**2025-3-7 STATS707**

**英文原文 & 中文翻译**

1. 英文原文（段落1）：

what is common between the a and B? that is the intersection, right? so what is it? so let’s 5, that’s right. let’s actually write this. so a is 1, 3, 5, and B is 5 and 6. so what is comment is really just five, okay? so, and once you write that down, it supports you straightforward. so the probability is one of six, okay?

中文翻译（段落1）：

A 和 B 之间有什么共同点？那就是交集，对吧？所以它是什么呢？好的，让我们来看 5，没错。我们实际上写一下：A 是 1、3、5，B 是 5 和 6。那么它们的共同点其实就只是 5，对吗？一旦你把它写下来，就很直观了。所以概率是六分之一，对吧？

2. 英文原文（段落2）：

now what is probability of a union B? okay, now what is a union B? so the formula for probability of a union B is on the slide, but it is basically a probability of a plus probability of B minus probability of a intersection B. so really again, the way to look at it is if you have a Wen diagram, I’m sure all of you have seen this before: if this is a and this is B, then this is the intersection, a intersection B, and so a union B is basically this whole thing, but we don’t want to be logged in this twice, so we add a and B and then subtract what is on. that’s the main thing.

中文翻译（段落2）：

现在我们来求 A 并 B 的概率？好的，那什么是 A 并 B 呢？幻灯片上有 A 并 B 的概率公式，它基本上就是 P(A) + P(B) - P(A 交 B)。同样地，如果用韦恩图来看的话，我相信你们都见过这个：如果这是 A 而这是 B，那么这里是它们的交集 A 交 B，于是 A 并 B 就是包含这两部分的整个区域，但我们不想把交集的那一部分计算两次，所以我们把 A 和 B 的概率相加，然后再减去它们的交集，这就是主要内容。

3. 英文原文（段落3）：

okay, so yeah, we have 1/2 plus 1/3, sorry, minus 1/6. so what would that be? so that’s five over six minus 1 over six. tell me if I made a primary school error here, but… okay, so that’s what 3 separate… yeah. now the main thing is — and I often illustrate it myself — okay, so the main thing you have to remember with probability is you can make silly mistakes, and normally when I’m doing this in class I always do at least one mistake, okay, so I’ll illustrate it naturally, but keep an eye out, okay? because for me, nobody’s going to cut my marks, okay, but yeah, so anyway, remember that. so that’s simple as that.

中文翻译（段落3）：

好的，是的，我们有 1/2 加上 1/3，对不起，然后减去 1/6。那会是多少？那就是 5/6 减去 1/6。要是我在这里犯了小学算术错误，请提醒我……好的，那是 3……对。现在主要的问题是——而且我自己经常会去演示这一点——好的，在概率里你必须记住，可能会出现一些愚蠢的小错误。我通常在课堂上演算时都会至少犯一次错误，好吧，我会自然而然地演示出来，但要自己留意哦。因为对我来说，没有人会扣我的分，对吧？但无论如何，记住这一点。就这么简单。

4. 英文原文（段落4）：

now, while we are there, we can also talk about what is a complement and what is B complement, because those are the things we need as well. so a complement is basically the even number, because what is not a? so if it’s not odd, then it has to be an even number, and so this will be 2, 4, and 6. and then B complement, of course, is less than 4, or less than and equal to 4, by the way, okay, so this is 1, 2, 3, 4. so we can also do probability of a complement and probability of B complement. probability of a complement, can anybody tell me what would that be? yeah, so one minus probability of a, that is exactly right, okay, so which is in our case one… yeah, mouth. and informative B complement would be two or three… 1 - 1/3. okay, so there you have it. okay, those are all the basic high school formulas for probability, okay.

中文翻译（段落4）：

现在，在这里我们也可以谈谈 A 的补和 B 的补是什么，因为我们也需要用到它们。A 的补基本上就是偶数，因为 A 不是偶数的话，那么就一定是奇数，如果它不是奇数，那就必须是偶数，所以这里就是 2、4、6。然后 B 的补当然是小于 4，或者说小于等于 4，顺便说一下，也就是 1、2、3、4。所以我们也可以求 A 的补的概率和 B 的补的概率。A 的补的概率，有人能告诉我是多少吗？对，就是 1 - P(A)，完全正确。好的，在我们这个例子中……是的，对。然后对于 B 的补的概率，大概就是 1 - 1/3。好的，就是这样。好的，这些都是高中概率的基础公式。

5. 英文原文（段落5）：

now what is probably not exactly high school necessarily is the conditional probability, okay, and that is essentially where we stopped last time. so let me also see if I can show the slides on the other screen. okay, so now one thing you should note is: would be one of these two projectors get recorded, okay, so the lecture recording is going to just include that, okay, so that projector is the one that gets reported. just remember that, so if you have not attended a lecture and have done something, then whatever is on this screen will not be seen in the recording. okay.

中文翻译（段落5）：

接下来，可能不完全是高中水平的内容是条件概率，对吧，这基本上是我们上次停下来的地方。好的，我再看看能不能把幻灯片投到另一块屏幕上。好的，还有要注意一点：这两个投影机中的某一个是会被录下来的，所以课程录制只会包括那一台投影机的内容。记住这一点，如果你缺席了某次课或者做了什么，那么这个屏幕上显示的内容是不会出现在录播里的。

6. 英文原文（段落6）：

so conditional probability. so what we have — that’s the formula, and this is very easy if you remember a simple trick. so what does this actually mean? this means probability of a given that B has occurred, so you already know that B has occurred, and then what is the probability of a? so B has already happened, what is the probability of a? so that’s the conditional probability. and the formula is really probability of a intersection B, divided by probability of B. now, what is the trick? well, the trick is you divide by the event on which it is conditioned, okay, as simple as that. so if I wanted to do probability of B conditioned on a, B given a, the numerator doesn’t change, but now my denominator is a, because that is what is sufficient, okay. so it’s very easy to remember.

中文翻译（段落6）：

那么条件概率。我们有的——就是那个公式，如果你记住一个简单的小技巧，这很容易。它究竟意味着什么？它表示事件 B 已经发生的情况下，事件 A 的概率是多少。也就是说，你已经知道 B 发生了，然后问 A 发生的概率是多少，这就是条件概率。而公式实际上就是 P(A ∩ B) / P(B)。那么技巧是什么？技巧在于你要用所给定的那个事件做分母，就这么简单。所以，如果我想算在 A 已经发生的情况下 B 发生的概率（记作 B|A），那么分子不变，但现在分母变成了 P(A)，因为这就是条件所在。所以非常好记。

7. 英文原文（段落7）：

so now that we have… so what a given B will be in our case is really what is the probability that a is an odd number given that it is greater than 4? so we already know it is greater than 4, and then I’m asking you: okay, given that it is greater than 4, what is the probability it is an odd number? okay, so we will do that now. so probability of a intersection B was 1/6, okay, just tell me if I make a mistake. probability on B is 1/3, so that should be 1/2, okay, and which again makes sense because there are two numbers which are greater than 4 — 5 and 6 — of which one is odd, so one number two. okay, so that makes perfect sense. okay, and then we can do the other way around. so given that it’s an odd number, what is the probability that it will be greater than 4? okay, again, there are only three odd numbers 1, 3, and 5, and only one of them is greater than 4, which is 5, so the probability is going to be one of the three in this case. very straightforward, but we can still do it using the same logic: 1/6 divided by probability of a, which is 1/2.

中文翻译（段落7）：

现在我们有了……那么在本例中，A|B 就是：给定它大于 4 的情况下，它是奇数的概率是多少？我们已经知道它大于 4，然后我要问你：好，如果它已经大于 4，那么它是奇数的概率有多大？好的，我们现在来做这个。A ∩ B 的概率是 1/6，对吗？若我说错请纠正。B 的概率是 1/3，所以那就应该是 1/2。对，这也符合直觉，因为大于 4 的数只有两个——5 和 6，其中 5 是奇数，所以只有一个是奇数，得到 1/2。好的，这就说得通了。然后我们也可以反过来：如果已经知道它是奇数，那么它大于 4 的概率是多少？再一次地，只有 1、3、5 这三个奇数，而其中只有 5 大于 4，所以概率就是三分之一。在这里也很直观，但是用同样的逻辑，我们也可以用 1/6 去除以 A 的概率 1/2 来计算。

8. 英文原文（段落8）：

now obviously, these are really straightforward. this is really straightforward example, so it’s just very easy, but the ideas don’t change. so the examples you do in the workshop or in the assignment will not be that trivial case. you’ll have to figure out. so the main thing there for all those examples would be actually figuring out what the sample space is, because that’s where you start, and then defining correctly, defining what event a is and what event B is, and then the rest will be straightforward because the formulas are accepted. okay, so that’s that.

