**2025-3-4 STATS707**

**英文原文 & 中文翻译**

1.

English:

this course we work basically, so um, and then i’ll start with the first, uh, prophec, and um, as so did you guys all get the email or the message?

i sent on a discussion.

have you got a message from me, maybe, i don’t know, last week Thursday or something? yes? yeah, yes?

中文翻译：

这门课我们主要就是这样运作，嗯，然后我会先从第一个，呃，prophec 开始，嗯，那么你们都收到我发的邮件或者信息了吗？

我是在讨论区上发的。

你们有没有收到我发的信息，也许，我不确定，上周四之类的？有吗？是吗？有吗？

2.

English:

great, okay.

so that, i think, the only message i sent pretty much,

so if you have a red deck message, what i say is that we will do kind of high brave thing in this course, okay?

so first week i’ll do both lectures, but in next week, i only do a workshop on Tuesday, and i’m giving you a full time table for the whole semester, right? so all of you. so that’s the first thing, right?

中文翻译：

好的，太好了。

所以我想，那是我几乎唯一发过的信息。

所以，如果你们看过那条信息，我的意思是我们会在这门课里尝试一种混合形式的方式，好吗？

所以第一周我会讲两次课，但下周我只在周二开一个研讨课，而且我会给你们整个学期的完整时间表，对吧？你们所有人。所以这是第一件事，对吗？

3.

English:

so if you’re clear with that, there are lots of spaces around here.

okay, right, so what i do is, let me run through the basic slides i bought, and then i’ll take you on panels and also show you around very fine things.

okay, so this is the sex seven or seven, just in case you’re in the wrong classroom or something.

中文翻译：

如果你们对此都清楚了，这里有很多座位可以坐。

好的，那么我要做的是，让我先带你们浏览一下我带来的基础幻灯片，然后我会带你们去看一下论坛，也展示一些非常好的东西。

好的，这是 sex seven or seven，以防你们进错教室之类的。

4.

English:

okay, phone case introduction to sex. so this is me, titania, the base way to pronounce my name, if you don’t know how to, is try for China here, Tanya, okay?

so check Tanya, and i’m in there on the third floor of building two three, there’s my email. so in terms of how to get in touch with me, the email is the best mode. i’m usually fairly responsive on the email, i try and get back to you as soon as i can. so you need to come and see me, please just email me. i really enjoy helping students one on one. so yeah, just email me or for anything else, just say.

中文翻译：

好的，电话壳（phone case）简介 sex。我是 Tanya，如果你不知道怎么念我的名字，最简单的方法就是试着把“China”换成这里的“Tanya”，好吧？

所以就是 Tanya，我在二三号楼的三楼办公，这里是我的邮箱。至于怎么联系我，邮件是最好的方式。我通常会比较快地回复邮件，尽量尽快回复。如果你需要来见我，请给我发邮件。我非常喜欢一对一帮助学生。所以，是的，有什么事就给我发邮件，或者别的事也可以和我说。

5.

English:

okay, so just coming back to the structure, just drink, drink, and we got two sessions a week, Tuesday and Fridays. you know, you already here, but as i just said, we want to run it in a hybrid manner. i have sent the whole timetable on a discussion, so we will have a look at that. and what we’re going to do is some lectures, many couple of lectures at the first week, definitely in person, and then mostly online. but then we’ll have a workshop. now, what is a workshop? well, a workshop is really a place for us to kind of, i’ll give you maybe some problems and maybe give you some questions to work on these problems. and these questions are what i want you to learn, okay?

中文翻译：

好的，那么回到课程结构，先喝口水，我们每周有两次课，周二和周五。你们现在已经在这里了，但正如我刚才所说，我们想以一种混合的方式进行。我已经在讨论区上发了整个时间表，我们会一起看一下。我们要做的是在第一周进行一些线下授课，确定地面对面上几次课，然后大部分在网上。但是我们会有一个研讨课。什么是研讨课呢？其实研讨课就是一个地方，我会给你一些题目，也许给你一些问题，让你动手去做。这些问题正是我想让你们掌握的，好吗？

6.

English:

um, i will also try and make so that these questions that you work on during your workshop are related to the kind of questions i would want you to know in your base, okay? so, um, the workshop is really mean, uh, really meant to enhance the learning, because online lectures, it’s very easy to just go through all the slides, watch the whole video, and kind of feel that, okay, you understood, but actually you haven’t. you haven’t understood anything, okay?

中文翻译：

嗯，我也会尽量让你们在研讨课上做的问题，和我希望你们在考试或测验中掌握的问题类型相关，好吗？所以，嗯，研讨课的真正目的，其实是为了增强你们的学习，因为线上课程很容易就是把所有幻灯片看一遍，整个视频看完，然后感觉自己好像理解了，但实际上你并没有。你其实什么都没懂，对吧？

7.

English:

so only when we get down to maybe some specific questions, some problems, i think that’s actually when the learning happens. so we’ll try and do that. also, mind you, i’m trying this format for this course for the first time, okay? so only last year, this course was more than a hundred man. i’ve tried hybrid picking two years back during Kobe for a different course, but this is i’m trying, so, you know, i will play by ear as we go. i’m going to deal with the workshops from scratch, so just so you know, hopefully, i’m able to keep up with them. okay, so that’s what the workshops will be. now, the workshops do have some mobs, we’ll come to that. um, now, as i mentioned earlier, keep checking your messages via, you know, so i will try and post it on a discussion. i’ll show you what i mean by a discussion, but essentially, you should receive it as a message.

中文翻译：

只有当我们真正开始做一些具体的问题、练习题时，我觉得那才是真正学习发生的时刻。所以我们会去做这些东西。另外注意一下，这是我第一次用这种形式来教授这门课，好吗？去年这个课人数还超过了一百。我在两年前的疫情期间曾经在另一门课上尝试过混合式，但这次是新尝试，所以我会边走边看。研讨课我会从头来带，所以你们知道就好，希望我能跟得上进度。好的，这就是研讨课的定位。现在研讨课确实会有一些评分比例之类的，我们之后会说。嗯，再提醒一下，我前面说过，要记得查看你们的信息，就是通过那个，你们知道的，我会尝试在讨论区发出来，我待会会展示给你们怎么在讨论区看，但基本上，你应该会把它当作一条消息看到。

8.

English:

a discussion is a forum that you guys can use for asking questions this way, so it actually works more like, um, you know, kind of a social media, just specifically for this course. um, so feel free to ask questions. uh, if i need to respond, i respond as very, like, you know, i do that all the time, and also, you can reach out to me. i can reach out to all of you. so that’s, i show you what i mean, okay? so that’s what you want to use.

中文翻译：

讨论区就是一个论坛，你们可以用这种方式来提问，所以它其实更像是，嗯，就像一个社交媒体平台，只是专门用于这门课的。嗯，所以随时可以提问。呃，如果我需要回复，我会回复，你知道的，我经常这么做，而且你们也可以联系我，我也能联系你们所有人。我到时候会展示给你们看，好吗？所以这就是你们要用的地方。

9.

English:

okay, um, so that’s kind of the broad outline of what you are going to learn. how many of you, ideally all of you, have done at least maybe some one, you know, at least one course in statistics at some point? is that correct? yeah, right? okay, now you might have forgotten almost everything, but hopefully, yeah. so the purpose of this course is to refresh your number three, okay? so you can say that whole thing is we, you know, most of it you should have at least come across if you have any tax of your learning. more importantly, uh, there is especially the first couple of points, estimation, hypothesis testing, even probability. they’re very deep subjects, so, uh, what is unique about statistics is not just mathematical, but it’s actually quite philosophical as well, okay?

中文翻译：

好的，嗯，所以这就是你们要学的主要内容的大纲。理想情况下，你们所有人应该至少学过一点统计学，比如至少修过一门统计学的课，对吗？是吗？好吧，也许你们差不多都忘了，但希望还记得点什么。所以这门课的目的就是帮你们回顾一下基础，好吗？所以可以说，你之前接触的大部分内容，这里都至少会覆盖，如果你对以前学过的东西还有印象的话。更重要的是，尤其是前面那几块内容，估计、假设检验，甚至概率论，都是很深奥的主题，所以，嗯，统计学的独特性在于它不仅仅是数学，也有很多哲学方面的东西，是吧？

10.

English:

and the concepts are not necessarily easy. they don’t seem, they are not what they see, okay? so, uh, that’s what i’ll be focusing on, okay? uh, you’re not going to make this very mathematical, although i believe most of you have done enough math step in it, so shouldn’t be a problem.

中文翻译：

而且这些概念不一定简单。它们看上去没那么难，但其实并不像看起来那样简单，好吧？所以，嗯，这就是我要重点放的地方。好吧？我们不会让它变得很数学化，尽管我相信你们大多数人已经修过足够的数学，所以这应该不是什么大问题。

11.

English:

so that’s exactly what i was talking about. the basic aim is to give you all the important concepts in statistics that you need so that you can then do further stats courses. so this is really like bridging courses, statistics knowledge. you guys, uh, we are expecting you to have done at least some, uh, programming, so, you know, they are going to teach you how to do programming, um, so what we use r. if you haven’t used r, it’s pretty easy to learn. we will, you know, go through it anyway. uh, everything will be available on canvas, and as i said here, the emphasis is on practical statistical inference.

中文翻译：

所以这就是我所说的。我们的基本目标是给你们提供需要的所有统计学重要概念，这样你们就能进一步学习更高阶的统计课程。这真的是一门“桥梁课”，来补充统计知识。你们基本上，我们也希望你们至少做过一些编程，所以你知道，他们会教你怎么编程，我们这里用 R。如果你从没用过 R，也不难学。我们反正会带着你走一遍。嗯，所有的材料都会放在 Canvas 上，就像我说的，这门课会侧重于实际的统计推断。

12.

English:

so my style of teaching is where i always try to emphasize on concepts, on understanding why you’re doing something, not just telling you, okay, this is what case, doesn’t make any sense. so i will always try and highlight those concepts, how i teach. and of course, as you can also see, i can’t talk while, you know, standing instead. i will always motor off, so you will get used to that. okay.

中文翻译：

我的教学风格是，我会尽量强调概念，强调为什么要做这些，而不仅仅是告诉你，这一步该做什么，没什么意义。所以我会一直去强调这些概念，这就是我授课的方式。当然，你们也会看到，我没法一边站着一边不说话，我会一直滔滔不绝，所以你们会慢慢习惯的。好的。

13.

English:

what is, yeah, this is all competition introduction d course. we are using r straight away. as i said, we expect you to learn, shouldn’t be a problem, but anyway, you feel free to come to me if you have any questions, i don’t mistake. we are going to use fairly basic r code anyway, and all my codes are available on canvas. so there’s also basic sense of playing more room on canvas, so are you point of actually very soon. there’s also plenty of help on the internet, so compose the problem there.

中文翻译：

这就是，嗯，对，这些都算是课程简介。我们会直接使用 R，就像我说的，我们希望你去学，这不应该是个问题。但无论如何，如果你有任何问题，欢迎随时来找我，我也不介意。我们所用的 R 代码也都比较基础，而且我所有的代码都放在 Canvas 上。Canvas 上也有一些基础的、可扩展的空间，所以你们很快就会看到。网上也有很多帮助资源，所以你也可以在那里找答案。

14.

English:

okay, so this is the important base, okay? so we bought 50 percent for the two tests, so there are two tests which are compulsory and in person, so you have to be here. the tests will happen during the lecture times, the message that i sent the other day, there are exact dates for the test, move the test. so the first test is not until after the mid-semester break, so in your favorite, i think, um, and the second test is the final legal tune also, the first legal to final legal teaching, okay?

中文翻译：

好的，这里是非常重要的基础信息，好吗？我们有两次测试，共占 50% 的分数。这两次测试是强制性的，并且必须现场参加，所以你要到场。测试会在上课时间进行，我之前发的信息里也给了测试的具体日期，移动了测试日期。第一场测试是在期中假之后才进行，我想是在你们的节选时间之后，嗯，第二场测试则是接近最后正式课结束的时间，也就是在第一周的正式课和最后那周的正式课之间。好的？

15.

English:

so those two days. um, there are four assignments, uh, the assignments typically two, three weeks, but with the sign dates, i don’t specify exact dates necessarily because it also depends on how far we are going. so it depends on the speed at which we’re going. so, but i will, of course, let you know at least two weeks in advance, so you will have enough time, and then off are you there, okay? so for assignments, ten percent for each of them, so that’s forty percent marks, two percent of course, made some basic era here, shouldn’t be twenty percent, should be twenty five percent each, but you pay that, okay? so that means they’re each twenty five percent, by the way, and then the in-class workshop is one percent for each workshop, so if you attend, you will get those marks, and not compulsory for you to attend, but of course, it’s good if you attend, and the boss mark is the givers of overall. any questions so far? yeah.

中文翻译：

所以就是那两天。嗯，还有四个作业，嗯，作业通常会给你两到三周时间，但是提交日期我不一定会精确指定，因为这也取决于我们的进度。所以，这取决于我们讲课的速度。不过我当然会提前至少两周告诉你，这样你就有充足的时间，然后你就可以交。好吧？所以这四个作业分别占 10%，一共 40% 的分数。呃，这里有个小错误，其实不应该是 20%，应该是 25% 各占，但就这样吧，好吗？所以它们分别占 25%。然后课堂研讨环节，每参加一次拿 1% 的分数，如果你来，你就能得到这些分数，不是强制要求你来，但当然，来上就更好。整体加起来就是期末的总成绩。到目前有什么问题吗？有吗？

16.

English:

so that’s a rough timetable for the four assignments, as i said, i mostly follow those, maybe a little bit here and there. i have given indicative dates also in my message, i think those are updated, okay, and the tests, let us say typically one hour test, okay? so our lecture kinds of two hour slots, but the test will only be for one hour, and it’ll be just here, so it’ll be just see if it’s like this, and yeah, it’s just an in-person test, both of them. again, i will show my amounts, actually know the exact dates. usually the tests have a combination of multiple choice and short answer questions, as i said, i focus on understanding, so i would ask you to explain to me what a specific concept means, okay?

