Contents

1 When and How Will You Die?

> It is difficult to make predictions, especially about the future.
>
> — Niels Bohr (probably)

In our first Big Question, we began to look at individual differences between people or what statisticians call variation within a population. If there is no variation—like in the bizarro world where everyone orients their toilet paper in the "under" orientation—then there is nothing to talk about, at least not statistically speaking. There is, however, considerable variation in health outcomes and human lifespan. Lots to talk about there. In our next Big Question, we ask "when and how will you die?" and "what, if anything, can you do about it?"

What kind of question is, "when and how will you die?" Well, according to some of my colleagues, it's a morbid question. Feelings aside, we might say that it sounds like a prediction question, since it's about the future. So to explore this big question, we will need to understand what it means in general to make a forecast about some future event. We'll also find it useful to distinguish between predictions that are or are not explanatory. Most efforts in health sciences attempt to explain relationships between behavioral and genetic factors and health outcomes. In particular, they try to understand causal effects. So in this chpater, we will also try to understand causal explanations more generally.

1.1 Not Quite Death, but, um... Rain?

Perhaps its a good idea to warm up, before we face the grim reaper. What does it mean to say there's a 30% chance of rain tomorrow in New York? Does it mean that it will definitely rain in 30% of the city (say, Brooklyn), but not in the other 30%? Or that it will rain for 30% of the day (say, from 8am-3pm). Here are some possibilities to consider:

a) It will definitely rain in some parts of the city but not in all of them
b) It will definitely rain for some part of the day in all of the city
c) It will definitely rain for some part of the day in some of the city
d) It may or may not rain anywhere in the city at any point in the day.

Read here for an explanation of what meteorologists *probably* mean

1.1.1 Stochastic vs Deterministic relationships

> Sometimes when I say definitely, I mean probably. Like if I say, I'm definitely going to do something about all of this clutter on my desk. But when I really mean business, I say deterministically. It definitely sounds more serious.

Meteorologists—scientists who model the weather—cannot tell us deterministically about weather events. A **deterministic** description of an event would be something like, if I let go of the umbrella I am holding

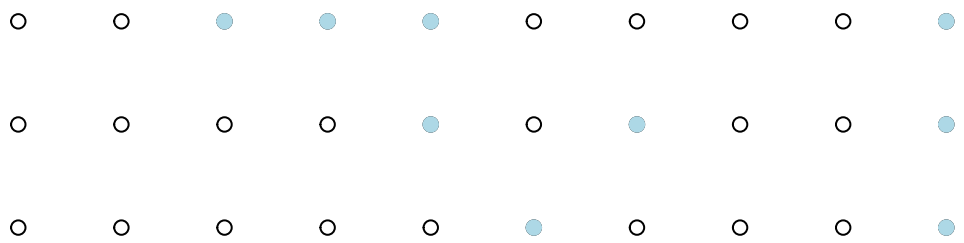


Figure 1: Rain (filled, blue dots) in 9 out of 30 possible worlds. It does not rain (hollow circles) in the other worlds.

in my hand, it will fall to the ground. If A then B. No exceptions. Weather events are **stochastic**. They have an element of randomness, like tossing a coin or rolling a die. So, just as we can say that a coin has a 50% chance of coming up heads—assuming it is a fair coin—we can make statements like there is a 30% chance that it will rain tomorrow. Stochastic is another word for random, but I prefer it because the word "random" is often used casually to mean weird or unusual (as in, "that's random!") Although we can make only probabilistic statements about random, or stochastic, events, that doesn't mean we can't speak usefully about them.

1.1.2 Ensembles

One way to think about the 30% chance of rain is to imagine that our experience in the world is one possibility in a multiplicity of possible worlds. See, I told you this idea of multiple alternate universes was going to be important! Imagine that there are 10 possible worlds, indistinguishable from ours in terms of the laws of physics, and that tomorrow it will in fact rain in 3 of them. To the great being-who-knows-all-things, which 3 may well be known. However, to us mortals who merely live in the world, we don't know which one of these possible worlds is the one we live in. Nevertheless we are capable of imagining these different potential outcomes. As you just did.

