Inequalities in Disease, Lifespan and Mortality

36-313, Fall 2022

6 October 2022 (Lecture 12)

Contents

1 Highlights of the previous lecture

We have just looked at how to measure differences in rates, proportions or probabilities of outcomes. The general setting was that there was some binary variable Y_i for each individual, with $Y_i = 1$ representing some kind of "success" (admission to a school, hiring, graduation, promotion, getting a loan, etc.), and $Y_i = 0$ representing the corresponding failure. The main goal was to estimate the rate or probability of success for every group defined by a variable X, perhaps adjusting for or controlling for covariates $Z, W, \ldots$.

If everyone in the sample has the same probability $p = \mathbb{P}(Y = 1)$ of success, and successes are independent (or at least uncorrelated), then the natural or direct estimator of p is the sample or empirical proportion,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The reason $\hat{p}$ is a sensible estimator is that (i) it's unbiased,

$$\mathbb{E}[\hat{p}] = \mathbb{E}[Y] = \mathbb{P}(Y = 1) = p$$

and (ii) its variance shrinks towards zero quite fast, like $1/n$:

$$\mathrm{Var}[\hat{p}] = \frac{p(1-p)}{n}$$

Thus as $n \to \infty$, $\hat{p} \to p$.

It is not necessary to have individual-level data on successes or failures; it's enough to know the number of successes N, since $N = \sum_{i=1}^{n} Y_i$, and hence $\hat{p} = N/n$.

If we are dealing with multiple groups defined by X and/or covariates, the most straightforward thing to do is calculate a separate sample proportion for each group.

As an alternative, we can lean yet more heavily on the facts that

$$\mathbb{P}(Y = 1) = \mathbb{E}[Y] \quad (1)$$

$$\mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X] \quad (2)$$

$$\mathbb{P}(Y = 1|X, Z) = \mathbb{E}[Y|X, Z] \quad (3)$$

etc., etc., to turn the problem of estimating the conditional probability of success into the problem of regressing Y on covariates. This is essential when some of the covariates are continuous numbers, and often *useful* when there are a very large number of discrete covariates. The reason it's useful in the many-discrete-covariates situation is that then the groups defined by the intersection of covariates tend to have very small sample sizes, but groups which share many but not all covariates tend to have very similar rates, and regression lets us "partially pool" information from similar groups.

Once we have (estimates of) success rates across groups, we can calculate differences (or ratios or whatever) across groups, and begin to do statistical inference to get measures of uncertainty in those differences, test whether they might be zero or exceed thresholds of concern, etc.

2 Rates of disease

All of this applies straightforwardly to looking at the rates of different medical conditions, especially diseases[1], across a population or groups within a population. The biggest wrinkle is now that "success", $Y_i = 1$, represents individual i *having* the disease or condition. This is not what we would ordinarily think of as "success", but it makes a bit more sense, perhaps, if we adopt the viewpoint of the disease.

At this point we need to distinguish between two similar but distinct kinds of rates for disease. One is called the **prevalence** of the disease, and is simply the probability that a random member of the population has the disease at a particular time (e.g., the probability that a random person in Allegheny County had Covid-19 at 12:01 am on Thursday, 7 October), or, more realistically, has the disease at any time during a particular interval (e.g., the week beginning 4 October). The natural estimator is simply the number of cases divided by the population size, in symbols $\frac{N_t}{n}$. The **incidence** of the disease, on the other hand, is the probability of *coming down with* the disease during that time interval, naturally estimated by the number of *new* cases divided by the population size[2], in symbols $\frac{\nu_t}{n}$, where I'm using ν_t (with ν "nu" for "new") to count the number of new cases between time $t-1$ and t.

Prevalence is *generally* higher than incidence; the exceptions are somewhat unusual[3].

To sum up these definitions, a disease's prevalence is (basically) the probability of *having* it, and its incidence is (basically) the probability of *getting* it. Both are probabilities or rates, so we know how to estimate them, the same way we'd estimate admissions rates or hiring rates.

At a large scale, both disease incidence and disease prevalence show a lot of variation across social groups. Plenty of diseases, for instance, are age-specific: there are twenty-year-olds with bad knees, heart disease, diabetes, arthritis and/or cancer, but all of these are conditions which become more common as we age. Some of this is due to unrepaired wear-and-tear on the body[4], some due to continued environmental exposures, etc., and some perhaps due to social inequalities. Similarly, differences in disease rates across racial and ethnic groups, across sexes, across educational levels, can all be subject to the same kind of analysis with control variables, and distinguishing between correlates and mechanisms, that we've previously examined.

2.1 Spatial variation in disease rates

One issue which is especially important in understanding disease rates is that of spatial variation. This shows up for the kind of "successes" we looked at in the last lecture as well, but it's even bigger for disease, for

[1] I am going to say "disease" to be definite, but this kind of reasoning gets applied to lots of things which are not, strictly speaking, diseases, like accidents of various kinds, homicide, etc.

[2] You might wonder why the denominator is the total population size, rather than the population size *minus* the number of old cases, $n - N_{t-1}$. This is largely a matter of convention among the epidemiologists, which I'd guess reflects the fact that for a long time they've mostly worried about conditions which *don't* affect large fractions of the population.