中文翻译：

这就是关于四个作业的大致时间表，就像我说的，我大致会照这个来，也许在某些地方做一点微调。我也在信息里给了你们一些预估日期，我想那些日期是最新的，好吧，关于测验，通常来说都是一个小时的测验，好吗？我们一节课是两个小时，但测验只考一个小时，就在这里进行，就像这样。对，就是现场测试，两次都是。然后我会再拿出我的笔记，确认具体日期。通常测试会是多项选择和简答题的组合，就像我说的，我强调理解，所以我会让你解释给我听某个概念具体是什么意思，好吗？

17.

English:

my style of question is a bit different than some others, um, and of course, i’ll give you enough information, a lot of information, okay?

中文翻译：

我的出题风格和一些其他人有点不同，嗯，当然，我会给你充足的信息，非常多的信息，好吗？

18.

English:

um, so these are just some of the reasons why people take this course, okay? now many of you might be doing data science or information, information things like that, so, you know, just need some stats for a lot of stats, and so this is what you’re doing. so i think before i begin, i want you to just think about it a little bit and convince yourself why you’re doing this course, okay? it’s not for English, but make sure that you understand yourself how this is going to be important, and this is going to be important, i will tell you that, because we, of course, live in an age of data, so i don’t need to say that.

中文翻译：

嗯，所以这只是一些人们为什么会选这门课的原因，好吗？你们中许多人可能在学数据科学或者信息类专业之类的，所以你们就需要很多统计知识来支撑，因此你们在这儿。我想在开始之前，让你们先稍微想一下，说服自己为什么要修这门课，好吗？这并不只是为了英语，而是要让你们明白这对你们来说为什么重要，而且它确实很重要，我会告诉你们，因为我们生活在一个数据时代，这点无需多言。

19.

English:

so at least for next six months, there are jobs for people to do data analysis. maybe in next six months, everything will be ai. we don’t know, okay? but at least for next six months, you need people to do it. ah, jokes aside, you know, we need people who have skills, ah, to understand how data can be used in reckless, right? and strategistics is really important because it’s all about uncertainty, it’s all about decision making uncertainty, um, and you know, so even for ai, you have progressive machine learning, uh, which requires statistical skills to quantify uncertainty, okay? so statistics is really, really useful.

中文翻译：

至少在接下来的六个月里，会有很多工作需要人来做数据分析。也许再过六个月，一切都会被 AI 取代，我们也不知道，好吗？但至少未来六个月我们是需要人的。啊，开个玩笑，你也知道，我们需要那些有技能的人，呃，懂得如何合理利用数据，是吗？而统计学非常重要，因为它的核心是处理不确定性，处理不确定性决策，嗯，所以即使在 AI 领域，比如渐进式机器学习，也需要统计技能来量化不确定性，好吗？所以统计真的非常有用。

20.

English:

okay, so i’m looking for a class rep. i think you guys are old enough to know what a class rep is. basically, you get to eat some free, as a, i think a few times in the semester, they have meetings where you, you know, go to the meeting, anything, any feedback that you have from your classmates, you provide their feedback, and their feedback tells us it’s a good course. so, of course, it looks good on your cv, but it probably won’t look better on the cv if you’re undergrad, okay? anyway, we need a class rep. anybody wants to volunteer? nobody wants to volunteer here?

中文翻译：

好的，我现在要找一名班代表。我想你们都已经足够大了，知道班代表是干什么的。基本上，你会有机会免费吃些东西，因为我想在学期中会有几次会议，你会参加这些会议，把同学们的意见收集好，然后反馈给学校，学校的反馈会帮我们改进课程。所以，这当然会让你的简历好看一点，但如果你还只是本科生，可能也没差很多。好吧，不管怎么说，我们需要一个班代表。有谁愿意自愿吗？没人想自愿吗？

（由于篇幅极长，以下将继续保持相同的“英文原文在上，中文翻译在下”格式，直至完整呈现。）

21.

English:

but i’m going to randomly get somebody there. come on, guys. gosh, you’re the least enthusiastic class i ever had. usually the least one person. okay, how volunteer? you volunteer? if someone wants to join, amazing. does anybody want to join? you’re no? congregation? you know, the same event? you may find in the real job, right? that would be correct. okay, cool. okay, but see me after the thoughts, i would, before. okay, thanks. what’s your name? Siri? Jim? Jeff. Jeff, okay. so Jeff, yeah, could you please just stand up, so everybody knows who you are? okay, this fabulous, fabulous, genuine class. okay, thanks, thank you.

中文翻译：

那我可能会随机点一个人了。拜托，伙计们。天啊，你们是我见过最不热情的一届了。通常至少会有一个人站出来啊。好吧，谁来自告奋勇？你要吗？如果有人愿意加入，太好了。还有人想一起吗？你不愿意？集体？嗯，你知道的，和现实工作中很像，是吧？可能确实如此。好的，酷。好吧，那就下课后来找我一下，好吗？谢谢。你叫什么名字？Siri？Jim？Jeff。Jeff，好吧，Jeff，对吗？能不能请你站起来让大家认识一下？好的，这个非常棒，非常真实的班级。好的，谢谢你。

22.

English:

okay, so, um, that’s, we’ve done the housekeeping part. so let me go to the canvas and show you what we want. so first and foremost, when you go to canvas, you have all these things that you can see on the left, in case some of you are new to canvas, you understand on the basic menu for you to move around. i am on the pinocchio video page, and you can see there is an open recording there already for you to see, okay? so that’s called lesson one, because what we are going to teach today is lesson zero and lesson two, okay?

中文翻译：

好的，那么，嗯，我们已经做完了这些课程管理方面的事。接下来让我打开 Canvas，给你们看看我们要做什么。首先，也是最关键的，当你打开 Canvas 时，你会看到左边有所有的菜单，以防有人对 Canvas 不熟悉，你可以通过基本的菜单操作来移动。我现在在 pinocchio video 页面，你可以看到那里已经有一个开放录制的视频给你观看，好吗？那个叫 lesson one，因为我们今天要讲的是 lesson zero 和 lesson two，好吗？

23.

English:

so lesson 0 is part of, i mean part of lesson 0, we already done, which is the kind of housekeeping, and the next bit is some introduction to r, which is really just basic stuff, and then i’m going to do with lesson 2, which is like the main topic, which is a probability theory. lesson one is really just introductory to give you a bigger context of how statistical data analysis process works. so i bought that recording. so what you will say essentially for next week especially, is whatever you need to do, those lessons we have here, here, okay?

中文翻译：

lesson 0 其实是，一部分内容我们已经讲了，比如刚才那些课程安排之类的事情，接下来会有一些 R 的介绍，都是很基础的内容，然后我会开始讲 lesson two，就是主要内容，概率论。lesson one 其实就是介绍性的，给你一个更大的背景，看统计数据分析的流程是怎样的。我录了一个视频在那里。所以尤其是下周，你基本上需要做的，就是在这里把这些课件、视频都看完，好吗？

24.

English:

so you, and i will probably also message you, you know, make sure you do these lessons before you turn up to the workshop, okay? so before you turn up every workshop, i would expect you to have done maybe two or three lessons. as you can see, these lessons are not going to be a too hard. a lesson typically fifteen, twenty minutes, maybe thirty minute lessons, okay? so it’s basically like watching a YouTube video, kind of, so next idea. any questions on that? okay, so i will let you know every week what i expect you to have watched before you show up in the workshop.

中文翻译：

所以你——而且我也会给你发信息，提醒你在研讨课前一定要先学完这些课程，好吗？也就是说，每次研讨课前，我都希望你先完成两三个课时。正如你所见，这些课时并不难，一节课可能也就十五、二十分钟，也许三十分钟左右，好吗？所以它就像你在看个 YouTube 视频之类的，很简单。你们有什么问题吗？好的，那么我每周都会告诉你，在来研讨课之前该看哪些内容。

25.

English:

um, there won’t be that much of a problem for the first workshop, because we’re actually going to do the lectures in person, okay? but from the following week, that would be important. okay, so that’s that. if you go to modules here, that typically has everything that you really need. so if you just go down, you have some resources and general info. so this housekeeping, the first set of slides that we just went through, those slides already here. and then that lesson one is the other slides for the video recording that i just showed you, okay? and then these are the slides for what we’re going to talk about today, and maybe on Friday, okay?

中文翻译：

嗯，第一周的研讨课问题不大，因为我们确实还要在课上进行线下讲解，对吧？但是从下周开始，这就很重要了。好，这就是这个。如果你到 modules 这个页面，你基本上能找到你需要的一切。只要往下看，你就能看到一些资源和常规信息。所以像 housekeeping，我们刚刚讲过的第一组幻灯片已经在那里了。然后 lesson one 是刚才我提到的那个视频录制对应的幻灯片。好的？然后这些是我们今天要讲的内容，也许到周五会继续用的内容，好吗？

26.

English:

so as we progress further, the slides will appear here, and then this following section, you have the r codes and dataset. so as i said, all my codes and data are available on datasets, so you already have some. so that’s the introduction to r, which we want to do now, okay? so again, as we progress further, you will see more and more appearing. if at any point you feel, or if you see that actually what is meant to be there is not there, then just email me, maybe i’ve forgotten to put it up, so please remind me.

中文翻译：

所以接下来随着课程进度，这些幻灯片都会放在这里，然后在这个板块，你能看到 R 代码和数据集。就像我说的，我的所有代码和数据都会放在那里，你现在已经能看到了。比如说这是对 R 的介绍，我们现在要做的，好吗？然后同样地，随着我们课程推进，你会看到越来越多的内容出现。如果你发现或者觉得有些应该出现的东西并没出现，那就给我发邮件，也许我忘了上传，请提醒我一下。

27.

English:

okay, so that’s that. and then the other important thing is this discussion thing. why does it not work? okay, so i will leave my student view. so i was in a student view, which means i could see exactly what you see. maybe that’s why it doesn’t work. okay, so the discussion, ah, this is what i was talking about. so you have, you can create a thread. so i believe you should all be able to create a new thread, okay? so please go and check it out. this is very brand new, okay?

中文翻译：

好的，这些就是内容。然后还有一个重要的东西就是这个讨论区。为什么它现在打不开？好吧，我先退出我的学生视角。我之前用的是学生视角，所以看起来和你们一样，也许这就是为什么它现在没反应。好的，这是讨论区，啊，这就是我提到的。你在这里可以创建一个话题，应该所有人都可以发起一个新话题，对吗？所以请你去试一下。这是个全新的东西，好吗？

28.

English:

so one day, last year, we used to use something or the other, so they have just decided to move to a different platform, which is a discussion. so this is my first time using it as well, okay? uh, so, uh, i guess we just have to play around and, uh, learn it, but i believe you should all be able to post a thread, and when you post a thread, uh, we can all see it, or, or if i think you must, you might be able to even, uh, post a thread, uh, and rest just to me, for example, i don’t know, uh, should be possible. anyhow, so this is how i would be communicating with you guys.

中文翻译：

之前，去年我们用的还是别的工具，现在他们决定换到这个新的平台，就是 discussion。所以这也是我第一次用，好吗？嗯，所以，嗯，我们只能慢慢摸索，我觉得你们应该都能发贴，一旦你发了贴子，我们都能看见，或者如果你想只给我看，可能也可以，我不确定，理论上应该行。不管怎样，我主要就会通过这里和你们沟通。

29.

English:

so this is my message that i sent five years back, uh, and these had, uh, this had all the details i was talking about. so just to be clear, the first test is actually second of may, Friday, and then the second test is Tuesday, third of June. as i said here, assignment one is either due Friday, twenty first, or Monday, twenty fourth, and assignment two is due in pretend. again, as i said, i might play around with these dates if you’re going slower or faster, depending on how we go, but that goes a number of guidelines, and i haven’t put dates for assignment three and four yet. i will do that later. okay, uh, that’s kind of the basic stuff, okay?

中文翻译：

这是我五天前发的消息，嗯，这里面包含了我提到的所有细节。需要澄清的是，第一次测验其实是在 5 月 2 日，星期五，第二次测验是在 6 月 3 日，星期二。就像我在这里说的，作业 1 要么是 21 号星期五交，要么是 24 号星期一交，而作业 2 在此之前交。就像我说的，如果我们进度变快或变慢，我可能会调整日期，这取决于课程节奏，但这里列了一些指导方针。我还没给作业 3 和 4 定日期，之后会补。好的，这些就是基本信息，好吗？

30.

English:

so if i go back to modules, so let’s go back to see the deal. so here, okay, so getting started with r and r studio. so how many of you have not done r at all? okay, so quite a few here. okay, so this is for you guys, okay?

中文翻译：

如果我回到 modules 这里，我们来看看。就是这里，好，这里是 getting started with r and r studio。那么，你们当中有多少从来没用过 R？好的，看起来有不少人。好，这就是给你们的内容，好吗？

31.

English:

uh, getting started with r and r studio, this is just a general introduction information. we booked up, this is not updated, so this would be a year old, so maybe some of the stuff is not exactly accurate, but it should be more or less accurate, so it just gives you a basic overview of what you need to run r. so r is a software which was, so r is one of the, i would say, one of the most widely used statistical softwares in the world, okay? so you have, i think, you have r, and you have a few others, but r is one of the most universal ones.

中文翻译：

呃，getting started with r and r studio 这里是一份通用的介绍信息，我们做好了，但这个文件没有完全更新，因为这是一年前写的，所以里边有些内容可能没有非常准确，但大体是对的。它会给你一个基础概览，告诉你运行 R 所需要的东西。R 是一款软件，应该说是全世界最广泛使用的统计软件之一，好吗？比如，你有 R，还会有其他几个类似的软件，但 R 算是通用性最强的之一。

32.