It didn't have to be 10 worlds, of course. That was arbitrary. If we imagined thirty worlds, it could rain in nine of them, as I've represented in Figure 1. I did this by making thirty circles and coloring in 9 of them at random. Since I like to pull back the curtain every once in a while, I will even show you the code I use to generate this simple figure.

```
norain <- cbind(rep(1:10,3), rep(1:3, each=10))  # start with a 10 x 3 grid of points
rainworlds <- norain[sample(1:nrow(norain), 9),]  # choose (sample) nine at random, using the sample()
plot(norain, xlab="", ylab="", ylim = c(1,3), axes = FALSE, asp = 1) # plot the points
points(rainworlds, pch=19, col="lightblue") # color in the nine
```

1.1.3 Degree of belief

There is another way to think about 30% as a probability. Suppose a meteorologist said to you, I'm 30% sure it is going to rain tomorrow. And you say back, "Oh, you mean that, say there are really 1000 alternate universes out there, that in roughly 300 of them, it will rain tomorrow?" And the meteorologist says, "I have no idea what you're talking about. There is only one universe, and I'm not totally sure what will happen tomorrow, but I put the chances of rain at 30% *[walks away slowly towards the door]*."

For your meteorologist friend, 30% represents a degree of belief. Importantly, the degree of belief is subjective. Here it is attributed to a meteorologist, which might make you take it more seriously than if your Uncle Bob said the same thing (unless Uncle Bob is actually a meteorologist). Anyway, degree of belief is subjective. Which doesn't mean it's arbitrary or just a matter of opinion. When it comes to forecasts, some people or some forecasting models are going to be right more often than others. More on that later.

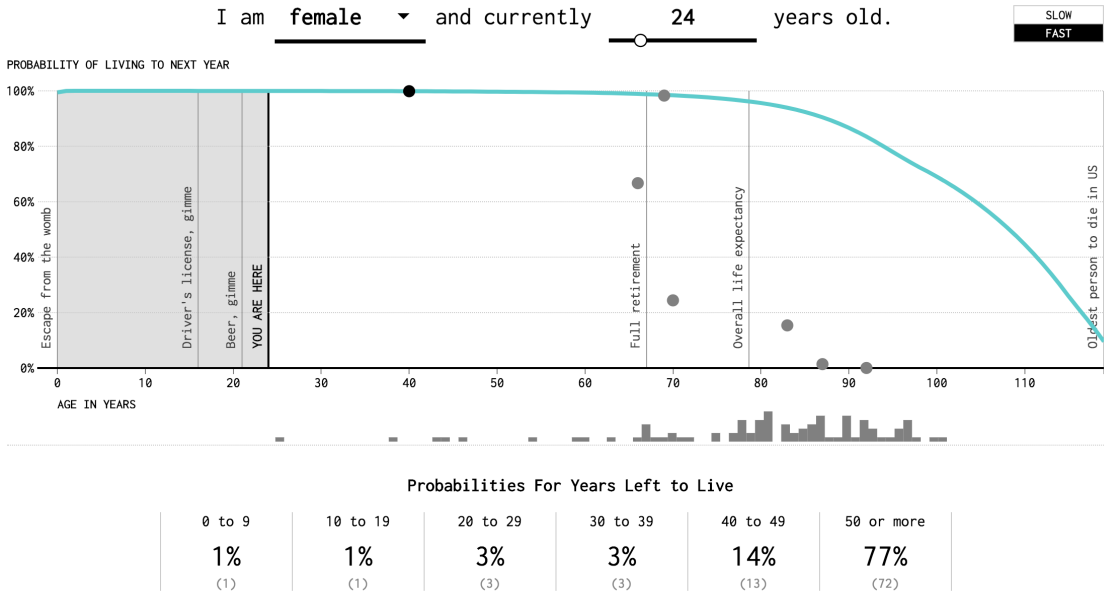


Figure 2: Screenshot of interactive data visualization

1.1.4 Decisions

Aside from subjectivity, which is a thorny topic among statisticians, there is really no *practical* difference between the interpretation of 30% probability as a frequency of occurrence in an ensemble of possible worlds or as a degree of belief about this world. It won't change what you do about it.

If you take this forecast of rain seriously, you have decisions to make. It could be whether or not to take an umbrella with you when you leave the house tomorrow, or whether to cancel your plans to have a barbecue outside. These decisions may not seem very high stakes. The worst case scenario is that you (and others at your barbecue) get wet. But other decisions you have to make on a daily basis can have more serious consequences for your health or even your life. You often have to make those decisions based on probabilistic and maybe subjective information.

1.2 Death

End of warm-up. It's time to talk about when you will die.

I highly recommend this data visualization called Years You Have Left to Live, Probably. Here is a screenshot, although it's not nearly as interesting when you can't interact with the simulation and watch the little balls drop.

```
include_graphics("../images/YYHLTLScreenshot1.png")
```

This visualization does a number of things. The most salient feature is probably the dropping balls. Each one represents a possible future outcome. This is exactly like an ensemble of alternate universes. As you watch the balls drop, you think to yourself, "ah, nice, I lived to be 92" and then moments later, "ooh, harsh! I died at 39!"

