7
Lumping

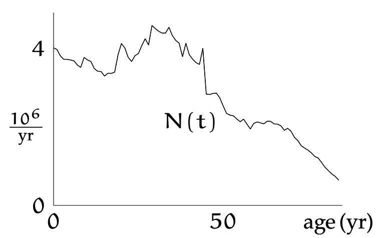
The symmetry chapter (Section 3.1) introduced the principle of invariance: 'When there is change, look for what does not change.' However, when you cannot find any useful but unchanging quantity, you have to make one. As Jean-Luc Picard often says, 'Make it so.'

7.1 Estimating populations: How many babies?

The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau [33] publishes the number of people at each age. The data for 1991 is a set of points lying on a wiggly line $N(t)$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N(t)\,dt. \qquad (7.1)$$



Global comments

I like how this section starts with a very easy example and then goes to a more complicated one and then gives a more difficult practice problem.

7

Lumping

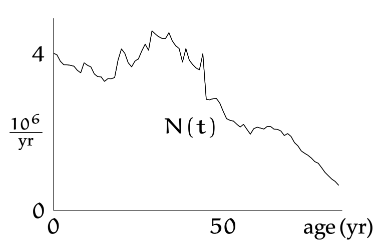
The symmetry chapter (Section 3.1) introduced the principle of invariance: 'When there is change, look for what does not change.' However, when you cannot find any useful but unchanging quantity, you have to make one. As Jean-Luc Picard often says, 'Make it so.'

7.1 Estimating populations: How many babies?

The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau [33] publishes the number of people at each age. The data for 1991 is a set of points lying on a wiggly line $N(t)$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N(t)\,dt. \tag{7.1}$$



great file name

this seems like it should go with divide and conquer at the beginning of the semester

Read Sections 7.1 and 7.2 for Sunday's memo. The final (fifth) page has a fun problem to think about, which we'll look at on Monday. Have a nice weekend.

I am still not convinced or do I understand the whole concept of lumping?

I just got deja vu here...wasn't this line used in a previous reading?

I think so... and he's used it in class a few times.

Yah...he uses it a LOT in class. Sometimes its use is rather confusing, but it makes sense here.

Might as well call it, "As Sanjoy Mahajan says..."

Yeah this has been used a ton of times it's kind of a theme of the class

This seems like a very unspecific question and thus requires more difficulty in approximating...it would help to ask maybe how many babies are born in __ time, or how many people between 0-2 are in the US?

Um, I think you're reading too much into this. This is just the title, and a little subtitle to go alongside the main topic of estimating populations. The actual question will be addressed in the text.

That's the very first step that's taken in the paragraph: defining the parameters of the problem for estimation.

I've been asked this exact question as part of consulting case interviews before, I really like seeing examples like this because they are so applicable to what I want to do in the future!

True, but wasn't this method a little too simple and also very sensitive to the guess you made based on everyone dying at 75 years?

I'd say that the method is not at all too simple...esp. for a problem you encounter in interview where you need to thing quickly. Having a simple (but accurate) method is extremely valuable.

7
Lumping

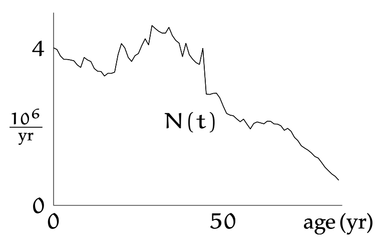
The symmetry chapter (Section 3.1) introduced the principle of invariance: 'When there is change, look for what does not change.' However, when you cannot find any useful but unchanging quantity, you have to make one. As Jean-Luc Picard often says, 'Make it so.'

7.1 Estimating populations: How many babies?

The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau [33] publishes the number of people at each age. The data for 1991 is a set of points lying on a wiggly line $N(t)$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N(t)\,dt. \tag{7.1}$$



This is a good example problem.

the first example OF LUMPING is... you have yet to define what you are actually talking about in this chapter

> This is a good point. I got to the end of the baby section and was thinking "So lumping is making integrals boxes?" I'd like to have a bit of an explanation of what is going to happen with lumping so I know what to look for in the coming sections.
>
> I felt the same way. I read up to page 3 and realized I STILL didn't know what lumping was. I had an idea, but it was never specifically stated.
>
> I think the introduction should define it immediately.

This just sounds very odd.

I would change "For definiteness" to "lets"

> I agree this sounds strange.

this phrase is a little odd to me. perhaps try: "in order to do this, we must first define a range of ages for which the child is considered a baby"

> Or you could just say, "For precision, ..."
>
> > I like both "for precision" and "lets". Alternately, we can word this as "Let's define a child as a baby..."

7 Lumping

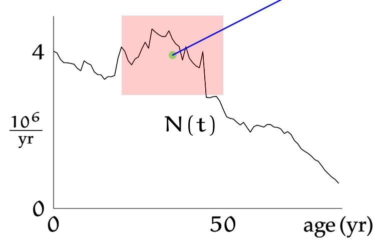
The symmetry chapter (Section 3.1) introduced the principle of invariance: 'When there is change, look for what does not change.' However, when you cannot find any useful but unchanging quantity, you have to make one. As Jean-Luc Picard often says, 'Make it so.'

7.1 Estimating populations: How many babies?

The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau [33] publishes the number of people at each age. The data for 1991 is a set of points lying on a wiggly line $N(t)$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N(t)\,dt. \tag{7.1}$$



I like seeing these problems that are from all the way back when we did the diagnostic. It gives me a good sense of the progression of the class and makes me see how much I've learned when comparing to my old methods.

I agree. There were many problems on the diagnostic that had me absolutely stumped as to how to even start them.

But as we encounter tools that allow us to answer them, it's like a 'light bulb' moment.

I'm still looking forward to the golf ball problem!

Haha yeah, me too! I totally did this problem differently. I used the number of people in the US and dived and conquered to get to the number of babies (guessed how many were able to have children, and used the average of 2 children per couple).

Agree, I remember looking at the irregular spring Hooke's ;aw problem at the beginning of this class and being so confused. It was awesome being able to actually arrive at the answer this time around, you really feel like you're making progress.

Very relevant topic this time of year

Yep, I hope everyone has turned in their forms!

haha, this reminds me I need to fill in my after I finish my two tests this week.

This paragraph could use a better introduction, explaining what's going to be done with the data. I expected a simpler method to be explained first.

Is this the baby boom?

Well I guess the data is from 1991...

Quoted from next comment (posted @10:37):

For this data in particular, the sharp rise and steady-high numbers would represent the Baby Boomer generation. (the sharp jump just to the left of this box)

The sharp decline just shy of 50 would most likely be due to the number of lives lost during WWII and the Vietnam War.

Very interesting... You seem to be right

7
Lumping

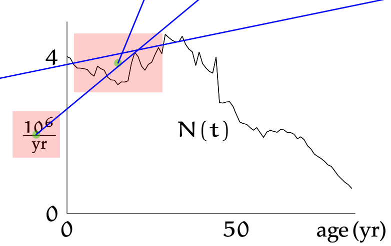
The symmetry chapter (Section 3.1) introduced the principle of invariance: 'When there is change, look for what does not change.' However, when you cannot find any useful but unchanging quantity, you have to make one. As Jean-Luc Picard often says, 'Make it so.'