English:

but how many of you knew that r was stand-up right here? so r was developed in the department of statistics here. so here i look, i’m probably the least interested person in r in my department, but you know, we have people who are rejects first, so actually manage the whole r thing, okay? why am i the least interested person? but i’m least interested in any programming language, because for me, it’s just a tool, right? i’m more interested in the concepts. so i just do the code that i need, okay?

中文翻译：

但是你们当中有多少人知道 R 就是在这里开发的？对，R 是在我们这儿的统计系开发的。所以我看着，其实我是我们系里对 R 最不感兴趣的人，但你知道，我们有些人是 R 界的大神，他们真的是在管理整个 R 的项目，好吗？为什么说我是最不感兴趣的？因为我对任何编程语言都不算特别感兴趣，对我来说，这只是个工具，对吗？我更感兴趣的是概念。所以我就只写我需要的代码就行了。

33.

English:

but, ah, you, that shouldn’t deter you. you need to learn coding. i’ve done, okay? i’ve done a lot of coding that i needed, okay, so i can choose what i need to do now, okay? so on that note, so that’s r. let me actually go to r, and let’s get started. so r was developed in the early nineties, okay? so long time back, and it’s probably one of the earliest open source versions for statistical analysis anyway, okay?

中文翻译：

但是，嗯，这不应该阻止你们去学编程。你们要去学编程。我也学过，好吗？我学过很多我需要的编程，所以现在我可以选择我自己想做的事，好吗？基于这个，回到 R 上，让我们真正开始吧。R 是在 90 年代早期开发的，好吗？那是很久以前了，而且它大概是最早的开源统计分析软件之一，好吗？

34.

English:

and back then, we used to have paid packages. one of them was called ace, so ace was a language, and so what these guys said is, okay, well, we can develop something very similar to ace, but make it open source. and so what’s the letter next to s is r. so they call it r, and so there you have it, okay? now, r is, of course, the back end. what we actually use is r studio. r studio is a nice front-end, uh, an editor provider, everything, uh, that allows you to, uh, run codes, okay?

中文翻译：

在那个时候，我们常用的是一些付费的统计软件包，其中有一个叫 S（注：原文听起来是 “ace”，但大多情况下是指“S”，可能和S语言有关——此处尽量尊重英文原文），所以这些人就想，说好啊，我们可以开发一个跟 S 很类似的东西，但是是开源的。而在字母 S 后面就是 R，所以他们就叫它 R，于是它就诞生了，好吗？现在 R 当然是后端，我们真正用的是 R Studio。R Studio 是一个很好的前端，一个编辑器界面，提供各种功能，让你能运行代码，好吗？

35.

English:

so you go make those codes. but you get the idea, okay? so basically, it’s actually going and opening my previous codes, so you get four panels in r studio. so what i mean to say is, when you install r first time, you might need to install two separate things. you might need to install r, and then you might then need to install r studio separately. sometimes, just installing r studio, it comes with r. if that happens, that’s great. but otherwise, you need to install two different things, okay?

中文翻译：

所以你写好代码就能运行。你明白了吧，好吗？基本上，它会打开我之前写的代码，然后在 R Studio 里会有四个面板。我的意思是，当你第一次安装 R 时，你可能要安装两个东西：先装 R，然后再装 R Studio。有时候，你只要装 R Studio 就自带 R，这种情况下就好，如果没有，就得分别安装，好吗？

36.

English:

so i’m going to be able to use r studio. r is just in the background if you need, okay? so what i’ve got here, i’ve got four different panels. this is, of course, the editor. you can have all your editors here. this is interesting also. so, you know, whatever, you guys know what the console is. this is that moment. so any objects that you created, variables that you created, they will be here. uh, that’s a model object, uh, for one of my clients, face models, uh, absolutely nothing to do with this course, but, uh, just either of saving at some point, uh, and so let’s just, uh, load it up, okay?

中文翻译：

所以我这里可以用 R Studio，R 在后端提供支持就行，好吗？现在我这里有四个面板。这当然是编辑器，你可以把所有脚本编辑器都放在这里。这也挺有意思。你们都知道控制台是什么，对吧？就是这个窗口。任何你创建的对象、变量都会显示在这里。呃，这里有个模型对象，嗯，是我为一个客户做的 face models，和这门课完全没关系，只是随手存的。嗯，让我把它载入一下，好吗？

37.

English:

uh, the reason i’m showing you this is because this is how typically r is structured. so you will create, uh, and we will do this in the second half when we work with regression, so we will create a regression object, and that object will contain all the elements that are, you know, part of that object. so lots of information, different types of information. so in this case, you can see, uh, that i’ve got this model, but it contains all these objects, um, you know, items inside this object, okay?

中文翻译：

呃，我之所以给你看这个，是因为这展示了 R 的典型结构。你会创建一些对象，比如在本课后期我们做回归分析时，你会创建一个回归对象，这个对象里会包含所有与其相关的元素。很多信息，不同类型的数据。所以，在这里你看到，我有这样一个模型，它内部包含了所有这些东西，嗯，全都打包在这个对象里，好吗？

38.

English:

uh, lots of them, okay? so this has basically all the information that you need, uh, for, from that pocket, basically, that object, okay? so, uh, this is our stack, and so on, so forth, okay? so that’s how it looks, uh, if i had any plots, they would show up here, uh, but of course, i also show you how to, uh, open them in a separate window, so that it is easy to save and work on, okay?

中文翻译：

嗯，非常多。它基本上包含了你需要的所有信息，就像一个打包好的对象，好吗？所以，这里是我们的栈，等等，就这样一层层的。好的，通常，如果我有任何绘图，它也会显示在这里。当然，我也会教你怎样把它们打开在一个单独窗口，这样你就能更方便地保存和处理，好吗？

39.

English:

so, um, let’s open what we need to. base, which one does that? how is the color scheme, by the way? so this is the color scheme i, you know, i prefer, because it’s nice and cool on the eyes. it doesn’t open the exact color scheme, of course, but you can go and change that. so the reason i was asking is i wanted to make sure that you can clearly read, okay? so at any point, if you can’t read or if the font is too small, just let me know.

中文翻译：

那么，让我们打开我们需要的东西。base，哪个来着？顺便问一下，你们觉得这个配色怎么样？这是我比较喜欢的主题，因为看着比较柔和。当然它不一定是默认的配色，但你可以自己去改。之所以问这个，是想确保你们能看清，好吗？如果你随时看不清，或者字太小了，就告诉我。

40.

English:

okay, so, uh, this is on canvas, so you guys literally can open it now, and you can run it, uh, with me if you want to, uh, and if it is very self-explanatory. so, uh, the way you put comments in r, uh, is using this hash. so as you can see, i’ve used hash to sort of, you know, create a doc of what i’m doing, uh, as you can see, uh, you know, although i, you know, you know, i keep moaning about coding, uh, i’ve actually had a really good playing coding, because i actually comment my code, no, i use all the good practices for coding, but you do the same, okay?

中文翻译：

好的，这些都在 Canvas 上，所以你们现在也可以打开，想跟着我一起跑就可以跑，而且里面的说明也很清晰。呃，在 R 中添加注释就是用这个井号。所以你看，我用了井号标记来说明我在做什么。呃，你看，虽然我总是抱怨写代码，但其实我自己写的代码还挺规范的，因为我会加注释，也会用一些代码的最佳实践。你们也要这样，好吗？

41.

English:

so you save your codes with the comments, so that next time when you need something, so assignment two, you might need something that you use in assignment one, if you haven’t integrated well, then you know what the text says or not, okay? right, so defining variables, it’s just as simple as that. so a simple detail, y equal to 17, whatever, z is x times y, you know, this is just super basic stuff, okay, and the way you run this is, you know, so it’s a line-by-line compilation, so you can just run one line at a time. if i press ctrl+enter, it runs.

中文翻译：

你要带着注释保存代码，这样下次你需要其中的内容时，比如作业二时，你可能需要用到作业一里写的东西，如果之前没整理好，你根本不知道哪里在干什么，对吧？好的，所以定义变量，就这么简单。比如一个简单的示例，y=17，随便，z=x\*y，你知道的，这都是非常基础的东西，好吗？然后在 R 里执行的方式是这样的，每行都可以独立运行，我按一下 Ctrl+Enter，它就跑了。

42.

English:

uh, there are other ways to run, for example, uh, this button here for you to press, uh, of course, okay? um, you guys have figured that out. so, um, i’m just, wingtenter, linker, because it’s easier. so x, y, z, whatever, uh, and the if you just type x, it will print it here in the console, and then this is just defining it. all right, so an array of numbers, we define it using command c, which means combine all of these elements together in one object, and then mean, uh, mean, and, uh, s, c, standard division, uh, these are all just ready-made commands. summary will give you a summary statistic.

中文翻译：

呃，还有其他方式可以运行，比如，你可以按这里这个按钮，当然可以，好吗？嗯，你们自己摸索就行。于是，我现在就简单演示一下 x, y, z 之类的，如果我只输入 x，它就会在控制台里打印出来，然后就算定义好了。好，至于一组数字数组，我们是用命令 c 来定义，也就是 combine，把这些元素组合到一个对象里，然后 mean（均值），呃，mean（中值？实际上 mean 就是均值啦），还有 s, c, standard deviation（标准差），这些都是内置函数。summary 会给你一个摘要统计。

43.

English:

so if i, uh, i do summary if you want, it’s gonna give me, um, minimum, the first quantile, median, mean, third quantile, maximum, okay? just the basic stuff. so just looking at that output here, if i use the summary command, you get all that output. what is missing from that output that you would need if you’re looking at a statistical information or statistical summary of the value? using the mean, the median, minimum, maximum, q1, q3. ah, does anybody know what the first quantile and the third quantile, what does that mean?

中文翻译：

所以如果我输入 summary，如果你愿意的话，它会给我最小值、第一个四分位数、中位数、均值、第三个四分位数、最大值，好吗？就是这些基本信息。比如在这里，如果我用 summary 命令，你就能看到所有这些输出。对于统计信息或者统计汇总，你觉得少了什么？我们只有均值、中位数、最小值、最大值、Q1、Q3。嗯，有人知道第一个四分位数和第三个四分位数分别是什么意思吗？

44.

English:

no? so the first quantile is the 25th percentile, so a quarter of values are below that. third quantile is the 75th percentile, okay? um, so what you really have is you have the minimum, you have the 25th percentile, so with one quarter of the values are, ah, so that’s below 2.5, the median, what is median? median is also the 50th percentile, the midway point, okay? so, uh, so you have the 25th, the 50th, the 75th, and then you have the max, okay, and you have the mean.

中文翻译：

不知道吗？第一个四分位数其实是 25% 分位，也就是有四分之一的数值在它下面。第三个四分位数是 75% 分位，好吗？嗯，所以这里你实际上得到了最小值、25% 分位（也就是四分之一的数据落在它下面）、中位数（它是 50% 分位，中间位置），好吧？然后是 75% 分位和最大值，还有均值。

45.

English:

so what is missing? variance. variance, that’s right, isn’t that right? so variance or standard deviation, okay? what does that tell you, any idea? what does the variance tell you? or even this? the distribution of the dataset, the variability in the dataset, yeah, so that’s very important, okay, so we are dealing with uncertainty, so you need to know the standard deviation or the variance, so it doesn’t give you that, okay, so just have sum, okay, plot will just plot it, but there are other ways to plot, we will go to that, okay.

中文翻译：

那么缺了什么呢？方差。对，方差，是这样吗？所以方差或者标准差都没有。你知道它能告诉你什么吗？方差能告诉你什么？或者说标准差能告诉你数据集的分布情况，它的波动有多大，对吧？这非常重要，因为我们要处理不确定性，所以需要知道标准差或方差，但这里 summary 并没有给你。所以我们只有这些基本指标。好吧，plot 这个命令只是画图，但还有其他画图方法，我们一会儿会说。

46.

English:

now i’m going to start using a dataset, which we will also do in our lessons later, so i’m reading a csv file, so this is just importing a csv file, and this option here is a file, choose, open records, close records, that basically opens a dialog box, so you can literally go and choose the file, which is much easier, rather than having to provide a whole path of the file here, okay, and header equal to true basically means r that the first row is just headers, so pretty self-explanatory, okay, so when i run that, it’s gonna open another box, which then allows me to go and find my way through.

中文翻译：

现在我要开始使用一个数据集，我们之后的课程也会用到。同样地，我会读取一个 csv 文件，这就是导入 csv 文件的方式，这里用到了 file.choose()，然后用括号括起来，这样做就会弹出一个对话框，让你自己去选文件，比在代码里写完整路径要方便很多。好，然后 header=TRUE 表示告诉 R 第一行是表头，含义很明显，对吗？所以当我运行它，就会出现一个窗口，让我自己去找文件路径。

47.

English:

so, okay, and then as you can see, i’ve got the school over here, okay, i will actually use another feature, which is one of my favorite features, so that you guys can, ah, you know, it’s very useful, ah, so i forgot to clean my environment, so there’s all these things which i don’t need, so how do you clean your environment? well, there’s a room here, nice, you know, just press that, clean your environment, so clean, of course, the same thing here, usually quite useful when you’re starting something and you don’t want to have that pleasure.

中文翻译：

这样，好，然后你看，我这里有了 school 这个对象，好吗？实际上，我还会用到另一个功能，这是我最喜欢的功能之一，你们也会觉得很有用。嗯，我发现我忘了清理一下环境，这里还有一些我不需要的东西。那怎么清理环境呢？你看这里有个小扫帚图标，挺好的，你只要点击它，就能清空你的环境，是不是？所以就清理掉了。通常在你开始一个新的项目时，这很有用，你不想有多余的东西在那里。

48.

English:

so i have to now, of course, choose the file again. okay, so what is this? let’s look at the state of it, so it opens here, okay, and, oh, the laptop, um, so, you know, the screen is not that big, but, um, you get the idea, so because we specified that the first row is headers, it has taken those first rows as header names, so school, roll, so this is actually data of New Zealand high schools, okay?

