功能的一些软件了，也就是说我们如果使用 r 的话它的最后的这个功能也相对来说是比较朴实的。你可以用它做很多事情包括你可以用它建个网站。如果大家有看过我们课题组的网站的话就是 huchuanpeng.com 就是用 r 做的，是我们助教之一胡孟真做的。

#### 强大的社区、众多的教程

我们有一个非常强大的社区。意味着有很多人教你做各种各样的事情，也就是说你如果你想做什么东西，99.9% 的情况下你不需要自己去真的去从原理到实现全部实现全部去做一遍。而是去搜索前人是怎么解决的。比如说你要就要做 meta analysis 你就搜索 meta analysis 视频，然后你能得到很多的教程，这时候你就去找一个好的教程就可以了。或者比方说我们要做混合线性模型，你就搜索一下肯定又会得到很多教程。

#### 课程意义

这门课的意义是什么呢？我们这门课的最主要意义在于让初学者从完全不会到能够不害怕使用 R。这是我们这门课的最大的意义。在心理学研究当中 R 语言也是慢慢地变得越来越流行，像我们在有一些比较新的期刊，他会发表很多教程性的文章，就是专门教大家如何去使用各种各样的新的方法。其中有多篇关于 R 语言的使用的。

## 2.2 R 语言使用的示例展示

我们已经讲完了第一部分关于为什么要学习的内容，希望大家在听完后仍然能够有学习的动力。接下来，我们将简单展示一些使用 R 语言的代表性情况。你现在看到的是其中最具代表性的情况之一，即遇到错误的情况。例如，在使用 R 语言时，实际上碰到错误的概率几乎是百分之百的，而即使你非常熟练，仍然可能出现很低级的错误，比如漏掉一个字符或者反引号。我们将在之后讲到数据类型时介绍不同数据类型需要对应的一些符号。

### 2.2.1 数据清洗

在数据清洗方面，我们一般会使用 dplyr，它是 Data science 里面非常常用的一个功能，需要进行各种数据转换、分组等等操作。但是数据清洗通常是数据分析中耗时最长的一个过程，即使是简单的数据也需要花费相对较长的时间进行清洗。虽然在使用 SPSS 的过程中，我们已经形成了一个非常快速的数据分析思维，但是在使用计算机语言

进行数据分析时，这个过程完全不同，需要花费大量时间进行数据清洗。即使是行为学数据，甚至是简单的反应时数据，也需要进行相对较长时间的数据清洗。

这里面可能有几个图没有截过来，在数据科学中，无论你从事哪个领域，完成一个数据分析项目的时间通常会包括从数据清洗到最终分析以及报告撰写。其中，数据清洗通常会占用至少 60% 的时间，这个过程可能需要反复查看和修改。传统的做研究的方式并不习惯分享数据，因为整理和清洗数据需要很长时间。即使你的文章已经发表，清洗数据所花费的时间也可能被认为是浪费的，尽管有时会带来间接的回报，如其他人可能会重新使用你的数据或发现你的分析的可重复性很高。

### 2.2.2 ggplot2 画图

另一个耗时的任务是画图，尽管这些图看起来很漂亮，但需要不断地调整和修改。有时，研究者会花很长时间去完美地绘制图表，而这些时间可以用来完成其他任务。经常会有研究在这个社交媒体上，他自己比方说做统计分析花了一分钟，然后画图画了两天的时间，就是不断的去调，并不是他不会画图，而是他总是觉得画出来这个图不满意，然后就不不断的调整，不断调整最后发现，时间就没了，并且呢你也会发现呢，就是当你掌握了不同的这个，画图这个方法以后呢你会不断的去想，我能不能找到一个更合适的方法，去对他进行更好的一个可视化，我没有来得及把我自己之前的一个画的图的这个历程贴上去，我刚开始就是最简单的这个，跟我们在，Excel 里面画的那个图是一样的，就是一个直方图上面加一个 error bar，这是我一开始用来画的用 APA 的格式，后来就变成那个带散点的，再后来变成了那个 raincloud，然后就是 raincloud 加 box，然后加这个 distribution，然后最后又回到了这个散点的，加上这个主题，就是 groups 这些。所以其实这个画图的时间需要大家当你熟练了以后也需要适可而止，差不多能够传达你的这个信息就可以停止了，要不然的话，这个画图是提高是没有止境的，有的也可能可以看看有的时候也看到，就比方说一些好的期刊，比如像经济学人他们涉及到数据的时候，像比方说，我们说 Nature、Science 或者 PS 对吧，nature Communications，当他们涉及到数据的图的时候，那都是非常漂亮的，包括他们配色包括的比例各方面，他实际上都是经过了，就是专业的人士进行调试的。那么有的时候其实我们要想要达到这个类似的效果的话，也需要花很长的时间去做这些细节的工作。但假如我们只是按照心理学传统的做法，Excel 的那种一个 bar 图加一个 error bar 的那种非常快基本上几行代码就可以搞定。

### 2.2.3 心理学数据分析与结果汇报

针对于各种各样学科、心理学的这个数据的分析的包，也有心理学的研究者开发的包，例如蔡华杰老师的 PLUS 包，该包专门针对心理学数据分析，并包含很多有用的功能，非常适合心理学学生使用。例如，对于做 T 检验的研究，可以使用该包的功能，我们可以用他的这个 T-test 对吧，它可以将结果输出成为一个简单的三线表格，并且可以直接在命令中加入指令将结果输出成为一个 Word。这个 word 文档里面就有这么一个表。你就可以直接复制打开到你的文档里面，这个是非简单的。他也告诉你，你的零假设是什么，我们这个单样本的 t 检验，那么他告诉你这个假设就是双侧的，然后这个叫做均值，他不等于 700，那么类似的可能我们还有其他的，比方说我们用这种配对样本 t 检验对吧，他同样的也可以得到这种非常适合我们输出的这个结果，而且他最近也把这个 face back 进去了，所以我们大家可以看到这里报告的信息肯定是比 spss 更加全面的。

## 2.2.4 Regression

那么另外就是关于这个 regression 这里面我们其实没有使用 BruceR，如果我们使用 BruceR 的话，他有一些专门的回归模型，然后我们这里有个简单的回归模型，把回归模型直接做一个这样的输出，Bruce 也是可以的。还有这种，这里应该也是一个简单的回归模型，还有就是我们也可以使用这个 SEM，这个地方的话，我们可能要使用一个新的包，就不是 bruceR，当然 bruceR 他也把那个，如果没有记错的话，把 SPSS 那个 process 整合进去了，这里我还没有尝试过，可能我们后面可以一起来探索。

## 2.3 课程安排

我们这学期的课程的有三个原则：第一个就是**即学即用**，我希望我们在课堂上教的这些代码，大家都能够在自己的数据分析中使用而不是我们在这里学了一遍之后，然后自己需要重新去学，或者这里面学的代码，都是大家用不上的，这个我们尽量避免。因为大家时间都有限，如果能够帮助大家省一点时间的话我觉得或者少走一点弯路的话，减轻大家的心理负担。

第二个就是**在做的过程中学习**，比如说我们当我们前面的两节课，第一节课我们是介绍对吧，第二节课我们就是教大家怎么去安装，然后去了解这个里面的各种各样的功能，以及数据导入，以及各种各样的情况，基本上第二节课以后，我们就开始就直接给大家讲这个代码。我们需要在做的过程中学习。