中文翻译（段落8）：

当然，这些例子非常直观，这是一个相当简单的示例，所以非常容易。但核心概念是不会变的。因此，你在研讨课或者作业里要做的例子就不会这么简单了，你需要自己去琢磨。对于那里的所有例子，最关键的事情是先找出它的样本空间，这就是第一步，然后正确定义事件 A 和事件 B，再用公式就可以很自然地得出结果了。好的，就先这样。

9. 英文原文（段落9）：

now what we can also do is we can go one step further and talk about a condition where a and B will be independent, okay. now, what do we mean by a and B being independent? so independent, just from an English word, right? so a and B are independent. that means that, you know, what happens to a is not related to what happens to B, or the vice versa. and so how we define this in probability terms is basically saying that the probabilities don’t depend on each other. the probability of a intersection B is simply probability of a times probability of B, and that’s the definition of independence.

中文翻译（段落9）：

接下来，我们还可以再往前走一步，讨论一下在什么条件下 A 和 B 会相互独立。好的，那么 A 和 B 独立是什么意思呢？独立，从英文上来讲，就是说 A 的发生与 B 的发生互不影响，或者反过来也一样。在概率领域，我们把它定义为：如果 A 和 B 独立，那么 A 交 B 的概率就等于 P(A)×P(B)，这就是独立性的定义。

10. 英文原文（段落10）：

now, how do I get that definition? well, basically, if a and B are independent, what do you expect? if a and B are independent, then I would expect that probability of a given B is just probability of a, because a doesn’t depend on B, okay, or vice versa, probability of B given a will be just equal to probability of B. so if either of those are true, or most of them are true, then all you have to do is substitute this in the formula here. so probability of B given a is just probability of B. so I get probability of B equal to probability of a intersection B, divided by probability of a, which then implies that probability of a intersection B is equal to probability of a times probability of B. it’s in chronological, okay. yeah.

中文翻译（段落10）：

那么我们是怎么得到这个定义的呢？其实很简单。如果 A 和 B 独立，那么我们会期待什么？我们就会期待在 B 已经发生的情况下，A 的概率仍然等于 A 本身的概率，因为 A 并不依赖于 B，或者反过来也一样。若满足这些条件，那么把它代回前面的公式即可。例如 P(B|A) = P(B)，就说明 P(B) 等于 P(A ∩ B) / P(A)，于是我们便得到了 P(A ∩ B) = P(A) × P(B)。这其实是顺理成章的，对吗？

11. 英文原文（段落11）：

so that’s how the concept of independence can be defined in terms of probabilities, right? so if you observe — so in a real example, you know, we distance those in the workshop — if you observe that probability of a given B… actually a intersection B actually happens to be probability of a times probability of B, then that implies that a and B are independent, okay. sometimes the independence is obvious based on the context. sometimes it is not obvious, in which case you can use this formula to figure out if it is independent, okay. so they are equivalent. so these two and this are basically just equivalent, because if you use this, you get this. if you use this, you get this.

中文翻译（段落11）：

这就是我们如何用概率来定义独立性的。也就是说，在一个实际的例子里——比如我们在研讨课里会讲到——如果我们发现 A 交 B 的概率刚好等于 P(A) 乘以 P(B)，那么就意味着 A 和 B 独立。有时根据语境可以很明显判断它们是否独立，有时不明显，这时你就可以使用这个公式去检验它们是否独立。它们其实是等价的，这两个条件和那个结论都是等价的，因为你可以用这边推那边，也可以用那边推这边。

12. 英文原文（段落12）：

and then the last thing is talk about the Bayes theorem. so Bayes theorem, by the way, Bayes theorem is named after a guy called Reverend Thomas Bayes, who was in 1760s, so 250 years back, okay, and the theorem was published after he died, because his friends published, so anyway. the whole Bayesian statistics or Bayesian methods, which are used quite heavily not only in statistics but also in machine learning — your Bayesian machine learning — have all started from Bayes’ theorem. so it’s a fundamental theorem, but it’s actually just as simple as conditional probability, okay, and I’m not going to prove it, but it’s pretty straightforward.

中文翻译（段落12）：

接着最后我们再说说贝叶斯定理。贝叶斯定理是以一个叫托马斯·贝叶斯的牧师命名的，他大概是 1760 年代的人，也就是大约 250 年前。这个定理是在他去世后由他的朋友们发表的。不管怎样，现在整个贝叶斯统计或贝叶斯方法，不仅在统计学中用得多，在机器学习里也大量使用——像贝叶斯机器学习——都始于贝叶斯定理。它是一个很基础的定理，但其实也和条件概率一样简单。这里我就不做证明了，不过它非常直接明了。

13. 英文原文（段落13）：

but what it really says is probability of a given B, okay, can also be expressed as probability of B times probability of… sorry, sorry, oh okay. so probability of a given B can be expressed as probability of B given a times probability of a, divided by probability of B. now, what is happening really? all we are doing is in this formula here, okay: so probability of a given B is intersection B divided by probability of B. now, this intersection B is replaced by using this formula. so in this formula, product of intersection B is product of B given a times product of a, so we just do that and then divide that. it’s as simple as that, you know, just… I mean, you don’t need… you know, I’m not going to ask you to prove it, although I could actually. anyway, lipstick in the air, but it’s as straightforward as that.

中文翻译（段落13）：

那么它真正说的是，P(A|B) 可以表示为 P(B|A) × P(A) 再除以 P(B)。那背后到底发生了什么？其实就是把原先公式里的 A ∩ B 替换了一下而已。在原先的公式里，P(A|B) = P(A ∩ B) / P(B)，而 A ∩ B 可以用 P(B|A) × P(A) 来代替，所以代进去之后就得到了这个形式。就这么简单。其实你不需要我来证明，虽然我也可以证明给你看，但这已经足够直观了。

14. 英文原文（段落14）：

but the main point is that this is a, what this allows you to do is it allows you to switch probabilities or switch conditional probabilities, okay. so you can derive the conditional probability of a given B simply by knowing the conditional probability of B given a, and then products of a and a. this is really helpful, and what you will do in the workshop is you will use this for a medical exam, okay. so for example, if you know that the vaccine works, say, 1%, 95% of the times, then what is the probability that a person who has a vaccine will still get analysis, okay, which was very relevant in the COVID, for example, times, okay. and so all those calculations are really based on just Bayes’ theorem, okay. so this is very, you know… so you can flip the condition probability. so you know one condition probability, then you can deduce the other, okay. but actually, it has a lot many more uses. in fact, a whole branch of statistics is based on Bayesian inference.

中文翻译（段落14）：

但关键是，这个定理让你可以“翻转”概率或条件概率。你可以只知道 B|A，再加上 P(A) 和 P(B)，就推算出 A|B。这非常有用，而你们在研讨课里会用它来处理一个医学检查的例子。比如，如果你知道疫苗的有效性是 95%，那么接种疫苗的人依然感染某种病的概率是多少？在新冠疫情中，这类问题特别相关。这些计算都是基于贝叶斯定理的。它很有用，你可以翻转已知的条件概率，从而得出未知的那个。其实它还有更多的用途，整个统计学中有一个分支叫贝叶斯推断，就是基于它的。

15. 英文原文（段落15）：

so we’re going to talk about statistical inference next time, and one whole area of statistical inference is based on Bayesian inference. it has lots of uses, but the reason you guys need to know this as a bare minimum is because this is used in Bayes classifiers and, you know, basic classification machine learning. so remember that, okay. and we will use this a lot in the workshop next week, okay.

中文翻译（段落15）：

我们下次会讲到统计推断，其中有一整个领域都在用贝叶斯推断。这东西用途很多，但你们至少需要知道它，是因为在贝叶斯分类器以及一些基础的机器学习分类里都会用到它。记住这一点。好的，下周我们在研讨课里也会大量使用它。

16. 英文原文（段落16）：

and one final thing is you can also use these conditional probability formulas to find the total probability, okay. so this is your Bayes’ theorem, and then the most important, but also the final important bit, is how you can calculate total probability. and what do I mean by that is, if you only know the conditional probabilities — so if you know probability of B given a, and you know the probability of B given a complement, then by knowing the probability of a and probabilities of a complement, you can denote the probability of B. this is also a formula we’re going to use a lot in the workshop. if you carefully look at the Bayes’ theorem, and if you carefully look at the total probability formula that I’ve written, they are all simply derived from the conditional probability, okay. you’re just moving things around and then you get it, okay. so that is why these things are important.