中文翻译：

所以我现在当然要再选一次文件。好的，这是什么？让我们看看它的状态，这里打开了，好，然后，哦，这是台笔记本电脑，所以你知道，屏幕不太大，但你能看出来，因为我们指定了第一行是表头，所以它把第一行当作表头名称，比如 school, roll，这实际上是新西兰中学的数据，好吗？

49.

English:

so i think this dataset has almost all the high schools in New Zealand, or maybe a large chunk of high schools in New Zealand, ah, so you have the name of the school, and then just roll, which means how many students there are in that school, and the number of teachers, the region is a name, and the region has a code number, okay, and then the student to staff ratio, and then which island it is, North Island or South Island, okay, actually, let me open this thing in existing, so we can have a better look. this is much bigger, okay, so you can see the name of the school, the school, and on the show, okay, heaven, look at this, take a second, okay, you know, just what’s on the screen.

中文翻译：

我想这个数据集几乎包含了新西兰所有的中学，或者至少是一大部分中学。你可以看到学校的名字，然后 roll 代表这个学校里有多少学生，还有教师人数，region 是这个学校所在的地区名，并且还有一个代码，然后是学生和教职员工之比，以及它在哪个岛——是北岛还是南岛。对吧？好，让我把它在现有窗口里打开，这样看得更清楚。这更大一些，好，你能看见学校名称、学校、然后有 roll，好的，看一下，你先大概扫一下屏幕。

50.

English:

so let me explain a couple of things. so the roll, of course, is the number of students, okay, and then your number of teachers, and then your student to staff ratio. so student to staff ratio is how many students do you have for every teacher, okay, so ideally, i think it should be around fourteen or fifteen, okay, that’s the lifeline, so they aim to have about fourteen or fifteen students for each teacher actually.

中文翻译：

让我解释几个地方。roll，当然就是学生数量，好吗？接下来是教师数量，然后是学生与教职员工的比率。所以学生与教职员工的比率指的是每个老师对应多少个学生，对吗？理想情况应该在 14 或 15 左右，这就是一个标准线，所以他们的目标就是每位教师大约对应 14 到 15 个学生。

51.

English:

what they aim is to have enough teachers so that they have one teacher every for them fifteen students, okay, and so you can see that there is, of course, variation here, okay, most school schools seem to be around that number, but some are much higher, like sixteen here, some are lower, thirty sixteen here, even lower, so those of you who are new to New Zealand, some of these schools are rural schools, which means they’re in an area where there are very few people, uh, and so very few schools, and therefore, you know, you typically have, you know, maybe slightly more teachers, uh, sometimes there is also a shortage of teachers, so you could be gathered around, but, you know, um, that then see that variation, and you can see, so, for example, this one here, there are just eighty-one students in that quote, high school, okay?

中文翻译：

他们的目标是有足够多老师，这样每 15 个学生就配一个老师，好吗？所以你能看到这里确实有些差异，大多数学校都在那个范围左右，但有些要高很多，比如 16 或者更多，有些要低一些，30 多，16 所在的这里更低。所以对你们这些对新西兰不太熟悉的人来说，有些学校在偏远地区，意味着那里人口很少，呃，也没有多少学校，因此一般会有相对更多的教师，有时也会缺老师，所以分配不均。嗯，你能看到这种差异，比如说这个学校，这儿只有 81 个学生，好吗？

52.

English:

so this is, uh, an high school, so the high school starts year nine, so nine, ten, eleven, twelfth, thirteenth, so five years of high school, okay, so five years, and only eighty students, on the other hand, you have some schools here about nine hundred students, so these are, you know, bigger centers, uh, if you go to, uh, schools in Auckland, or Hamilton, or Christchurch, one of the biggest cities, you will probably see a thousand schools as well, so there’s a lot of variation, okay, and then the school to staff ratio, of course, keeps varying.

中文翻译：

所以这是一个高中，从 9 年级开始，9、10、11、12、13，一共五年，好吗？五年就只有 80 个学生。另一方面，也有些学校有 900 个学生，这些就是更大型的，比如在奥克兰、汉密尔顿、基督城这样的大城市，你可能看到超过一千个学生，所以差异是很大的，好吗？学生与教职员工的比率自然也会一直在变动。

53.

English:

so thanks the dataset, so this is what i was saying, you know, eleven hundred, thirteen hundred, fifteen hundred students, they’re all schools and all, okay, so does anybody have any questions about this data? what’s the meaning of the column d, sorry, on d, yeah, oh, so that’s the region, uh, it’s just a number, so there are two big, so region is indexed and is the number,

中文翻译：

这些就是数据，所以就像我说的，有的学校有 1100、1300、1500 个学生，各种规模都有，好吗？对这份数据有什么问题吗？你问那一列 d 是什么意思？抱歉，哪一列？对，哦，那是 region，它只是一个编号，所以 region 这个字段是由序号标识的，

54.

English:

um, one thing that strikes me is, why is the teacher not like a whole number? why is a teacher not a whole, that’s exactly why i asked you that question, because i wanted somebody to point it out. why is a teacher not a whole number? can you make a fraction of a teacher as well? some teacher working full-time, some teacher working part-time, exactly, some teacher working part-time, right, so that’s basically why, but i really want, this is exactly why i wanted you to, yeah, okay, because when you look at the data, you need to, you know, look at it carefully, you need to know, you know, have these questions, right?

中文翻译：

嗯，有一点让我注意到的是，为什么教师这一列不是整数？为什么老师不是一个整数数量？正是因为我想让你们提出这个问题，所以才问的。为什么会有小数呢？难道老师还能切分？对，有的老师是全职，有的老师是兼职，对，有些老师是兼职，对吗？所以这就是原因。但我真想说，这就是为什么我要让你们真正去看数据，因为当你观察数据时，你得仔细想，你得提出这些问题，对吧？

55.

English:

okay, why, um, and then, of course, sometimes there is a valid explanation, which is fine, uh, but sometimes it could be that there is actually a problem in the data, you know, you can’t have 42.1 teachers, for example, well, in this case, you can, okay, uh, but, okay, so that, that’s important, okay, so that’s, that’s basically the dataset, right, and we are going to look at this data, um, a lot for the, at least some of the lessons in the, uh, first few weeks, so how many, we’ve got about 350 or high school, okay, right.

中文翻译：

好的，为什么是这样，嗯，然后当然有时会有合理的解释，这没问题，但有时也可能是数据本身有问题，比如你不能真的有 42.1 个老师，对吗？不过在这个例子里其实是合理的，好吧。所以这些就是关键之处，好吗？这就是这个数据集，我们在接下来几周至少一部分课程里会一直用到它。大约有 350 所高中，好吗？好。

56.

English:

so let’s go to, let’s go back to R, basically, what i did is i have imported this dataset in R, it’s called roll, so roll is now the dataset object that contains the data, and so i’m just showing you a few basic things that you can do with this dataset. what i like to do is i like to create variables that represent each of the important columns in the dataset, so that is easier for me to access that information, rather than, um, so what i’m doing, for example, is i’m defining students to be the second column in, i guess, the roll, uh, the first row, of course, is, so you specify row and column, uh, i’m not talking about row, so that is empty, i only don’t consider, what, the column, so basically, i’m saying that the whole of second column is students, fine.

中文翻译：

好，让我们回到 R。基本上，我现在做的事就是把这个数据集导入到了 R 里，变量名是 roll。所以 roll 是包含所有数据的对象，现在我展示一些你可以对这个数据集做的基本操作。我喜欢先创建一些变量，代表数据集中每个重要的列，这样之后要访问某一列信息时会更方便。比如说，我这里做的，就是把数据集中第二列（我猜是 roll 的第二列）定义为 students，呃，第一行什么的，你指定行列的时候，如果是空，就表示所有行，我只是要这一列。所以就是说，整条第二列都归到 students 里，好吧。

57.

English:

so once i define students, teachers, ratio, and region, which are the four main variables, i can just work with these variables much easier, so that’s what i personally do, you don’t have to do it, there are different ways of doing it, so one of the things in R is that if you, for example, and i encourage you to go online, find resources to learn R, if you haven’t done any R, and you will see that if you check multiple resources, likely you will come across multiple ways of doing the same thing, okay, literally, there are always ten different ways of doing anything, especially now, so this is just how i do it, okay?

中文翻译：

所以，当我定义好 students、teachers、ratio、region 这四个主要的变量之后，我就能更方便地使用它们了。这是我个人的做法，你也不一定要这样做，还有别的方式。在 R 里常常如此，如果你网上搜搜 R 的资源，尤其当你初学 R 时，你会发现处理同一件事可能有好几种方式，好吗？基本上一件事可以有十种方法来做，尤其是在当下。所以这只是我的一种做法，好吗？

以下内容衔接上一次对照翻译的末尾（即序号 57）继续进行。仍保持「英文原文」在上、「对应的中文翻译」在下的上下结构，无任何删减或总结，只添加必要的标点符号。段落编号将从 58 开始延续，以保证衔接完整。

58.

English:

so just going to run all of that, okay, and then, because i’ve got these, variable names, i can now use the summary statistics on soldiers. so this would be very interesting, because this is velido, okay, so let’s have a look, so you can see that we have this statistics, which looks like this, okay, so have a look, and tell me, what can you infer about this data, like this video storage, just looking at that information, which is the summary strategistics, so specifically, i’m asking you to infer about the distribution of variable.

中文翻译：

好了，我就运行所有这些，好吗？然后，因为我已经有了这些变量名，所以我现在可以对这些数据做摘要统计。这会很有意思，因为这是 “velido”，好，我们来看一下，你可以看到我们有了这样的统计结果，看上去是这样，好吗？看一看，然后告诉我，你能从这些数据里推断出什么，比如这个 “video storage”，只看这些信息，也就是摘要统计。更具体来说，我是想让你们推断一下这个变量的分布情况。

59.

English:

so have a look, and tell me, what can you infer about the distribution of variable. specifically, i’m asking you to infer about the distribution of variable. skewed to the right, skewed to the right, why, the median and mean are not the same, i can’t hear you so, okay, okay, that’s one, okay, so is basically saying that max is way too much bigger, basically, okay, so that’s a really good observation by the way, okay, so max is really much bigger than rest of the data.

中文翻译：

看看，然后告诉我，你能从这个变量的分布里推断出什么。具体来说，我想让你们说说这个变量的分布形态。有人说偏右偏态，为什么呢？因为中位数和均值不一样？我听不太清，好，好，对，这是其中一个原因，对吗？这基本上就是说最大值远远大于数据的其他部分，好吗？这其实是个很好的观察，对，非常好。最大值确实比其他数据高出很多。

60.

English:

so your 75th percentile is 800, max is 2000, so that then see that there is a tail on the right, like maybe a long tail on the right, so that’s a good point. anybody else? so see, the mean and the median are not the same, okay, which one is higher, the mean is higher than the median. the median is always the 50th percent of the data, right, so 50% of the observations are below it, 50% are above it, so you know that 50% of the schools have a roll up to 589, and the rest 50% have 589 or above.

中文翻译：

所以当你的 75% 分位是 800，而最大值却是 2000，这就说明右边有一个尾巴，可能是很长的右尾，这是个好点子。还有别的吗？你看，均值和中位数并不相同，好吗？哪个更高？均值比中位数高。中位数始终代表数据的 50% 分位，对吗？也就是说有 50% 的观测值在它下面，50% 在它上面，所以你知道有一半的学校学生数在 589 或以下，另一半在 589 或以上。

61.

English:

okay, so, um, the mean is higher, what does that mean, what would that mean, yeah, presence of outliers, presence of outliers, yeah, sure, that’s a great point, mean is influenced by whether there are outliers or not, mean is influenced by extreme values, okay, so if you have extremely small or extremely large numbers, they would influence the mean, they won’t influence the median, because median is always the middle point in terms of how many observations are below the data and how many observations are above, okay, so that, again, that actually is consistent with this finding.

中文翻译：

好的，那么均值更高，这意味着什么？有人说是因为有离群值，对吗？对，离群值的存在，对吧？是啊，这一点不错。均值会被离群值影响，而中位数不会。均值会受极端数值的影响，好吧？如果数据中有非常小或非常大的值，它们会拉动均值，但不会影响中位数，因为中位数始终是基于有多少观测在下方、有多少在上方这个角度的中间点，对吗？所以，这和我们之前的结论是一致的。

62.

English:

okay, there are probably a few schools that have very large roll, okay, so that is influencing the mean to be larger than the median, also the max is really high, um, the other way to look at this is also just compare, um, the difference between the quantile, between the median, and between the median and the same effect, what do you see? so the 25th percentile is 353, the median is 589, okay, so the difference is 236, or wait, did i do that correct, i might do that correct, let me do that carefully, 353 plus 200 is 553, plus 36 is 589, so yes, 236, okay.

中文翻译：

好，也许有几所学校的学生人数非常多，这会使均值比中位数更大，而且最大值也的确很高。嗯，另一个看法是比较四分位数和中位数之间的差距，还有中位数和另一个指标之间的差异。你看到了什么？25% 分位是 353，中位数是 589，对吗？那么差距是 236，等等，我算对了吗？我再仔细算一下，353 加 200 等于 553，再加 36 等于 589，所以是 236，对吧，好。

63.

English:

whereas here, not 47, i think 147, let’s see, 353, 453, 553, oh, i see, i made a mistake: 353 to 589 is 236, not 147, okay? so that difference is 236, whereas 589 to 882 is almost 293, or something. so again, there’s a bigger difference on the right-hand side, which is consistent with that distribution being right skewed, okay, so let’s keep that in mind.

中文翻译：

而另一边，不是 47，我想是 147？我来看看，353、453、553，哦，原来我算错了，从 353 到 589 其实是 236，不是 147，好吗？所以那个差距是 236，而 589 到 882 差不多是 293 左右。所以你看到右侧的差距更大，这也说明分布右偏，好吗？要记住这一点。

64.