然后第三个就是**逆向学习**，逆向学习就是你先做，你先能够在哪里面实现这个东西，然后我在这个我演示的过程中，我用一个命令，比方说就是 t test 或者 f test 对吧，我能够得到这样的结果，大家首先说的就是你在你自己电脑上，或者说你在这个云计算平台上你能够实现这个功能，你能把这个代码抄下来，然后呢得到跟我一模一样的结果。然后你后面再去理解他，他到底为什么这样或者说如何去使用他，你先会做然后再去理解，这个对于学习代码来说，我觉得是很好的一个一个做法。因为如果你看到你一两本书之后，你其实没有写过几行代码这个实际上是非常的浪费时间的一个事情，尤其现在大家的这个时间也比较紧张对吧，然后呢我们的一个目标就是想要去压平这个学习曲线，因为如果大家知道学习，就是说以前大家认为这个 R 语言的学习曲线是非常陡峭的，就刚开始特别难，你要进步的话很慢，就是很难，也花很多时间。我们希望他尽量的快一点，然后可以慢慢的去后面的不断的学习。

那么参考教程的话，英文版的有这个 (Naverro, Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6.1), <https://learningstatisticswithr-bookdown.netlify.app>)，然后中文版的话，有一个叫做王敏杰老师的《数据科学中的 r 语言》，他实际上是一个公开的一个教程，那么大家可以把它当做参考书，因为这个书里面讲到了非常多的一些知识点。但是我们不会按照他的这个知识点进行讲解。然后另外一个呢是叫做 Tidiverse，张继星老师，给大家展示一下，另外一本书，这个我最近也买了。那他实际上跟我们的课堂的这个内容契合度是非常高的，因为 Tidiverse 就是我们最常使用的一个工具包，那么课程的安排的话我们就是，我们课程的内容的话其实就基本上跟我们的研究密切相关，因为大家都是研究生对吧，所以我们是希望：首先是跟数据分析这部分密切相关，然后等我把这部分解决之后，我们看到后面进阶的话，在设计实验的时候怎么把这个部分做好，那么在数据分析的过程中的话，我们就会有有一个完整的流程：我们从原始数据到清理，然后到数据的探索，统一分析，然后分析完了之后统一推断，然后把它结果去验证，然后撰写一个报告。

我自己本人也非常喜欢可重复性和开放科学，因为我觉得他是科研中很重要的一个方面，我们后面也整合了一些有关我们如何跟他人进行协作的内容，帮助他人一起共同的合作来完成任务。还有如何保证我们的可重复性，如何采用一些更加先进的计算机技术，来帮助我们更好的保证我们的这个计算的可重复性。然后以及如何直接能够从代码到数据，生成一个 PDF 文件或者 word 文档，能够生成一个直接可以提交的一个版本。

### 2.3.1 课程大纲

我们不会按照传统的这种介绍 R 的方法：先介绍 R 里面有什么数据，有什么对象，有什么语法规则，这些通通都不介绍。我们怎么介绍？就是直接数据拿过来，我们要怎么用，怎么分析，第一步做什么，那我们看一下这个，那么第一步是什么？我们从第二章来，就是安装，假如我们现在有了一个数据，我们需要用 r 语言，我们第一个问题就是把 r 安装到自己的电脑上面去，如果有使用的是 windows 系统的话，最好把自己的用户名改成英文或者是拼音，不要用中文作为用户名，因为 R 语言它是英语的使用者开发的，所以说他的这个编码的话可能会对中文不是很友好。我们之前碰到过一个问题，就是当我们使用中文作用户名的时候，可能没有办法画图。

然后我们会后面帮助大家解决一些安装中的问题，然后的话我们会介绍安装之后的各个界面的介绍。然后呢我们也会介绍如何更加方便的使用 r，如果我们是使用原生态的 r 的话他会非常的难用，他相当于只是给我们提供了一个引擎就是做 R 计算的一个引擎。我们还要需要有一个写代码的一个界面，更加方便我们进行交互，那么我们会使用 Rstudio，那目前的话，Rstudio 应该是使用最广泛的，当然越来越多的人也使用 Microsoft 的那个 VS code，但是我们使用 R 语言，他其实也是挺方便的。

然后我们也介绍他的这个各个界面以及，如何开始我们的这个数据分析，那么现在是从最基础的开始，软件的安装，然后呢告诉大家我们现在拿到一个数据之后，我们如何去把这个数据导入进去，那么在 SPSS 里面大家知道，就是一个 File 导入，那么在里面呢我们除了这种方式以外，我们有其他方式，我们也会进行批量的导入，我们不仅要导入一个数据，我们可能需要导入多个数据，导入完数据之后呢我们就有东西了。

我们的数据进入到里面了，那么我们从现在就开始认识 R 里面的这些数据，就是 R 里面的这个对象，然后我们会以这个为基础来讲解 R 里面的各个对象有什么特点。

首先我们要讲解数据和处理，因为这个过程有点复杂，所以我们会分成两次来讲解。我们会首先讲解单个对象在 R 中如何操作，一个一个的我们在 Rstudio 中都是可以看到的，然后讲解一些运算规则。这时候，我们就把 R 的一些基础知识放在这里了。然后我们讲解每个对象的特点之后，会展开介绍一些函数和规则等等。然后我们就可以开始自己的处理工作了。我们会按照 tidyverse 的风格进行基本的操作，或者我们叫做比较管道的操作。在这个过程中，我们可能需要给大家演示一下。如果大家能理解了，我们就可以进入下一个阶段。但是，如果有些同学对编程没有基础的话，第四章和第五章可能会比较难理解。所以我们会多停留一段时间，确保大家能够理解。

然后我们已经导入了数据，接下来就可以进行一些探索。看一下我们数据长什么样，他有什么样的一个模式，特点，我们应该对他进行什么样的一些分析，这就是去了解我们的数据。在传统的心理统计学和 SPSS 中，不知道老师有没有要求大家查看原始数据。实际上在我们用做数据分析的时候，我们一定要去看，并且是在开始的时多的去看原始数据他到底是什么样子。我们不能只看统计指标，我们一定要看数据他到底是什么样子，这样的话就避免让我们发生一些非常非常基本的错误。

那么这个数据探索的话其实他包括两个部分。一个部分就是我们会去给他进行描述。第二个部分就是对他进行一些基础的一些可视化，所以探索数据部分其实已经包含了我们两个部分的一个知识点：一个是描述统计，另一部分就是最粗糙的可视化。因为这个时候，我们只需要自己看就行了，我们不需要给别人看，我们也不需要去把它做的非常精美，所以这个时候是最基础的这个可视化，然后的话我们就会接下来几章我们就会告诉大家如何用 r 语言实现大家常用的一些统计分析，因为这个可能是我们心理学常用的。

所以我们先展示一下，然后他的这个分析的流程是什么样的。这几个部分呢，其实应该是可以并行展开的，但是我们没法进行并行展开，所以我们只能依次的介绍。

然后呢，介绍完了之后呢我们会介绍一个目前来说在国际心理学界比较流行的一个做法，或者说大家在推荐的一个做法，就是看我们的这个结果是不是稳健的，那么这个时候比如说我们同样一批数据，我们采用多个方法来分析它，最后我们得到的结论是不是一致的，这个称之为 Multiverse。在心理科学进展上最近有一个文章，介绍这个 Multiverse 的。大家有兴趣话可以去看一下。