As the simulation runs, it also accumulates data in bins at the bottom, labeled "0 to 9", "10 to 19", and so on. (Recall the discussion of bins, frequency tables, and histograms in Section **??**.) Note that these bins represent ranges of years-you-have-left-to-live, not age-at-death. This may be confusing, because age-at-death is what is shown along the horizonatal, or x-axis, of the figure. Also, right below the x-axis, and corresponding to

age-at-death is a set of gray bars that grow as the balls drop. In the screenshot, the simulation has been running for a little while, so that the following counts have been accumulated.

bin	counts
0 to 9	1
10 to 19	1
20 to 29	3
30 to 39	3
40 to 49	13
50 or more	72

Notice that by the time this screenshot was taken, 93 balls had dropped. The visualization took the counts, converted them into proportions of total counts (e.g., $72/93 = 0.774$; $3/93 = 0.33$), and represented each of these proportions as a probability, expressed as a percent (e.g., 77%; 3%).

Another thing that you will notice if you play around a bit is that as the balls drop, the probabilities change. In the beginning, when the number of samples (balls dropped) is small, the numbers change rapidly and sometimes by a large amount. However, after a couple of hundred samples, the changes are much smaller.

By watching the balls drop on this simulation (which I, for one, find mesmerizing), you may actually be meditating on some profound ideas in statistics. Every time you restart the simulation, you begin the sampling process. Each sample is a **draw** from some distribution of possible life outcomes. Your future life bounces around in this distribution from sample to sample. And in the beginning, when you have only collected a small number of samples, the distribution itself seems unstable. For example, if you put in 24 as the current age and start the simulation in slow mode, the estimated probability of living 40-49 more years fluctuates a lot. However, as you accumulate samples, the shape of the distribution literally comes into view as a pattern among the gray bars just below the x-axis. As the sample size increases, the probabilities becomes more stable. Eventually, if you let it run long enough, you end up with the same values, regardless of how things started out.

Although we are now talking about probabilities about your remaining years left to live, the interpretation of probabilities is similar to that in our discussion of rain predictions. In the case of rain, there were only two possibilities, rain or no-rain. (A dichotomy!) In the death simulation, there are six bins, each of which represents a range of years. In the case of rain, we understood the meaning of a 30% chance (i.e., probability) of rain by imagining a large number of possible worlds, where it rains in 30% of them. Thus the probability was associated directly with a frequency of something occurring. This is known is as the **frequentist** interpretation of probability. In the case of death, we say you have a 77% chance of living 50+ more years if, in a large number of possible worlds, you live 50+ more years in 77% of them.

You probably realize that we don't get to see all of these alternate universes, even though we can imagine them. Therefore our probability estimates in many cases are based on things that we have observed happen to *other* people. For example, among 100,000 people that we do observe from the moment of birth, suppose 78% of them lived into or past their 70s. We convert that observed frequency into a probability for you. You could say that we treat the other people we observed as alternate-universe versions of you.

1.2.1 How does the death (simulation) work?

The Flowing Data animated visualization is based on data collected in "life tables", which can be found online from sources like the National Center for Health Statistics (NCHS) and the Social Security Administration (SSA). Different life tables are produced every year, as life expectancy continues to evolve along with changes in health science and nutrition. Figure 3 plots data for age-at-death (for Americans) as of 2010. There is a bar for each age from 0 to 120, and the height of each bar represents a count of deaths at that age per 100,000 people.

Figure 3: How long Americans were living in 2010

If you're like me, the first thing you notice in Figure 3 is that little spike at age 0, like a rattle sticking up at the end of a rattle snake's tail. It shows us that roughly 5 out of 1000 babies don't make it to their first birthday. After that, your odds get considerably better for a while.

Another feature that you may detect is that the distribution of age-at-death is not symmetric. It has a long tail to the left. Distributions like this are also called left-skewed.

So how does age-at-death relate exactly to the years you have left to live? Life tables are a bit of a strange thing. First of all, they are not tables of "raw data" for a sample of 100,000 people. Rather, they represent a summary of data from many more deaths. According to the SSA source, "the life table represents a hypothetical cohort of 100,000 persons born at the same instant who experience the rate of mortality represented by qx, the probability that a person age x will die within one year, for each age x throughout their lives."

Most of us don't think about our lives in terms of questions like, are we going to die this year? But that is technically how the life table works. The life table is a set of numbers—including deaths-at-age-x and expected-years-left-to-live-at-age-x—that are all derived from one initial set of numbers which represent *the probability that a person age x will die within one year*. If you're curious what that initial set of numbers looks like, I've plotted them in Figure 4.