7.1 Estimating populations: How many babies?

The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau [33] publishes the number of people at each age. The data for 1991 is a set of points lying on a wiggly line $N(t)$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N(t)\,dt. \qquad (7.1)$$



Why is there this dip and then rise in the number of people?

There are different numbers of people born every year, so it's possible for more people to be born in year x and then less people born in year x+k, which would account for the dip and then rise.

For this data in particular, the sharp rise and steady-high numbers would represent the Baby Boomer generation.

The sharp decline just shy of 50 would most likely be due to the number of lives lost during WWII and the Vietnam War.

It would be nice to see a date on when this data was taken - which census? That could explain the dim in people.

He does, it is the data for 1991 (so the 1990 Census)

true, but a title on the figure would also be nice.

When using data like a census? How far back is reasonable? 1991 was a decade ago, so using that data kind of seems unreasonable to me.

Guess he just couldn't find the figures from 2000. 1991 isn't so bad though.

These units kind of confuse me. I understand that if we take the integral we'll get the correct units, but does the census give the data as a rate. It seems more logical that they would just have the ages of people and a simple summation would do.

I agree, that is a bit weird. It makes the true nature of the graph look a little strange. Wouldn't the integral be different too if it is measured as a ratio?

Yeah I don't understand it either. The census takes data of how many people are currently in the country.

Is this necessary? We can see the chart on the side. Maybe reference that instead?

yeah i agree, this sounds a bit weird

I think addressing the graph itself would definitely be better. This is kind of out of context given that the rest of the text is professionally written and aimed at a fairly high level.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

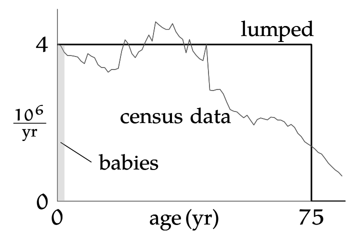
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \qquad (7.2)$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \qquad (7.3)$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

When computing the area, obviously it works out when multiplying. Intuitively, we're really saying "number of people per age group" where age group is a year.

It's interesting, though, I don't think I would have noticed it unless this question pointed it out!

Does it have to do with looking at a particular year? It works out dimensionally when integrating (people/year * dt= people)

I think it's because the bin widths are small enough (1 day) that this is considered a smooth curve (and therefore why we integrate, instead of taking a discrete sum, above).

People per year is more useful than people per day, or just 'People' if this were presented as discrete data with each bin width equal to one day.

Also, "people" would be somewhat constant, as theoretically births equals deaths, and so we would get no useful information out of that graph

Mathematically, it's because we're integrating over the time period 0 to 2 years. We defined a baby to be a person between 0 and 2 years old, and since we're integrating over this time period to get a unit which is number of people, the vertical axis must be units of people per year.

Perhaps, having the vertical axis in units of people/year will specify the rate of change of people in a particular age group. This could be more useful for interpretation since it will allow us to see changes in the rate, which could help with anticipating future effects before actual changes happen.

You might want to specify "the vertical axis" in this sentence. For the longest time I was really confused as to why the horizontal axis supposedly had units of 1/T when it was labeled with units of time...

I got caught up in this same thing - based on the previous sentence it could even say "why does this axis..."

You might want to specify "the vertical axis" in this sentence. For the longest time I was really confused as to why the horizontal axis supposedly had units of 1/T when it was labeled with units of time...

That sounds redundant. If anything, just change "the axis" to "this axis"

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

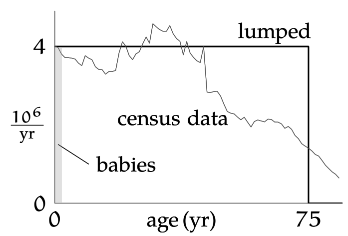
▶ *What are the height and width of this rectangle?*

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

▶ *Why did the life expectancy drop from 80 to 75 years?*

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

Yeah, I was actually thinking to myself while reading the above method that it was not practical, especially not in terms of approximations

Also, i'd specify what method is used. this is like the very precise standard method so having some sort of adjective, "this ___ method" would help

Yea I agree with that. When I originally approached the problem without reading, I thought that I could just do that and get the answer by subtracting they different y axis numbers each year. This is too simplistic.

You might want to explain what you mean by "desert island" and "back-of-the-envelope". I know what you mean coming from this class, but if other classes at other schools are going to be using this, then they might wonder why you're talking about desert islands.

I wonder why this would bother you if you were on a desert island to begin with. You'd think you'd rather spend time trying to get off the island than figuring out how many babies there are.

Got a lot of time on your hands I guess.

I think we've used these terms so many times throughout this class that mo matter where you come from if you've been paying attention they wouldn't bother you at this point.

This is a good explanation of the flaws of the brute force method.

This paragraph is awesome. It's one of those "this is why you're taking this course" paragraphs.

what does it mean to be integrated numerically

It just refers to finding the numerical solution of a particular curve

but we can simplify the integration using a similar, easy form

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

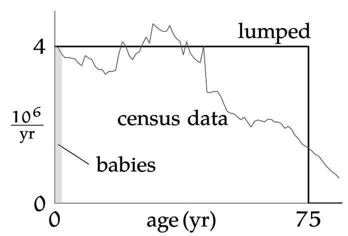
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

With today's computational tools, this isn't much of a problem–unless you're also referring to the "you're on a desert island" problem.

also, the idea of an exact numerical integration goes against the very idea of this class.

Numerical integration isn't really compatible with backs of envelopes.

I disagree. You can make a rough estimate by treating the trends in this graph as 3 lines...or even one plane and one slope

A rough approximation to the curve doesn't count as 'numerical integration', at least not in the sense he's using the term. http://en.wikipedia.org/wiki/Numerical_integ

well, wether or not it is much of a problem, it provides little insight and has minimal transfer value as stated, it's data specific, where it is about generality here.

I'm not sure how computing the number of babies can be translated to other problems, though...

agreed. I also think that integration in general, unless there is some sort of symmetry involved, isn't ideal when trying to approximate things.

why wouldn't you want to use data specific to this problem? I understand you're trying to teach something here, but this doesn't make sense to me.

I think this connects to the rest of his sentence about generality. When you are doing a integral, you want something that can be used for any problem not specifically a problem about babies. Also the integral may be significantly different in another time period. The 2010 census may yield data that will change the result and graph completely.

The focus of the class, and the methods we've been exploring, seems to be on portability and ease of reproducing methods of estimation. Data-specific solutions are the exact opposite of that and don't really teach you anything for your next encounter with a problem.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

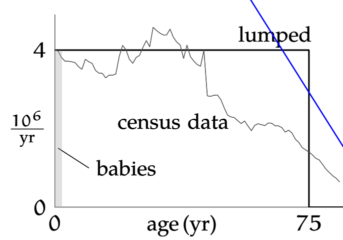
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

How can this have little insight when we compare our estimation this calculation for validity?

I believe this is mainly a reference to the point about generality.

General solutions (usually) provide more insight into the context surrounding the problem compared to exact solutions relevant only to specific data.

My problem with this statement is that integration is a very general technique for solving problems, and while this particular integral may not be useful for any other problems, neither will Sanjoy's lumping calculation, because both are particular to this problem. However both techniques can be widely applied to a number of problems with good results...