中文翻译（段落16）：

最后还有一点：你也可以用这些条件概率公式去计算全概率（total probability）。这儿就是贝叶斯定理，然后还有一个最重要、也是最后一个关键点，就是怎么去算全概率。什么意思呢？如果你只知道一些条件概率——比如你知道 P(B|A)，也知道 P(B|A^c)——那么再加上 P(A) 和 P(A^c)，你就能求出 P(B)。在我们研讨课里也会大量用到这个公式。如果你仔细看贝叶斯定理和我写的全概率公式，你会发现它们全是基于条件概率而推导出来的，只是把东西移来移去而已。所以这就是为什么这些知识点很重要。

17. 英文原文（段落17）：

so the conditional probability is really important because it leads to the Bayes’ theorem and total probability, which are very helpful in actually finding some useful probabilities in practice. so any questions so far? no? great, okay. so what we will do next is… so in fact, on this slide I’ve got an example which you can check. so the throw of a dice, I think we already did this bit, that’s fine, okay. so let’s quickly browse through the remaining slides that I already asked to read. I’m going to just highlight important bits, and then we will do some actual examples using R, okay, and we’ll, you know, look deeper into how we use those probability distributions, okay.

中文翻译（段落17）：

所以，条件概率非常重要，因为它引出了贝叶斯定理和全概率公式，这在实际中求一些有用的概率时非常有帮助。就到这里有没有什么问题？没有？好，那我们接下来要做的是……其实在这个幻灯片上，我有一个可以让你们检查的例子，比如掷骰子的例子。好像我们已经讲过了，那就没关系了。让我们快速过一遍我之前让大家读的剩余幻灯片。我只会强调一些要点，然后我们会用 R 来做一些实际例子，好吗？接着我们会更深入地看看如何使用这些概率分布。

18. 英文原文（段落18）：

so we have two types of variables, discrete and continuous, okay. I’m going to switch the screen, we don’t need the document camera right now, so… okay, good. okay, so we have seen that. so we’ve seen that we have two types of variables: discrete, continuous, okay. there are some examples here, we discussed those last time. the important thing is: how do you read them, okay. now we’re only talking about reading the plot of the probability distribution. the main difference is that if you have this simple example, a variable having a discrete variable which takes just the values, say, sunny or cost or shoulder between different types of weather, then it’s like a histogram basically, okay, which means the bars, the height of the bars, show the probabilities. so discrete probability distribution will look like this, where the height of the bars will show you the probability of that value, okay. so for example, probability it is sunny is simply the height of that bar, which is 0.3, okay. and all of these bars obviously add up to 1, because that’s the total probability, okay. so very simple to read. and the important thing to remember is that this is only true for discrete distributions.

中文翻译（段落18）：

我们先说两种类型的变量：离散型和连续型。我先切换一下屏幕，目前不需要用到实物投影仪……好的。行，我们之前也提到了：我们有两种变量，离散的和连续的。这儿有一些例子，我们上次讨论过了。关键是：怎样去理解它们的分布图？现在我们只谈如何阅读概率分布图。最主要的区别在于，如果这是一个简单的例子，比如你的变量是离散型的，只能取“晴天”“多云”“阵雨”等天气类型，那么它看起来就像直方图，条形的高度代表概率。离散型概率分布就是这种样子，每个条形的高度都是该值的概率。举例说，“晴天”的概率是 0.3，那就看那个条形的高度就是 0.3。所有这些条形高度加起来是 1，因为那是总概率。这样就很容易读。而要记住的是，这只适用于离散分布。

19. 英文原文（段落19）：

for continuous distributions, you cannot read them like this, okay. so let’s say height of a student. so what you have, again, what is common is that in both distributions, the x-axis shows the values the random variable takes. so in discrete, it’s only three. when you’re talking about height of a student, this is a continuous axis, okay. so I’ve just made up some numbers here, but basically there are infinitely many that you can show on the x-axis, because it’s a continuous scale. now, what does this show? this shows the probability density. the probability density is essentially the rate at which the probability is changing. so it’s the derivative, okay. so it’s the rate at which the cumulative probability is changing, okay.

中文翻译（段落19）：

对于连续型分布，你就不能这样去读。比如说，学生的身高。虽然这两类分布都有相同点：横轴代表随机变量的可能取值，但离散型可能就只有三个或几个值可取，而像身高这种是连续的。这里我举的数字是随意的，但它可以在横轴上取无数个值，因为它是一个连续的刻度。然后这个图表示的是什么？它表示的是“概率密度”。概率密度本质上就是概率随 x 的变化速率，也就是累积分布的导数。它就是累积分布增加的速率。

20. 英文原文（段落20）：

so it’s not very straightforward to interpret off the probability distribution when it is continuous, okay. so you shouldn’t interpret it like that. in fact, what we can say is that the area under this curve, which is the probability, you know, that’s always true, and all we can say is: what is the probability of x taking values in an interval, okay. so for example, what is the probability that in this case height is less than 170? well, that is just the area under the curve to the left of 170. why? but remember, we talked about this last time: if it’s a continuous variable, it takes infinitely many values, so the probability of any single value is zero. so you really can’t talk in terms of that. so all you can do is talk in terms of an open or closed interval of values, okay. so you can say: okay, but I can always find probabilities for a range of values, which is all you can do, and that is the area under the curve for that range, okay.

中文翻译（段落20）：

这就说明，当分布是连续型的时候，从概率分布图上直接读取就没那么简单了。所以你不应该像读离散分布那样来解读它。实际上我们只能说，这条曲线下的面积代表概率，这一点始终成立；然后我们能问的问题就是“x 落在某个区间的概率是多少”。比如，这个例子里，学生身高小于 170 的概率是多少？那就是曲线在 170 左侧这部分的面积。为什么这么说？回顾一下我们上次讲的：如果是一个连续变量，它能取无穷多个值，那么某个单点的概率几乎是零。因此我们根本没法谈某个精确值的概率，只能说某个区间内的概率，这也就是该区间下曲线对应的面积。

21. 英文原文（段落21）：

so that’s the main difference between the two types of distributions. this is just for illustration. if you had a probability distribution for a female height, then maybe it could look something like this, okay.

中文翻译（段落21）：

这就是离散型和连续型分布之间最主要的区别。这里的例子只是为了演示。如果你有一个女性身高的概率分布，它可能会看上去类似这样。

22. 英文原文（段落22）：

now two important characteristics are expectation and variance, okay. basically expectation is just the mean. now you have to remember this concept: expected value or expectation is just a technical term for the mean of a random variable, okay. so we all know what a mean is, and we write it as E of x, if x is the random variable. the expectation is expressed as E of x. what is the variance? well, variance, now you guys, some of you already know because we talked about it last time. so variance tells you the variability in the data, but really what is it, okay?

中文翻译（段落22）：

好，接下来有两个重要的概念：期望（expectation）和方差（variance）。从本质上讲，期望就是平均值。你要记住，期望或期望值其实就是随机变量的平均数的一个专业说法。所以我们都知道平均数是什么，也通常写作 E(X)，如果 X 是随机变量的话。那么方差是什么？有些同学可能已经知道，因为我们上节课提到过它。方差是用来衡量数据中变动性的，但它到底是什么呢？

23. 英文原文（段落23）：

so if you look at the slide carefully, have a look at the slide carefully and tell me, how would you interpret the variance? so what you should note is that the formula for variance, or it is defined, as the expected value of x minus mu squared. what is mu? mu is the mean, the expected value, okay. so basically it’s an expectation, okay, or an average, okay. but average of what? the average of x minus mu squared, which is the squared difference between x and the mean, okay. so what is the squared difference between x and the mean of x, the expected value? that is the variance.

中文翻译（段落23）：

如果你仔细看幻灯片，请告诉我怎么理解方差。你会注意到，方差的公式或定义是 E[(X - μ)²]。其中 μ 是平均值，也就是期望。所以它其实是一个期望或平均值，但平均的是什么呢？是 (X - μ)²，也就是 X 与其平均值之差的平方。这就是方差。

24. 英文原文（段落24）：

so if I say that the variance is 10, which means on an average the values are about 10 away from the mean, the squared difference, okay. that’s what it means. okay, so that’s how we interpret it. now that is why standard deviation is important, because standard deviation is the square root of variance. so right now in variance, we’re talking about the squared difference. now it’s kind of not straightforward to interpret, you know, we don’t talk in terms of squared differences. so if we take the square root, then we’re talking maybe more like the difference, okay. so that is why standard deviation is used more often, okay.

中文翻译（段落24）：

所以如果我说方差是 10，这就意味着从平均值算起，它们平方误差的平均水平是 10。我们可以这样来理解。正因为如此，标准差就显得很重要，因为它是方差的平方根。方差里我们谈的是平方差，这个不太直观，我们平时并不直接谈一个值和均值之间有多少平方差；所以取平方根后得到标准差，这就更容易理解一点。这也是为什么我们更常使用标准差。

25. 英文原文（段落25）：

now normal distribution, you guys have all heard about normal distribution ideas, yeah. so some just basic facts: normal distribution is nice and symmetric, centered around the mean, okay, and the tails go asymptotically down to zero, which basically means the probability of observing extreme values goes down, okay.