那么到这个时候的话，其实我们基本上统计分析就基本上，如果我们只做传统的这个数据分析的话，我们就其实就已经做完了。那么我们接下来就是，我们汇报结果，这个时候我们如何得到一个可发表的图像，或者可发表的是一个插图，那么这个时候我们会进一步的讲我们如何进行拼图，如何操作每一个每一个这个元素然后，如何把多个图拼到一起，让他更加的美观，以及他的比例等等。然后呢我们会讲一下，这个叫做文学编程 literature program 就是，实际上他就是把我们前面讲到——就是在这一章以前，我们讲代码，所有的都是直接代码，那么我们在这个地方的话我们就开始介绍如何把 r 代码和文字进行混合，也就是说我们同一个文档里面既有文字描述也有代码，这个代码还是可以运行的对吧，他生成的图片还可以直接插到这个文档里面。

我们就要介绍这个 Rmarkdown, 它是用 LaTeX 的语法, 所以我们会介绍一些最基本的 LaTeX 的语法, 帮助大家进行排版, 还有公式的撰写等等, 然后到这里的话基本上就是我们主流的部分, 一个人干活的话基本上就差不多了。但是我们知道, 现在的已经都不是一个人干活对吧? 所以我们要经常跟他人进行协作, 所以我们后面介绍就是如何跟合作者或者导师进行协作那么这个地方就涉及到两个, 一个是版本控制——其实一个人干活的话你版本控制也很重要, 为什么呢? 因为你可能前后代码有很多迭代对吧, 你有的时候可能删除一个功能, 但是你后来发现, 这个删除功能它是有用的, 但是如果你直接删除的话, 你就找不到了, 那么如何我们能够找到以前的版本, 这个其实是, 当大家写代码越来越多的时候很重要的, 另外一个就是多人协作对吧, 通过这个 Github 来进行多个人进行同时, 完成一个数据分析, 你完成这个 t test, 我完成 f test, 最后我们两个人, 进行一起合并, 这样的话我们能够更加快速, 更加有效的进行协调。到第 14 章的话, 我们相当于是一个研究做完了, 然后呢, 我们可以计划下一个研究, 这里我们会介绍一些心理学常用的方法, 包括 meta-analysis 元分析。元分析实际上就是我们把多个研究的这个效应量进行综合, 综合起来之后我们就有一个用来估计样本量的一个东西, 那么我们接下来就是做这个 power analysis。以及我们如何在没有任何数据的情况之下, 我们就可以把自己的分析数据的代码先写出来, 那么这就涉及到这个假的数据对吧, fake data 或者叫做 simulator data。

如果我们涉及到这个数据非常复杂的话, 我们可能会考虑如何进行变形处理。就是说我们的数据量比较大的话, 如果我们做模拟的话有可能他要花很长的时间。但是我们的 CPU 有多个核对吧, 我们是不是能够把多个核都用上来, 让他更加快速有效的进行这个模拟, 当然这个可能是我们看情况啊, 如果说有时间的话, 最后大家自己去搜索就可以了, 因为对于传统的行为学的这个研究来说, 除非我们用贝叶斯的这个混合线性模型, 要不然的话基本上都是在可预期的时间内能够完成的。然后最后的话, 我们基本上完成了整个分析。我们全部做完了对吧, 然后可能, 假如

我毕业了，那么师弟师妹他们如何来重复我的研究？或者导师如何能够重复我的研究？那我可能过了一段时间之后，这个版本有不断的更新的，那么别人如何能够还能够重复我的研究？这个时候我们如何把这个版本，这个包给记录下来发布？现在比较主流的这种计算机的方法就是容器技术。用这个帮助我们更好的达到这个 computational reproduce ability，当然这个地方我可能只会做点介绍，大家如果真的要完成这一点的话，可能也需要花点时间去琢磨。因为 R 语言他有一些包可能会涉及到比较系统里面比较底层的软件的交互，所以呢他可能——就是说依赖于你使用的什么包，如果你使用的包是比较主流的，比较常规的，那可能，你把它打到 docker 里面是很容易的，如果你涉及到的包可能是比较新的，有可能会他有很复杂的一个底层的一个编译的话，那你有可能就没有那么容易打包。

所以如果我们再回过头看一下，就是说我们整个教学大纲就是按照我们一个研究生，拿到数据之后会做什么一步一步的往下走，所以我们不会说很系统的去介绍 R 语言里的知识，我们就是碰到了什么我们就讲什么，那么我们会用什么数据呢？我们的数据就是我之前的公开数据。这里面包括了 2019 年在 scientific data 发表了一个数据，那么它的数据呢以问卷为主，那么还有原研究，我们后面会把这个发给大家。另外一个实验数据，也是我 2020 年发表的一篇文章，我们就会从这些数据开始，然后一步一步的进行处理。

那么当然如果大家说 OK，我想就在这个学习的过程中，顺便把我的数据给处理了，那我就把我自己的数据拿过来可不可以呢？也是可以的，你可以尝试，然后你碰到什么错误。如果说需要问的话呢也可以问，但是不能太超纲了。

我们这里面大家可以注意到一点，就是我们对统计方法的讲解没有涉及到很深，因为我们是 R 语言课对吧，所以是以 R 语言本身的操作为主，就是比方说大家这个里面设计的很复杂，我有一个链式中介模型要试试，或者什么调节模型。如果说你碰到那种技术性的问题，我们可能不会回答因为这个完全超纲了，但你说我碰到了一个数据导入的问题，这个没有问题，我们可以解答。因为现在是说一个是知识一个是操作，我们教的主要是操作，因为实际上如果你懂了，这种比较复杂的 SEM 的知识的话呢，你要操作起来很简单，可能就一两行代码就能够解决了或者复杂一点就是十几行代码就能够解决了，但是，如何设置这个模型，如何解读模型的输出，这个不属于我们 R 课的内容，而是属于你的统计知识的内容，大家清楚了吗？所以我们这个课，你可以把自己数据拿过来，但是如果你的数据非常 specific 分析的话，我们可能不在我们这个课程所覆盖的范围之内，但是可以帮助大家去把前面这个部分解决。你可能原来完全不知道如何使用 R，那么我们现在就教你怎么把这个数据导入怎么使用，做基本的数据分析。

那么到后面，你可能很容易就找到适合你的这个数据分析的工具包，那么你需要去阅读这个工具包相关的知识点，去把它用于自己的数据。有时候我们会以这个原始数据——两个原始数据，一个是问卷数据，一个是反应时的数据为基础，然后一步一步的去走完整的过程，那么中间大家可能会有一些要抄写代码的地方，这个就是大家可以在课堂上实现。那么我们可能会需要，为了保证大家能够不断的进行这个讨论和反馈，我们需要分组。那我们可以加到群里面，然后后面需要进行分组，大家小组长最好能够负责带一些插线板之类的，这样的话就保证每个人都能够有电，在上课的时候有电来进行后续的操作。

我的想法就是最开始的时候我们就分组，那么大家可以比方说根据自己的兴趣，或者是我没有兴趣的话，我们可以由助教来进行随机的分组，那么分组之后呢大家比方说碰到了问题大家相互讨论，我不知道大家以前在做数据分析的时候是不是相互讨论，但是在写代码的时候，相互讨论是一个非常有用的一个东西，为什么呢？可能你写了一段代码之后你犯了一个非常小的错误，然后导致你的代码没办法运行，你自己看不出来为什么，因为你太熟悉了这个代码，但是别人来看的话可能一眼就看出来了。