English:

(as you can see, i’m not really spending your time telling you about the r command, okay, that’s how i will teach, but the important thing is how do you get the data, i mean, that’s the really important thing. anybody can learn R, in fact, these days, AI can do better R than you can, i guess, so it’s not about you doing R, it’s about you understanding what is happening.)

中文翻译：

（你也注意到，我并没有花时间详细讲 R 的命令，好吗？这就是我的教学方式。真正重要的是如何获取并理解数据，我觉得这才是真正关键的。任何人都能学 R，其实现如今，AI 写 R 代码可能比你还快，所以问题不在于你会不会写 R，而在于你对正在发生的事情有没有理解。）

65.

English:

so the second thing is, uh, there’s so much information, uh, even in that basic line of output, if you know what to look for, and that’s what i will want you to do it, uh, when you do these course, okay, so keep your eyes open. any other questions or comments before we go further?

中文翻译：

因此，第二点是，即使在那些基础的输出中，也有非常多的信息，只要你懂得怎么看。这也是我希望你们在学这门课时去做的，好吗？多留意细节。我们继续之前，还有没有别的问题或意见？

66.

English:

okay, so as we discussed, this doesn’t tell you the important aspect which is the variation, so i’ve got the standard deviation as well, so the standard deviation is 376, what does it mean, while you compare the standard deviation against the mean, it’s almost half or maybe more than half of the mean, so that tells you that there’s quite a bit of variation in the data, okay, compared to mean, and again, that is consistent, because we already know there’s a lot of variation, so no surprises.

中文翻译：

好，如我们讨论的，这些摘要统计没有告诉你一个关键点，就是变异程度。所以我这里还算了标准差，标准差是 376，这意味着什么呢？当你把标准差和均值相比，大概是均值的一半，甚至可能超过一半，这就说明这个数据里存在相当大的波动，好吗？和均值比起来也不少。而这和我们之前所知道的情况一致，因为我们已经知道这个数据的差异度挺大，所以这并不让人意外。

67.

English:

and then we can do the same thing for the teachers, and you can see that, but in this case, the mean and median are kind of a bit closer to each other, so maybe this distribution is slightly more symmetric, but you can still see that the max is quite a bit higher, okay, what is unique here though, okay, have a look at the first quantile, the median, and the third quantile, what do you notice for teachers?

中文翻译：

接着我们对教师这一列也做了同样的事，你能看到，在这个列里，均值和中位数相对更接近一些，所以它的分布也许更加对称，但你还是能看到最大值相当高，好吗？不过这里有什么特别之处呢？你看看第一个四分位数、中位数和第三个四分位数，在教师这个变量上你发现了什么？

68.

English:

the difference between first quantile and median is 15, exactly the same as the difference between the median and the third quantile, so there’s perfect symmetry, okay, so that tells you that the 50th, the 25th, and the 75th are evenly spread out, okay, so it is the central part of the data is E symmetry, we don’t know about the data, or that is the central part is symmetry, okay, so that indicates to you that this is, you know, more symmetric than students.

中文翻译：

第一个四分位数和中位数之间的差是 15，和中位数到第三个四分位数之间的差也正好是 15，所以是完全对称，好吗？所以这说明 25% 分位、50% 分位和 75% 分位在中部分布得非常平均，好吗？所以数据的中间部分很对称，也许外部部分我们不知道，但至少中心部分挺对称，好吗？这表明这个分布比学生人数那个更对称一些。

69.

English:

you see what is instructed here, which is n a s, what does that mean, does anybody know? so if this is, you know, usually it means that n a is not applicable, uh, but basically it shows up when you have missing values, okay, so that tells me there is one missing value in the teachers, so for one of the schools, you don’t have the number of teachers, okay, so again, if you have missing values, it will show up here.

中文翻译：

你还看到这里有一个 “n a”，这是什么意思？有人知道吗？嗯，通常 n a 表示缺失数据（not applicable 或者 not available），当有缺失值时就会显示在这里，好吗？所以这告诉我，在教师这一列里有一个缺失值，也就是说有一所学校教师数量是空的。好吧，所以如果你有缺失值就会出现这种情况。

70.

English:

and then the standard deviation for teachers is about the same with respect to the mean, so it’s about half of its mean, so still quite a bit of variation, okay, and then we look at the ratio, the important aspect, because as i said, we expect the ratio to be close to fourteen or fifteen, uh, then you can see that the mean is 14.67, the median is 14.97, so the median is slightly greater than the mean, but i think, not sure if the distribution is spiced enough or not, though, okay, you can see there is a difference of 1.5, there’s a difference of like 2.2, right, so again, bit more difference on the right-hand side, but we can look at it, um, better with the plot also.

中文翻译：

然后教师这一列的标准差和它的均值相比，大致也是一半左右，所以仍然有不小的波动，好吗？接着我们看 ratio，那是很关键的一个指标，因为正如我说的，我们希望 ratio 大约在 14 或 15 左右。嗯，那么你就能看到这里的均值是 14.67，中位数是 14.97，所以中位数比均值稍微高一点，但我觉得还不太确定分布够不够明显地倾斜，好吗？你看有个差距大约 1.5，另一个差距 2.2，对吧？所以右边一侧差异稍微大一点，不过我们还可以用图来看看。

71.

English:

the other thing is we can also, uh, find quantiles, um, so just any quantiles. in this case, i’m looking at the 5th percentile and the 95th percentile, so the quantile command will give you any percentiles that you need, and because i need two percentiles, i’m clubbing them with the function c, so the c is a generic function in R that you can use to club multiple values together, okay, so even when you, so you can insert that within a command, like i’ve done, to tell R that i need these two percentiles for this variable ratio, okay, and so, yeah, the 5th percentile is 10.75, the 95th is 17.47, gives you a better idea for the distributions.

中文翻译：

另一件事是我们也可以去算一些分位数，比如任何分位数。在这个示例中，我要看 5% 分位和 95% 分位，所以 quantile 这个命令能给你需要的任何分位数。而因为我要两个分位数，我就用 c 这个函数来把它们组合起来。c 在 R 里是一个通用函数，可以把多个值放在一起，好吗？所以就像我现在做的，在一个命令里告诉 R 我需要 ratio 这个变量的 5% 和 95% 分位，好吗？结果是 5% 分位是 10.75，95% 分位是 17.47，这能让你对分布有更好的了解。

72.

English:

however, always better to plot. now this, so if i just do plot teachers versus students, okay, it will show up here, okay, you can see nice, all those straight lining relationship there, but it’s, you know, this is a small window, so it’s always better we plot it on a separate window, and you can do that using the x11 command, get some x11 plot, will essentially open a new window.

中文翻译：

不过呢，最好还是画个图来看。就拿这个来说，如果我简单地画一下教师人数和学生人数的散点图，好吗？它会显示在这里，好吗？你能看到一个看起来比较直线的关系，但你知道这是个小窗口，所以最好能在一个单独的窗口里画图。你可以用 x11 这个命令去做，x11 plot 大致就是打开一个新窗口来显示。

73.

English:

if so, when it comes to your assignments, i’ll ask you to submit your assignments in a pdf format. you can even do that using R markdown file, um, but you can also just use a word file. so you can type your answers, paste your output, you know, make it, you know, in a quarter fashion way, uh, and then convert it to pdf. so if you’re doing that, then getting your plots in a separate window like this is very helpful, because it makes it a nice big window that you can save as a jlk file or png file, and then you can insert, okay, so very helpful.

中文翻译：

在作业方面，我会让你们把作业做成 pdf 文件提交。你也可以用 R markdown 来做，不过你也可以用 word 就行。你可以把答案打好，把输出贴上去，弄得整齐一些，然后转成 pdf。如果你这样做的话，把图画在单独的窗口里就会很方便，因为它会是一个大一些的窗口，你可以保存成 jpg 或 png，然后嵌进去，好吗？这很有帮助。

74.

English:

while we are on this, how does this plot look, what can you say about what does this plot tell you, basically, sorry, are the two variables are in linear regression, a linear relationship, linear regression, okay, that’s right, basically, so the main point is, it’s telling you the relationship between the two variables, okay, and as she created, it looks like, you know, you could fit a linear regression with a straight line into it, so perfectly linear relationship, almost, okay, so, and that’s not surprising, because we are expecting that all schools aim to have a school-to-staff ratio about fourteen or fifteen, which is kind of the same, so if they have more students, you would expect them to have more teachers, okay, so, you know, this is what you expect to say.

中文翻译：

既然说到这张图，你觉得它看起来怎么样？它告诉我们什么？基本上，是不是这两个变量看起来呈线性回归关系？一个线性关系，对吗？对，没错。主要是，它告诉你这两个变量之间的关系。像你说的，看起来你可以用一条直线拟合做线性回归，几乎是一个完美的线性关系，好吗？这也不奇怪，因为我们猜想所有学校都想把师生比维持在 14 或 15 左右，所以如果学生多，就意味着需要更多老师，对吗？这就是你会看到的东西。

75.

English:

okay, but that’s that plot, uh, you can also, because we talked about, uh, the shape of the distribution, uh, the best way to do that is use the histogram, uh, i put an open test to them in a separate window, so i’m just not going here right now, so you can see that the student-to-staff ratio is not that symmetric at all, you can see there’s a tail, but the tail is on the left, not on the right, okay, so the student to staff ratio summary statistics was a bit confusing, because the median was smaller than the mean, yet q3 was further away from the median, so it was conflicting, so you can’t, you know, you couldn’t get a clear picture, but obviously, a picture speaks six thousand words, so when, you can always plot.

中文翻译：

好的，这就是那张散点图，嗯，我们还谈到过分布形状，对吧？看分布形状最好的方法是画直方图。呃，我把它放在另一个窗口里了，所以这里就不演示了。你可以看到学生和教职人员的比率分布其实并不对称，你能看到它有一个尾巴，但这个尾巴是在左边，而不是右边，好吗？所以之前我们看那些摘要指标时，会有点困惑，因为中位数比均值更小，但 Q3 又比中位数更远，所以会有矛盾，所以单靠那些数字你看不清楚。但显然，看图就一目了然，一张图胜过六千字（笑），所以你想看就随时可以画出来。

76.

English:

um, so yeah, uh, basically, uh, if you expect a school to staff ratio to be fourteen, fifteen, uh, then you can see that a bunch of schools are actually struggling to do that, uh, that’s one day, uh, plot is okay, you can also get a box plot, which is also usually good, and then a box plot by region as well, so i’m actually going to do this in a separate window.

中文翻译：

嗯，是的，基本上，如果你预期学校的师生比在 14、15 左右，那么你能看到好些学校其实还达不到。这是一方面。你可以用 plot 来看看，还可以画一个箱线图，这通常也很有用，而且可以按地区来画。我现在就把它放在一个单独的窗口里演示。

77.

English:

okay, let’s have a look at that, just out of interest, what, what can you spot, so what you have is student to staff ratio for different regions in New Zealand, okay, and you can’t see all the region names, because there’s not enough space, so you can see some of the names, but, uh, yeah, you can see the distribution, basically, for each, um, region, now, what does a box plot show, does anybody know what a box plot shows, which values does it plot?

中文翻译：

好的，我们来看看，好奇之下，我们能看出什么？你现在看到的是新西兰不同地区的师生比，对吗？你不会看到所有地区的名字，因为地方不够，所以只能看见部分，但你能看到各个地区的大致分布。现在，我想问，箱线图展示的是什么？有人知道吗？它绘制的是哪些数值？

78.

English:

it shows max, max and min, and the, on the first quartile, on the second quartile, yeah, almost right, okay, so basically, it plots the first quantile, so the box is the lower end is the first quantile, so 25th percentile, the upper end is the third quantile, 75th, but this is not the mean, this is the median also, okay, so you have basically the 25th, 50th, 75th, that’s the body of the box, and then the whiskers, these lines, they don’t necessarily go to the maximum, but they don’t, they basically go to where the main chunk of the data is, okay, now if there are outliers outside this main chunk of the data, they are shown as points, so what you see here, for example, uh, is that main, most of the data is from they say, i don’t know, about 18 point something, uh, to about 11 point something, so that range, but there are four schools which have a unusually lower school to staff ratio, so they are outliers, so they shown as points, okay.

中文翻译：

有人说显示最大值、最小值，还有第一四分位、第二四分位？对，差不多。箱线图的箱子部分，底边是第一个四分位数，也就是 25% 分位，上边是第三个四分位数，也就是 75% 分位，中间有根线表示的是中位数，不是均值，好吗？所以基本上，箱子的主体包含 25%、50%、75% 这几个位置。然后两条须线（whiskers）并不一定会延伸到最大值，它们基本上只到主要数据范围的边缘，好吗？如果在这个范围外还有离群值，就会标成一个单独的点。比如，你能看到这里有四所学校的师生比特别低，算是离群值，所以它们就以点的形式显示出来。

79.

English:

so that’s how you read a box plot. what you can say, just looking here, is that obviously every region is kind of different, okay, what’s the, so looking at that plot, which region do you think has highest median and lowest median, Siri, do you think the lowest, yeah, the lowest and the highest, the lowest be the west coast, the lowest is the west coast, okay, and the highest, so there’s Auckland, and there’s this Waikato, which one was, yeah, it’s similar, right, so does anybody know possibly why, basically, what does it mean, it means that in Auckland and Waikato, so Auckland is here, and Waikato is maybe slightly lower, but almost there, the median student to staff ratio is much higher, it’s almost 16, whereas in west coast, the student to staff ratio is almost 12.

中文翻译：

这就是解读箱线图的方式。仅仅看这张图，你就会发现不同地区差异很大。那我们问一下，哪个地区的中位数最高？哪个最低？Siri，你觉得最低是谁？对，最低的是西海岸（West Coast），好，最高的呢？奥克兰 (Auckland) 和怀卡托 (Waikato) 之间哪个好像更高？看起来相差不大。有人知道为什么吗？其实，这意味着在奥克兰和怀卡托，比如奥克兰在这儿，怀卡托略低一点点，但差不多，它们的师生比中位数更高，几乎达到了 16，而在西海岸却只有 12 左右。

80.