[3] Think about how the numerator for prevalence changes: $N_t = N_{t-1} + \nu_t - \rho_t$. (That is, the number sick at time t is the number sick at time $t-1$, *plus* the number getting sick between $t-1$ and t, *minus* the number removed between $t-1$ and t. If the amount of time that passes between measurements is short compared to how long people typically have the disease, then $\rho_t < N_{t-1}$ and so $N_t > \nu_t$. But if the progress of the disease is very rapid, or the time interval between measurements is very long, then more people might have acquired the disease *and passed through it* than have it at either time $t-1$ or time t. To use a contemporary example, Covid-19 lasts a few weeks (one way or another), so if we measure time in days or weeks, incidence does indeed have to be less than prevalence. But if we measured time in intervals of 3 months (a "quarter"), the number of people who *got* Covid-19 and either got better or died during that period might well be higher than the number who still had it. (It's true that if we look at the *cumulative* prevalence over the quarter, the number of people who had Covid-19 at any point during those months, that has to be at least as great as the number of people who came down with it during those months. But then there's a big gap between cumulative and instantaneous prevalence.) *Usually*, though, people try to collect data on a given disease fast enough that incidence will be less than prevalence.

[4] This raises the question of *why* the body doesn't always repair itself. One theory is that genetic variants which increase fitness in the young, but have bad effects later on, will tend to be favored by evolution if the problems they cause don't show up until their carriers are past their reproductive years (or, in social species, their help-out-the-kids years). Indeed, this could explain the origin of "reproductive years" in the first place, since even an organism that didn't age would always be subject to some risk of accidental death, so variations that increased fitness early but decreased it later could be evolutionary favorable in expectation, and so selected for. This is an interesting but necessarily rather speculative theory. TODO citations to Medawar 1952, Williams 1957.

two reasons. One is that lots of diseases have environmental causes, which are spatially correlated. (People have known since ancient times that some areas are more healthy than others.) The other is that lots of disease are infectious, which *also* produces spatial correlations. But space is big, which means that even large national surveys, when spread out over a whole country, often have only a few individuals sampled from any given location. Special techniques have been developed for estimating disease rates over space which take these issues into account. There are, once again, issues about control and adjustment — it is often no accident who ends up living next to the toxic waste dump.

Dealing with this properly would involve going over a lot about spatial statistics which we don't have time for. So I will pacify my conscience by just noting the issue, and offering some pointers to an exemplary paper (Kafadar 1996) and to a standard textbook (Lawson 2006), and plugging the special-topics class "Data Over Space and Time" (36-467/667).

3 Survival Curves, Life Tables, and Mortality Rates

We will now switch gears a little bit, from thinking about disease rates to thinking how long people live. Let's think about total lifespan as a random variable, say T. It's conventional to look at the upper or complementary[5] cumulative distribution function of T:

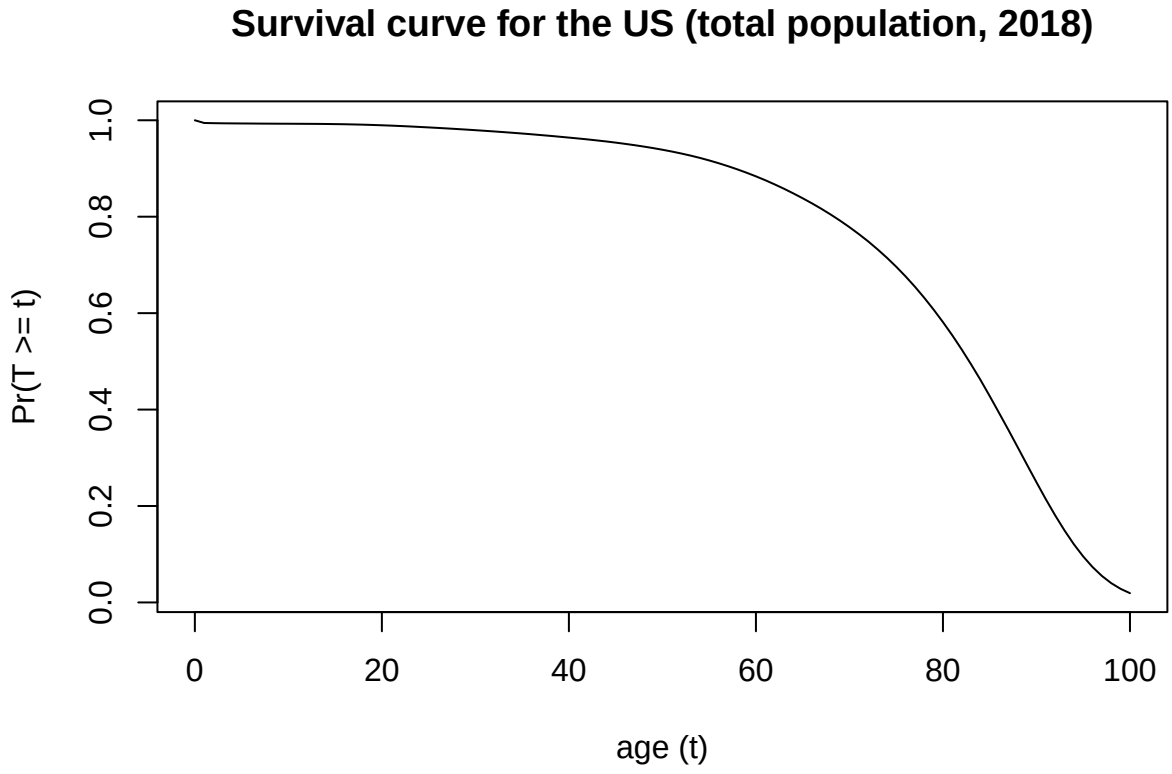
$$\mathbb{P}(T \geq t) = \mathbb{P}(\text{survive to at least age } t) = \overline{F}(t)$$

Especially in this context, this is often called the **survival function** or **survival curve**.