> But see, that's not true. Integration is not a general technique, not in the slightest. When you are done with the problem, all you have learned is how to integrate that equation – which is great for solving the problem, but not useful for understanding the limits and framework of the quantities. The lumping technique can be applied to any integration, as well as various other uses I'm sure we'll cover as we explore the unit. They're just different classes of problem-solving.

Huh? If you're solving a problem with mathematics, shouldn't the math be specific to the problem?

I think he's saying that you want a way to solve the problem that doesn't matter on the specific data in the problem. You want to find a mathematical technique that will give you the solution of any data.

> I agree with the second post, but I also think this sentence doesn't really help to illustrate his point. Integration is a very general procedure, and one that we could apply to this problem particularly by integrating the Census data.I think it would be more worthwhile to make a different case for why we might want to use a lumping technique instead of straight up numerical integration, since both are valid ways of approaching the problem.

>> I think what he's saying here is that in this case we're doing a "numerical integration" for this very specific curve. And we should be more interested in a general solution we can apply to more than just this exact situation. so he approximates the curve as one rectangle

the rectangle inherently has a large "transfer value" then?

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

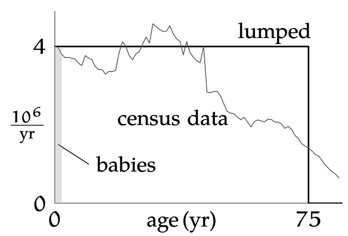
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \qquad (7.2)$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \qquad (7.3)$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

How does the estimation we did using lumping have transfer value? (Not that it doesn't, but I want to know what exactly is transferable to better understand why this was brought up as the third reason)

I think there is an easier way to go about the problem. It is common knowledge that the US population is roughly 300 million, the average household is about 3 people, distribute parents in each household by age distribution, and estimate how likely a couple in each age group will have a child that year...

answered my previous comment. I got a bit ahead

A rectangle? why specifically a rectangle? wouldn't it be more accurate as some sort of parabola curve?

Don't we lose the strong skew expressed by the data?

more accurate, yes. but a rectangle works well & is much easier to do the math for.

doesn't it kinda look like it should be a triangle instead- did you choose rectangle because it is so easy?

But this doesn't deal with the desert island problem either, although it helps with the integration.

Agreed, this solves the issue of not having integration tools on a "desert island", but you definitely wouldn't have the Census data and therefore any integral to try to approximate.

But it does help if we at least have a rough guess about the total US population and life expectancy (as he shows below). Then you can "integrate" your own estimated curve. I agree though that the wording in this section does make it sound like we're working directly off of the census data and not on our own estimates.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

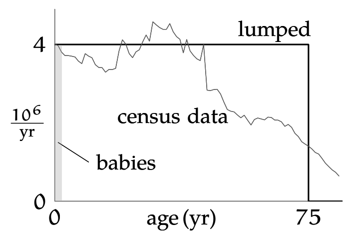
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

This is definitely an approximation method that we've all learned about from calculus

This approximation makes the assumption that there are equal numbers of people of each age though, correct?

Nope, it doesn't because its still using this ratio form...if I'm not mistaken.

I suppose if you're only looking at one section, you can choose the average of that section only as opposed to the average of the entire curve.

Right, the more sections that you lump, the more accurate your approximation will become. This is cool because this is exactly how we learned to do integrals in high school, taking the limit of the number of lumpings on a curves area.

Yea, I like how each of these methods we're learning all relate back to something that we learned in the beginning of our classes. I remember that the first thing we learned in Chemistry before the equations was solving using dimensional analysis and then this brings us back to calculus

It is a very simple, useful method.

I would assume you would make the initial value of the curve the height of the rectangle. right?

why are we using life expectancy rather than just looking at the region between 0-2 years?

We're using what we feel are two "known facts" to estimate something we don't know as well: the height of the rectangle.

The two knowns are: total population (area of the rectangle) and life expectancy (width of the rectangle)

Then we divide area by width to obtain height.

Now that we have height, we multiply it by 2 years to get the area spanned by babies, giving us # of babies.

To answer your question more, we ARE looking at the region between 0-2 years. But we are choosing to do so by first approximating the graph as a rectangle. To do that, we needed to do what I just said above.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

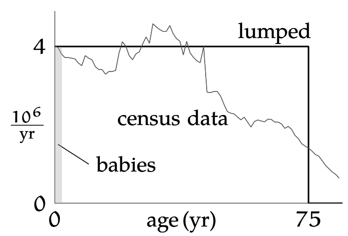
▶ *What are the height and width of this rectangle?*

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

▶ *Why did the life expectancy drop from 80 to 75 years?*

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

Since we use 75 anyway for the calculation later on, can't we just say that the life expectancy is roughly 75 years to avoid confusion?

Well, if he's trying to point out that you can fudge your numbers a bit, starting out with 75 doesn't provide a starting point for that.

thanks

so it seems like this works ok for the early ages, but if we were trying to estimate the number of elderly it might be way off.

Can we assume this? The curve slopes down pretty far by the time we reach 80

As mentioned above though that was due in large part to WWII, so today, the curve would not fall so abruptly at 50, it would fall at 60, and in the theoretical future, would fall at 80

I think we're basically figuring that 80 is right around the center of the curve (if we take life span as a sort of bell-like curve).

I would actually estimate it to be a little less.

I don't have a good sense of population growth in the US, so how long will we be able to use 300 million for future back-of-the-envelope calculations?

Should you consider how many babies a typical person has?

I think you could use that information as another method to get to a (hopefully) similar solution

Yes you could do it this way, but you'd have to think about what size the baby-having subsection of the population would be, how many babies each has on average, etc. These probably aren't as easy numbers to estimate as average life expectancy or US population.

I agree and the point of this exercise is to show us how to effectively lump things together, such as the graph.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

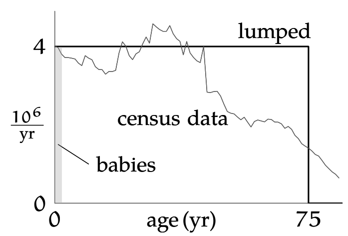
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3\cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4\cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0\ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4\cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8\cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980\cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

is there a way other than this to estimate the height? I still don't fully understand what this unit represents.

The height could be seen as an estimate of the number of people at a given age. What's important is that by integrating the graph, we get a total population. Since the graph is represented as a rectangle, then a simplified assumption of height is area/x-axis.

It may be helpful to include this in the description of height and maybe even to describe why finding the height is important before we do it. It seems like a lot of people get confused when they aren't explicitly told why we're doing something .

I thought we were using 80 years?

I think he changed it when he realized that 80 doesn't divide 300 evenly.

It might be a good idea to check out case and point- it's a book about case interviewing that approaches this exact approximation in a slightly different manner, it might be useful to see an alternate- perhaps even less complicated way to do it.

It took me a while to understand what this was asking- I don't really think it needs to be addressed

I think it definitely does need to be addressed. In one sentence, he says the life expectancy is 80 years, but then he turns around and uses 75 years in the calculation. I think some explanation is necessary, otherwise I know I would be confused.

Yeah, this is very confusing, my understanding is it dropped to 75 just so that the math would work out nicer.

I think the wording of the question is a little awkward. Maybe "Why did we use 75 instead of 80" would work better?