中文翻译（段落25）：

现在说到正态分布，你们应该都听说过吧。先给一些基本信息：正态分布是对称的，以平均值为中心，然后它的两条尾部渐近于零，也就是说，出现极端值的概率会越远越小。

26. 英文原文（段落26）：

in fact, what is very specific about the formula of the normal distribution? so you guys know all the probability distributions have a formula, like the mathematical expression, right? yeah, so the normal distribution also has a mathematical expression, which is called the Gaussian, or sometimes Gaussian function, and it has this specific property that for any normal distribution, regardless of what the mean or the variance is, 68% of the data is always within one plus or minus one standard deviation. so that’s that, and then 95% is between plus or minus 2. so this is why we use 95% a lot in statistics, okay.

中文翻译（段落26）：

实际上，正态分布的公式有什么特殊之处呢？大家知道，每一种概率分布都有它的数学表达式，对吧？正态分布也有一个叫作高斯函数的数学形式。它有个独特属性：对任意一个正态分布，不管它的均值和方差是多少，总有 68% 的数据落在均值加减一个标准差的范围内。然后 95% 的数据落在均值加减两个标准差的范围内。因此我们在统计中经常用到“95%”这个概念。

27. 英文原文（段落27）：

so if you come across 95% confidence intervals, 95% prediction intervals, things like that, it’s because they’re all originated from using normal distribution, and if you do two standard deviations around, then you get a 95% interval, okay. now if you do three standard deviations away from the mean, that basically contains almost all the data, 99%. okay, so there’s very little that is outside, okay.

中文翻译（段落27）：

所以，如果你碰到 95% 置信区间、95% 预测区间之类的概念，它们都是基于正态分布而来的。因为如果你在均值上下加减两个标准差，就能获得一个大约涵盖 95% 数据的区间。如果加减三个标准差，则几乎可以囊括 99% 的数据，外面只剩下很少的一部分。

28. 英文原文（段落28）：

so normal distribution is great if it is applicable, because it is very well known, it’s mathematically in a very good form, you can do lots of nice things with it, okay. and just to give you a better idea of how normal distribution works, normal distribution is, you know, defined by its mean and variance, or standard deviation, say, okay. how does… what does… how does the mean affect? well, if you change the mean, if you move the mean, you know, increase it or decrease it, you’re really just shifting the distribution, okay. so for example, this distribution has a mean of 0, the dashed line distribution has a mean of maybe 1, okay. so you literally just shift.

中文翻译（段落28）：

如果一个场景符合正态分布，那就很棒了，因为它是研究最透彻、数学形式也很便利的分布，所以你能用它做很多事情。为了让你对正态分布有更好的理解，我们知道它由均值和方差（或者说标准差）来定义。那均值起到什么作用呢？如果你更改均值，比如把均值从 0 改成 1，那么你只是把整条分布曲线平移了。举例说，实线上那个分布均值是 0，虚线这个分布的均值可能是 1，它就是整体搬了一下位置。

29. 英文原文（段落29）：

if you don’t change the standard deviation, they will look exactly the same shape, just shifted. on the other hand, if you don’t change the mean, but just change the standard deviation, then it changes the shape. well, the shape stays the same, it just becomes more, you know, spread out or more shrink, whatever you want to call it, okay. so that’s how they look. essentially, the tails of the distribution become heavier if you increase the standard deviation, which means the probability of observing something extreme kind of goes up, because you have a higher variation, okay.

中文翻译（段落29）：

如果不改变标准差，它们的形状就完全相同，只是位置不一样。相反，如果你不改变均值，只调标准差，那么曲线的形状会发生改变——准确说来，同样的“钟形”会变得更宽或者更窄。简单地说，如果你增大了标准差，分布的尾部就更“厚”，也就是说出现极端值的概率变高了，因为整体的波动范围变大了。

30. 英文原文（段落30）：

so normal distribution is great when you can use it, but in many cases you actually can’t. it’s not the right distribution. so on that note, let me get back to document camera and let’s see how we’re doing. okay, so let’s now consider an example of toss of a coin, okay, so very simple. what is the sample space? when the sample space, it just heads or tails, and of course you’re not even… but really heads or tails, okay.

中文翻译（段落30）：

因此，在能够使用正态分布的场景里它确实很好，但在很多情况下它并不适合。说到这里，让我先回到这个实物投影仪看看我们要做什么。好，让我们来考虑一个投硬币的例子，特别简单。它的样本空间是什么？就是正面或者反面。当然你可以说还有别的情况，但其实核心就是正面或反面，对吧？

31. 英文原文（段落31）：

now what do you know about a coin? if you toss a coin, typically what do you expect the probability of heads and probability of tails to be? more or less equal, right? that’s what we expect, okay. it’s not necessary, okay, so we expect that probability of heads and tails is half and half, but that’s not necessary. so let’s just say that probability of heads is some value, let’s say y, okay. so now here y is used as a symbol, it’s not the actual value of pi as a constant, okay, it’s just a symbol, okay.

中文翻译（段落31）：

那么我们对硬币知道什么？如果你抛硬币，一般会认为正面和反面的概率各是 50%，对吗？这是我们的预期，但不是一定如此。有时硬币可能不公平。所以我们就设正面出现的概率是某个值，叫它 y，好了。这里 y 只是一个符号，不是代表圆周率 π，别混淆，就当作一个普通符号。

32. 英文原文（段落32）：

so just to highlight, y is still between 0 and 1, and it’s just a symbol for, you know, your probability of heads, okay. so next, say the probability of heads is y, what will be the probability of tails? 1 minus y, great, okay, good. now let’s say you observe a heads, then of course the probability of a will be y, okay. now let’s define event B, when you observe a heads and you toss again and you observe a heads again, okay. so what will be the probability of B? yeah, 2 y… y squared? y times y, great, okay, y squared, okay, good idea.

中文翻译（段落32）：

强调一下，y 介于 0 和 1 之间，只是用来表示出现正面的概率。于是正面的概率是 y，那反面的概率就是 1 - y。好的。现在假设我们第一次抛出正面，事件 A 的概率就是 y。然后我们定义事件 B：第一次是正面，再抛一次又是正面。那 B 的概率是多少？对，两个 y 相乘，也就是 y²。好的，很好。

33. 英文原文（段落33）：

why? because, look what you’re doing: we multiply the probabilities, probability of a times probability of B, okay. basically, this is saying that the tosses are independent, okay. so we assume it, which is mostly true, but not necessarily, because if a person has a coin a weird way of tossing a coin and tosses again, maybe the probabilities are not independent, okay. but so I just wanted to highlight that fact, that, you know, we have… you’re making these assumptions, so you have to be careful about it. okay, but yes, let’s assume that they are independent, so y times y, y squared, okay.

中文翻译（段落33）：

为什么呢？因为你是在把两次正面的概率相乘，也就是 P(A)×P(B) 这种思路。这其实就是在假设两次抛硬币是独立的。我们一般都这么认为，但也不排除某人有特殊的抛法，让第二次的结果和第一次有关联。不过这里我们就默认它们独立，所以就是 y×y = y²。

34. 英文原文（段落34）：

now let’s define an event C, which is heads, heads, and tail. what will be the probability of C? shouldn’t be that hard. yeah, y squared is y… what, after that? we haven’t assumed that y is half. so probability of a tail is 1 minus y, okay. so in general you get the idea.

中文翻译（段落34）：

现在定义一个事件 C，包含三次投掷的结果：正面、正面、反面。那它的概率是多少？这不难，对吧？对，前两次正面就有 y²，然后下一次是反面，就乘以 1 - y。当然我们没假设 y 就是 0.5，所以就得保留这个表达式。大体上你能看出我们在做什么了。

35. 英文原文（段落35）：

so let’s do one more example: let’s say heads, heads, tails, heads, tails, okay. so what is the probability of d now? so there are three heads and two tails, right? so y cubed, 1 minus y squared. why? what do you mean it’s still equal to y? the results… why? ah, no, ah, look: each of these, each of the heads is probability y, right? so you’re assuming independence again, so y and y and 1 minus y and y and 1 minus y, so y cubed times 1 minus y squared. make sense?

中文翻译（段落35）：

那再举一个例子：正面、正面、反面、正面、反面，这样一串共有 5 次投掷。那这个事件的概率是多少？这里一共有 3 次正面、2 次反面，对吧？所以应该是 y³ × (1 - y)²。为什么呢？你可能会说还等于 y 吗？不对，每次正面都是 y，再乘以每次反面都是 (1 - y)，假设它们独立，所以最后就是 y³ × (1 - y)²。能理解吧？

36. 英文原文（段落36）：

yeah, I hope it makes sense, but you can try this out, okay. so what do you get in general, okay? so let’s say that event E is you have x number of heads, so x is the number now, and y tails, again, y is also the number, then what will be the probability of E? so you have, yeah, exactly, that’s right, so y to the power x times 1 minus y to the power y, okay. in general, let’s say you have n tosses, okay, and you have x heads out of those n tosses, okay. then probability of, you know, basically x heads… so let’s say you have heads-tail in some order, okay, will be y raised to x times 1 minus y raised to n minus x, because heads and then the remaining is x, okay.