3.1 Survival curves

Let's establish some basic properties of survival curves:

- The curve starts at 1: $\overline{F}(0) = 1$
- The curve tends to 0: $\overline{F}(t) \to 0$ as $t \to \infty$
- The curve is non-increasing: $\overline{F}(t+h) \leq \overline{F}(t)$ for any $h > 0$
  - In fact the curve is *usually* decreasing, $\overline{F}(t+h) < \overline{F}(t)$. This is true for any biological survival curve I can think of, but I suppose there might be exceptions.
- The curve determines the probability distribution
  - If t is discrete, then $\mathbb{P}(T = t) = \overline{F}(t) - \overline{F}(t+1)$
  - If t is continuous, then the pdf $f(t) = -\frac{\overline{F}}{dt}$ (why the minus sign?)
  - Since T is non-negative, $\mathbb{E}[T] = \int_0^\infty \overline{F}(t)dt$ (see Lecture 2, sec. 6.2, "Expected values and CDFs"; I will leave working out the corresponding discrete formula as an exercise.)



[5] If we think of T as a continuous variable, we can say $\mathbb{P}(T \geq t) = \mathbb{P}(T > t)$, and this really is the complement of $\mathbb{P}(T \leq t)$.

3.1.1 Life expectancy

Life expectancy is simply the expected value of T, perhaps conditional on certain other events. Unmodified, it usually means life expectancy at birth, which is

$$\mathbb{E}[T] = \sum_{t=0}^{\infty} t\mathbb{P}(T = t)$$

We can of course condition T on other variables, such as membership in a social group and/or covariates,

$$\mathbb{E}[T|X = x, Z = z] = \sum_{t=0}^{\infty} t\mathbb{P}(T = t|X = x, Z = z)$$

We can *also* condition T on having lived to at least a particular age, say t_0:

$$\mathbb{E}[T|T \geq t_0] = \sum_{t=t_0}^{\infty} t\mathbb{P}(T = h|T \geq t_0)$$

This gives us life expectancy at age 18, or 25, or 65, or anything else we want, depending on how we set t_0.

We can also look at what's called the "expectation of life" at a given age,

$$\mathbb{E}[T - t_0|T \geq t_0] = \sum_{h=0}^{\infty} h\mathbb{P}(T = t_0 + h|T \geq t_0)$$

which is the number of *additional* years, beyond t_0, someone can expect to live.

3.2 Life tables

Among actuaries, demographers, etc., it's conventional present the information plotted in a survival curve as instead a table, called a **life table**. The idea is that in this table we will track how many people remain alive at each age, from an original (imaginary) population of a certain size, conventionally 100,000 people. So a table might look like this:

Age	Surviving
0	100,000
1	99,435
2	99,399
3	99,372
...	...
44	95,603
...	...
t	100,000 $\times \overline{F}(t)$
...	...

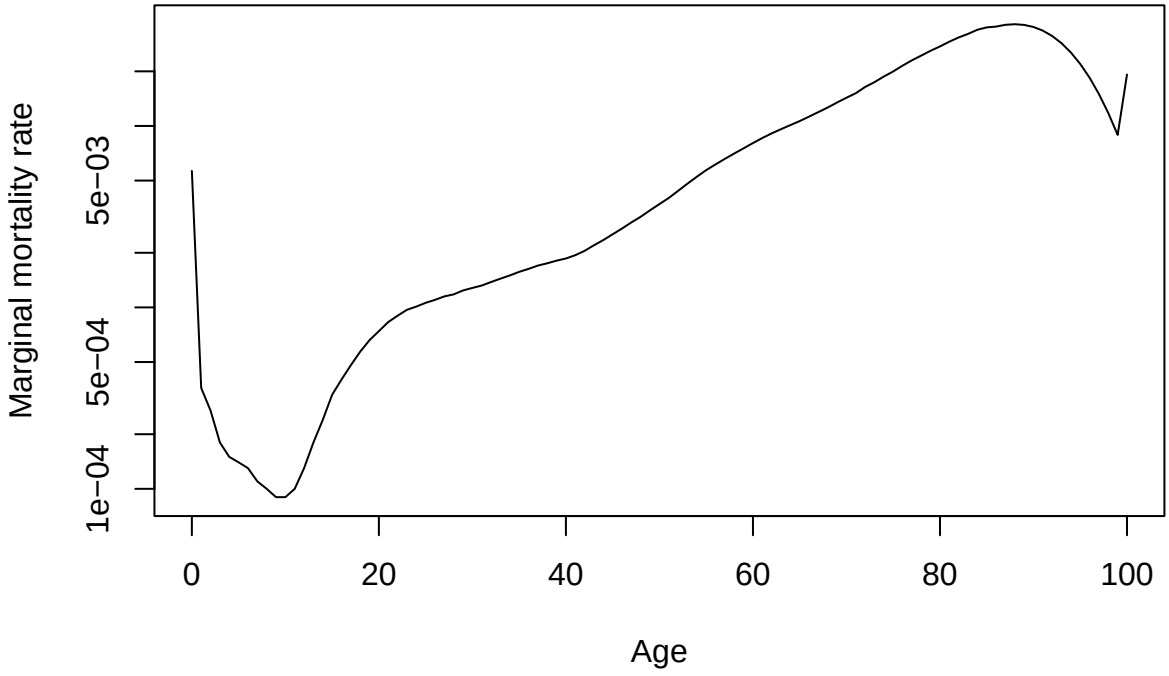
(I have taken these numbers from the life table for the whole US population in 2018 [https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/NVSR/69-12/Table01.xlsx], as you worked with in the after-class exercise.)