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

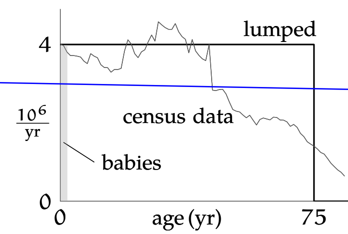
This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

▶ *What are the height and width of this rectangle?*

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

▶ *Why did the life expectancy drop from 80 to 75 years?*



Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.

Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

I feel like this should be said earlier - "instead of 80 years, we will use 75 years because...".

Is this because we take out the 2 years for the babies? Or is it just to make our calculations easier.

I agree with the first person it isnt really clear what is going on here

I think its to make the resulting numbers easier to deal with. Also, It can't really be mentioned earlier, because until we know the population and see the division, we don't really know that 75 will produce a better number than 80.

Taking 2 years out for the babies would only make sense at all if that 300 million excluded children younger than 2 years old. This is just to make calculations easier.

We regards to the original comment: It may be difficult to move this explanation any earlier because he needs to set up the height calculation in order to explain why 75 years is more convenient.

I don't agree; it changed, and was immediately explained in the next line. I think that the small change is actually being used to teach another lesson here also (hence, the subsection for it).

we should have said this in the paragraph above, so that people dont get confused when they see the 75 as I was.

agreed. i didn't get confused but i didn't notice the change either. so it was more of a double-take

I agree as well - seems little reason to explain 80 if we're just going to change it around later.

or you could've used 81

It's a little hard to see that the babies rectangle is grayed in. Maybe make it darker?

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

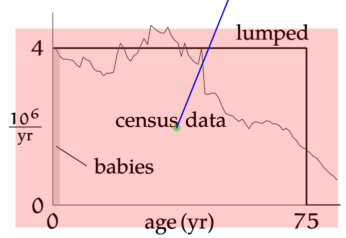
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

Note that you're plotting a calculation based on 2008 population on top of 1991 census data (when the population was 258 million – 14% less!).

Other than being bad practice, it makes it look like we chose an average that was too high for the whole population (it's pretty clear that the population doesn't integrate to be the area of the box).

Wow, good point. I was wondering why the box seemed to be so much larger than the curve it was supposed to approximate.

It's partly that, and could also be a couple of other things related to our life expectancy approximation.

First, there is some error in our estimate.

Second, using life expectancy as an estimator for population distribution ignores time in a couple of ways. Time has an influence in both changing population growth rates and changing life expectancies. Changing growth rates means more babies each year than the last.

An increasing life expectancy – and this is where it gets tricky – would mean that the true average life expectancy of the population (still measured from birth, but at the time of each person's birth, not in 1991 for everybody) would be lower than this estimate 75 or 80 years.

I don't, however, think these are significant factors, and I do think our analysis is a good approximation. Mostly I'm just bothered by the plotting of 2008 data on a 1991 graph, especially because (as in the Exxon example) a truer representation would help Sanjoy make his point even more convincingly.

You're right, the population in 1991 was less than in 2008. However, I think the reason he chooses to use the 2008 value of 300 million is because it's an approximation. If you actually used the 258 million value from 1991, you would probably round that up to 300 million to make the calculations easier anyway. Yes, it's 14% less, but I think in the approximation world, if it's in the same factor of 10, then it doesn't really matter.

A glance at the area covered by each shape makes it seem as though the rectangle should be a little smaller.

I had the same thought, it seems to work better with the same height but a width ending at age 70 or so instead.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

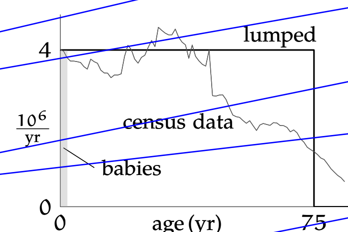
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3\cdot 10^8}{75\,\text{yr}}. \quad (7.2)$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4\cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0\ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4\cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8\cdot 10^6. \quad (7.3)$$

The Census Bureau's figure is very close: $7.980\cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

"and it might even cancel the lumping error": i feel like that stuff like this happens a lot in this class. and it's always a little too serendipitous for my comfort.

I somewhat agree...why we change the number makes sense, only I probably would have chosen 70 or 75 to start with, not 80. Numbers always magically working out make me think that the answer was already known...and now we just tweak things slightly so they fit

Although this kind of stuff happens a lot, there are also a fair number of examples where he says "the actual answer is 3 times greater" or "this approximation is off by a factor of 10 because.." which leads me to believe he's not using the answer for the approximating.

This is pretty true, but when we introduce error I'm still uncomfortable with the idea that "this error will hopefully cancel out the other error" without any explanation of how cutting 1/16 off of one value fixes the lumping error.

I agree that it's not ideal to try to make up for previous errors by making more error, but sometimes when we make a guess and we know we're low, but don't know by how much (as in we know we're low, but if we try to guess a higher value we might go over, etc), then it might make sense to try to overestimate something else.

Can you just make that generalization?

I feel like this sentence doesn't really add in terms of clarification.

I don't see how using 75 years "makes" the height 4*10^6 yr^-1.

Is their a logical way to get this on the y axis or is it just eyballing and guessing?

I think this problem is a very good example of "making it so".

I don't think making the box width 75 was necessary for this calculation

I agree. we should have chosen a box that fit the 0-2 year demographic better, and used that instead.

Yep, that seems unnecessary for htis demographic.

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

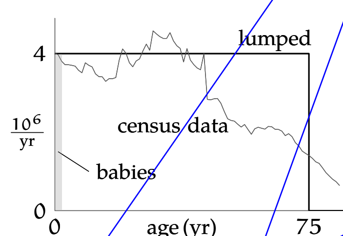
▶ *What are the height and width of this rectangle?*

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \qquad (7.2)$$

▶ *Why did the life expectancy drop from 80 to 75 years?*

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \qquad (7.3)$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

I don't know if height is the right descriptor... (especially because infancy isn't labeled 'width', nor do I think it should be). I don't personally have a better suggestion, but this seemed off.

Maybe plotting the infancy 'slice' of the rectangle on the graph would make what this calculation is doing more clear.

Or maybe just calling it 'infancy width' or something.

that is really amazing how close it is, even looking at the curve there is a lot of error and it seems like the rectangular approximation would overestimate most ages by a lot, but it works well for babies. if we were trying to estimate the number of people between 50-60 would we use the same approximation or how would it change?

Why do these errors tend to cancel out so often?

It's important to understand what each error does to the problem. By lumping, we are counting the later years more then they should be but when we drop the age from 80 to 75, we are eliminating from the later years. It's important to know in which direction each error affects the answer.

This is impressive. I can think of a few other ways to estimate this number, but none could come nearly that close.

Does reducing the age range from 80 to 75 tend to add people just to the end of the distribution or is it more evenly distributed?

No matter how many times it happens, I am always amazed at the calculations ending up so close to the actual values, and I usually expect it each time at this point too.

is there a way to tell if the cancel each other out or multiply to make it worse?

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

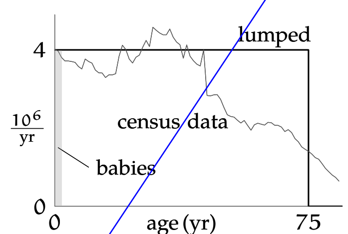
▶ *What are the height and width of this rectangle?*

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

▶ *Why did the life expectancy drop from 80 to 75 years?*

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

But couldn't the fudging have just as easily made the error greater? It feels like you just got lucky that the life expectancy you chose happened to cancel nicely with the lumping.