Looking at Figure 4, you can say that the probability of dying within one year gets higher as you grow older, which comes as a surprise to no one. If you're under 65, say, that probability doesn't even feel that high. It's less than 0.01 or 1%. The probability that you will die *this year* only passes 50% after age 100. That's reassuring, right?

Well, don't get too optimistic. Your chances of dying every year may be small, but every year is another draw from this morbid lottery. If your chances of dying were 1 out of 2000, then in 2000 universes, you died in one of them. In the other 1999, you live on to another year, but then you have to press your luck again. This happens every year, and the chances slowly get worse.

But what if you wanted to know your chances, at birth, of dying in your 60s, that is between 60-69. For now, we will try to answer this question using only the life table and assuming that we know nothing else about

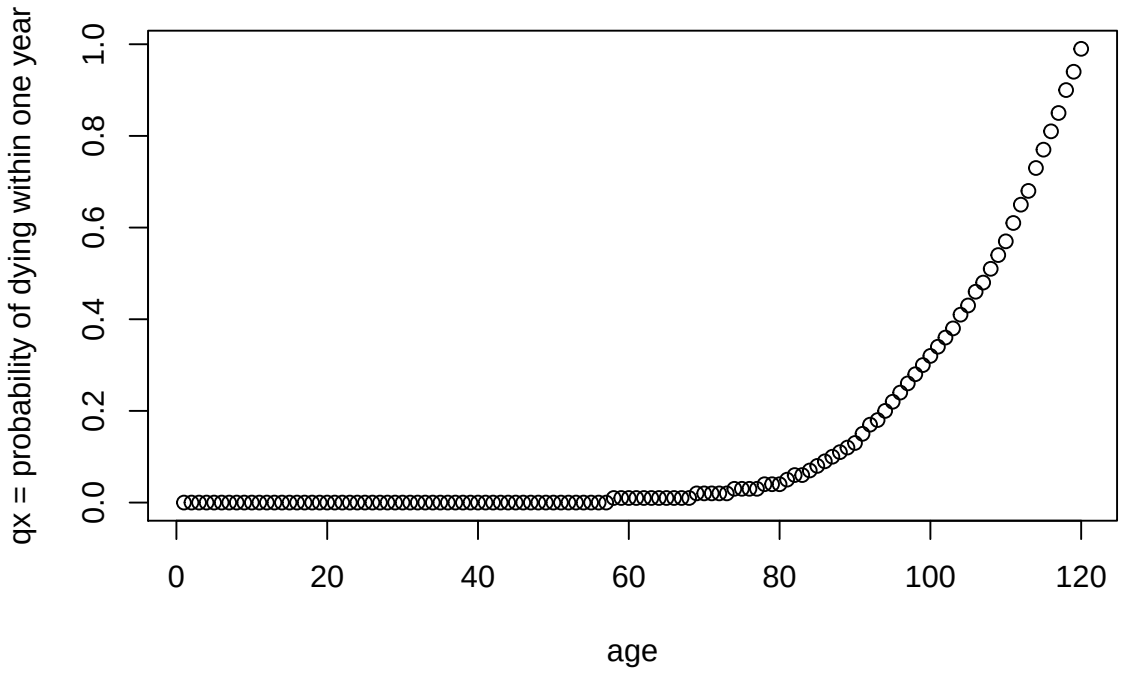


Figure 4: Mortality rate per year of age

you. The rows of the life table corresponding to this age range are these

This is a lot of numbers. Recall that each qx is the mortality rate for age x, the probability of dying within one year of age x. So should you add up the qx-values for each age in the interval 60 to 69? Maybe pause here to think about this question for a moment before reading on.

Here is a partial answer. You can die at 62 and you can die at 64, but you can't die at both ages. In that sense, it was okay to add the probabilities of these events because they are **disjoint**, i.e., they can't both happen and you are interested in whether any one of them does happen. However, if you add up these probabilities, you will still over-estimate the probability for a different reason. Can you guess what you've left out?

Here is the rest of the answer. You've left out the fact that these probabilities assume that you have already made it to 60, and there's a chance (at birth) that you won't.

To answer the original question, you want to add up the following probabilities:

```
(Probability of making it to 60 and then dying at 60) +
(Probability of making it to 61 and then dying at 61) +
... +
(Probability of making it to 69 and then dying at 69) +
```

How do you figure out the probability of making it to 60 without dying? It sounds a little bit like a riddle whose answer is "one year at a time." Indeed, to make it to 60 without dying, you need to not die every year for the first 59 years of your life.