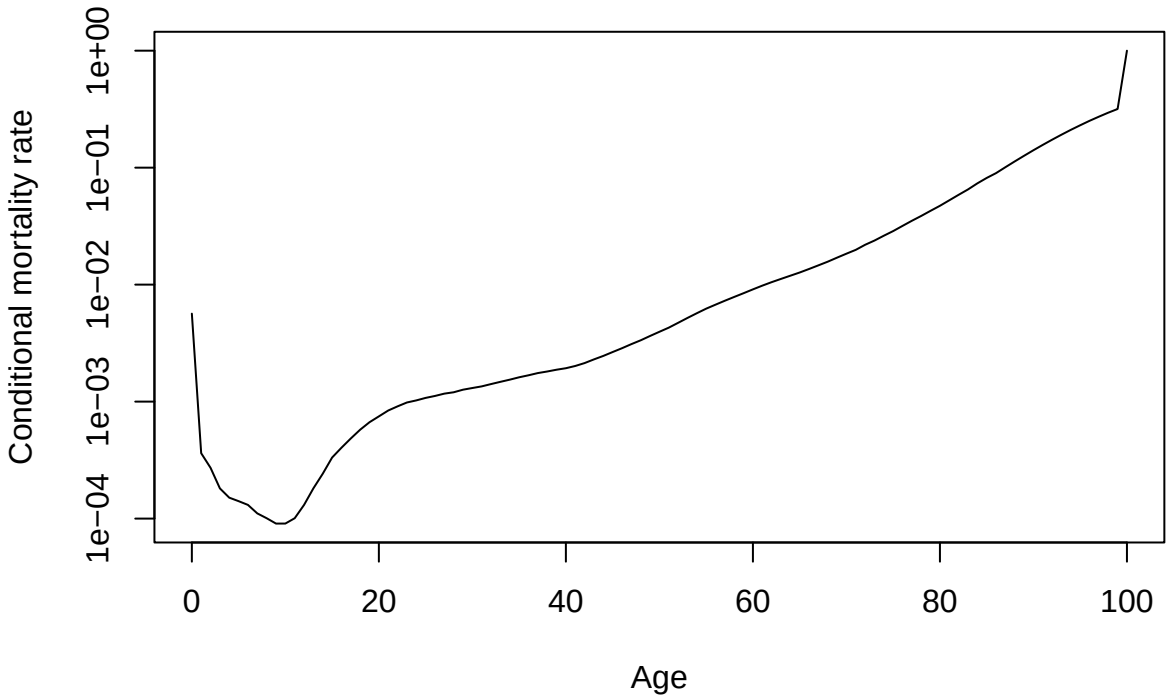
Actuaries like to work with life tables because they simplify certain sorts of calculation. For instance, if you want to find the marginal probability of dying at a certain age, you just take the difference between successive

rows. There is also a neat way to calculate life expectancy, and the additional expectancy of life, from the table, which you can see illustrated in the CDC life table for 2018. However, there is clearly no information in the life table that isn't also in the survival function (and vice versa).

3.3 Death rates

There are two ways we can define the death rate at age t. One is the unconditional or marginal probability that we reach age t, but no higher age, i.e., $\mathbb{P}(T = t) = \overline{F}(t) - \overline{F}(t+1)$. (If we're working in continuous time, this would be the probability *density* at $T = t$.) The other is the probability that we die by age $t+1$, *given that* we have reached age t: $\mathbb{P}(T = t+1 | T \geq t)$. The latter is often a more appropriate notion of the immediate hazards to life. We can estimate this age-specific mortality rate simply by tracking everyone who is age t for a year, seeing how many of them die before the year is out, and dividing by the number we started with. (Some people would divide by the number left at the middle of the year.) Sometimes you see this aggregated across a small group of years, say ages 41–45.





(The apparent uptick in the marginal probability of death from 99 to 100 is because everyone who dies at or above age 100 is lumped together in that last bin.)

As discussed in class, age-specific mortality rates start out comparatively high for infants, then (under modern conditions) plunge down to their minimum in childhood, climb a bit for young people, and then begin a slow, steady rise which accelerates as we pass from middle age to old age, steepening as we go. That first year of life is risky, more so than any year we'll face again until about fifty or so.