I think the decision to fudge the life expectancy downward was not random, but done for 2 reasons. One, to make division easier, and two, to cancel the lumping, which overestimated the population to 300 million.

There was some logic to picking a lower age and minimizing error. Looking at the distribution of the curve, it would be a mistake to use 85 instead, since you'd be making your approximation worse. Even without the plot, if you know that the average lifespan is 80, you want to pick something lower than that since the distribution is higher at the younger ages.

So basically lumping means making broad but reasonable generalizations?

This seems like one of the most basic forms of estimating, like back to the beginning. I don't quite see how its "lossy" though, since we aren't necessarily throwing things out, we just ignore them like we did at the outset of the class and with random constants always.

I saw it as "lossy" since we have ignored all the details presented by the U.S. Census. By showing the graph, we know that N(t) is some complex function. Thus, we ignore the complexities of the problem to form an estimation.

The errors canceling out almost perfectly seems like serendipity to me.

It's not entirely random; it's about finding the most important pieces of information and using them to find your solution in an easy way.

Yeah, I see lumping as using common sense approximations–what we've been doing at the beginning of the course. For example, estimating volume of a sphere as a cube for easy math. Except now, that sphere is some complex shape, and we have no choice but to approximate.

it is safe to assume that this rarely happens when applied to most problems though, correct?

Problem 7.1 Dimensions of the vertical axis
Why is the vertical axis labeled in units of people per year rather than in units of people? Equivalently, why does the axis have dimensions of T^{-1}?

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it—lump the curve into one rectangle.

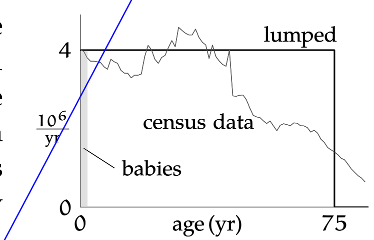
What are the height and width of this rectangle?

The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population—conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3 \cdot 10^8}{75\,\text{yr}}. \tag{7.2}$$

Why did the life expectancy drop from 80 to 75 years?

Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4 \cdot 10^6\,\text{yr}^{-1}$.



Integrating the population curve over the range $t = 0 \ldots 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim \underbrace{4 \cdot 10^6\,\text{yr}^{-1}}_{\text{height}} \times \underbrace{2\,\text{yr}}_{\text{infancy}} = 8 \cdot 10^6. \tag{7.3}$$

The Census Bureau's figure is very close: $7.980 \cdot 10^6$. The error from lumping canceled the error from fudging the life expectancy to 75 years!

I like starting with this simple example... it makes the concept much easier to grasp without having to fight through a bunch of physics.

I agree, it's really given me a good idea of what lumping is all about too.

I agree. It was simple, clear, easy to follow, and helpful in understanding lumping.

Problem 7.2 Landfill volume
Estimate the US landfill volume used annually by disposable diapers.

Problem 7.3 Industry revenues
Estimate the annual revenue of the US diaper industry.

7.2 Bending of light

The fundamental principle of lumping is to replace a complex, changing process by a simpler, constant process. Let's apply the method far beyond mundane concerns about the number of babies, using lumping to revisit the bending of starlight by the sun. Using dimensional analysis and educated guessing (Section 5.4), we concluded that the bending angle is roughly GM/Rc^2, where R is the distance of closest approach (here, the radius of the sun), and M is the mass of the sun. Lumping provides a physical explanation for the same result; it thereby helps us make physical predictions (??).

So once again imagine a beam (or photon) of light that leaves a distant star. In its travels, it grazes the surface of the sun and reaches our eye. To estimate the deflection angle by using lumping, first identify the changing process. Here, the changing process is the angle that the light beam makes relative to its original, undeflected path; equivalently, the photon falls toward the sun as would a rock. This deflection angle increases slowly after the photon leaves the star, increasing most rapidly near the sun. Because the angle and position are changing, which means the downward gravitational force is changing, calculating the final deflection angle requires setting up and evaluating an integral – while carefully checking items in the integral such as the number of cosines and secants.

In contrast, the lumping approximation is much simpler. It pretends that the deflection is zero until the beam gets near the sun. Gravity, in this approximation, operates only near the sun. While the photon is near the sun, the approximation pretends further that the downward acceleration (toward the center of the sun) is a constant, rather than varying rapidly with position. Finally, once the beam is no longer near the sun, the deflection does not change.

Babies probably use a diaper a day, right? Or more?

Definitely more than 1/day.

Many more. I would guess about 4.

Being off by a factor of 4 in this case would be considered good by our standards, right?

These are really good problems, they build off the answer we just got so you have an easy place to start your estimation.

yeah i agree..also i've heard that the question about estimating the US diaper industry is often asked as an estimation problem during interviews!

After having solved the earlier problem, this problem seems relatively easy to solve- it should just involve mutiplication by a few factors.

This topic came back.

This should have been said much earlier

Agreed, I feel like in most sections, we dive right into some problem or example, and we don't spend any time describing the concept that we're trying to learn. This short explanation should have been put before the baby example.

Agreed.

Eh, I kind of like it here.. the baby estimating example is a nice way to flow into this so when he talks about lumping you have already been introduced to it

I also agree with the last opinion. It's nice to get a feel for the concept first with an example before defining it. Just like when you teach someone a card game: you jump right in and then teach the detailed rules as you go along.

I agree, this would have been a good sentence for the introduction - before we launched into any examples of lumping.

It's nice to see that this has been stated here though - I too think it should be said earlier, but it should not be removed. Repitition = Importance!

yeah i agree...I completely understood the first example but initially I didn't quite understand what lumping was exactly.

Problem 7.2 Landfill volume
Estimate the US landfill volume used annually by disposable diapers.

Problem 7.3 Industry revenues
Estimate the annual revenue of the US diaper industry.

7.2 Bending of light

The fundamental principle of lumping is to replace a complex, changing process by a simpler, constant process. Let's apply the method far beyond mundane concerns about the number of babies, using lumping to revisit the bending of starlight by the sun. Using dimensional analysis and educated guessing (Section 5.4), we concluded that the bending angle is roughly GM/Rc^2, where R is the distance of closest approach (here, the radius of the sun), and M is the mass of the sun. Lumping provides a physical explanation for the same result; it thereby helps us make physical predictions (??).

So once again imagine a beam (or photon) of light that leaves a distant star. In its travels, it grazes the surface of the sun and reaches our eye. To estimate the deflection angle by using lumping, first identify the changing process. Here, the changing process is the angle that the light beam makes relative to its original, undeflected path; equivalently, the photon falls toward the sun as would a rock. This deflection angle increases slowly after the photon leaves the star, increasing most rapidly near the sun. Because the angle and position are changing, which means the downward gravitational force is changing, calculating the final deflection angle requires setting up and evaluating an integral – while carefully checking items in the integral such as the number of cosines and secants.

In contrast, the lumping approximation is much simpler. It pretends that the deflection is zero until the beam gets near the sun. Gravity, in this approximation, operates only near the sun. While the photon is near the sun, the approximation pretends further that the downward acceleration (toward the center of the sun) is a constant, rather than varying rapidly with position. Finally, once the beam is no longer near the sun, the deflection does not change.