Note that, while death can occur in only one year of your life, to survive into your sixties you need ALL of the following to be true: NOT dying at 0 AND NOT dying at 1 AND ... NOT dying at 59. The probability of each event (not dying in each year) is independent, and the probability that all of them happen is the product of the individual probabilities.

```
Probability of NOT dying at 0 *
Probability of NOT dying at 1 having made it to 1 *
```

Table 1: Life Table

Age	qx	lx	dx	L	Tx	ex
60-61	0.008732	88745.98	774.97	88358.50	2051875	23.1
61-62	0.009335	87971.02	821.18	87560.42	1963516	22.3
62-63	0.009983	87149.84	870.00	86714.84	1875956	21.5
63-64	0.010715	86279.84	924.46	85817.61	1789241	20.7
64-65	0.011568	85355.38	987.39	84861.68	1703423	20.0
65-66	0.012586	84367.98	1061.84	83837.06	1618562	19.2
66-67	0.013763	83306.15	1146.57	82732.86	1534724	18.4
67-68	0.015057	82159.58	1237.07	81541.05	1451992	17.7
68-69	0.016380	80922.51	1325.52	80259.75	1370451	16.9
69-70	0.017756	79596.98	1413.34	78890.31	1290191	16.2

... *
Probability of NOT dying at 59 having made it to 59

Since in any given year, you either die or don't die, these two probabilities must add up to 1, so having gotten to any age x, the probability of surviving it is (1-qx). Now we can take the product of (that is, multiply) all of the survival probabilities (1 - qx) for each x from 0 up to age 59. (I will include the code here. The data table I have loaded from the National Center for Health Statistics is called "lifetableNCHS").

```
prod(1-lifetableNCHS[1:60,"qx"])
```

```
## [1] 0.887458
```

You may notice that this probability had already been calculated for you in the life table, but it had been presented slightly differently as column lx, which is the number of persons (in a cohort of 100,000) surviving to exact age x. If we multiply our rate by 100000, we get 88745.8, which (up to a rounding error) is the same as the number in Table 1.

Okay, so now we are ready to complete the probability calculation. Recall we wanted to add up ten things: Probability of making it to 60 and then dying at 60, etc. We know that the probability of making it to age x is the same as the value of column lx in the table divided by 100,000. And the probability of dying is qx. So we need to multiply these two numbers in each row and add them up.

The result is 0.1056. An American child born in 2010 has a 10.5% chance of dying in their 60s (and a 20.7% chance of dying in their 70s).

So, we've figured out how to do that. And we're almost ready to move on, but it is worth noticing something. The product of the value qx and lx in each row of the life table is the value dx, which is the number of deaths at age x (or between x and x+1). So when we multiplied and added before, we were really just adding up the number of deaths (dx) at ages 60-69 and dividing by 100,000.

Now hopefully that makes sense to you that this should give us the answer we were originally looking for, namely what are the chances, at birth, of dying in your 60s. We could have looked at our hypothetical cohort of 100,000 people all born at the same time and asked: how many of them will die in their 60s. Well, that would be the sum of the dx-values, namely 10562. It wouldn't be a probability, though, unless we divided it by the total number of people (100,000).

So we've shown that we can answer our particular question two different ways:

A) Computing the total probability of your making it to 60 and then dying at 60 *or* making it to 61 and dying at 61 *or* making it to 62 and dying at 62 etc. up to age 69.

or

B) Computing the overall proportion, out of 100,000 people, who die in their 60s.

Table 2: Bizarro world

	chunky	smooth
over	0	23
under	17	0

A = B in this case. An important property of mathematical sciences is that you can arrive at the same answer in different ways. Maybe that sounds like a waste of time, but I view it as one of the most reassuring things about math. If you try something two different ways, and you do *not* get the same answer, then something is probably wrong.

1.3 Some facts about Probabilities

A lot of books would have tried to establish some basic facts about probability up front. (See, for example, OpenIntro Stats, chapter 3). There is a sound logic to setting up foundations like that. But in this book, I've taken the strong position that ideas should be driven by questions. So I've tried to reason through the example above without setting up any foundations. Nevertheless it's a good time to recap some of what we established about probabilities. We will also introduce the most basic notation `P(A)` for the probability that event A happens. For example, event A can stand for "you die at age 64" or "it rains in New York tomorrow."

- When possibilities are disjoint, or mutually exclusive, the probability that either one of them happens is the sum

```
P(A or B) = P(A) + P(B)
```

An example of this was dying at age 62 or dying at age 64.