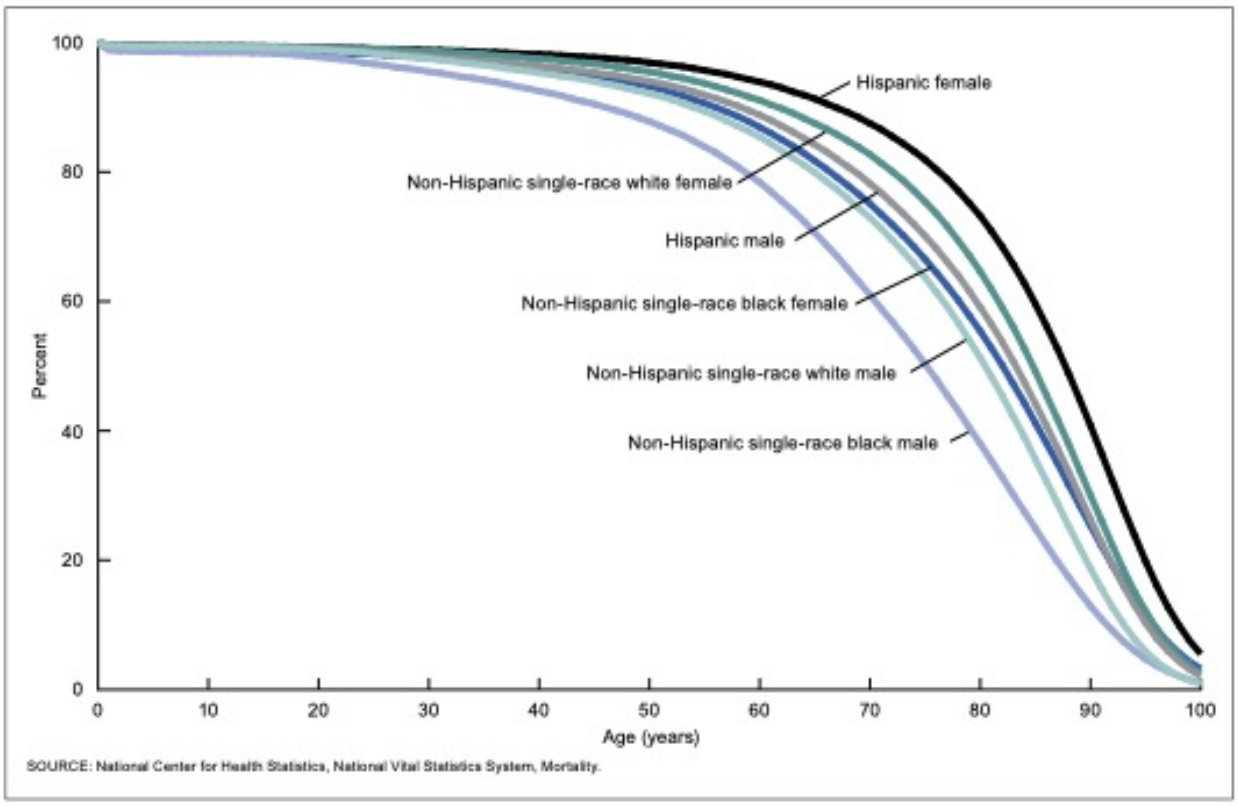
4 Basic facts about inequality in survival curves and life expectancy

4.1 Inequalities

- Women live longer than men, on average. (This is true in almost every country, and the apparent exceptions are places with *really* bad data.)
- More-educated people live longer, on average, than less-educated people.
- In the contemporary US, black people are shorter-lived than white people, who are shorter-lived than Hispanics.

The interactions among these are complicated. (Black women live longer than white men, but not as long as Hispanic men.) You will look at some of those complications in Homework 6, but in the meanwhile, here are survival curves subdivided by race and sex for the US in 2018 (Arias and Xu 2020, Fig. 4, p. 7):