English:

so does anybody know that, sorry, population, yeah, so west coast, exactly, so west coast is one of the most, uh, you know, maybe one of the least, uh, populated parts in New Zealand, whereas, uh, Auckland and, uh, next Waikato are probably the most, uh, heavily populated, okay, so, uh, that’s really what you see, so you have them, basically, i don’t know, two-thirds of the population is between important Hamilton and Toronto so, this frankly, okay, so, um, something like that, so, um, so obviously the regions with higher population are up here, um, and, you know, least populations typically have, uh, lower ratio, because they still need the teachers anyway, okay, so you can see some interesting information, okay.

中文翻译：

有人知道为什么吗？有人说人口？对，西海岸就是如此。因为西海岸是新西兰人口最少的地区之一，而奥克兰和怀卡托是人口密集度最高的地区，好吗？所以，这基本就是你看到的情况，可能全国有三分之二的人口都集中在奥克兰、汉密尔顿、陶朗加这些地方，大体如此。所以那些人口更多的地区，师生比自然更高，而人口很少的地方往往师生比更低，但他们还是需要老师，对吧？所以你能看到一些有趣的信息，好吗？

81.

English:

we don’t need to go too much deeper here, but you get the idea, this is how you should look at plots and different types of plots. okay, so just out of curiosity, let’s look at, uh, histogram of students, this is what we discussed at the start, and as you have correctly pointed out, um, you know, that we had three clues to tell us that the data was not symmetric, in fact, it was, as, and you had a long tail on the right, and that’s exactly what you see, okay, so no surprises there.

中文翻译：

我们不需要在这里过度挖掘，你明白大概意思就好，这就是你该如何看待各种图表。好，再比如说，出于好奇，让我们看看学生数量的直方图，这就是我们一开始讨论的内容。正如你们指出的那样，我们有三条线索表明这个数据不是对称分布，实际上是有右偏，而且带着一个长尾，而这正是你会在图里看到的，好吗？所以这也不令人惊讶。

82.

English:

um, if i take a square root transformation of students, then the transform data looks nice and symmetric, okay, so this is something we will do in the second half, we take log transformation or square root transformation if you want to make data symmetrical, or make it roughly symmetrical, um, so just to show that it works sometimes, it works, okay, so if i wanted, um, to assume, like, a normal distribution, normal distribution is symmetric, ah, then, of course, if we’re involved on students, because students data is not symmetric, okay, so i can’t use a normal distribution, but if i do, so let me go back and explain that.

中文翻译：

如果我对学生数量做一个平方根变换，那么变换后的数据就会变得看起来比较对称，好吗？在下半学期里我们会做这件事，如果你想让数据看起来对称一点或大体对称，可以做对数变换或者平方根变换。只是想给你们演示一下这个方法有时的确有效，好吗？所以如果我想假设它是正态分布，正态分布是对称的，但原始的学生数据并不对称，对吗？所以我不能直接用正态分布，但我要是做了变换，比如……让我回到这里解释一下。

83.

English:

so this is how the students are distributed, that’s more like a normal, because normal distribution is, so, what would i do, well, one way, it is to try and transform the variables, so if i take the square root of students variable, then i get a nice symmetrical distribution, so then i can use a normal reference, okay, we will do this in the second half of the regression, okay.

中文翻译：

这就是学生数量的分布，那就更像一个正态分布，因为正态本身是对称的。我会怎么做呢？做法之一就是对变量做变换。如果我对 students 这个变量做平方根变换，我就能得到一个更对称的分布，这样我就可以用正态参考来处理，好吗？我们在回归分析后半部分会讲这个。

84.

English:

uh, so i’m creating another dataset for roll, which is the data frame, so this is a way to create a data frame data.frame, okay, and then you can just input the variables, roll, and that frame, and then doing this, basically what i’m doing is, again, there are zillion ways of doing things, the aggregate command is one of the ways to find the mean or the median or any other summary for the whole dataset, so that’s what i’m doing, so i’m doing by region, okay, i’m creating mean, uh, student-to-staff ratio by region using that command.

中文翻译：

然后呢，我又给 roll 创建了另一个数据集，叫 data frame（data.frame），这是在 R 里创建数据框的一种方式，好吗？然后你可以把各个变量，比如 roll 之类的都塞进去。然后我在这里其实是再做一件事，你要知道在 R 里做同一个事情方法很多，aggregate 命令就是其中一种方法，它可以帮你计算整个数据集的均值或中位数或者别的汇总指标。所以我现在在做什么呢？我按地区去算学生和教职员工的平均比率，使用的就是那个命令。

85.

English:

so i’ve just put that command there, it’s not that we’re going to need that command a lot, but i’ve just put a random set of commands in this audio, just essentially to give you a sample of things that you can do, uh, just to get you started, okay, um, so yeah, you know, i guess, if you just, you know, if you haven’t done any R, and quite a few of you haven’t, uh, then if you just focus on some of these commands and learn them, the next step really will be, you know, getting on work, uh, with R, okay, so that’s that code, it is on canvas, so i try that out, there’s also a pdf sheet there which has essentially the questions, the worksheet for you to try out, this code probably does most of that worksheet, okay, but if there’s anything else, then you can try that out.

中文翻译：

我把这个命令放在代码里，不是说我们会经常用它，而是想随便给你看一些示例，告诉你还能做些什么，以便让你快速上手，好吗？所以对你来说，如果你几乎没怎么用过 R，而你们很多人可能确实没有，那你只要先掌握这些最基础的命令就好，之后就可以开始慢慢使用 R 做一些实际工作了。好的，这些代码都放在 Canvas 上，你可以去试试。那里还有一份 pdf，里面有一些题目，是给你练习用的。我的这些代码基本都能覆盖那份练习的大部分需求，好吗？如果还有别的，你也可以试着自己写写。

86.

English:

okay, any questions so far, okay, so what i’m going to do is i’m going to now maybe do, so it’s exactly two o’clock, that’s great, so my plan is to teach up to 2:30, so we’re not going to do the whole two hours, okay, so i’m going to assume that you will need a break to probably go half an hour earlier, right, okay, good, so let me get going with the probability distribution.

中文翻译：

好的，到目前为止有没有问题？嗯，那我现在打算，也许我们就……现在正好两点，很好，我计划只上到 2:30，所以我们不会占满整个两小时，好吗？我假设你们需要早一点休息，提早半小时结束，是吧？好，那我现在就开始讲概率分布。

87.

English:

so the probability, so, ah, why do we need to learn about probability, what do you know about, sure, you all come across the club, yeah, yeah, it’s about chance, and what does it do with chance, so you quantify the chance of something, great, okay, that’s basically what probability does, okay, i think we’re pretty much done, because thanks for probability, okay, probability is a way to quantify uncertainty, okay.

中文翻译：

所以先说概率，好吗？为什么我们需要学概率？你对它有什么印象？当然，你们都见过这个概念，对吧？对对，它跟机会、可能性有关，那它干什么呢？它就是去量化某件事发生的概率，很好，好，这基本就是概率要做的事，好吗？我想我们已经基本说完了（笑），因为概率的本质就是用来量化不确定性。

88.

English:

and that’s, and it is basically the only way of properly, mathematically, quantifying uncertainty, okay, there are some other approaches to quantify uncertainty, but they’re not really mathematical, so it’s pretty much the uniform and major way of quantifying uncertainty, and because statistics is all about quantifying uncertainty, making decision under uncertainty, all of statistics, i mean, this is the heart of the engine model, the engine of statistics is probability, so, um, you need to know what is the basics, so as i said, this is a bridging course, so, you know, if you haven’t done any stats, or you’ve done some stats but taken a long time back, i mean, typically, the half-life for information storage is greatness after the exam, what does that mean?

中文翻译：

而且，从数学角度来说，概率是唯一能够严谨地量化不确定性的方式，好吗？当然也有别的方式来量化不确定性，但它们未必是数学化的。所以概率基本算是最主要、最通用的方法。因为统计就是在处理不确定性，在不确定性下做决策，所以我想说的是，整个统计学都是在用概率这个引擎。嗯，所以你需要知道它的基础知识。就像我说的，这门课是个补充课程，所以如果你之前没怎么学过统计，或者学过但很久了，一般来说，学过的东西都会衰减，对吧？考试后信息的半衰期通常都很短。那这是什么意思？

89.

English:

so half-life means how long it takes, ah, for 50% decay, okay, so what i say is half-life for information storage is 20 minutes after the exam, so once your exam is done, 20 minutes later, you already forgot about, right, um, so if you made some stats courses maybe two years back, two years back, or whatever, i don’t know, okay, you can skip it, so that is why, you know, we are doing this, because this is just the basics of everything we talk about.

中文翻译：

半衰期是指，嗯，50% 的东西消失所需的时间，好吗？所以我经常开玩笑说，考试结束后 20 分钟，所学的信息就已经衰减了一半。也就是说，考完试后 20 分钟，你可能都忘了自己学了什么，对吧？嗯，所以如果你两年前上过统计课之类的，现在都忘光了也不奇怪。这就是我们要补这些基础的原因，因为这是一切讨论的根基。

90.

English:

so having said that, because all of you have done some stats, this is more like a refresher, so i don’t want to really be teaching everything from the scratch, like your first student, you’re not, so i’m really going to point you to different things more as the refresher memory, okay, if you do have any questions, please, please stop me and ask me, okay, because i’ll be very happy to explain, okay, so, um, let’s get going, we’re going to talk about all these topics, so we’re going to start by talking about some basic mathematical things about probability theory, i, random variables, and what are the basics, and also probability and so on, and then we also have a good overview of probability distributions, uh, because, uh, that’s how jatch you can use, okay.

中文翻译：

不过，说到这儿，因为你们都或多或少学过点统计，所以更像是复习巩固，我也不想从最基础的地方一步步重新教，你们已经不是初学者了。所以我会用一种快速回顾的方式，如果你们有什么问题，请你们一定要随时打断我提问，好吗？我很乐意解释。好的，那我们开始吧，我们会谈到这些主题：首先一些概率论的基本数学概念，比如随机变量、一些基础规律，然后再谈到概率以及后面的内容。接着我们还会快速回顾概率分布，因为那是我们在很多地方要用到的。

91.

English:

so let’s start by, because we’ve more learned about probability distributions, so before we go to probability, let’s start about in ourselves, what are probability distributions, we already talked about it, like, you know, i was asking you, okay, what is the distribution of students, what can you guess about the distribution of students, so you already know what i’m talking about, but it is important to ask what is exactly, what is, and why do we need it, so probability distributions are really mathematical functions that help you assign probabilities to all possible values that are random variables, okay, so you have a random variable that will take lots of different values, and so it is a mathematical function which assigns, which is a value, okay, but if the variable behaves in such a way, then there is so much probability that it will get this value, so much probability, data, and so on, okay.

中文翻译：

那么我们就先从概率分布说起吧，因为我们后面还会学更多内容。我们先不直接讲概率，而是先搞清楚概率分布是什么。其实我们之前已经提到过了，比如我问你学生人数的分布是怎样，你们就已经有点概念。但还是要正式问一下：到底什么是概率分布？为什么需要它？概率分布其实就是一种数学函数，帮助你为随机变量所有可能的取值分配一个概率，对吗？所以你有一个随机变量，它可以取很多不同的值，那么概率分布就是一个数学函数，会给出“如果变量是这样，那么它出现的概率是多少，如果是另一个值，概率是多少”，等等，好吗？

92.

English:

that’s kind of a very vague way of explaining what a variety distribution is, it’s not mathematically accurate, but it serves the purpose, okay, now, there are two types of distributions, discrete and continuous, why do we need them, well, because we have to understand uncertainty associated with random variables, and i have talked about this uncertainty thing a lot, okay, statistics is about quantifying uncertainty, what are random variables, okay, have you wondered how random variables are different from just variables, i mean, we all learn variables in math, say, a, y, so on, so forth, what makes it random variable.

中文翻译：

这当然是一个比较概括的说法，并不是非常严谨的数学定义，但大体是这个意思。现在我们知道，分布有两大类：离散的和连续的。为什么我们需要区分它们？因为要理解随机变量所伴随的不确定性，而这一直是我们强调的：统计就是量化不确定性。那么随机变量到底是什么呢？你有没有想过随机变量和普通的变量有什么区别？我们在数学课里都会见到变量，比方说 x，y 之类的，可是随机变量是怎么回事？

93.

English:

so let’s have a look, okay, so what is a variable, so variable can take different values, that’s the name, right, variable can take different values, so there are many examples, but the main point here is that the value that a variable will take is deterministic, so, meaning, say, variables, especially in the statistical context, a variable is just something that will, you know, take different values, but in a deterministic fashion, okay, so it is going by an order or by a fixed formula, so for example, time is a variable, okay, but it is deterministic, you know the rate at which we measure time, okay, so no surprises, thankfully.

中文翻译：

我们来看看。什么是普通的变量？变量可以取不同的值，对吧，这是它的定义。但是这里的关键是，普通变量的取值是确定的，也就是说，它是基于某种确定性的规律或固定公式来变化的。举个例子，时间就是一个变量，对吧？但它是确定的，因为我们知道计时的速率，对吧？这没什么可惊讶的。

94.

English:

now, what is, yeah, this is all maybe, sorry, think that BMI, okay, so BMI has a fixed formula, weight over height, uh, once you know the weight, once you know the height, there is no uncertainty, uh, conversion of currency exchange, that’s, again, the deterministic, there’s a fixed rate at what you can point off, and then you can use that to convert one currency from another, that rate, okay, so that, those are the examples of variables, but what we are specifically concerned about in statistics are random variables, and random variables are variables for which you cannot say exactly what value they will take, okay, you can only talk with those probabilities, okay.