- A special case of this addition rule applies when one or the other MUST happen. For example, in logic, either something happens or it doesn't happen. Either A or NOT A. Since these possibilities are disjoint:

```
P(A) + P(not A) = 1
P(not A) = 1 - P(A)
```

An example of this was the probability that you do not die at age 0. We found it by subtracting out the probability that you will die from 1.

The last fact we used is

- The probability rule for **independent** events that BOTH occur is the product of the individual probabilities of each event occurring.

```
P(A and B) = P(A) * P(B)
```

We used that to figure out how you survive by not dying every year. Notice that I've snuck in the word independent (well, I snuck it in boldy, so it wasn't that sneaky). There is an intuitive reason why it is important to make a distinction about independent events.

In the last chapter, we said that two events (we were talking about responses to questions) are independent if knowing about one of them does not give you any information about what the other one might be. But remember bizarro world where the toilet paper orientation and peanut butter preference were deterministically related, and specifically everyone is either under-chunky or over-smooth? I've reproduced this result in Table 2. If I told you that 53% of the total population prefers smooth, then what proportion of the total population prefers smooth AND likes to over-hang? Also 53%. What proportion prefers smooth AND under-hangs? 0!

In bizarro world, toilet paper orientation and peanut butter preference are NOT independent, because knowing one of them DOES give you information about the other.

```
P(tp = over AND pb = smooth) does NOT equal to P(tp = over) * P(pb = smooth)
```

This will become even more clear in the next section.

1.4 Conditional Probabilities

Recall that we would NOT have gotten the right answer to the probability of dying in your 60s if we added up the mortality rates qx for all ages x in [60-69]. (Exercise: verify this.) Rather, we had to multiply these numbers first by the probability of living to age x. Another way to say this is that the mortality rate qx was actually a **conditional probability.** It was the probability of dying at age x *on condition that* you have survived to age x. To be absolutely clear, we are measuring x in whole numbers, like birthdays, but we don't mean dying on your xth birthday. Rather, we mean dying anytime between turning age x and turning x+1. We need a special notation to distinguish conditional probabilities. We write,

```
qx = P(You die at age x | You survived to age x)
```

and we read this as "qx is the probability that you die at age x given that you survived to age x" or as "qx is the probability that you die at age x conditional on your surviving to age x." These are equivalent, but they differ from

```
P(You die at age x)
```

which is the **unconditional** probability that you die at age x. This is also different from

```
P(You die at age x AND You survived to age x)
```

which is called the **joint probablity** of the two events. We calculated exactly this joint probability above when we wanted to add up the probabilities that you die at some point in your 60s. The way we computed the joint probability for each year was by application of this general rule for conditional probabilities

```
P(A and B) = P(A|B) P(B)
```

which we read as "the probability of both A and B happening is equal to the probability of A conditional on B multiplied by the probability of B." Note that this rule *always* holds. That's because what I've called the general rule is equivalently just the definition of conditional probability. For example, I could have written it this way:

```
P(A|B) = P(A and B) / P(B)
```

This is just a rearrangement of the formula, but we have a tendency of seeing whatever is on the left side of an equation as being defined by what is on the right.

As far as death is concerned, the following are all true:

```
P(die at x AND survived to x) = P(die at x | survived to x) * P(survived to x)
P(die at x AND survived to x) = qx * P(survived to x)

qx = P(die at x AND survived to x) / P(survived to x)
```

where in the second line I substituted the mortality rate qx for the conditional probability that defines it. In the last line, you can see how the mortality rate could be estimated from data if you actually observed a whole bunch of people. You would count how many of the die at age, say, 62, and divide that number by the number who survived to age 62. You can also probably see why the following is true:

```
P(survived to x | die at x) = 1
```

That is, if you died at 62 then you must have survived to that age. That may seem too obvious for words, but it helps to show clearly that for conditional probabilities, it is not generally true that P(A|B) = P(B|A).

Considering toilet paper in bizarro world, we can see explicitly why the rule for joint probabilities of independent events `P(A and B) = P(A) * P(B)` did not hold. The conditional probability relationship always holds, but independence is a special case. We can see what it is now:

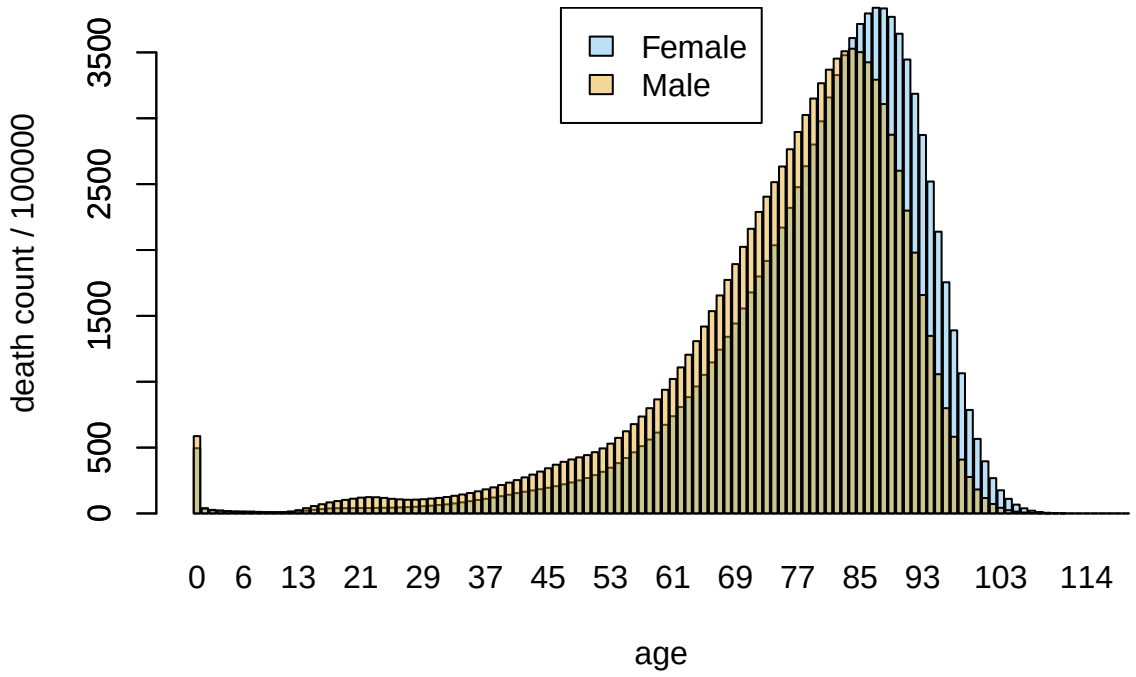


Figure 5: Deaths by age for male and female (2010)

```
P(A and B) = P(A|B) P(B) = {only in special cases} = P(A) * P(B)
```

Thus, when A and B are independent, it must be true that

```
P(A|B) = P(A)
```

which reads as "the probability of A conditional on B is equal to the probability of A (regardless of B)." Another way to say this is that no matter what we know about B, it doesn't tell us anything informative about A. But that was NOT true in bizarro world, where knowing peanut butter preference told us EVERYTHING about toilet paper orientation. If A is the probability that a person is an over-hanger, and B is the probability that they prefer smooth peanut butter, then it is not true that

```
P(tp = over | pb = smooth) = P(tp = over) ## NOT TRUE in bizzaro world
```

which would be the case if these observations were independent. Rather,

```
P(tp = over | pb = smooth) = 1
P(tp = over | pb = chunky) = 0

P(tp = over AND pb = smooth) =  P(tp = over | pb = smooth) * P(pb = smooth) = P(pb = smooth)
```

1.5 Conditional Death

Earlier I said we would use the life table to answer questions about when you will die assuming nothing else about you. Now, you might be aware that life expectancy is not the same for males and females. Indeed, there are separate life tables for each sex. I've plotted the death column dx from both tables in Figure 5. Females are shown in light green bars, and males using pink. Unfortunately for the males, their mortality rate is higher not only in their later years, but even in their late teens and twenties.

```
sum(lifetableNCHS[81:101,"dx"])/sum(lifetableNCHS[,"dx"], na.rm=T)
```

```
## [1] 0.5749251
```

```
sum(lifetableFemale[81:120,"dx"])/sum(lifetableFemale[,"dx"])
```

```
## [1] 0.59967
```

```
sum(lifetableMale[81:120,"dx"])/sum(lifetableMale[,"dx"])
```

```
## [1] 0.4680319
```

Suppose I

Check your understanding

P(tp = under | pb = smooth) = ? P(tp = under | pb = chunky) = ?

Using the

```
sum(lifetableNCHS[61:70,"dx"])
```

```
## [1] 10562.34
```

1.6 Bayes Rule

Conditional probabilities may be easy to define, but they are probably not intuitive to most people. Even experts make mistakes when reasoning with conditional probabilities. Consider the following scenario:

1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammograms. 9.6% of women without breast cancer will also get positive mammograms. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

A) 90.1%

B) 70.4%

C) 28.2%

D) 7.8%

E) 1.6%

```
P(A and B) = P(A|B) P(B)
P(A and B) = P(B|A) P(A)
```

1.7 Bayesian Networks

A whole computational framework known as Bayesian networks has been established to make it easier for computers to help us with these problems. Bayesian networks are named for Thomas Bayes, who also put his stamp on Bayes' rule.