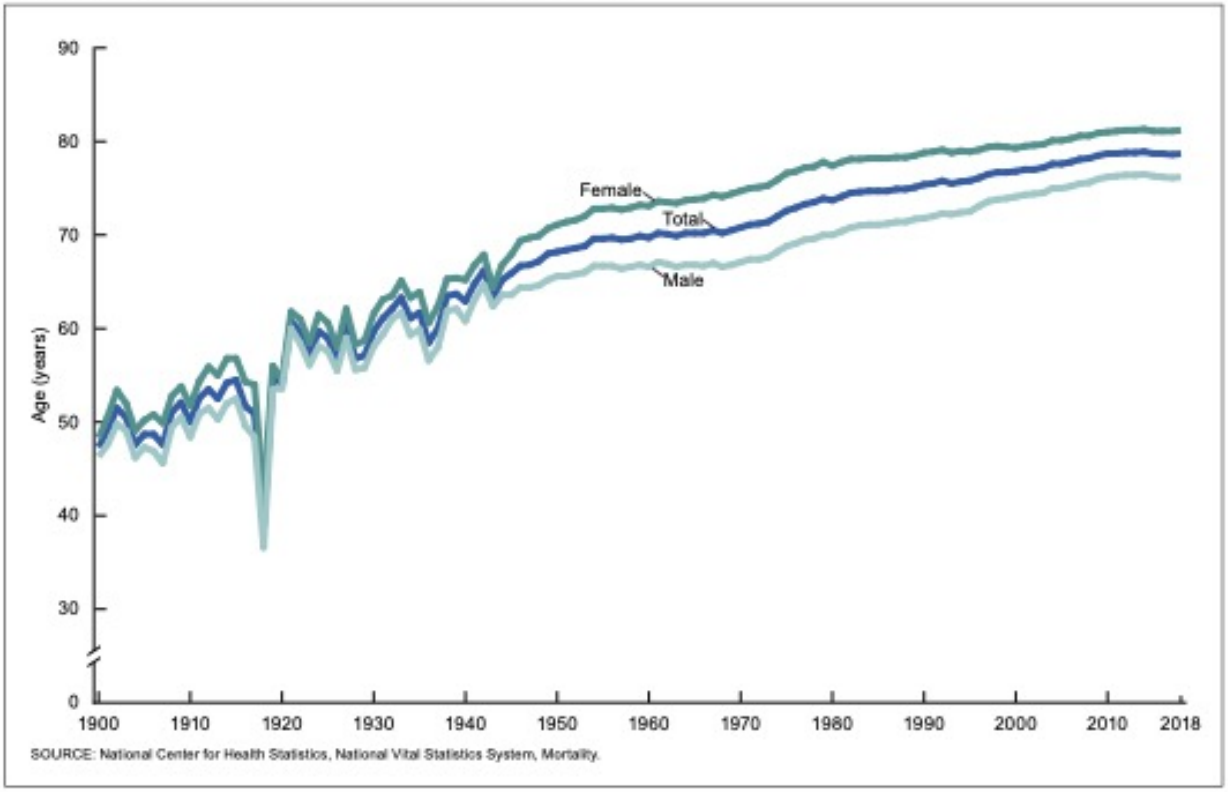
Figure 4. Percentage surviving, by Hispanic origin and race, age, and sex: United States, 2018



4.2 Lengthening lives

In developed countries, life expectancies have been increasing for pretty much all groups, pretty much all of the time, since around 1900 (and even before, for some countries). Here, for instance, is a plot of life expectancy since 1900 in the US (Arias and Xu 2020, Fig. 1, p. 5):

Figure 1. Life expectancy, by sex: United States, 1900–2018



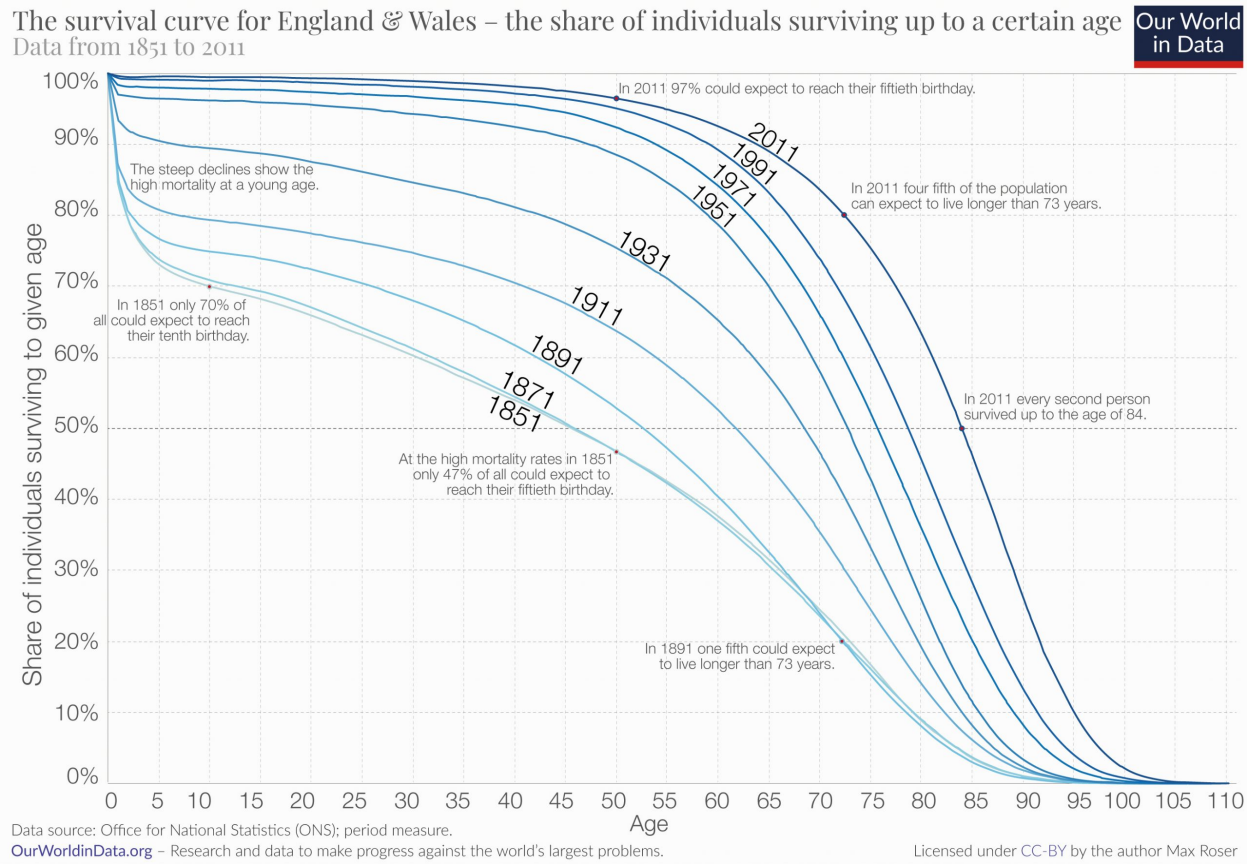
4.2.1 Shortening lives from catastrophes