中文翻译（段落36）：

好的，希望能说得清楚，不过你可以自己试着去做。那么一般情况下会得到什么？假设事件 E 是说我们有 x 次正面，记得这里的 x 代表次数，而 y 次反面……当然，我用 y 有点混淆了，但实际上就是另一个数。然后你问它的概率是多少？就是 y^x × (1 - y)^“尾数”（不过要小心符号不混用）。更通用地讲，如果我们抛 n 次，出现 x 次正面，那么不管它们以什么顺序出现，单个序列的概率是 y^x × (1 - y)^(n - x)。对吧？

37. 英文原文（段落37）：

so this distribution is a Bernoulli distribution, okay. so the first distribution on your list here, Bernoulli, is that distribution. so it’s so easy, okay. well, of course, this is the easiest distribution, but you can get the idea how the probability distributions are derived, okay. so you define a random variable, you try and understand how it behaves, okay, and how it can give you different probabilities. so with this function, like you can say, just starting from the very basics and using just the assumption of independence, we could derive this mathematical form, which is applicable for any number of tosses of coins. doesn’t have to be 1 or 2, any number of tosses, with some number of x, some number of tails, you have the probability distribution, okay. that’s what the probability distribution is. that’s how they come about. basically Bernoulli is for a specific sequence, which is this. in general, if you’re not interested in a specific sequence, but just interested in: what happens if I toss the coin n times? how many different ways? I don’t care in which order I get, but what is the probability that I’ll have x number of heads out of n tosses? then you get a slight different distribution, which is a binomial distribution.

中文翻译（段落37）：

这样得到的分布就叫伯努利分布（Bernoulli distribution）。你们清单上的第一个分布就是伯努利分布。这很简单吧。当然，这是最简单的分布之一，但它能让你明白概率分布是怎么来的：你定义一个随机变量，然后去理解它的行为，以及它如何产生不同的概率值。只要从一些最基础的假设（比如独立性）开始，你就能推导出像这样的数学形式，适用于任何次数的硬币投掷，不管是 1 次还是更多次，出现 x 次正面和若干次反面的概率都能算出来。这就是所谓概率分布的来源。严格来说，伯努利分布是针对一次试验或特定序列的；但如果我们不是关注具体序列，而是只想知道“n 次投硬币里恰好有 x 次正面”的概率，而不在意次序，这时你就会得到一个稍微不同的分布，称为二项分布（binomial distribution）。

38. 英文原文（段落38）：

so probability in general, can you see this, is of x number of heads out of n trials, okay, regardless of the order, okay, so this is regardless of the order in which it occurs, is n choose x, which means number of possible combinations of getting x heads out of n, times y raised to x, times 1 minus y raised to n minus x, which is the same, okay. so that’s it, it’s the same, but you’re just multiplying by the number of combinations, okay, the number of ways in which you can have x heads, and which is the same, and that’s a binomial distribution, okay. so this distribution is called the binomial distribution. so this is Bernoulli, this is binomial, they’re quite related, just slightly different, okay.

中文翻译（段落38）：

因此，假如你想知道在 n 次独立试验里恰好出现 x 次正面的概率，那么不管具体顺序如何，公式就是 C(n, x) × y^x × (1 - y)^(n - x)。这里 C(n, x) 代表从 n 次投掷中选出 x 个正面的组合数，然后乘以单个序列的概率 y^x × (1 - y)^(n - x)，这就是二项分布。所以这个分布就叫 binomial distribution，它与伯努利分布紧密相关，但又略有不同。

39. 英文原文（段落39）：

so that’s just… so you have a question now? so that’s just, you know, I just wanted to give you a simple example to show, you know, what probability distributions actually are like, how do they look, and how do they come about, because it’s very easy to just read a formula, you know, you can say, oh this is the formula, but if the formula doesn’t really talk to you, then you don’t really understand it, okay. whereas now, at least for these two distributions, you can see how they may derive, okay.

中文翻译（段落39）：

就这样……你现在有什么问题吗？我只想通过这个简单例子来说明概率分布实际上是什么样子的、它们是怎么来的。因为你很容易就直接拿到一个公式，说“哦，这就是公式”，但如果你不了解公式背后的推理，你就不会真明白它。而现在，至少对于伯努利和二项分布这两个例子，你能看到它们是怎么从基本原理推导出来的。

40. 英文原文（段落40）：

again, all these distributions, these are people: Bernoulli, binomial, Poisson, exponential — these are kind of common distributions. also gamma distribution, beta distribution — these are not people, just names. and then if you’ve done some hypothesis testing — we will do that, okay — you will come across t distribution, F distribution, chi-square distribution, all that. so there are quite a few commonly used distributions, okay.

中文翻译（段落40）：

再说一下，这些分布比如伯努利、二项、泊松、指数，通常都是用某些人的名字来命名的，还有伽马分布、贝塔分布就不是人的名字。不过如果你做过一些假设检验——我们到时候也会讲，你就会遇到 t 分布、F 分布、卡方分布等等。这些都是常见的分布，有相当多。

41. 英文原文（段落41）：

so any questions so far about what I did with the binomial or Bernoulli? yeah, one, good. can you get the main idea? so this thing is not about… okay, so some examples now, in short time I’ll show you the code I used to run that, okay, but this is just a visualization. so what I’ve got here is one of the distributions, which is gamma distribution, okay. I’ve plotted three different gamma distributions. now we’ve seen how two normal distributions look, okay, so we looked at how two normal distributions look. now you have three gamma distributions. what do you see?

中文翻译（段落41）：

到目前为止有没有对二项和伯努利分布还有什么问题吗？好的，有一个问题？能理解核心概念就好。这里不再展开了……行，那接下来我们看一些例子，稍后我也会给你们展示我运行代码的方法，现在先给大家看看可视化的效果。比如这里我有一种分布，叫伽马分布（gamma distribution）。我画了三个不同的伽马分布。我们之前也看过两个正态分布是什么样子。现在你看到这三个伽马分布，有什么感觉？

42. 英文原文（段落42）：

with normal distributions, they all had the same shape; the only thing that was different was the mean was different or the variance was different, okay, so they were either shifted or they were just more spread out or less spread out, but the shape was the same. what you see here is that the shape can vary a lot, okay. so what I’m trying to say is the non-normal distributions are not necessarily of the same shape. like the same distribution can take very many different shapes if you change the parameters, okay. so you have this highly skewed distribution here, maybe slightly less skewed, and this is almost like a symmetric distribution. they are all gamma distributions, just changing the parameters.

中文翻译（段落42）：

在正态分布的例子里，我们发现它们形状都一样，只是位置和宽度变一变，比如均值不同或者方差不同，于是曲线要么平移、要么压缩或拉伸，但整体还是那个“钟形”。可你在这里看到的是，伽马分布的形状可能差别很大。也就是说，非正态分布并不必然保持统一的形状；对同一个类型的分布，如果你改变它的参数，形状会大不一样。你看这里，一个非常偏斜，一个稍微好一点，还有一个接近对称，但它们都是伽马分布，只是参数变了而已。

43. 英文原文（段落43）：

another example is three different beta distributions, okay. so what do you notice that is different here? there is no tail? sorry, there is no tail? no tail — which one has no tail? is it just a medium line, is it just a medium line? no, it is a distribution. so a good point, that is one of the things I wanted somebody to say, that just asked the question, okay. so you can’t actually easily see the third distribution here, because the third distribution is just a flat line. have you come across that type of distribution before?

中文翻译（段落43）：

还有一个例子是三个不同的贝塔分布（beta distribution）。你看它们有什么不同？有人说没有尾部？怎么，没有尾部？哪个是没有尾部的？那条平的线？它其实也是一个分布。非常好，你问到这是我希望有人提出的问题。你可能在图上不太容易看出第三条线，因为它就是一条水平线。你见过那种类型的分布吗？

44. 英文原文（段落44）：

has anybody come across that type of distribution? well, it is called a uniform distribution, okay. that’s also another distribution. so uniform distribution is essentially like a flat line over the whole range. basically means all probabilities are equal, okay. so you can actually get that flat distribution also using beta. so beta can produce so many different shapes, okay.

中文翻译（段落44）：

有人见过那种分布吗？它叫做均匀分布（uniform distribution）。也是一种分布。它在整个区间上的高度都是一样，意思是每个值的概率都相同。所以在贝塔分布里也能产生这样的平直线形状；贝塔分布能生成非常多样的形状。

45. 英文原文（段落45）：

now, what is, do you notice here that is different from what you’ve seen before? look at the range. the range is 0 to 1 for all of them. so beta distribution is only defined over the range 0 to 1, okay. it’s not defined over anything outside of that range. also, if I go back to the previous plot, the gamma distributions, they all start at zero. they’re only defined on the positive side of the axis, not on the negative side, okay. so these are the points I wanted to show, and that’s kind of summarized in these two slides.