所以呢大家可能就可以形成小组，在上课的时候练习的阶段，大家就可以相互来进行这个相互进行检查，同时呢我



们也有几位助教，我们分组讨论的时候呢，助教就会跟我们一起来解决。

我们现在有一个助教在线上没来，蔡镇在这里，胡孟真在这里，田彩玉也在，还有我们这边有两个本科生同学也非常积极的愿意协助，一个是柏松石，然后一个是孙禾嘉，然后他们可能主要是不一定就是能够解决大家的问题啊，就是有些其他的问题是他们帮我负责，大家不要觉得有压力啊。本科生来当助教了是吧，太可怕了。

我们会后面会有一个我们课程的 PPT，完成之后会把它上传到这个 Github，为什么我有这么多助教呢？有一部分助教是会把我们上课的录音进行一个，就讲的这个东西进行一个文字的整理，然后让他更加的符合逻辑，把它变成文字稿。

有些同学或者老师，特别想学，但是好像没有时间过来，你觉得参加小组讨论也太麻烦了，后面就可以看这个文字稿。或者这个书也可以电子书也也是可以的。这大概就是我们这个课程的安排。

这个课程的整个大纲啊，相比上学期的话是进行了重大调整的，也不知道效果怎么样，大家跟我就一起一起体验。如果大家感觉也很糟糕的话，也可以继续跟我反馈，我后面在不断的改进。我们去年的结构是跟今年不一样的，去年结构就是按照传统的计算机编程语言的方法来进行讲解：先讲他的基础知识，再讲如何应用到心理学，但是讲他的知识的时候就会比较枯燥，很多同学就听着听着可能就觉得听不懂了，有可能就心情就到了谷底，坚持不下去了。今年我们希望大家尽量不要出现这种情况。在这个过程中，不断的把自己的每一个小的问题解决，如果大家看我们这个大纲的话，我们每一讲都是一个问题的，所以我们每一节课希望能够解决一个问题，那么在这个解决问题中的话，有的时候我 PPT 讲的内容不一定能够完全解决你的问题，因为你有可能在你的电脑上去解决这个问题的过程中你碰到了新的问题，那么这个时候我们有很重要就是小组讨论。我和助教来一起帮助大家解答疑惑，这个基本上就是我们这个课的安排。

### 2.3.2 成绩分配

那么选课的同学，可能就会涉及到这个成绩的问题了，那么我们一般来说的话就是出勤，只要大家来的话，就有 10% 的分数，第二个的话就是小作业，我们会给大家三个小的作业，然后帮助大家去练习课堂上的一些问题，那么小的作业的话最好大家是单独完成，或者是通过小组讨论之后，自己每个人提交一个代码。

最后的话会有一个大的作业，我们一开始不就分组了吗，每个人会需要选用比方说公开数据或者我们给的数据，做一个完整的这个数据分析，然后生成一个 PDF，并且能够在课堂上进行汇报。最后的话，大家需要提交的作业就是一个 r 代码文件.rmd 和一个生成的 PDF 文件。

那么我们会检查这个代码文件是不是真的能够生成这个 PDF，如果生成不了的话那可能会稍微扣一点分，然后还有就是大家需要以小组的形式汇报一次，这是大作业。

## 2.4 如何学好这门课

如何学好这门课？我觉得最关键的就是不要害怕这个课程，这课程其实没有大家想象的那么难，为什么呢？因为我们不需要成为 r 开发者，我们也不需要懂很多很多 r 里面代码，我们只需要成为一个合格的调包侠就可以了。别人开发的包我们能够合理的使用，这就是我们这门课的目的，大家如果说有志于要成为大神，比方说我要开发某一个

包，或者是我后面所有的东西都用 R 解决，包括给自己建一个网站什么的，这个东西不在我们的这个课程范围之内，也说大家能够用 r 语言，第一个就是消除 r 语言的畏难心理。能够用 r 的这个生态里面的一些包帮助大家解决问题，这就是我们这门课要达到的一个，让大家入门的一个目标那么当然我们也希望通过这个入门让更多的人，能够成为长期的这个 R 语言的使用者，能够长期的在这个 community 里面活跃。甚至有一天能够为他人答疑解惑，或者是明年的时候来给我当助教。

那么我们这个 r 语言，虽然说不直接讲统计知识，但是还是会涉及到一些统计知识的，那比方说我们讲的主要是操作，那么一个这个函数或者是一个代码，他的运算出来这个结果如何解读，其实需要一些统计知识的，如果你完全不懂的话，那你肯定是很困难的。

第二个就是敢于尝试，一个就是说大家一定要去不断的犯错，另外一个就是说你可以借助一些比较新的一些东西来帮助你解决这个问题，我们有些助教就是，非常熟练的使用 ChatGPT。假如你没办法使用 ChatGpt，是不是还可以使用 Bing 的相关功能？它也非常的强大。我们最好是用英文进行搜索，因为相对而言，英文的这个社区要比中文的社区要强大很多，把你的问题用英文描述出来，然后在 Bing 里面搜索，99.9% 的这个情况之下你都能够找到答案，如果你的问题描述是正确的话。你也可以很简单就把那个报错最关键的地方放到这个搜索框里面，基本上你也能够找到答案。R 语言他再怎么统计，他也本质上也是一门计算机的编程语所以他还是有编程的这些成分在里面，那么编程他最大的一个特点就是你要需要去勇敢的尝试，不断的犯错，犯的错越多的话，学习的越多，尤其在课上。假如说你犯了 200 个错误 300 个错误，那么后面自己分析数据的时候可能还是会碰到这些错误，但你就知道怎么解决了。所以只有多犯错才能多学习。

然后的话，我们需要以计算机的这种思维方式思考，就是计算机他是非常非常机械的，你告诉他什么，你输入什么指令，他就给你什么结果。所以如果出了错误，一定是我们的指令或者是哪个地方出错了，所以很多时候你需要把这个逻辑想清楚。特别是你有一个比较复杂的分析问题，你要想我第一步做什么，第二步做什么，第一步的这个输入变量是什么，输出变量是什么，这个输出的变量，他能不能进入到下一步作为输入。简单的这种机械的操作的思考，能够帮助我们去使用这个 R 语言。

然后第三个当然就是我们在小组的讨论中一定要相互的帮助，我有一些朋友，他是学习计算机本科的，他们的快乐之一就是上课的时候，相互 debug，相互帮助，因为写代码写多了之后你会发现他是一个非常有及时反馈的一个事情，你输入一行代码，他给你一个正确反应非常开心，输一行代码错了，能够解决了也非常开心。但有一种情况对于初学者来说，就是你犯了一个错误，结果你两三天解决不了，就非常的头疼。那这种情况的话，如果是有人能够帮你的话呢，就能够其实极大的促进正反馈。还有一个就是，遇到比较复杂的代码的话他确实有可能起作用了，但是你可能也不知道为什么，他不起作用你也不知道为什么，但是我们需要尽量减少这种情况。

其实当你越来越熟悉某一个编程软件之后，大部分的错误你都是知道为什么，那不知道大家有什么问题吗？

我去年给大家说的時候，会有一个很强烈的情绪的变化，就是学 r 的基础知识会慢慢情绪会变得很差，因为你发现自己又听不懂，然后经常犯错。今年的话我希望大家稍微平缓一点，这样那种不断波动的一个状态比较好。如果没有问题的话我们这节课就到这里。