There *are* events which interrupt this steady increase. The most important are wars and epidemic diseases. You can see the influenza pandemic that began in 1918 in the plot above. The US has been fortunate that its 20th and 21st century wars didn't kill enough Americans to have a big impact on our life expectancy. If we could extend this back to the 1860s, we *would* see the impact of the US Civil War. If we looked at other countries, the wars of the 20th century would definitely show up — Russia, say, or France or Germany or China[6].

[6]Beyond wars and epidemics, the other kind of catastrophe which shortens life expectancy at the level of an entire population is famine. This is not a problem in democracies, for reasons explained long ago by the great economist Amartya Sen (Sen 1999). (At least, not under modern conditions where there's global trade in food.) If we were took look at non-democratic countries, however, we'd certainly see the "collectivization" famine in the Soviet Union in the 1930s, or the famines brought on by the Great Leap Forward in China in the 1950s, in their historical life expectancies.

5 The demographic transition and changing survival curves

While it's necessarily hard to get mortality data and survival curves for historical populations, we do know something about them, and what we know is not pretty. Here, for example, from the excellent website "Our World In Data" is a compilation of historical survival curves for England and Wales, starting in 1851:



You can see that survival to age 20 was only about 65%. That is, roughly a third of all those born did not reach adulthood. This is a fact with profound consequences for the history of our species. Suppose that a population is going to just reproduce itself, and not shrink. Then every adult woman needs to have, on average, at least two children who reach reproductive age and go on to have children themselves. Take reproductive age to be 20. (It may have been a bit lower, but comparatively-poor diets delayed puberty and the onset of fertility.) A 65% survival rate means that there need to be, on average, $2/0.65 = 3.1$ children born per woman. If some women don't have children (whether by choice or necessity), those who do need to have even more to make up the difference. If, say, 10% of women don't have children at all, the mean number of children for the rest has to go up by a factor of $1/0.9 = 1.1$ to compensate, so we're at 3.4 children per reproductive female. Moreover, we're assuming that having offspring who survive to reach age 20 is enough, but that might not give *them* time to reproduce, which calls for even more offspring initially. (Child-birth itself was very hazardous for mothers.) We are also making no allowances for failed pregnancies, since the survival curve starts at birth. We are easily looking at something like *at least* four pregnancies per reproductive woman just to keep the population constant.

The upshot of all this arithmetic is that when survival curves were as bad as they were in England and Wales in 1851, most women who could reproduce at all *had* to have lots of children, in order for enough of their children to reach adulthood to reproduce the population. (This in turn implied spending a lot of time pregnant, trying to get pregnant, or, for lack of alternatives, breastfeeding.) And England and Wales in 1851 were actually, by the standards of world history, a favorable time and place for survival. For many long ages, all over the world, the fact that so many children died so young dictated that families had to *try* to be large in

order to reproduce themselves at all, and so essentially all cultures have spent millennia encouraging fertility.

This had tremendous consequences when survival curves began to change, especially, at first, at young ages. You can see from the figure that the 1871 curve is essentially the same as the 1851 curve, but that by 1891 there is already a difference of about 10% in the probability of reaching age 20, and it had gone up by about 20% by 1931[7]. If almost 90% of children will live to reach reproductive age, replacement fertility is more like 2.2 than 3.1. But initially, people kept having large families, only, unlike all previous ages, the children lived. The result was a population boom, accompanied by people learning that they didn't *need* large families. (Indeed, in industrial, education-dependent societies, they could give their children better lives by having fewer of them and putting more into each child.) Also, women discovered that they might like to do things other than constantly raise children[8].

This pattern — declining infant mortality plus a cultural norm of high fertility leads to a population boom *and* to declining fertility — is something that has been observed all over the world; social scientists call it the **demographic transition**. Different countries have gone through the demographic transition at different times; England was one of the first, with its population going from 8.6 million in 1800 to 30 million in 1900 to 41 million in 1950 [https://ourworldindata.org/grapher/population-of-england-millennium]. Many of the countries which have fully gone through the transition (e.g., Japan, parts of western Europe) now have lifetime fertility rates below replacement level, which also has all kinds of consequences.

[7]The most important causes for the initial upward movement of the survival curve, especially for the young, was improved nutrition from getting enough to eat, followed by such notable, and widely-opposed, public health measures as "sewers" and "clean drinking water", and then early vaccinations and pasteurizing foods. More complicated medical treatments didn't really start to become effective until well in to the 20th century.

[8]In case it's not obvious that I'm simplifying and exaggerating for rhetorical effect: I am simplifying and exaggerating for rhetorical effect. But it's no coincidence that while people had been proposing political and social equality between the sexes for thousands of years, those ideas didn't *go* anywhere until the 19th century.