中文翻译（段落45）：

现在，你还能观察到跟前面不同的一点吗？你看它们的取值范围，都是在 0 到 1 之间。因为贝塔分布只定义在 0~1 这个区间内，而不会出现在区间之外。另外，如果你回到前面看伽马分布的图，你会发现它们也是从零开始，只在正数一侧有定义，没有负数部分。这些就是我想展示给你们的，之前的两张幻灯片也做了总结。

46. 英文原文（段落46）：

so normal distribution is always symmetric, nice, same shape, everything, but other distributions are actually nothing like that, okay. the t distribution we will come across later in hypothesis testing is kind of close to normal, because it’s symmetric and very similar to normal, 0-1, but aside from that distribution, all other distributions are not symmetric mostly. they are often skewed, and they can take very different shapes depending on the parameters you use, and they will be defined on different ranges. they’re not like normal distribution can be defined on the whole real axis; not other distributions.

中文翻译（段落46）：

因此，正态分布总是对称、光滑，形状看起来也都差不多。但是其他分布就未必如此。有些分布，比如我们以后会在假设检验里遇到的 t 分布，它和正态分布很像，因为它也对称，而且接近正态 0~1 的形状。但除此之外，许多分布并不对称，往往是偏的。而且它们会根据参数不同，呈现出非常多变的形态，并且它们的定义域可能不一样——不像正态分布那样可以在整个实数轴上定义。

47. 英文原文（段落47）：

and also normal distribution is the only distribution which is defined using mean and variance, okay. so remember mean and variance is like the basic thing we talk about in stats, but it’s actually only useful for normal distribution. all of the distributions are defined in different parameters, okay. so you have degrees of freedom, or shape, shape-scale, and other things. so different types of parameters for different distributions, okay. but we all can talk about the mean and variance of those distributions as well, but they’re not defined using mean and variance, okay. they’re defined using different parameters. so this is just to show that, give you at least some idea about how other distributions look, because you do need to have that. it’s not just about normal distribution.

中文翻译（段落47）：

还有一点，正态分布是唯一一个用均值和方差来定义的分布。我们总是谈均值和方差，好像它们是最基础的概念，但实际上它们只对正态分布格外管用。其他很多分布会用别的参数来定义，比如自由度、形状参数、尺度参数等等。每种分布都有不同的参数。虽然我们也可以讨论它们的均值和方差，但那并不是直接用均值和方差来定义的。我要说这些是想让你们对非正态分布也有个印象，因为我们不能只盯着正态分布。

48. 英文原文（段落48）：

and so then there is an example of Bernoulli, so Bernoulli slide, so now hopefully when you read this, you will make… it will make more sense. and here I’ve used the symbol p, rather than y, so in class I used y, but it’s the same thing, and I’ve given you the mean and variance for both Bernoulli distribution as well, so you can find the mean and variance, but it’s not defined using them directly, okay. and then the binomial, again, the mean and variance for binomial: that’s n choose… so that shouldn’t be p by that, that should be n choose x, I made a mistake. so that thing should be… I’m not sure if I can change that right away. yeah, that’s the right one, okay. makes sense, right, should be n choose x, okay, so that’s that. and then I’ve also said Poisson distribution, how you get it, so you can read about it, but again, it could now make more sense.

中文翻译（段落48）：

然后这里又举了一个伯努利分布的例子，所以当你读到这部分时，希望能更好理解一些。在课上我用了符号 y，这里用的是 p，但其实是一样的。我也给出了伯努利分布的均值和方差公式，你可以自己去算算看。不过要注意，这个分布并不是用均值和方差来定义的。同样地，二项分布的均值和方差我也给了，比如说 C(n, x)……哦，对，我这里还打错了一个地方，本来应该是写 n choose x 才对，我暂时先没法改，不过大致就是这么回事。然后我也提到了泊松分布，你可以读一读，这样或许也能理解得更深一些。

49. 英文原文（段落49）：

poisson distribution is used a lot for rare events, such as number of earthquakes, number of burglaries, and things like that. number of earthquakes, for example, a lot of these kinds of things, and I’ve given the sample space, the expected value, now what you notice here is that you define using a parameter lambda, which is the rate parameter, so rate at which certain things happen, and that rate parameter is both the mean and the variance of the distribution, okay. so the mean and… so if you have a Poisson distribution, it will have the same mean and same variance always. so I’m just showing you how these different distributions, actually how different they can be from normal distribution, okay, just so that you have a more general idea of what mean and variance means, and what probability distributions actually mean and they used for different purposes, okay. so some kind of data, you only need that type of distribution, you can’t use a normal.

中文翻译（段落49）：

泊松分布常用来描述罕见事件的出现次数，比如地震数量、入室盗窃事件数之类的。像地震这种，就是典型的泊松过程。我也给出了它的样本空间和期望值。你会发现，泊松分布用一个参数 λ 来定义，λ 代表某件事发生的速率（rate），而在泊松分布中，这个速率 λ 同时也是它的均值和方差。如果你使用泊松分布，那么它的均值和方差始终相等。我之所以展示这些不同分布，就是想让你们看看它们和正态分布有多不一样。这样你就能对均值、方差以及概率分布在不同场景下的意义有一个更广泛的认识。有些场合你只能用那种特定分布，不能套用正态。

50. 英文原文（段落50）：

and then exponential, okay, so exponential gives you the time between two events. so if you had an earthquake and you are expecting another earthquake, what is the time between those two? so it’s related to Poisson distribution, but Poisson distribution can say how many, exponential can say what’s the time between the two. it’s also used a lot by the way in optimizing resource allocation, so for example if you go to a supermarket, you have a number of checkout operators, and they keep changing that, but how do they know? so you can model how many checkout operators you need if you knew how many customers you’re likely to have, using exponential distribution and also assumptions, okay. so all these distributions have, you know, a lot of good uses. so I’ve got an example here as well, I think, okay, and then you can argue gamma distribution, so there’s a formula for that, so you get an idea, okay. so that’s the summary of distributions, basically what I said, okay.

中文翻译（段落50）：

然后是指数分布（exponential）。它通常用来描述两个事件之间的时间间隔。比如经历一次地震后，再下一次地震到来的时间间隔可能符合指数分布。它和泊松分布是相互关联的——泊松分布关注在一定时间里发生多少次，而指数分布关注的是两次事件之间的时间长短。它也被大量用于资源调度的优化中。比方说，你去超市，结账柜台的人数需要多少？他们总在调整。那他们是怎么知道要几个呢？你可以用指数分布去建模排队顾客到达的间隔时间，再结合一些假设来进行算。总之，这些分布都有其各自的用途。我这里也有一些示例可以看看，然后还可以讨论伽马分布之类的，它也有自己的公式。好啦，这就是分布的概述，大概就是我说的那些。

51. 英文原文（段落51）：

so this is what we’re going to do now, which is actually do some real data exercise… well, not real data, but some exercises, okay, so first and foremost, now I’ve made these things active on cameras, so the two different two articles, you can say, one is called probability\_distributions\_for.Rmd. now what is Rmd? Rmd is an extension for an R Markdown file, which means you can basically write a blog using R, okay, and that blog can be either an HTML file or a PDF, or even a Word file, I think, okay. so you can specify what kind of document you want, and then you have to knit it to run it. so how many of you have used R Markdown for… fainting type? some of you have, okay. so in many of our courses we actually ask students to prepare assignments using R Markdown, okay, but in case you haven’t come across this, this is a code on Canvas, have a read to it, you know, pretty easy to understand, but just another way for you to learn R, okay, and what you can do. so when I knit it, it will produce this document.

中文翻译（段落51）：

接下来咱们要做的就是一些基于“真实”数据的练习……好吧，其实不一定是完全真实的数据，但总之是一些习题。首先，我已经在 Canvas 上放了两份素材文件，可以说是两个文档，其中一个叫 “probability\_distributions\_for.Rmd”。那什么是 Rmd 呢？Rmd 是 R Markdown 文件的扩展名，意思是你可以用 R 来写个文档或博客。它可以生成 HTML 文件、PDF，甚至 Word 文件之类的。你可以自行指定想要输出哪种格式，然后用“knit”来运行。有谁用过 R Markdown 写文档的？有人用过吧。在很多课程中，我们都要求学生用 R Markdown 来做作业。如果你没见过，可以去 Canvas 上看看相关内容，其实很容易理解。它只是让你有另一种方式学习 R，以及如何把代码、结果和文字写在一起。当我点击“knit”时，它就会生成这样一个文档。

52. 英文原文（段落52）：

do you want to install? yes. so it might also ask you questions like that, you know, it requires… you need to install certain packages, or it’s not updated, so usually if you just click yes or whatever, it does its thing. so what I’ve done here is, of course, you can specify the title and all that stuff, and you specify your codes inside these three kind of marks, and whatever you type here is usually shown as a text, okay. so you have a lot of texts like your blog, and then snippets of codes that you can show in between. you can also include some plots and things like that. so I’m not sure if it is going to run now or not. oh, I thought it did that. okay, so I actually have the HTML file, so I can just show you that file. there is some problem running that R Markdown here, but you, you know, that shouldn’t be a problem for you guys, okay.