This is the intro to lumping, it shouldn't be in the second example.

It seems like this is the main principle for a lot of the things we are learning in this class.

This was shown earlier, but this really solidified it. It might be better to introduce it somewhere earlier

so from these examples seem to suggest that lumping is like using the riemann sum approach as opposed to the integral. They are both sums, although the integral approach has more accuracy, its easy to make bars and add them up as you do in a sum.

I like that analogy, that really helps me visualize what how lumping estimates things.

I like that you acknowledge that your target audience is nerds.

Mundane could just refer to the earthly nature of babies, not that they aren't important.

Either way, great wording in this paragraph.

I also like that this parallels and example of bending light from an earlier reading

will we be required to memorize this for the final? or will it be given

I highly, highly doubt he would ever require us to memorize something like that.

I am wondering the same thing. From what I've seen on the psets we'll have to memorize it. Unless we can use a cheat sheet.

It might be nice if you put the same diagram as you had in the earlier chapter repeated here since we read about that a while ago

I agree, it's always nice for these types of problems to get a picture. I think visualizing the problem in the correct manner is a huge part of being able to approximate.

well assuming this was in a book of some sort, you could always return to the section, since he did refer to it.

Problem 7.2 Landfill volume
Estimate the US landfill volume used annually by disposable diapers.

Problem 7.3 Industry revenues
Estimate the annual revenue of the US diaper industry.

7.2 Bending of light

The fundamental principle of lumping is to replace a complex, changing process by a simpler, constant process. Let's apply the method far beyond mundane concerns about the number of babies, using lumping to revisit the bending of starlight by the sun. Using dimensional analysis and educated guessing (Section 5.4), we concluded that the bending angle is roughly GM/Rc^2, where R is the distance of closest approach (here, the radius of the sun), and M is the mass of the sun. Lumping provides a physical explanation for the same result; it thereby helps us make physical predictions (??).

So once again imagine a beam (or photon) of light that leaves a distant star. In its travels, it grazes the surface of the sun and reaches our eye. To estimate the deflection angle by using lumping, first identify the changing process. Here, the changing process is the angle that the light beam makes relative to its original, undeflected path; equivalently, the photon falls toward the sun as would a rock. This deflection angle increases slowly after the photon leaves the star, increasing most rapidly near the sun. Because the angle and position are changing, which means the downward gravitational force is changing, calculating the final deflection angle requires setting up and evaluating an integral – while carefully checking items in the integral such as the number of cosines and secants.

In contrast, the lumping approximation is much simpler. It pretends that the deflection is zero until the beam gets near the sun. Gravity, in this approximation, operates only near the sun. While the photon is near the sun, the approximation pretends further that the downward acceleration (toward the center of the sun) is a constant, rather than varying rapidly with position. Finally, once the beam is no longer near the sun, the deflection does not change.

I'm not quite sure how it's a "physical" explanation. I don't really see the physical part...

I think it means physical like making the population into a box like the babies problem. Population curves don't really mean much at first glance in terms of the birth rate, but we know rectangles and have a good intuitive grasp for them.

Isn't lumping still a kind of an abstraction? Physical makes me think we are talking about something more literal.

Definitely have no idea how this is possible.

are you saying that we are deriving the same result again or using it in other ways (i.e. lumping) to show that that is correct?

What are these doing here?

Probably a missing reference (citation, or possibly figure). This is how my missing references show up in latex documents.

I don't really understand what "changing process" means here.

I see what you're saying here, but I don't think the sentence is very clear.

Agreed, perhaps a diagram would show this better. I still have trouble thinking about the starlight grazing the sun with and without the pull of the sun. I guess it's because if I imagine light that just grazes the sun on its way to me, but consider the pull of the sun, then that light no longer just grazes the sun, instead (I think) it gets pulled in more so I wouldn't see it.

I see what you're saying here, but I don't think the sentence is very clear.

You usually break up your blocks of text with diagrams or figures. Your writing is easy to follow, but when I read on a computer screen large chunks of text are hard to process.

agreed. i'd like a figure somewhere on the page. you have two on the next page, maybe one can be brought up?

I feel like a diagram here would be really helpful. I'm a little lost as to what variables we're dealing with here.

Problem 7.2 Landfill volume
Estimate the US landfill volume used annually by disposable diapers.

Problem 7.3 Industry revenues
Estimate the annual revenue of the US diaper industry.

7.2 Bending of light

The fundamental principle of lumping is to replace a complex, changing process by a simpler, constant process. Let's apply the method far beyond mundane concerns about the number of babies, using lumping to revisit the bending of starlight by the sun. Using dimensional analysis and educated guessing (Section 5.4), we concluded that the bending angle is roughly GM/Rc^2, where R is the distance of closest approach (here, the radius of the sun), and M is the mass of the sun. Lumping provides a physical explanation for the same result; it thereby helps us make physical predictions (??).

So once again imagine a beam (or photon) of light that leaves a distant star. In its travels, it grazes the surface of the sun and reaches our eye. To estimate the deflection angle by using lumping, first identify the changing process. Here, the changing process is the angle that the light beam makes relative to its original, undeflected path; equivalently, the photon falls toward the sun as would a rock. This deflection angle increases slowly after the photon leaves the star, increasing most rapidly near the sun. Because the angle and position are changing, which means the downward gravitational force is changing, calculating the final deflection angle requires setting up and evaluating an integral – while carefully checking items in the integral such as the number of cosines and secants.

In contrast, the lumping approximation is much simpler. It pretends that the deflection is zero until the beam gets near the sun. Gravity, in this approximation, operates only near the sun. While the photon is near the sun, the approximation pretends further that the downward acceleration (toward the center of the sun) is a constant, rather than varying rapidly with position. Finally, once the beam is no longer near the sun, the deflection does not change.

i don't understand this phrasing.

Me either.

I think he means "also assumes"

This seems to me pretty intuitive, its the way we learn everything, first lumped into the basics, then gradually with more and more variables

Although this makes sense to me, it still hasn't reasoned far enough for me to know how to complete the problem.

is this different than the assumptions from last time- I am just remembering the drawing on the board or did it not matter since we were looking at dimensions?

From the diagram it makes it seem as though the force of gravity acts once, to deflect the point, and then stops acting on it. Am I misunderstanding that?

Nevermind, it's explained later.

Isn't this (using a constant) how we use gravity on earth?

I like this paragraph. It makes the similarity between how we're solving this problem and how we solved the last problem very clear - we're simply making a rectangle.

I agree...the contrast helps a lot in these building examples comparing multiple methods.

agreed, as an engineer, I really appreciated this paragraph.

I'm having trouble seeing how lumping is different from just making general approximations

It feels like the same approximations we did when we divided and conquered things like the MIT budget–there we used approximations that were lossy in the sense that we neglected or pooled things.

They definitely seem very similar.

yeah I agree..would it be possible to include an example in the text where the difference between divide&conquer and lumping is more noticeable? or are they always this similar?

Problem 7.2 Landfill volume
Estimate the US landfill volume used annually by disposable diapers.

Problem 7.3 Industry revenues
Estimate the annual revenue of the US diaper industry.