# Chapter 3

## 第二讲：如何开始使用 R

### 前言

数据分析的出发点是解决问题，即需要有一个问题来指导数据分析的过程。对于社会科技或心理学背景的同学来说，面对像工科、计算机或统计等领域的同学时，可能会感到技术不够强，但是随着经验的积累，他们会发现技术必须为问题服务。因此，关键在于了解要解决的问题，拥有问题的意识和思维，这对从事研究工作的同学来说尤为重要。提出一个好的问题需要不断思考与导师讨论。最终，学习 R 语言或其他语言，或做一系列数据分析工作的目的都是为了回答一个特定的问题。问题的重要性本身决定了数据分析的价值。

除了研究问题以外，实践问题也是数据分析的重要应用。例如，解决实际的问题，如交通分流，设计道路等，以及辅助决策，特别是在疫情期间，行为科学在公共决策中的作用受到了重视。通过大量的行为研究数据，可以帮助政府或决策者做出决策，这也是在公司中的应用。因此，数据分析的出发点是解决问题，这是最重要的，也是学习数据分析和学习各种工具的目的。因此，要让数据分析工作可重复和更好地帮助解决问题，不能忽略这个最根本的出发点。

### 3.1 要解决的数据分析问题简介

#### 3.1.1 第一个数据

为了帮助大家更好地学习如何用 R 语言解决问题，我们提供了两个示例问题，这两个问题是我在之前参与过的项目中遇到的。这两个问题都对于心理学背景的同学都是比较熟悉的，其中一个是对问卷调查的研究，是有关人类社会关系和体温调节的问题。在我读博士期间，我参与了我的合作者的这个研究项目，他现在在法国一所大学有自己的 lab。他最关心的是人的体温和社会关系之间的关系。他提出了一个理论，认为人和哺乳动物有一个很苛刻的一个生存的环境，我们的核心温度必须要维持在某一个很窄的范围之内我们才能够生存，我们人类其实是比一些动物生存的环境是要求更苛刻的，比方说一些动物他会冬眠，这个外面的温度到了一定程度他就找个地方就躲起来冬眠，人类没有这个机制，所以人类一直要在一个恒温的条件下，或者相对恒温条件生存。那么在哺乳动物里面发现哺

乳动物会形成一个群，形成一定的社会关系，社会关系帮助大家调节这个体温，Hans 当时感兴趣的一个问题就是：在人类社会上是不是实际上也存在这种机制？当人有一个比较好的一个社会网络之后，这个社会网络本身，他会帮助我们去调节我们的体温。有一些类似的有趣研究，比如握着热咖啡杯的人会让人觉得更加温暖和友好，孤独的人更喜欢洗热水澡等。这些研究都有一些扎实的数据支持，但并非所有的研究都有强有力的证据。Hans 想验证社交关系是否能够调节体温这个问题，于是进行了一项大规模的跨国实验，收集了来自十几个国家的数据。他首先进行了一个 *piror study*，然后收集了数据。我们会将这些数据分享给大家。这个数据集来自 12 个国家，包含了 1,500 多个人的信息，其中包括很多问卷，如怀旧程度、对家的依恋、主观压力感受、自我控制和新陈代谢和社交网络质量。这个数据集最初是从国外收集的，Qualtrics 在线问卷被下载下来，和问卷星之类的网站下载的格式是类似的，现在以 CSV 和 SAV 格式提供。这个数据集可以被用于做数据分析。在这门课程中，我们会假设自己是 Hans 课题组的同学，获取了数据后，我们将学习如何进行后续的数据分析。这个过程从数据导入开始，一步一步进行。我们最终可能想要解决的问题是如何预测核心体温，其中一个重要的分析方法是监督的机器学习，我们将使用一个常规的算法，条件随机森林 *conditional random forest*。我们将寻找能够预测核心体温的变量。在图中，红色线以上的变量对预测核心体温有很强的相关性，红色线以下的变量预测效果较弱。虽然我们测量了很多变量，但通过这个机器学习算法，我们发现只有部分变量与核心体温有比较强的相关，我们可以在这个基础之上进行进一步的挖掘。里面能够预测核心体温的有一个 *complex social integration* 指标，是社交网络的一个指标，我们后面会完全复现他的这个过程，最后希望感兴趣的同学，可以在这地方得到这个指标。也有一些传统的做问卷分析的一些结果。比方说这个调节作用：他发现是否处于亲密关系其实是对这个关系是有调节作用的。对于进行整套分析的研究，首先需要将原始数据输入到 R 语言中，然后进行数据清理和描述性统计，例如数据质量、问卷信度、均值、标准差等统计指标。此外，也可以进行探索性的数据分析，例如相关矩阵和 *running force* 等方法，来探索变量之间的关系。最后，我们需要呈现研究结果，包括 *condition to Reinforce* 的结果和 *mediation* 的结果，并将其整合到可视化报告中。这部分内容实际上原研究没有进行过。作者当时使用 R 语言编写了代码后，直接采用了传统的方式来准备手稿。在本节课中，我们可以使用 Rmarkdown 来完成整个数据分析，并生成 PDF 文稿。当时，由于很多工具还不够成熟，因此我们有机会做得更好。我们这里的这个数据分享流程可能只是一个大致的流程，里面可能还会有一些小的细节，例如如何去假定变量之间的关系，当然我们进行的是别人研究的复现，通过复现去了解 R 语言数据分析中的流程，提高自己的技术。

### 3.1.2 第二个数据

这个实验是一项简单的认知实验。在认知心理学中，我们通常会在实验室中进行背景教导，然后进行一些简单的按键反应。这个任务可以分为几个阶段。首先是练习和学习阶段。在这个阶段，我们会呈现几何图形（正方形、三角形和圆形）和三个人物标签（好人、普通人和坏人），然后让你建立联系，例如将三角形和好人联系起来。接着，你需要判断屏幕上呈现的图形和人物标签是否匹配。这是一个非常简单的任务，只需要看一眼就能够记住图形和标签之间的关系。练习二三十次之后，你就能够熟练地完成这个任务了。

接下来，我们会进行一系列的任务，包括匹配任务和分类任务。匹配任务需要做出一个决策，即图形和标签是否匹配。它的反应时间非常短，大约在 800 到 1100 毫秒之间。分类任务需要判断图形是好人还是坏人，然后按相应的键。这个任务也需要在非常快速的时间内进行反应。

我们的实验目的是一个二乘二的实验，包括人物标签（自我和他人）和笑架（反派）。我们只有两个条件，非常简

单。在分类任务中，我们可以采用两种标准：自我和他人，好和坏。在学习过程中，我们需要将它们组合起来，形成四个人物标签和四个集合图形。在 BASE 学习中，我们学习四个人物标签和四个集合图形之间的配对关系，完成匹配任务和分类任务。分类任务包括按身份和家族进行分类。这是一个相对简单的实验任务。

在研究中，我们想要了解简单的即时集合图形是否会影响知觉上的反应。我们发现，给集合图形打上标签后，会影响反应时间。虽然总体上来说，好的自我反应最快，好的他人的反应也不错，但是坏的自我和他人人都比较慢。这是一个总体的趋势，但并不是每个集合图形都是这样的。我们使用这种方式来绘制图形，既能看到总体水平上的结果，也能看到个体差异。这样，在推断时就不会过度推断。