中文翻译（段落52）：

它会问你是否要安装一些东西，你就点“yes”就行，有时它需要一些额外的包或者要更新版本。然后在这个 R Markdown 文件里，你可以指定标题、作者之类的，在文档中用一段三重反引号把 R 代码包起来，就像写博客一样，把文字和代码片段掺在一起，也可以包含图表之类。至于现在我不确定是否能立刻运行……哦，我本来以为能行……没关系，我有一个生成好的 HTML 文件，可以展示给你们看。这里遇到点兼容问题，但对你们来说不应该有什么麻烦。

53. 英文原文（段落53）：

so this is what I was trying to compile, usually should be very straightforward. okay, ah, so you can see that I’ve written this out. now, these are Markdown files, so I’m not going to… I’ve just made this into a PDF file or something, or just slides, okay, but I decided to do an R Markdown, so that the code is there, you can actually use it and learn to do that, which is one idea, but also it is a really nice, convenient way of showing the output, snippets of code, everything together, rather than copy-pasting it and, you know, things like that.

中文翻译（段落53）：

这就是我想编译的东西，通常是很容易的。好了，你可以看到我写的内容。这些都是 Markdown 文件，我本来也可以直接转成 PDF、或者做成幻灯片，但我选用 R Markdown 的原因是：这样可以把代码直接放在里面，你可以看到并且学习怎么写，这也是其中一个目的。同时，这种方式也让展示输出结果更方便，不用到处复制粘贴，可以把所有内容整合在一起。

54. 英文原文（段落54）：

so okay, so what I’ve got is how do you use probability distributions in R. now, for most distributions, you have four commands in R which allow you, which allow you to sample from the distribution, find the probability distribution, find the cumulative probability, so the cumulative distribution, and also the quantiles, okay. and these are called like dnorm, pnorm, qnorm, and rnorm. this is for normal distribution, by the way, okay. so these functions will allow you to do these different things with those distributions. so I’ve given some example: so let’s simulate a hundred values from normal distribution with mean 10, standard deviation 1, okay. now how do you simulate? so this is where you need the rnorm, which is basically r stands for random, okay. so basically simulate randomly from this distribution, that’s basically what rnorm is, okay.

中文翻译（段落54）：

好的，那么我这里要说的是如何在 R 里使用各种概率分布。对于大多数分布，R 中都有四个常用函数可以做：从分布里抽样、查看分布的密度函数、查看累积分布、以及求分位数。比如正态分布就对应 dnorm、pnorm、qnorm、rnorm，这些函数就能分别完成你需要的操作。我举个例子，比如我们要从均值为 10、标准差为 1 的正态分布里模拟 100 个值。该怎么做？要用到的就是 rnorm，其中 r 代表“random”，也就是随机数生成的意思。它就是用来从指定的正态分布里模拟抽样。

55. 英文原文（段落55）：

and so what I’ve done here is, anybody knows what set.seed does? or if you don’t use it in R but in some other coding language, do you know what kind of thing set seed is? so how does simulation, or randomization, work when you have algorithms, okay, which simulate, randomly simulate numbers basically? now those algorithms need a seed. if you provide the same seed, it will simulate exactly the same numbers every time you run it, okay. this is a common idea in coding when you’re simulating something or generating random numbers, okay.

中文翻译（段落55）：

这里我还用了一个 set.seed，大家知道这是什么吗？如果你不用 R，用其他编程语言也许见过类似的东西。它是干什么的呢？当你用某个算法来模拟或随机生成数字时，需要一个种子。如果你传入的种子相同，那么每次运行产生的随机数序列就会一模一样。这在编程里做模拟或生成随机数时很常见。

56. 英文原文（段落56）：

so what this does is, you don’t need to provide this command, but if you don’t provide this command, by default, it will produce a different set of 100 values every time I run it, okay, because just randomly it is going to simulate different numbers every time. but if I provide the same seed, then it’s going to generate exactly the same numbers. so I did it here, I normally don’t do it, but I did it here because I wanted the same output to be produced every time I run it, because I’m going to use it for teaching, so I don’t want different output, okay. that is why, but also again, the reason I put it here is this is how you get to learn different things about R, okay.

中文翻译（段落56）：

所以你并不是必须要使用 set.seed。如果你不设定种子，每次运行代码时它都会生成一组不同的 100 个随机数，因为每次都是一个新的随机过程。但如果设了相同的种子，每次运行就会得到完全相同的结果。我在这儿加上 set.seed 是因为想保证每次我演示时都得到同样的输出，用来教学比较方便，不然数字老变。还有一个原因是，通过这个例子也能让大家学习 R 的一些小技巧。

57. 英文原文（段落57）：

so as I said, I’m not going to teach you how to use R, but I will point these tricks, and the codes are available, so that’s a good way for you to learn different things, okay. so I just wanted to clarify what that was. okay, and then what I’m doing here is I’m using rnorm command. rnorm command has three different inputs: the first one is how many simulations you want, so how many values you want to simulate, in this case I specify 100, and then the next one is the mean, and then the standard deviation. so I’m basically saying simulate 100 random values from the normal distribution with mean 10 and standard deviation 1, and store it in x. so x will contain 100 values, okay. and then I use the summary command, if you’ve seen, to look at what is the summary statistics for x.

中文翻译（段落57）：

正如我说的，我不会在这里系统讲怎么用 R，不过我会提示一些技巧，而且代码都给你们了，可以自行摸索。所以我这里先解释一下这一行在做什么。我调用 rnorm 函数时，传了三个参数：第一个是要生成多少个模拟值，这里是 100；第二个是均值；第三个是标准差。我就指定均值=10、标准差=1，然后把结果存到 x 里，于是 x 里就有 100 个模拟值。接着我用了 summary(x) 来查看 x 的统计概要。

58. 英文原文（段落58）：

now what do you expect to see in this summary statistic? what should be the mean, what should be the median? so for a normal distribution, the mean is of course mu, whatever you specify, and the median is also the same as mean. that’s one of the characteristics of a normal distribution, okay, the mean and median are always the same, okay, because it’s nice and symmetric. so what do we expect to see is mean and median to be equal to 10, because that’s what we specified, okay, which is total R to simulate values from normal distribution with mean 10. what do you see? it’s close to 10, it’s not exactly 10, okay, because they’re just a sample. see, this is an example here, the population parameter is the population mean is 10, but what we actually have is just a sample of data, 100 values. so the mean of those 100 values is not going to be exactly 10, okay. so just remember that.

中文翻译（段落58）：

你觉得 summary 会给出什么结果？它会显示均值、最大最小值、中位数等等。对于一个正态分布，如果我们的均值设为 10，那么理论上它的均值就是 10，并且因为正态分布是对称的，中位数也应该是 10。但你猜我们得到的结果是多少？它不会恰好是 10，而只是接近 10，因为我们拿到的是一个样本，不是总体。总体参数确实是 10，但在 100 个样本里算出来的均值可能会有点偏差。这一点要记住。

59. 英文原文（段落59）：

and if you do this, repeat this code, and you don’t use the seed, okay, so if you use the same exact code, which means you also have the seed, and then you do this, you should get exactly the same like this, okay, but if you don’t use the seed and simulate it, you will get different answers, because it won’t generate different random numbers, okay. so try this out. so try this out in your time, the code is on Canvas. use it and, you know, use it without the seed, and then see if it does different things, okay. so in my case, that’s what I get. and then this is the histogram to see how the data looks, okay. so again, you expect the data to be nice and symmetric because it is from the normal distribution, but it is actually not exactly symmetric, right, why? because it’s a random sample, okay.

中文翻译（段落59）：

如果你多次运行同样的代码，而且用到了相同的种子，那么你每次得到的那 100 个值都一模一样，所以统计结果也一样。但要是你不用种子，每次运行就会得到不同的随机数序列，所以也会得到不同的结果。你可以自己试试看。代码已经上传到 Canvas，自己在空闲时运行，用或者不用 set.seed，然后比较结果有什么不同。就我这里举的例子，我得到一组数据，再画出它的直方图。我们知道它大体上应该是对称的，因为来自正态分布，但实际上会有些不完全对称，因为只是一个随机样本嘛。

60. 英文原文（段落60）：

and then you can do the boxplot and so on, okay. you can do it. now what am I doing here is a really useful bit, okay. so I’m saying, okay, let’s look at the first simulated value. so the first value… so I’m just saying, okay, give me the first value of x, because x is 100 values, okay. so that is 9.43. so 9.4395 is the first simulated value. what I can do is I can find a density for that value using the dnorm, okay. so the dnorm function will give you the density, okay. so I got 9.4395, the density of that, if it’s a normal distribution with mean 10 and standard deviation 1, you need to specify that, okay, then the density is 0.34. it’s just the probability density.