7.2 Bending of light

The fundamental principle of lumping is to replace a complex, changing process by a simpler, constant process. Let's apply the method far beyond mundane concerns about the number of babies, using lumping to revisit the bending of starlight by the sun. Using dimensional analysis and educated guessing (Section 5.4), we concluded that the bending angle is roughly GM/Rc^2, where R is the distance of closest approach (here, the radius of the sun), and M is the mass of the sun. Lumping provides a physical explanation for the same result; it thereby helps us make physical predictions (??).

So once again imagine a beam (or photon) of light that leaves a distant star. In its travels, it grazes the surface of the sun and reaches our eye. To estimate the deflection angle by using lumping, first identify the changing process. Here, the changing process is the angle that the light beam makes relative to its original, undeflected path; equivalently, the photon falls toward the sun as would a rock. This deflection angle increases slowly after the photon leaves the star, increasing most rapidly near the sun. Because the angle and position are changing, which means the downward gravitational force is changing, calculating the final deflection angle requires setting up and evaluating an integral – while carefully checking items in the integral such as the number of cosines and secants.

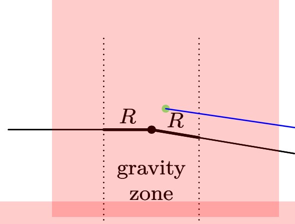
In contrast, the lumping approximation is much simpler. It pretends that the deflection is zero until the beam gets near the sun. Gravity, in this approximation, operates only near the sun. While the photon is near the sun, the approximation pretends further that the downward acceleration (toward the center of the sun) is a constant, rather than varying rapidly with position. Finally, once the beam is no longer near the sun, the deflection does not change.

would small angle approximations be considered a type of lumping? i feel like it would be, and i wanted to see if i'm understanding "lumping" correctly.

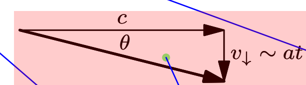
also. i think using biot numbers to determine whether or not the lumped parameter model can be applied, from 2.005, would be lumping as well?

Yeah. It's definite used in this example later. And it definitely feels like it belongs in this unit.

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



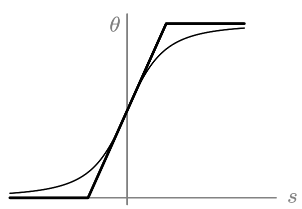
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \tag{7.4}$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



. . . the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

this graph makes it much more clear

Could you include the sun in this picture with all of the places that R should appear? I think that would make the paragraph easier to understand. I had to read it a couple of times to decide what you meant by "R."

I agree. I thought the 'dot' was supposed to be the sun at first, as if the light was bouncing off the sun like a billiard ball. But then i realized it would have to be horribly out of scale...

I agree as well. I think a full diagram of sun, R, gravity zone and where you are measuring the deflection from.

Is it always best to use numbers we have for numbers that are arbitrary. I like it, keeps things simple.

This approximation still seems really arbitrary to me, although I guess it depends on how much it affects the final solution

I agree. This seems like a random choice.

I don't think it's that random, R is a property of the planet or star we are looking at so clearly it will have an effect on our gravity zone. Also by using R we can easily do some scaling later if our results do not seem to make sense

GM/cˆ2 is also a length. I would argue that this is also important. The slower something is moving, for example, the further away 'significant' deflections would start to accumulate.

True...except I believe we are talking about photons, so they are all moving at the speed of light

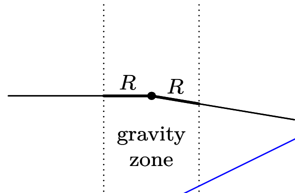
So? GM/cˆ2 being invariant in this problem doesn't make it irrelevant. Granted it is much smaller than R (since it equals R*theta and theta is small), but that's not particularly convincing, I don't think.

I like this diagram. It helps me organize my thoughts quite a bit.

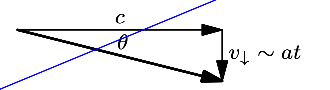
Yeah, it's interesting using a triangle where the units are all velocities instead of distances.

Yea, you can use triangles for many different units. Aerospace uses velocity triangles a lot for trajectories and anything w/ relative velocities and positions when compared to things such as the wind or other disturbances

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



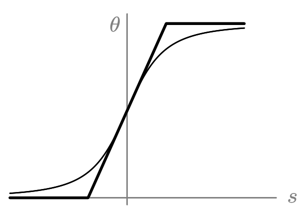
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \qquad (7.4)$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



...the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

So i guess the "lumping" comes from lumping the total angle of deflection that is experienced over time into one angle that happens instantaneously. This makes it trigonometrically easier to evaluate.

Yeah I think that's correct.

Lumping the curve into a triangle

it would I think make more sense to put the diagram down here, so people don't wonder why there is a single kink instead of a smooth curve

isn't the diagram the one above? it clearly shows a bent line.

I don't think that is the diagram above. I am also a bit confused by the wording in general.

Seems like a lot of this is just proportional reasoning. Looking forward to seeing more problems that really rely on lumping.

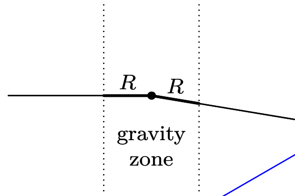
this is a little confusing

Why don't you say length 2R when you know that's closer the actual length the light travels through the zone (it can't be anything less)? I understand we're concerned with dropping unnecessary complexity, but a factor of two doesn't seem too complex to me, and it might get us closer to the actual answer?

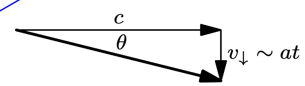
Although I'm still not quite sure on the whole lumping thing, this is a good explanation of 7.2, and a nice connection to previous things we've done. Very easy to understand.

not totally following this logic

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



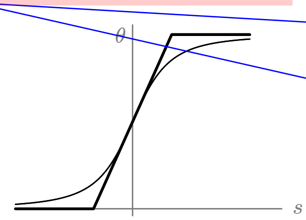
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \tag{7.4}$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



. . . the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

Where did this equation come from? For me it's a lot easier to follow these descriptions of mathematical processes if you list them as a series of equations like you do with 7.4 below and reference the equations that you used to derive the one you're listing.

this just comes from velocity = distance / time (and so time =distance /time). The distance is R, and c is the speed of light, so time = R/c.

Yea, this is just a basic velocity equation which I don't think needs a series of equations to lead up to it, but it may help to reiterate that c is speed of light just so that people know this is a velocity.

This t=R/c comes from the usual velocity equation. we have to find this because we know acceleration a=GM/(R^2) and velocity is acceleration*time. so by multiplying the 2, we can get velocity

You might want to mention that the small-angle approximation you're using is tan(O) (O). Most people will think of cosine when hearing small angle approximation.

small angle approximation: sin(x) x, cos(x) 1. It follows that tan(x) x.

it might not necessarily be intuitive that the deflection angle will be small

I agree, I wouldn't have jumped to the tan approximation by reading that.

Still good to point out. NB proves its usefulness here.

I had to re-read this section a couple of times, but it was well worth it!

This section was very clear and I always love to see a previous section revisited with a different method of approximation.

This is actually pretty cool.

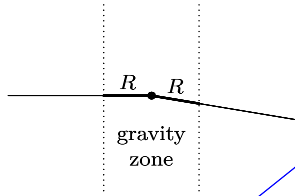
I think it's important to note that these guesses are random and so we won't always get results that are the same as in other methods.