右边的图是 deep prime，是信号检测论中的一个指标。信号检测论是心理学背景课程中的一个重点，也是考研的重点之一。在心理物理学中，信号检测论是一种常用的数据分析方法，用于认知心理学研究。它可以计算 deep prime，即敏感性，这比准确率更能判断信号是否敏感。我们将匹配条件视为信号，不匹配条件视为噪音，以此来计算信号检测论。对于非匹配任务，如分裂任务，我们也可以计算类似的标准和数据。我们使用了一个名为 Drift Fusion Model 的简单计算模型来分析数据，它是用 Python 工具包制作的，但也可以在 R 中使用。导入数据可能有些复杂，因为每个 Base 的数据都有三个文件，需要合并和清理数据。在数据分析部分，我们需要使用 Aloha 计算 T-test，并进行数据清理和可视化。我们还将使用另一个演示数据作为 R 语言的基础，来进行后续讲解。如果使用传统的方式，我们需要使用 Excel 处理数据，使用 SPSS 进行统计分析，使用 Excel 或 PS 进行图形绘制和美化，最后使用 Word 文档进行写作。PS 是认知做视觉的人必备的工具之一。如果使用 R 语言，我们可以在 R 语言中完成所有工作，从 Tinyverse 开始进行数据清洗和工具分析，使用 GGPlot2 进行图形绘制。我们将使用 Markdown 或 Papilla 工具来输出报告的话。在 Brusa 中，我们特别关注 T-test 方向分析和多重比较等内容。由于 Brusa 对心理学数据分析进行了特殊优化，因此非常方便。正如我们之前提到的，它可以保留所有分析过程，并且可以通过代码直接重复分析。分析的代码和方法也非常灵活，新的方法也容易共享。

## 3.2 如何安装

这节课的主题是数据分析，这是研究生需要学习的知识。我不会在此深入探讨技术，而是开始讲解如何操作。首先是安装过程，可以在百度或必应中搜索 R 语言官方网站，然后在该网站中下载安装程序，该程序适用于不同的操作系统和版本。对于 Windows 系统而言，安装过程比较简单，但对于 Mac 系统，安装过程可能会稍微复杂一些。Mac 系统可能会有两个版本，分别为 ARM64 和 PKG 版本。ARM64 版本是一种新型的芯片架构，适用于苹果芯片的电脑。使用本地化的中文语言安装 R 语言会更加方便，以避免编码问题。现在，我们开始安装 R 语言吧！首先，你需要下载 R 的安装包，你可以选择从官方网站下载，或者从国内的镜像站点，如清华镜像站。下载完成后，运行安装程序。在安装过程中，你可以选择是否自定义一些安装选项，但是默认选项通常已经足够了。最后，点击完成按钮，等待安装完成即可。安装完成后，你可以打开 R 语言的界面，看到官方的导航界面，其中有许多可供选择的选项。为了演示启动 R 语言的过程，我们将点击“Download R”按钮。进入后，你将看到一个“R for Windows”的基础包，下载并安装即可。为避免出现问题，请确保选择英文路径作为安装目录。完成安装之后，你可以开始实际学习操作。在此我也想提醒大家不要把安装目录放在中文文件夹里，因为我们使用的是来自国外开发者的国外软件，基本上是以英文为主的，如果使用中文可能会出现编码问题。

在上方你可以看到，此处默认选择了中文作为系统显示语言，大概是系统语言设置为中文后自动采用了中文显示。

请问各位组员，现在都可以正常地看到这个界面了吗？刚刚我检查了一下，大部分的人都没有遇到任何问题，但是有一位同学看到了一些警告，猜测是语言设置引起的。对于那位出现警告的同学，可以将软件的信息全部切换到英文，允许我给大家展示一下，这是我第一次通过这种方式进行操作。以前我通常通过写代码、搜索相关软件然后在搜索框中打开软件来实现，强行改变所使用的语言环境被称为硬改。相对来说，这种方式更有效。

请问各位注意到我的操作路径了吗？它应该是在“C 盘”的“users”文件夹下，接着是 R 语言的安装目录。例如：在“Program Files”中。然后我们要打开文件并使用 TXT 工具修改文件“language”。可能需要管理员权限才能执行操作。这可能是必须的，让我再试试，如果还不行，我们就先暂停，以后再使用某些软件完成这种修改。还是不行吗？好吧，那我们来看看下一个问题。

针对另一个问题，我认为以后可能会遇到同样的情况。就是，你可能会看到一些红色警告的字体。在 R 语言中也会遇到这种情况。这里有两种类型的警告：一种是称为“警告（warnings）”，另一种是“错误（errors）”。对于那些被警告的代码，我们需要仔细检查，看看是否会对我们的运行造成严重的影响。如果没有明显的影响，那么就可以忽略。然后，我们可能会遇到一些问题，例如，遇到一些非 UTF-8 编码的语言编码。这意味着我们不是使用 Unicode 编码方式。为了解决这些问题，我们通常使用 UTF-8 编码方式，这是中文中非常常见的编码方式。在大部分计算机系统中，我们都可以使用 UTF-8 编码方式，因为它是国际上通用的编码方式。

### 3.3 如何方便使用？Rstudio 的安装与界面介绍

大家都已经成功安装了 R 语言，现在看到的是 consol 控制台界面。早期的 R 语言使用这个控制台进行输入和操作。比如，我们可以输入一个简单的命令，例如：`a <- rnorm(100)`，在变量 a 中存放了一个分布。控制台支持用户输入任何语句，比如 `demo()`，`plot(x,y)`。（这将自动打开一个窗口显示绘制的图形。）

对于没有编程经验的同学来说，这个可能是一个不太熟悉的点。为什么我们在这里输入这个东西呢？为什么会跳出这个东西呢？有时候会出错吗？例如，我们输入 a 会得到一个数字，但是如果输入 b，他会告诉你没有这个对象，同样地，如果输入 c，也不会有任何效果。这里的一个难点是，为什么有些输入会被计算机执行，但是有些不会？还有就是，为什么有些输入会产生图形效果，有些则不会？这实际上是我们和计算机交互的一种方式，这与过去使 spss 软件的方式有所不同。在使用 spss 软件时，我们通过点击屏幕来进行交互，并从菜单中选择选项。这被称为 GUI，即图形用户界面。然而，当我们使用 R 语言或其他语言时，我们主要使用的不是这种图形界面。我们只有一个输入代码的窗口，通过在这里输入代码与计算机交互。这是我们最常用、最需要的程序。对于其他语言也是一样。很少有人真正使用这种代码式的交互，因为它不太方便。

这样的交互方式的缺点是，我们不知道哪些变量保留在内存中。当我们进行一系列操作时，我们不知道哪些操作是我们刚才输入的。它们存在于计算机的缓存之中。如果我们输入了许多变量，如 a、b、c、d 等，我们无法回忆起这些变量的细节。因此，当我们学习 R 语言时，我们希望有一个更好的地方来编写代码。我们不希望一次只写一行代码，然后再运行。因此，我们需要一个更友好的代码编辑器，而不仅仅是控制台。虽然控制台是与 R 进行交互的窗口，但有时不够友好。### Rstudio 的安装与界面介绍 {#2-Rstudio-install}

幸运的是，我们可以使用 R Studio 来完成这项任务。R 我们需要安装好 R 之后，再安装 R Studio。有两个版本可以选择：Windows 版和 Mac 版。安装时，我们需要将其安装在非中文目录下。安装后，我们可以选择 64 位系统。使用 R Studio 的最大好处是我们可以一次编写一长串代码，然后执行它们。