中文翻译：

再举个例子，比如说 BMI，它有个确定的公式——体重除以身高的平方。只要你知道体重和身高，就没有什么不确定性。还有货币转换，也是确定的，因为汇率固定在那里，你可以用它把一种货币换成另一种。对吧？所以这些都是普通变量的例子。然而在统计里我们关心的是随机变量。随机变量是一种你无法确切说它最终取什么值的变量，只能用概率来描述它取某个值的可能性。

95.

English:

you cannot say for sure what value you will take, and that’s the definition for a random variable, okay, so the value that it takes is governed by a random mechanism, there can be different types of random mechanisms, so in some cases, like if we are, for example, collecting a sample or a survey, then there is a random mechanism, which is the chance of something being selected in x site, if there is a random chance, so for example, if i randomly select a sample of ten of you, um, then i don’t know exactly what i’m going to get, i can’t predict it with certainty, because who gets selected depends on the chance, okay.

中文翻译：

你没法提前精确地知道它会是多少，这就是随机变量的定义，好吗？它的取值受一种随机机制支配。而这种随机机制可以有不同形式，比如，如果我们在做抽样或调查，那么就有一个随机机制来决定谁会被选进去。如果我要从你们当中随机抽十个人，那么我并不知道结果里会是谁，我无法确定地预测，因为谁被选到是一个概率事件，对吗？

96.

English:

and so this is true for all the surveys that you see, so for example, before elections or anything is, they ask people who are you going to vote, and then you get all these outputs, right, okay, so this is what the survey shows, there is uncertainty around those systems, okay, and in fact, in most cases, they will tell you what is the margin of error, okay, so like some way of quantifying that error so in those cases, the uncertainty is because we don’t know exactly who is included in the sample and therefore what value they will provide, and if you were to do this again, okay, you will have maybe some other people in that sample, they will provide different values, so you get a different answer.

中文翻译：

这在所有的调查中都是适用的，比如选举前或其他时候做民意调查，问人们投谁，然后你得到了各种结果。对吧？所以这就是调查的结果，但它周围有不确定性。实际上，大多数情况下，他们会告诉你一个 “误差范围”，好比说某种方式来量化这个误差。这些不确定性都源于我们不清楚具体是谁进了样本，以及他们会给出什么答案。如果你再做一遍抽样，也许就会抽到别的人，他们给出不同的回答，你最后得到的结果也就不一样。

97.

English:

okay, so that’s a random mechanism. now, there are other random mechanisms, in fact, everything, almost everything, you come across in your life is governed by a random mechanism, so a simple thing as in, okay, when is the bus going to arrive, you might have an expected time of arrival, but when exactly it’s going to arrive is always subject to many different factors, okay, or what is an outcome of a match, a sports match, or an exam, for example, right, you can’t really predict these things, you can have some probability distribution that will help you understand what you can expect, but you never know for sure.

中文翻译：

这就是一个随机机制。实际上，生活中几乎所有事都由某种随机机制在起作用。比如说，公共汽车什么时候到站？我们可能有个预估时间，但它到底什么时候来，还会受到很多因素影响。又或者一场球赛的结果如何，一次考试结果如何，也都无法准确预测。你可以基于概率分布去做一些预判，但永远无法百分百确定。

98.

English:

okay, you can understand how does a person behave soon, you know, there’s just almost everything, okay, so, um, that is why you need to see, because almost everything you come across is governed by a random mechanism, okay, so the only way to understand a random variable is to quantify the probabilities, that’s all we can really do, okay, so we can then make informed decisions or informed estimates based on understanding that probability distribution, okay, so that’s the best you can do.

中文翻译：

你也可以想想一个人会怎么行为，这也有不确定性，几乎所有事情都是如此。所以，这就是为什么我们要研究概率分布，因为你遇到的几乎所有事情都会有随机性。我们唯一能做的，就是给这些随机变量赋予概率，通过量化这种概率来理解它。这也是我们能做的全部了，然后基于对这个概率分布的理解，我们可以做更有依据的决策或者更好的估计。这就是我们所能达到的最好方式。

99.

English:

and so probability distributions are really these mathematical functions which allow you to assign these probabilities. now, you’re not going to go into the maths, but you’re going to be, uh, you’re going to certainly look at how do we read the probability distribution, like a lot, okay, how do we interpret it, okay, as i say, uh, different types of random variables, discrete and continuous, discrete is where, uh, the variable can only take, you know, limited number of different values, so toss of a coin, throw of a dice, these are discrete random variables.

中文翻译：

因此，概率分布本质上就是这些数学函数，帮助你给各种可能值分配概率。我们不会深入数学推导，但会大量去看怎样理解概率分布，好吗？怎样解读它。就像我说的，随机变量也分离散型和连续型。离散型就是变量只能取有限个值，比如掷硬币、掷骰子，这些就是离散随机变量。

100.

English:

which part year would devote discrete random variables, but many of the things are continuous, we have the list on the slide. now, in practice, one little word to note, for example, let’s look at, say, a height of a person, or a weight of a person, now, in reality, you’re measuring them in some units, so height, for example, in centimeters, right, or maybe millimeters, so it’s kind of discrete, because we can’t actually measure, you know, to arbitrary decimal places, but because there are quite a few, many possible values, these variables are still considered as continuous.

中文翻译：

我们这里提到过一些离散随机变量，但很多现象其实是连续的。幻灯片上有列举一些。需要注意一点：比如说身高或体重，实际上，你会用某个单位来测量，比如说以厘米或者毫米来量。从这个角度，它也算是离散的，因为无法测量到无限小的小数位。但由于可能的取值还是非常多，我们在统计里依然会把它当作连续变量来看。

101.

English:

so although it’s going to be like, okay, so many centimeters, or 120.5, whatever, it’s still, you know, considered continuous, so in practice, if a variable contains anything more than a handful of different values, we probably consider as continuous, it’s just more than wait, okay, now, some basic notation, i’m sure you have come across this kind of notation in my stats before, but basically, we use capital letters X, Y, Z to denote random variables, and then you have subscripts, so of course, Xi will be, you know, the income of the ith individual, for example, or Xt will be, you know, the temperature at time t, and so on, okay.

中文翻译：

尽管看上去是某个确定的数值，比如多少厘米，或者 120.5 厘米，但我们仍把它视为连续的。实际上，如果一个变量能取的不同数值多到一定程度，我们就把它视作连续了，比方说体重也一样。现在再看一些基本的记号，你们在统计学里肯定见过：我们通常用大写的 X、Y、Z 来表示随机变量，然后会有下标，比如 X\_i 代表第 i 个人的收入，或者 X\_t 代表在第 t 个时间点的温度，等等。

102.

English:

now, there is a distinction we make, so if you go into statistical textbooks, for example, this is important, sometimes you make a distinction between small letters and capital letters, the capital letters are the random variables, and the small letters, you know, the actual value they take, okay, so small letters are observed values of random variables, which are fixed, so random variables are random variables until you observe them, once you observe them, then it’s a data point, that’s what the data point is, a data point is an observation of a random variable, okay, so then that is fixed, so the small letters indicate that fixed, okay, so just some small minor, but important text, okay.

中文翻译：

同时我们还会区分大写和小写字母，这在一些统计教材里非常常见。大写字母表示随机变量，而小写字母表示它所取到的具体值，也就是观测值。一旦我们真正去测量了这个变量，就得到了一个观测值，这个观测值是固定的，因为它已经发生了，而在观测之前，还是个随机变量。所以小写字母通常用于表示那个已观测到的、固定的值。嗯，这是一点小常识，但也很重要。

103.

English:

so that’s what this slide is about, what another slide, which is a notation for projected distributions, so typically, we use f of x for continuous distributions, sometimes g of x for discrete mass function, okay, so there is a difference, the district distributions are called probability mass functions, and continuous are probability densities, there is a nice mathematical difference, but we will look at the interpretation side of it, um, and then, of course, you have the specific distributions, so normal mu sigma, basically the normal distribution with mean mu, uh, standard deviation sigma, um, and then we use the tilde symbol to denote that a random variable follows a certain distribution, so if x follows a certain distribution f x, we say x tilde f x, um, so z follows a normal with mean zero, standard deviation one, okay, and then for multiple variables, uh, we would say Xi follows x for i, for i from one to n, just kind of open the notation, okay.

中文翻译：

这一张幻灯片就是关于这些符号的介绍，还有一张可能讲的是分布的记号。通常我们用 f(x) 来表示连续型分布，有时也用 g(x) 表示离散的概率质量函数（PMF）。是的，离散分布一般叫做概率质量函数，连续分布叫做概率密度函数，两者在数学上是有差异的，但我们主要看它们的意义解释就行了。然后关于具体分布，比如正态分布记作 N(μ, σ)，表示均值 μ、标准差 σ 的正态分布。我们还常用 “~” 这个符号来表示“服从”，如果一个随机变量 X 服从分布 f(x)，我们就写 X ~ f(x)。举个例子，如果 Z 服从均值 0、标准差 1 的正态分布，就写 Z ~ N(0,1)。对于多个变量，比如说 X\_i 服从某分布 X，就写 X\_i ~ X，i 从 1 到 n 之类的，用来表示有 n 个同分布的变量，好吗？

104.

English:

um, everybody happy with that, yeah, so this is just really laying out things that you may come across, okay, what are the basic rules of probability, either you all know, but it’s still important that we put it up, so probability can only be between zero and one, okay, so the total probability is always one, and the minimum is zero, so when you have a discrete distribution, you will see that all the bars, like a histogram, all the probability is zero to one, but continuous distribution, we measure the probability by area under the curve, okay, i’m going to talk about that here.

中文翻译：

嗯，你们对这些有没有问题？好，这主要就是给你们展示一些常见的符号和写法。接下来讲一些最基本的概率规则。你们大概都知道，但还是要强调一下。概率只能在 0 和 1 之间，对吗？总概率和是 1，最小是 0。比如在离散分布里，如果你画柱状图，所有柱子的高度总和就是 1。而在连续分布里，我们用曲线下面积来表示概率，好吗？我会在这里说明一下。

105.

English:

now, this board is important, so for a continuous variable, because it can take infinitely many values, so that’s the definition, continuous variable can take infinitely many values, the probability that it will take any specific value is always zero, okay, so for continuous distributions, probability that it will take any specific value is always zero, okay, so let’s learn the basic stuff, this is really stuff that you have learned in maybe high school as well, mostly in high school also, okay.

中文翻译：

接下来这点很重要：对连续型随机变量来说，由于它可以取无限多的值（这是连续型变量的定义），所以它在某一个确切值上的概率为 0。好吧？也就是说，在连续分布里，随机变量恰好等于某个具体数值的概率是 0。我们先掌握这些基本概念，你们在高中可能也学过，主要就是这些东西。

106.

English:

but so we have, we start defining probability by talking about events, so events are like, what are the possibilities, what are the possibilities, heads or tails, these are the possibilities, and what are the possibilities, uh, and then we can talk about the complement intersection, reunion, and so forth, and then basic law of probability, but important thing that i’m, you know, the reason we’re doing this is because i want to talk about conditional probability eventually, okay, and that’s the base here, which is also values for in many other ways, okay.

中文翻译：

我们是怎么开始定义概率的？我们先说 “事件”。事件就是那些可能的结果，比如掷硬币，会是正面还是反面，这就是所有可能结果。然后我们可以讨论事件的补集、交集、并集等等，然后讲一些概率的基本定律。但我们这样铺垫是因为我们最后要讲条件概率，好吗？这是我想要说的重点，而且在很多场合都非常有用。

107.

English:

so before we can define probability, we need to define events. i’m going to go quickly through this, but for example, if you’re tossing a coin, then the coin can be heads or tails, so you have two units, if you have either heads or tails, so these are very subtle, you don’t have using capital letter, so it says an a, b. now, events in, we are said to be dependent if they are unrelated, which means value one doesn’t depend on the value of color, there is a particular way of figuring out if they are dependent or not, which is verification probability, which we won’t, okay, so let’s look at some examples, so throw of a dice, i can define multiple events, okay.

中文翻译：

所以在定义概率之前，要先定义事件。我在这里会快速带过。比如说，你掷硬币，那么结果要么正面，要么反面，所以有两个事件。你可以把它们记成 A、B 之类的。事件之间如果互不相干，就称作独立，意思是一个事件的发生不依赖另一个事件。有确定的方法来判断它们是否独立，比如用概率验证，虽然我们这里就不细讲了。然后举些例子，比如掷骰子，可以定义许多事件，好吗？

108.

English:

lots of different events, but if i’m throwing a dice, then an event could be that the outcome is an odd number, okay, so a could be a define that the outcome is an odd number, in which case that set a has three possible outcomes, one, three, or five, okay, because those are the only values of a dice that are odd numbers, and then you can have b, which is even number, for example, or you can define b to be the outcome is greater than four, so then b is a set of five and six, so typically an event has a set of possible values associated, and then we can define an event a, we can define this complement, we can define the intersection, b, union, b, and so on.

中文翻译：

掷骰子就能定义很多事件，比如说，事件 A 表示掷出的结果是奇数，那么 A 就包含 1, 3, 5 这三个可能值，因为骰子的奇数面就是这三个。你也可以把事件 B 定义成偶数，或者把 B 定义为 “掷出的结果大于 4”，那么 B 就包括 5 和 6。通常，一个事件就是由若干结果组成的集合，然后我们可以讨论事件 A、它的补集 A^c、A 和 B 的交集 A∩B、并集 A∪B，等等。

109.

English:

the another event, okay, there’s also something called another event, which is denoted by this symbol phi, what is another event, can somebody, did you guys see the questions, come across another event, no, what’s another event, another event is just a freak possibility basically, okay, which is probability dizzy, you know, so if you toss a coin, what if it lands on its edge, that’s another event, okay, the probability of that is zero, but it can happen, okay, so when we define, so now we’re talking about, so when we define a random variable, we define the sample space as in the set of all possible values it can take and typically we have all these very defined values, but we also have another event just to include everything else that we haven’t accounted for.