That's a good point...seeing as we have the answer already, lumping made this problem seem way simple. However, I feel like lumping would prone me to more errors

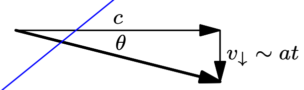
It is cool, and a good note of the robustness of these methods. Redundancy at its best. ;)

I like that with different methods, you can still reach the same conclusion in this class.

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



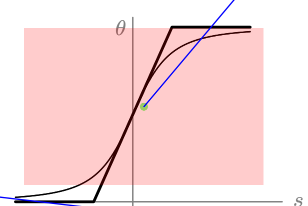
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \tag{7.4}$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



...the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

So accuracy in this case depends on what we choose for R, correct?

For the way we modeled the problem by neglecting the angle not near the sun. However how accurate can we get with a "better" R?

I'm not really sure, for me it's hard to conceptualize how to determine the accuracy in a lot of these approximations.

It also depends on the 'zone' we picked, for where gravity matters.

It was initially very unclear to me what this section meant – it took several readings to understand that this was position *along* the trajectory of the light, rather than some other measure of position.

I had to go back and reread a couple sections in this reading to appreciate what they were saying.

Looking at this diagram - this approach seems super useful for anything symmetric where the over and under estimates will cancel nicely.

This is a useful diagram. I think the addition of dashed lines at the angles would be nice. (dashed vertical lines).

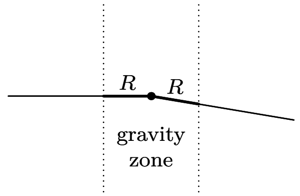
I agree, this graph definitely helps visualize the process of lumping and how well it works in approximations.

Absolutely. The darkened line is a good attempt, but some dashes would be even better.

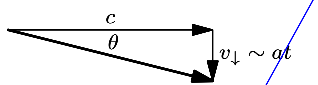
I agree–I think this diagram really shows how lumping can replace something complex (the smooth curve) with something simple (the straight curve)–this especially helps in problems like this where there are three distinct zones (outside the gravity zone, then inside, then out again), since each of these zones is in a sense separated. reminds me of divide and conquer!

This method of using piecewise straight curves to approximate quadratic looking curves is actually used a lot in my EE classes (bode plots, etc.) I guess those methods were kind of using lumping all along.

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



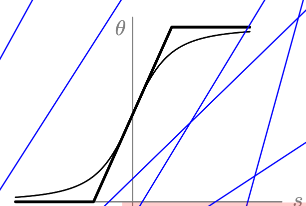
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \tag{7.4}$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



. . . the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

is gravity zone same as the near zone, beyond R?

Yes, it is the zone within distance R, where we consider gravity to have a meaningful effect on the photon.

lumping seems to just ignore units of measure instead of combining them in useful ways

Since you describe two regions of the graph, why not say the third as well? "At the end, theta remains constant, so the photon is traveling in a straight line but at its deflected angle."

I would like to see the results of this calculation involving lumping compared to the actual results of being integrated.

It almost feels like cheating. Almost.

Ha, yeah, it does. I wish I tried this more often during 8.01...

I've started using it in some of my other classes to get a feel for the answer before I really try it, and its really helpful!

weird switch in tense here

Yeah, grammatical correctness says it should be "adds."

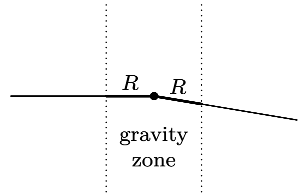
I like this explanation.

Good summation of the tools and there realtions

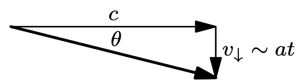
yeah, these help me to cement the concepts that I've read and see how the over-arching theme connects with particular cases and examples. I think these summations are really useful and retaining the information.

I agree, I think you should set aside sentences like this that give a good summation of the methods.

The problem then becomes one of estimating the deflection produced by gravity while the beam is in the gravity zone. But what is the near zone? As a reasonable guess, define 'near' to mean, 'Within R on either side of the location of closest approach.' The justification is that the distance of closest approach, which is R, is the only length in the problem, so the size of the near zone must be a dimensionless constant times R.



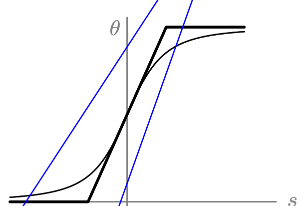
The deflection calculation is easiest at the location of closest approach, so assume that the bending happens there and only there – in other words, the beam's track has a kink rather than changing its direction smoothly. At the kink, the gravitational acceleration, which is all downward, is $a \sim GM/R^2$. The downward velocity is the acceleration multiplied by the time in the gravity zone. The zone has length $\sim R$, so the time is $t \sim R/c$. Thus the downward velocity is $\sim GM/Rc$. The deflection angle, in the small-angle approximation, is the downward velocity divided by the forward velocity. Therefore,



$$\theta \sim \frac{GM/Rc}{c} = \frac{GM}{Rc^2}. \tag{7.4}$$

The lumping argument has reproduced the result of dimensional analysis and guessing.

The true curve of θ versus position (measured as distance from the point of closest approach) varies smoothly but, as mentioned, it is difficult to calculate. Lumping replaces that smooth curve with a piecewise-straight curve that reflects the behaviors in and out of the gravity zone: no change in θ outside the gravity zone, and a constant rate of change in θ inside the gravity zone (with the rate set by the rate at the closest approach). Lumping is a complementary method to dimensional analysis. Dimensional analysis is a mathematical argument, although the guessing added a bit of physical reasoning. Lumping removes as much mathematical complexity as possible, in order to focus on the physical reasoning. Both approaches are useful!



. . . the crooked shall be made straight, and the rough places plain. (Isaiah 40:4)

Tell that to the math department when they take off points for "non-rigorous" thinking and reasoning...

that's why it's the math department! :)

this class is the art of approximation *in science and engineering*, where such approaches are not only allowed, but necessary!

Agreed! I think the math department wants to show the rigor behind such methods, so using the methods to prove the methods is a little silly.

I like this quotation, it's simple and humorous, yet speaks volumes about exactly how this class has taught us to approach problems.

I don't really like having scripture quotes in my textbooks, however.

I'd go further and say that I _really_ don't like having scripture quoted in non-fiction books not about religion

I love scripture quoted in non-fiction books about religion, I find it really funny and amusing; but then again I am not Christian.

This section reminded me of an interesting photon bending example that we see on a daily basis, although different principles are in effect. Every time the sun sets or rises, there is not yet a linear path from sun to that point on earth. You could estimate how much time there is between when light first reaches the earth and when there actually is a linear path.

Problem 7.4 Higher values of GM/Rc^2
When GM/Rc^2 is no longer small, strange things happen. Use lumping to predict what happens to light when $GM/Rc^2 \sim 1$.

Comments on page 5

It seems like this might result in the photon orbiting (or nearly orbiting) the incredibly massive object, rather than simply being deflected.

So you get a big swirl of light going around the sun?

I hope we go over this in class!

It is one of the definite items on the agenda for today.

I'll try this. Are we going to do this example in class? I think it would be interesting to go through it.

G=6e-11, C=3e8,C^2=1e17, Schwarzschild radius of earth is about 1e-2, mass of earth=6e24

GM/Rc^2=6e-28(M/R)=1

if theta gets large it forms a black hole. Was this the answer they are looking for?