打开了 R studio 之后，界面会有一些关于信息提示的内容。如果 studio 崩溃了，是否允许自动给你把报告发回给 studio，这是一个关于隐私保护的咨询同意。点 yes 也没有问题。现在我们可以看到，打开 RStudio 后，显示出的一个白色背景的界面，我们可以通过更改外观中的主题来改变界面颜色。在这个地方，工具中有一个全局选项，外观，你可以把它改成更专业的样式。现在界面变得专业很多了，这是我们通过调节一些参数实现的。当然，也可以根据自己的喜好来调整。另外，将界面背景改成灰色可以省些电，因为不用一直发亮光。

接下来讲讲界面的调整。我们的界面是由四个面板组成的，但是在刚安装 RStudio 时，我们看到的是三个面板的界面，界面可能还有点不一样。我们可以新建一个 project 来呈现一个新的面板。

现在我们有 4 个窗口，我们可以将它们平衡一下，即在 PPT 上面显示 4 个窗口。这 4 个窗口分别是脚本编辑区、控制台、环境和文件。脚本编辑区可以记录下我们想要写的代码，并且可以选择性地运行。我们也可以直接创建一个 R 脚本，然后输入代码。在输入代码的时候，RStudio 会提示我们，根据我们之前输入的字符来预测我们可能想要输入的代码。这样的话，我们就不用记住很多代码，可以更方便地完成代码。我们可以将所有的代码放到一起，形成一个脚本文件，一般以 R 结尾。我们可以用 RStudio 打开它，这样可以更好地编写 R 代码。

在这个 project 里，我们可以新建一个写代码的文档。在这个时候，会有一个没有命名的文件，叫做“Untitled1”，一般情况下我们会给它命名，比如我们可以给它命名为“readme”，作为保存项目相关信息的文件夹。我们也可以给它命名为“this is a testing file”。然后我们可以将它保存为 txt 格式，或者更好的 Markdown 格式，这个时候不会对代码执行产生影响。在 studio 中，我们看到了四个窗口，这四个窗口分别是这个脚本的编辑区，控制台，环境变量以及帮助。脚本的编辑区是一个可以记录我们想写的所有代码，并且可以选择性的运行某段代码的空间。如果我们创建的是一个 R 脚本，就可以看到它上面有 run 或者其他命令选项。提示输入代码的时候，可以根据之前输入的一部分字符来预测接下来可能会输入的内容，这有助于提高代码输入效率。我们可以将代码分类放到一起，形成一个脚本文件。当我们保存 Readme 文件后，一般会将其文件后缀名命名为 R，它本质上也是一种文本文件，可以使用文本编辑器来打开。

在文件夹里，我们可以找到我们保存的 Read me 文件和以.r 结尾的测试文件，它们实际上都是以 TXT 格式存储的，我们可以用 TXT 编辑器来打开它们。但是，为了更好地写 R 语言代码，我们一般会选择用 R Studio 软件来打开和运行这些文件。

### 3.3.1 如何方便使用？Rstudio 的安装与界面介绍

# Chapter 4

## 第三讲：如何导入数据

### 4.1 路径与工作目录

那么什么是工作目录呢？简单来说，工作目录就是 R 会在其中寻找文件的文件夹。如果你没有指定工作目录，那么 R 会默认使用你的用户文件夹作为工作目录。但是，如果你的数据文件不在用户文件夹中，那么你需要告诉 R 去哪里找这些文件。这就是为什么我们需要设置工作目录的原因。我们可以使用“`get.wd`”命令来查看当前的工作目录。当我们打开一个 R 项目时，它会自动打开项目所在的目录。

#### 4.1.1 相对路径与绝对路径

我们建议在新建数据分析项目时，将其放在一个文件夹中，并创建一个 R 项目。这样，每次打开项目时，它都会自动打开所在的位置。在讲解路径和工作目录的概念时，我们需要了解两个概念：**绝对路径**和**相对路径**。绝对路径指的是文件或文件夹在硬盘上的真实路径，而相对路径则是指相对于当前路径的位置。例如，在 `chapter3` 文件夹中，`data` 文件夹的相对路径为 `./data`。相对路径的好处在于，当我们复制整个文件夹时，相对路径的位置不会改变。相对路径是相对于当前路径的位置，而绝对路径则是相对于硬盘的起始位置。

如果我们有一个子文件夹叫做 `data`，那么它就是当前文件夹的子文件夹，也可以称为子目录，而当前文件夹则是它的父文件夹，也可以称为父目录。如果我们复制这个文件夹到另一个电脑上，它们之间的相对关系不会改变，但如果我们把它从 D 盘复制到 E 盘，那么绝对路径就会改变，相对路径仍然保持不变。因此，我们建议把所有文件都放在一个文件夹中，这样所有文件的相对路径都是相对于这个文件夹的，方便编写代码。如果文件夹放在不同的位置，每次引用或查找文件时都需要使用绝对路径，非常麻烦。

在 Windows 系统中，每个磁盘都有自己的盘符，例如 C、D、E、F 盘。绝对路径是从盘符开始一直写到当前文件夹，对于 Mac 和 Linux 系统，它们只有一个根目录，没有盘符的概念。我们最常用的是“用户”，比如我的用户名，然后有一个“主目录”，这就是用户的文件夹。当我们使用苹果系统或者 Linux 系统时，基本上很多时候都是基于这个“用户”文件夹或者“`users.html4c`”进行操作。除了“主目录”以外，它可能还有一个“下载”，就是你下



载的文件都放在哪里，还可能有一个“图片”，就是你的各种图等等。所以它就是在一个文件夹下面有不同的子文件夹，比如有“etc”、“用户”还有其他的“根目录”等等。我们主要就是基于“用户”，然后建立自己的一系列的文件。在 Windows 中，它可以分成三个盘，每个区都是自己的盘符，然后基于这个盘符你又可以重新定义一个新的路径。这就相当于路径的定义不太一样。我们刚刚也讲了相对路径和绝对路径。如果我们所有的分析都放在一起的话，那么我们用相对路径是更好的。就像我们写代码的时候，不要把 C 盘 D 盘什么这些一整串写进来，因为写进来之后你换一个电脑，这个 C 盘它就不是 C 盘了，它可能是变成 D 盘了，或者你换到苹果上面之后，它可能就没有 C 盘了，它就是“用户”开头了。

### 4.1.2 Here 包的使用

另外一个就是我非常常用的一个包，就是这个 Here 这个包。Here 这个包可以帮助你很快地去定位你的，因为它能够很快地返回你当前的新的目录。然后你打开了这个之后，我们就能够迅速地看到我自己就是说它在哪。它另外一个好处就是我们能够不用考虑我的系统是什么，然后我就直接能够把它的一系列文件夹都调出来。然后我先看一下我这里有没有 Here 这个包，好像还没有，现场直接安装。网速有点慢，OK，大家看到现在成功安装了。我们再看一下，它有一些警告的信息，但是它告诉了我们一个结果，就是把这个关于这个 Here，它我们现在这个文件夹在哪输出来。我们刚才说它有一个好处就是比如说，我们不知道这个当前的这个文件夹它到底是哪个，我们可以不用去写这个，它的符号跟 Windows 的地址栏是有区别的。我们的 Windows 系统中的地址栏和 R 中的地址栏斜线方向不同。如果我们直接复制 Windows 地址栏的地址作为绝对路径，可能会出现错误。为了避免手动拼接路径出错，我们可以使用 here 函数来自动补全路径。这个函数会自动将当前文件夹的路径补全，我们只需要在路径中加入文件夹名即可。例如，我们可以使用 here(“data”, “match”) 来获取 data 文件夹下的 match 文件。这个小技巧可以帮助我们避免一些错误，特别是在不同电脑之间切换时。