6 Hey, wait a minute (or a decade)...

It may have occurred to you to wonder how we can know the life expectancy of those now alive, or indeed the shape of the survival curve. There would seem to be two possibilities, both with serious drawbacks:

1. We can look at the **cohort** of all those born in a particular year (1851, or 1951, or 2001) and track how many of them are left in each succeeding year, getting a cohort-specific survival curve. This is accurate for that cohort, but only valid retrospectively. For those born in 2001, the survival curve today only goes up to age 20; for those born in 1971, only up to age 50. When it comes to saying how long today's 20 year olds will live, we don't really have *data*. So we have a cohort approach, but that doesn't let us extrapolate.
2. We can look at the **period** or **cross-section** of (say) 2021, counting how many people there were at each age at the start of the year, and then how many of them died over the course of that year, to get age-specific death rates for each age. We can then combine these to create an over-all survival curve or life table. This has probabilities for each age, but it doesn't reflect what actually happens to people as they age.

The truth is that the data we started with is a *period* life table / survival curve. There is no actual cohort of people which has experienced or will experience precisely this pattern of lifespans and mortality. But it does show the pattern of mortality confronting the population at a particular moment. We can be pretty sure that today's 20 year olds will, in 30 years, have a *different* death rate than today's 50 year olds, but it's hard to say exactly how different and in what direction. So extrapolating forward as though the future would resemble the present is a sensible, or at least widely-used, procedure.

Once you are willing to contemplate "synthetic" or imaginary populations, of course, there are all sorts of possibilities. You might, for instance, want to extrapolate forward various improvements in death rates. This can easily get very conjectural, and small differences in assumptions can compound to big differences in predictions several decades hence.

7 The "Deaths of Despair" Controversy

I said above that in modern, developed countries, it is very rare for life expectancy to decrease, and the exceptions are mostly attributable to particular catastrophic events. One notable exception has occurred in the United States in the 21st century, where life expectancy for (non-Hispanic) white people without college degrees has stagnated and even decreased, while it has continued to increase for other groups. While they were not the first to notice this, the economists Anne Case and Angus Deaton did pioneering work on this topic, helping to show that this was due specifically to an increase in death rates among the middle aged. They also drew attention to rising death rates from suicide, alcohol and drugs. They gave this trio the memorable name of "deaths of despair", and speculated that they were rising because of economic and political changes which adversely affected this group, either absolutely or in relative terms.

This has led to a substantial scholarly controversy, with a number of strands.

- Case and Deaton's initial study looked at death rates among 45–54 year olds. One worry was that the distribution of ages *within* that bracket had shifted, so that there were relatively fewer 45 year olds and relatively more 54 year olds, and older people have higher death rates. More careful studies showed that *some* of the effect was due to this change in composition, but not all or even most.
- Another was how much of the increase in death rates, and the decrease in life expectancy, could be explained *specifically* by the "deaths of despair".
- A third was whether the "deaths of despair" really belonged together, with critics suggesting that the changes in suicide and alcoholism were dwarfed by those due to drug abuse, specifically opioid abuse. The idea there was that the issue wasn't general despair and social alienation, but a drug epidemic. (One could of course ask *why* that population was susceptible to a drug epidemic at that time.)

Notice that these are all controversies about explanations, adjustments and mechanisms, just as we went over in Lecture 10.

The next homework will introduce you to one strand of this controversy, by having you read a later paper by Case and Deaton. Their book, Case and Deaton (2020), has a lot to say about how they have modified their positions in response to criticism, and how they have responded to criticisms.

8 Further reading

Survival curves, life tables and mortality rates are all part of the science of **demography**, which studies the dynamics of populations. This is a vast and intricate subject, intimately connected to statistics, but with its own details and (alas) jargon. Alho and Spencer (2005) is a very good introduction to demography for statisticians.

On the history of population and the demographic transition, Cipolla (1970) is older but brief, clear, and not too superseded by more recent research. Pearce (2010) is a readable journalistic book on the probable consequences of below-replacement-level fertility in an increasing number of countries around the world.

References

Alho, Juha, and Bruce D. Spencer. 2005. *Statistical Demography and Forecasting*. Berlin: Springer. https://doi.org/10.1007/0-387-28392-7.

Arias, Elizabeth, and Jiaquan Xu. 2020. "United States Life Tables, 2018." *National Vital Statistics Reports* 69 (12):1–45. https://www.cdc.gov/nchs/data/nvsr/nvsr69/nvsr69-12-508.pdf.

Case, Anne, and Angus Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton, New Jersey: Princeton University Press.

Cipolla, Carlo M. 1970. *The Economic History of World Population*. Fifth. Penguin.

Kafadar, Karen. 1996. "Smoothing Geographical Data, Particularly Rates of Disease." *Statistics in Medicine* 15:2539–60. https://doi.org/10.1002/(SICI)1097-0258(19961215)15:23<2539::AID-SIM379>3.0.CO;2-B.

Lawson, Andrew B. 2006. *Statistical Methods in Spatial Epidemiology*. Second. New York: John Wiley; Sons. https://doi.org/10.1002/9780470035771.

Pearce, Fred. 2010. *The Coming Population Crash: And Our Planet's Surprising Future*. Boston: Beacon Press.

Sen, Amartya. 1999. *Development as Freedom*. New York: Knopf.