中文翻译（段落60）：

你也可以画箱线图等等，这些都没问题。然后我这里想演示一个有用的小点子：我取 x 里的第一个模拟值，比如它是 9.43（具体数字是 9.4395 之类），然后用 dnorm 来算这个数的概率密度。dnorm 会根据给定均值=10、标准差=1 的正态分布，输出 9.4395 这个点的密度值。这里算出来大约是 0.34，那只是一个密度值。

61. 英文原文（段落61）：

what it means is the cumulative probability is changing at that point in that rate, so it’s a bit abstract, okay. now, what is the cumulative distribution at that? what is a cumulative distribution? do you guys know what the cumulative distribution is? so cumulative distribution gives you the cumulative probability, okay. so for example, this is important, okay. so cumulative distribution function, so this is the function, and this is denoted by x, okay. so it’s written as the function of x, capital x, random variable x, taking values less than or equal to small x. so what is the cumulative probability of x taking values? so for example here, let’s say I’ve got 10, the mean, standard deviation is 1, so if I have a normal distribution with mean 10 and standard deviation 1, what would be kind of this value here? so let’s say this is like, you know, three standard deviations away from the mean, okay, so this would be about 7, and this would be 13, okay.

中文翻译（段落61）：

这个数值意味着在那个点上累积分布增加的速率。确实有点抽象。那么，什么是累积分布呢？你们知道累积分布函数吗？它给的是累计概率。比方说，这很关键。累积分布函数（CDF）可以记作 F(x) = P(X ≤ x)。它表示随机变量 X 取值不大于某个 x 的概率。如果我们这里举个例子，均值=10、标准差=1 的正态分布，假设我们看一下 x=7，这是比均值小 3 个标准差的位置，而 13 则是比均值大 3 个标准差的位置。

62. 英文原文（段落62）：

now, we know that 99.7% of the data is here, okay, there’s only a tiny bit that’s in this way. so what I can do is I can say, okay, what is the probability of x being less than 7, okay? so x less than 7, but that will be basically zero almost, because there’s nothing below zero, okay, so sorry, okay, so because most of my data starts from 7. what is the probability that x is less than 7? that is also 0, okay, but what is the probability that x is less than 10? so let’s see if you guys notice. so what is the probability that x will be less than 10? sorry, about 50%, 1/2, yes. why? because it’s a mean and median. we have a normal distribution with mean 10 and median 10, okay, so 50% of the values are below it, so I wrote 0.5, I wrote 1. so basically this curve will go up, and it will be 0.5 at 10, and then it will keep going up, and it will be 1 at 13, okay, that’s the cumulative probability.

中文翻译（段落62）：

我们知道有 99.7% 的数据落在均值±3 个标准差之间，也就是 7 到 13 这个区间，其余在更远的极端。不过如果我想问 X<7 的概率是多少？几乎是 0（其实并不严格是 0，但非常小）。那如果问 X<10 的概率是多少呢？那大约是 50%，因为均值=10 的正态分布是对称的，所以有一半在 10 以下。这里我就写了 0.5 表示中位数在 10 的位置。再往右到 13，就几乎达到 1 了。这就是它的累积分布。

63. 英文原文（段落63）：

so probability of x being less than 7 is about 0, x less than 8 is somewhere here, x less than 9 is somewhere here, 10 is 50%, and so on, okay. so this is called the cumulative probability distribution, okay, and this is represented as capital F, the probability distribution function F of x, okay, that we talked about. so the PDF is the derivative of this CDF, that’s why this is probability density, okay. as I say, it’s the derivative, it’s the derivative of the CDF. anyhow, we don’t want to go into those details, but this is, like, really basic stuff you should know, okay, this is what we expect you to know because you’ve done at least one stats course in your undergrad, that’s the previous. okay. so but I know most of you have forgotten everything, so that’s why I’m trying to remind you, okay. so go and read about it, okay, I can’t, I don’t have time to teach everything, but go and read a book, okay.

中文翻译（段落63）：

所以 X<7 的概率大约是 0，X<8 大概在这儿，X<9 在这儿，到 10 刚好是 50%，等等。这就叫累积分布函数（CDF），用大写 F 表示，就是我们讲到的分布函数。而 PDF（概率密度函数）就是它的导数，也就是随 x 的微分变化率。我在这里就不多展开了，但这些确实是很基础的内容，应该在你们本科的统计课里都学过。如果你们大部分都忘记了也没关系，我现在做点提醒，但还是希望你们能回去翻翻书，好吗？我也没时间面面俱到地再教一遍。

64. 英文原文（段落64）：

so now how do you get this value? you can get that value using the pnorm function. so that’s what I’ve got here. so if you go here, the pnorm command will give you that probability, the cumulative probability, okay. so essentially it is giving you the probability of area under the curve, okay. and then the qnorm will give you the quantile, so what’s the 84th quantile, the 84th percentile, okay, of this distribution? it will tell you, well, x = something, for example, okay, things like that.

中文翻译（段落64）：

那你如何计算某个区间或某个点的累积分布值呢？就用 pnorm，这就是我想说的。如果你想知道某一点的累计概率，pnorm 就会输出那个值，也就是曲线下方的面积。再比如，你可以用 qnorm 去算分位数，比如问这个正态分布的第 84 个百分位是多少？它会告诉你 x 大约等于多少。

65. 英文原文（段落65）：

so what I want you to do, okay, is go to this document, this document is also on Canvas, okay, so go to this document and think about these questions. so I’ve actually asked you to think about how do you use qnorm, how do you use pnorm, and how can you use it for doing different things, okay, and we will use this in the workshop one Monday anyway, okay, but you guys need to have a good look at it otherwise it will not make any sense, okay.

中文翻译（段落65）：

所以我想让你们做的是，去看看我放在 Canvas 上的这个文档，然后思考里面的问题。我里边问了：qnorm 怎么用，pnorm 怎么用，它们能干什么？我们下周一的研讨课里也会用到它们，如果你们不先熟悉一下，就会觉得很难懂。

66. 英文原文（段落66）：

now I’ve also asked you to think about what is this? this is a probability distribution, I literally plotted a normal distribution with mean 10 and standard deviation 1. how do you plot this using the pnorm command? think about that, okay, or this is the CDF that I was trying to show here, okay, this is actual CDF. you can plot that using the pnorm command, okay, so think about that. you have the same kind of commands for other distributions, so if you have a binomial, you have the rbinom, dbinom, qbinom, pbinom, sorry, dbinom, and then Poisson, exponential. for almost all these standard distributions, you have these same set of commands, okay, so I want you to try these things out, okay, so for each of these commands, each of these distributions, try playing with it, try simulating from the distribution, try the p- command, the q- command, the d- command, okay, r- command, and then see what you get, okay, and see how you can use it for defining different probabilities. so try that out, and we will use this in the workshop.

中文翻译（段落66）：

我还让你们想一想这幅图：它是一个均值=10、标准差=1 的正态分布。我用了 pnorm 来画出来。你想想怎么用 pnorm 来画CDF图？或者类似地，pnorm 也可以帮你得到累计概率，dnom 也可以帮你画 PDF。对其他分布也是一样，比如你想要二项分布的函数，就有 rbinom（生成随机数），dbinom（密度/概率），pbinom（累积概率），qbinom（分位数）等等；对泊松、指数分布也是同样的思路。你可以尝试去操作、去模拟，然后看看得到什么结果。这样你就能理解该怎样用它们去定义或计算概率。请务必试一下，我们在研讨课会再深入应用。

67. 英文原文（段落67）：

any questions so far? for too many questions? okay. it’s only 1 hour 20 minutes, I should be going up to two hours, of course, it’s only Friday afternoon, but that is the reason what I’ve done in most of… if you look at my schedule, most of the times the workshop is on Tuesday, it’s only one or two Fridays that we’re using because, you know, I understand, by Friday afternoon you’re not exactly fresh, that’s fine. okay, but all good, okay.

中文翻译（段落67）：

到这里有没有什么问题？问题太多也行，哈哈。现在只过去了一个小时二十分钟，我本来可以上两小时的课。不过嘛，今天是星期五下午，我也知道这个时候大家都不是很精神，这也是为什么在我的课程表里，大部分研讨课安排在周二，只在一两个周五会这么安排，我能理解大家此刻的状态。好吧，那都没问题就继续。

68. 英文原文（段落68）：

so if you have no questions, actually we can stop because that’s all I wanted to do, okay, but yeah, make sure you go through the code, you know, study this, and then we’ll see you on the Tuesday or the lecture.

中文翻译（段落68）：

如果没有问题的话，我们就到此为止吧，因为我想讲的也差不多了。不过，请一定要把这份代码熟读一遍，搞明白，然后下周二或者在下次课上再见。