中文翻译：

还有一种叫做“空事件”，用符号 φ 来表示。有人知道这个吗？可能你们在题目里也见过。空事件其实就是那种非常小概率的情况，比如你掷硬币时，它万一立在边上，这个概率几乎是零，但理论上可以发生。所以在定义随机变量时，我们会定义一个样本空间，包括它能取的所有可能值，通常我们会列出那些主要结果，但也会有这样一个空事件来包含那些没有覆盖到的极小概率情形。

110.

English:

um, so let’s go to the next slide, so for example, um, so this is an example of the events a and b that were defined, and that complements, so a complement, b complement, intersection, b, a union, b, you guys can read this, um, okay, this is what i was talking about, so the sample space omega, so for every random vehicle, we have a sample space omega, which essentially is all possible things that can, all possible outcomes that are possible, okay, for that random vehicle, so if i’m tossing a coin, my sample space can be heads, tails, or the null event, okay, there is nothing else, okay, if i’m throwing a dice, okay, one, two, three, four, five, six, and, okay, the probability of a null event is zero, but it is typically included for consideration, okay.

中文翻译：

嗯，让我们到下一张幻灯片。这里就举了关于事件 A、B 的例子，以及它们的补集 A^c、B^c，还有交集 A∩B、并集 A∪B，你们可以自己看。嗯，这就是我所说的，样本空间通常记作 Ω（欧米伽）。对于每个随机试验，都有一个样本空间 Ω，包含了所有可能结果。如果我掷硬币，样本空间就是正面、反面和那个极端的小概率事件，也没别的了。如果我掷骰子，样本空间就是 1, 2, 3, 4, 5, 6，再加上那种空事件。空事件的概率是 0，但一般也把它放进样本空间里。

111.

English:

so the question is, what would be the sample space for something not straightforward, so let’s say if i ask you to define a sample space for height of a person, how would you go about it, any ideas, so the reason i ask that question is because height of a person is a continuous video, the examples i’ve discussed so far are simple discreet variables, okay, so you’re defining event spaces using sits in a continuous variable, you can’t define in revaluing a sit, so you will define an interval, so if so firstly, you need to decide what is the unit, right, is eight centimeters of meters of feet or whatever, and then you need to put a range around it, okay, so then of course, you can find information and put a range zero to two hundred centimeters or whatever, right, but it has to be an interval, okay.

中文翻译：

那么问题来了，如果是一个复杂点的情况，比如说定义一个人的身高的样本空间，怎么办？有什么思路吗？我之所以这么问，是因为身高是一个连续型变量，我们前面举的例子都是离散的。对于离散的情况，你可以很容易地把那些值列成一个集合，但对连续型变量，你没法一个一个地列出来，所以你要用区间来定义。首先，你得决定用什么单位，对吧？是厘米还是米还是英尺之类的？然后你给它一个范围。比如你知道身高范围大约在 0 到 200 厘米之间之类的，对吗？所以它得是个区间。

112.

English:

uh, is that a closed interval or an open interval, um, but that kind of stuff, so they say more, okay, um, so that’s, uh, the in location windows that’s, uh, those are the basic rules of probability, okay, so the probability of any event a is always between zero and one, probability of the sample space is always one, and probability of a null event is zero, now it’s important to understand what does probability one and what does probability zero mean, okay, a common misconception is that probability one means that event will always happen, and probability zero means event will never happen, that is not true.

中文翻译：

嗯，那它是一个闭区间还是开区间呢？就会有各种细节问题，对吧？这个就不多讲了。嗯，好，那么我们继续说概率的一些基本规则：任何事件 A 的概率都在 0 和 1 之间，整个样本空间的概率是 1，空事件的概率是 0。现在我们得搞清楚，概率为 1 和概率为 0 分别意味着什么。很多人会误以为概率是 1 就一定会发生，概率是 0 就一定不会发生，但这不对。

113.

English:

okay, a probability zero events happen, so we talked about the example of tossing the point, okay, it can land on this age, so it can happen, in fact, i would say that probability zero events happen all the time, can somebody give me an example, so let’s say you’re throwing darts, okay, now you have a dart board, so that dart board is your sample space, okay, if i was the person throwing the Dart, then the whole board will be my sample space, because my Dart can hold you there, okay, but assuming that the Dart board is your sample space, the board probability is one within the Dart board.

中文翻译：

实际上，概率为 0 的事件是会发生的。我们之前提到过，掷硬币时硬币立在边上，虽然概率几乎为 0，但可能还是会发生。实际上，可以说概率为 0 的事件一直都在发生。再举个例子，你玩飞镖，对吧？你有一个飞镖盘，那么飞镖盘就是你的样本空间。如果是我投飞镖，我基本能把整个飞镖盘都算进来，因为可能会投到任何地方，但假设这个飞镖盘就代表所有可能落点，那么在飞镖盘上的概率就是 1。

114.

English:

now when you throw the Dart, it is going to land at a particular point, but there are infinitely many points on the dart board, so probability of any particular point is zero, in fact, we said that for continuous random variables, they interactive any possible values, so probability that if we take any single value is always you so just by that logic, you can say that probability zero events happen all the time, because there’s so many continuous random variables, and they take some value, right, so yeah, so that’s just doing away with the misconception that for these zero events can’t happen, they can happen, they happen all the time, and by the same logic, probability one event may not happen, okay, think about it.

中文翻译：

当你把飞镖掷向飞镖盘，它会落在某一个确切的点上，但飞镖盘上有无穷多个点，所以任意一个点的概率其实都是 0。我们之前说过，对于连续型随机变量，随机变量可能取无穷多的值，所以单个确切值的概率是 0。但它依然会取到某一个具体值，对吗？所以从这个角度来讲，概率为 0 的事情一直都在发生。相对地，概率为 1 的事情也并不保证一定发生。想一想就明白了。

115.

English:

okay, so that’s how you, you know, that’s the way to think about probabilities, okay, so hopefully you have a good refresher of it. now how do we calculate probability, so there are two ways available, basically, but one way is the relative frequency, which is usually what is taught in schools or, you know, basic first-year courses, so you’re tossing a coin or you’re throwing a dice, okay, what are the total number of possibilities, and out of which how many are event a, that is the probability, so for example, let’s say you are throwing a dice, all six outcomes are equally possible, and let’s say event a is that, uh, the outcome is a long number, so there are three different ways in which it can be a long number, so it can be one, three, or five, so three out of six, and therefore the probability is three out of six, which is, okay, so you all know this, right, so this is the most simple relative frequency approach.

中文翻译：

好，这就是思考概率的方式，希望你们能借此回顾一下。接下来讲讲怎么算概率。基本上有两种方法，一种是相对频率法，通常在学校或者大一入门课里会先教这个。比如你掷硬币或掷骰子，所有可能结果有多少？事件 A 包含多少种可能？概率就是 A 的可能数除以总可能数。举个例子，你掷骰子有 6 个面，假设所有面出现的概率都相同，你定义事件 A 是掷出奇数，那奇数面是 1, 3, 5，共有三个，所以概率就是 3/6，也就是 1/2，对吧？这就是最简单的相对频率法。

116.

English:

this is it fully for simple problems, but in reality, where you have continuous variables or more complex relationships, you can’t do relative frequency approach, okay, so then we do modeling, so where we kind of either estimate the parameters of appropriate distribution, or we do something calibration, or time series, or other models, uh, which all have probability distributions in the background, so when you fit this model, and you predict, you can predict, uh, the probability of selling some equipment, okay, ah, so that’s the most common way of actually, you know, so modeling is the most common way of predict values, or in other words, probabilities.

中文翻译：

这种方法只适用于简单的问题，但现实中如果变量是连续的或关系更复杂，你就没法这么算了。所以我们会用建模的方法，比如说我们去估计某个分布的参数，或者做一些校准、时间序列模型等等，这些背后都用到概率分布。你在拟合了模型之后再做预测，你就能得到某件事情发生的概率，比如说卖出某种产品的概率之类的。嗯，这就是我们实际中最常用的方式，也就是通过建模来预测结果或预测概率。

117.

English:

there is another way, which is subjective probability, okay, subjective probability is something that you do all the time, okay, so if i ask any person, for example, what is the probability that you pass the course, okay, how will we answer that question, that’s a subjective probability, okay, because it’s just meaning what you believe you will have, okay, what is the probability that a terrorist attack will happen, it’s a subjective probability, okay, because there is no data that can find the probability of a unique single growth terrorist attack, so that’s a real life application, okay, um, and there are lots of such applications.

中文翻译：

还有另一种方法，叫做主观概率，也就是 subjective probability。它是我们经常在用的概念。举个例子，如果我问你，你通过这门课的概率是多少？你怎么回答？其实这是一个主观概率，因为它只是你对自己会不会过的信念。同样问“发生恐怖袭击的概率是多少”，这也是主观概率，因为没有数据能精确给出某一次恐袭的概率。现实里有很多这样的应用场景。

118.

English:

so in fact, in many cases, experts have to quantify their probabilities subjectively, okay, personally, that’s my idea of expertise, but we’re not going to talk about that, okay, but just so you know that, you know, we do have a lot of subjective probabilities that we do around, so probability is not just throwing coins and licenses, and you know, whatever, it’s a lot more, it’s basically everywhere, okay, that’s what i want you to really take from this course.

中文翻译：

所以在许多情况下，专家们必须主观地去量化他们对某些事情的概率。对我来说，这正是专家判断的意义所在。但我们这里就不深入这个话题了，只是让你知道，我们平时处理的概率其实有很大一部分是主观概率，并不只是抛硬币、掷骰子这些客观随机事件而已。概率的应用远远不止这些，它几乎存在于生活的方方面面。这就是我希望你们从本课中带走的理念。

119.

English:

okay, again, going back to the school maths, those are some basic formulas, if you’ve forgotten, because it’s been five years, okay, refresh your memory, and then some basic examples, so what, i’m not spending time on this, but of course, i’m expecting you to read this in your own time, and refresh your memory, because these are the basic things you should know, okay, and then we come to the conditional probability, which is what i want to talk a little bit about.

中文翻译：

好的，回到中学数学里学过的一些基础公式，如果你已经五年没碰了，可以复习一下。然后也有一些基础例子，我这里就不花时间细讲了，你们可以自己找时间看看，回顾一下这些基础知识。接下来我们要说的就是条件概率，这是我想稍微展开讲一下的。

120.

English:

so conditional probability is really important, okay, and what is conditional probability, it means that what is the probability of observing something, given that something else has already happened, so given that this is a cloudy day, what is the probability that it will rain, okay, so we already know it’s a cloudy day, but if we didn’t know it was a cloudy day, i can just ask you an unconditional, what is the probability tomorrow is going to be rain, okay, that’s an unconditional, but what is the probability that tomorrow is going to rain, given that you have so much, you know, so much humidity, that’s a conditional probability, okay, uh, or it’s a cloudy day, that’s a conditional probability, so that’s what this symbol, you know, so a given a.

中文翻译：

条件概率非常重要。它指的是：在知道另一件事已经发生的前提下，我们观察到某个事件的概率是多少。比如说，已知今天是阴天，明天会下雨的概率是多少？如果你不知道今天是阴天，你只能问一个无条件概率：明天会不会下雨？那就是另一回事。但如果给定了今天的湿度很高或给定了阴天的条件，那么明天会下雨的概率就不同了，这就是条件概率。它通常写成 P(A|B)，表示在 B 发生的条件下 A 发生的概率。

121.

English:

so what is the probability that the throw will be throw of a dice, will it be one, given that it is an odd number, it changes, okay, provided your ord number is one, three, or five, there can be three different ord numbers, what is the probability that it will be one, given that it is an odd number, is one third, because one out of those three, okay, so that’s what we mean here, a given b, and that is given by probability of a intersection b divided by probability of b, this is the main formula we’re going to study, okay.

中文翻译：

再举个简单例子，你掷骰子，问“给定结果是奇数，这个结果是 1 的概率是多少？”很显然，如果结果是奇数，那只可能是 1、3、5 三种，所以在这三种情况里有一种是 1，所以条件概率就是 1/3。这就是我们说的 P(A|B) = P(A∩B)/P(B) 这个主公式，好吗？这就是我们主要要学的东西。

122.

English:

so this is what we’re going to focus on on Friday, okay, so on Friday, i will actually do some real probability examples in class, uh, i would also use up our, um, programming to find some answers, do some probability, however, before you show up on Friday, i want you to go to the rest of these slides, okay, so forget about the Bayes theorem, i’m going to talk about this Bayes theorem on Friday, but then the later half, so this last, like, 70, 19 slides from slide 29 onward, so let me zoom out.

中文翻译：

我会在周五重点讲这个，好吗？周五我会在课堂上举一些实际的概率例子，也会用 R 程序来算一些概率。但是，在你周五来上课之前，我希望你先把剩下的这些幻灯片都看完，好吗？Bayes 定理这些先别管，我会在周五再讲，然后还有后面那部分，比如从第 29 页往后的那些，我给你们翻一下……

123.

English:

so there are 57 slides here, okay, out of which slide 29 is where we start talking about probability distributions, so i want you to just read everything from slide to, and design homeworks, because this is just how to interpret probability distribution, and what are the different types of distributions, and so this is just general information that you can, okay, um, so hammering, and of course, we will discuss this, uh, on Friday quickly, but i will focus on, uh, solving some probability examples, which will then be needed for the workshop on Tuesday, okay, so that’s how it goes, okay.

中文翻译：

这里一共有 57 页幻灯片，其中从第 29 页开始谈概率分布。所以我要你们在周五之前把前面的内容都看一下，包括到这儿的内容，还有关于如何解读概率分布以及不同分布类型的那些。这些都是通用的信息，供你们参考。然后我们在周五会快速讨论一下，但重点会放在解一些概率例题上，然后那部分知识也是下周二研讨课所需要的。大体上就是这样，好吗？

124.

English:

so are there any questions before we stop, everything good, okay, great, i’ll see you Friday.

中文翻译：

在结束之前还有什么问题吗？都还好吗？好的，很好，那我们周五见。