## 4.2 读取数据

工作目录是 R 与硬盘交互的接口，我们可以使用 getwd() 函数来获取当前工作目录的绝对路径。我们需要使用相对路径来调整子文件夹的位置。接下来，我们需要查看数据的相对路径和绝对路径。假设我们的工作目录是在电脑桌面上，数据存储在“data”文件夹中，其中包含两个子文件夹，一个是“match”，包含 178 个文件，另一个是“pengiun”，只有一个文件。相对路径中的一个点代表当前文件夹，数据在“data”文件夹中，“match”文件夹下有 178 个文件。如果我们要读取某个文件，必须完整地写出文件名。对于“pengiun”文件夹，它的相对路径是“data/pengiun”，假设它是一个简单的 CSV 文件。

### 4.2.1 read.csv 函数

现在我们知道了数据的相对路径，可以使用“read.csv”函数来读取数据。我们需要给数据一个名称，例如“pengiun.data”，并使用“read.csv”函数来读取数据。在函数中，我们需要指定相对路径。这里建议使用相对路径而不是绝对路径，因为绝对路径需要每次更改代码。

“`head = TRUE`”是函数中的一个参数，用于指定文件是否包含列名。如果文件的第一行是列名，则将其设置为“`TRUE`”，否则为“`FALSE`”。在统计学中，假用 `false` 表示，真用 `true` 表示。读取 `txt` 文件时，需要使用分隔符来分隔不同列的数据。为了更直观地了解这个过程，我们可以打开 `csv` 文件。虽然这是一个 `csv` 文件，但我们也可以使用 `txt` 打开，这样更清晰。但是，如果我们使用 `txt` 打开，数据看起来会很混乱，因为它包含换行符。在第一行中，从 `site` 到双行符号之间的所有内容都是第一行。我们可以让它不自动换行，但是每个人打开 `txt` 文件时可能会看到不同的行。数字 9.1 表示第一行的列数。每个字符都是一个列，而不是每个单词。因此，在读取时，我们需要将字符分割成不同的数据。在 `R` 中，我们使用逗号来分割。

因为 `txt` 文件包含换行符，所以它会识别行和列。行是通过每一行的数字来识别的，而列必须使用分隔符来识别。在这里，我们使用逗号作为分隔符。这样，我们就可以将数据读取为行和列。当我们遇到字符型数字时，需要将其转换为类别数据。在 `R` 中，我们使用负值来表示这个符号。我们可以使用 `data table` 包来方便地展示读取后的数据。为了方便在 `PPT` 中展示数据，我们可以使用 `head` 函数来查看数据，而不必亲自查看。我们已经读取了数据，按照逗号分隔后，第一行是列名，我们可以将其识别为表头。从第二行开始是数据，第一个列的第一个值是 1975，后面是 1995 等等。这是一个问卷数据，每个编号代表一个问卷和其 `items`。我们可以用不同的方式对其进行编号，比如 `uncertainty` 等等。对于常用的心理学数据文件，我们可以使用 `here` 来定义相对路径，然后使用 `import` 来读取数据。读取后，我们可以看到有 247 个变量。对于 `.out` 文件，我们可以使用 `read csv` 来读取，但是需要将 `separator` 参数设置为斜杠加 `t`，表示以 `tab` 为分隔符。在进行统计分析时，我们需要明确定义用于分割数据的符号。除了逗号和 `tab` 键，还可以使用分号、空格或分行符等多种方式。

`read_csv` 是最常用的数据读取函数，但实际上它是 `read_table` 的一个特殊形式。因此，我们也可以使用 `read_table` 来读取数据。但是，由于 `.out` 文件是我们自定义的一个 `txt` 文件，无法使用 `bruce-r` 的 `import` 函数进行导入。这种情况很常见，当你尝试使用不同的分隔符时，你需要清楚你的数据结构，以便正确读取数据。在编写代码时，我们必须注意，没有报错并不代表它一定是正确的。首先，我们需要读取一些数据，包括 `first path` 和 `normalized path`。然后，我们可以使用 `pumping data` 或 `match data` 命令来读取数据。读取数据后，我们需要检查数据是否正确读入。在读取 `.2` 文件时，我们只读取了一个文件，而没有将 100 多个语句全部读入。这是一个非常困难的任务，但是我们可以通过反复比较 `R` 界面和文件夹来理解当前工作目录、相对路径和绝对路径的概念。

完成这些操作后，`R` 中的环境应该发生了变化。在右上角的环境中，我们可以看到两个相应的数据 `match data` 和 `PenguinData`。`PenguinData` 是一个符号，我们可以用任何其他符合命名规则的名称来代替它，比如 `abcd`，比如 `xyz`。这里存储了一些可操作和创建的对象，我们称之为 `R` 对象。现在我们导入的对象的种类被称之为数据框，有关数据框的知识将在之后讲到。

### 4.2.2 命名

在命名时，应尽量简短，避免使用中文。如果发现文件目录没有问题，但仍然出现错误，有两种可能性。一种是拼写错误，例如拼写单词 `penguin` 时少了一个字母或交换了两个字母的位置。另一种可能是文件夹的大小写不匹配，例如使用大写字母开头的文件夹，但在输入路径时使用小写字母。在这种情况下，我们需要仔细检查并逐行对比当前工作目录和文件夹的实际目录是否匹配。最有效的方法是直接点击文件的属性，查看其目录和名称是否与输入的完全相同。

## 4.3 了解 R 里的数据（R 语言中的对象）

刚才提到的几个常用对象包括：向量、矩阵、数组、数据框和列表。其中，向量是一列数字或其他相同类型的元素，而矩阵是一个二维的数据结构，有长和宽，可以看作是将向量按行排列而成的。数组是在矩阵的基础上再往上拓展，变成了一个三维的数据结构。在 R 语言中，我们更多地使用数据框，它看起来像矩阵，但每个列可以是不同类型的元素。而列表则更加复杂，每个元素都可以是不同类型的对象。这些对象在编程软件中非常常用，比如 Python 和 Matlab。这是修改后的文本：有四种数据类型，它们各自具有独特的特点。其中，向量是将同一类型的数据排列成一列，内部数据必须相同。内部数据类型包括逻辑型（true/false 或 0/1）、整型数据（整数）、双精度和复数数据、字符和原始形象。这些数据类型有不同的操作方式。在心理学中，有一个因子向量，用于将数据分成不同的类别，例如男性和女性。因子向量对于分类数据的处理非常重要，例如不同的种族和民族。但是，将因子向量转换为数字时需要注意，因为它会按照级别进行变化。数字 1 对应着“女性”，数字 2 对应着“男性”。在转换过程中，我们将“女性”转换为 1，“男性”转换为 2。第三个和第四个数据都是 2，因此它们直接转换为 2。需要注意的是，我们不一定使用“女性”和“男性”，而是使用数字 1 到 7。如果我们将数字得分和实际数字进行比较，我们将使用“fact”来定义等级，例如 54321。如果我们直接将“fact”转换为数字，它将变成 12345，因为我们将其定义为分类变量并赋予了等级。如果我们想将其转换为对应的数字，我们需要先将其转换为字符，然后再将其转换为数字。需要注意的是，如果我们不注意这个问题，就会出现错误。

我们没有时间讲解数据框，但是我们已经解决了读取数据的问题。我们下节课再深入讲解这些知识。在下节课之前，大家可以查看王敏杰老师和张金星老师的两本书，这些书将更清楚地介绍这些基础知识。