1.8 Causality

Does eating meat cause heart disease? Does smoking cause lung cancer? What does it mean to say A causes B? First of all, this may sound like a philosophical question, and indeed the philosopher David Hume shed some important light on the question of how we conceive of causation. But this is a book on probabilistic thinking, not philosophy. So we are going to take a more pragmatic approach and focus on how we use the

concept of causation in everyday life. Nevertheless, it helps to first recall our distinction between deterministic and stochastic processes.

If I hit a porcelain tea cup hard with hammer and the tea cup breaks, we can safely say that hitting the teacup with a hammer caused the cup to break. We don't really feel the need to say that if you hit a teacup hard with a hammer, there is a 99.9997% chance that it will break. Even if that's actually true. And we don't feel the need to define "hard" in this case either. We use an example like a teacup and hammer when we want to focus on the common-sense big picture and not the details. And the big picture here says that hitting a teacup with a hammer deterministically causes the teacup to break. Let us also assert that if we do not hit the teacup, and it just sits there, then it will not spontaneously break. In the case of the physics of hammers and teacups, we feel that we know this much is true.

What about buying a lottery ticket? Does buying a lottery ticket cause one to win the lottery? Well, you certainly are not guaranteed to win the lottery if you buy a ticket. (In fact, your chances will be very low. The subject of making money is the next Big Question). But you can't possibly win if you don't buy a ticket. So, strictly speaking, buying a ticket does influence the probability of winning.

We've now dicussed two examples. In the first case (hammer and teacup):

- If A (hammer hits teacup) then definitely B (teacup breaks)
- If not A (hammer does not hit teacup) then definitely not B (teacup does not breaks)

In table form:

	Teacup breaks	Teacup doesn't break
Hammer hits teacup	Always*	Never
Hammer does not hit teacup	Never	Always

*pretty much; we're not splitting hairs here.

In the second case (lottery ticket):

- If A (buy lottery ticket) then maybe B (win lottery) and maybe not B (do not win lottery)
- If not A (do not buy lottery ticket) then definitely not B (do not win lottery)

	Win lottery	Do not win lottery
Buy lottery ticket	Rarely	Probably
Do not buy ticket	Never	Always

Now, let's pause for a moment and think about one of the questions we started with: does smoking cause cancer? Does it fit either of these two cases?

Unfortunately the question about smoking does not. It belongs to a yet another case.

In the third case (smoking):

- If A (smoke) then maybe B (cancer) and maybe not B (no cancer)
- If not A (do not smoke) then maybe B (cancer) and maybe not B (no cancer)

	Get cancer	Do not get cancer
Smoke	Maybe	Maybe
Do not smoke	Maybe	Maybe

Now I'm not saying that the chances of cancer are the same whether you smoke or not. That remains an open question so far as our present argument goes. But even thus far, we can see that the smoking causality

question, posed this way, invites some more questions.

How big a difference does there have to be between the cancer rates for smokers and non-smokers for us to be convinced that there is an association between smoking and cancer? And if there is an association between smoking and cancer, what would drive us to call this a causal relationship, to say that smoking causes cancer? Could causality go the other way?

Testing for an association between two variables

For a moment, let's focus on the first question: How big a difference does there have to be between the cancer rates for smokers and non-smokers for us to be convinced that there is an association between smoking and cancer?

Suppose that we go out and find a random sample of 1000 people for whom the following information is available: a) whether the person smokes (or has smoked in the past) and b) whether the person has ever been diagnosed with cancer. The beginning of our dataset looks something like this:

	Cancer?	Smoke?
Person 1	Yes	Yes
Person 2	No	Yes
Person 3	No	Yes
Person 4	No	No
Person 5	Yes	No

As a first step, you tabulate the data and get the following contingency table:

	Cancer: Yes	Cancer: No
Smoke: Yes	46	204
Smoke: No	93	657

Then, you use the table to estimate the following:

$$P(\text{Cancer}|\text{Smoke}) = \frac{46}{46+204} = 0.184$$

$$P(\text{Cancer}|\text{Not Smoke}) = \frac{93}{93+657} = 0.124$$

You might say that these numbers suggest an association (i.e., dependence) between smoking and cancer: Within this sample, a higher proportion of smokers were diagnosed with cancer than non-smokers. But is this enough of a difference to convince you that, if you went out and found 1000 new (random) people, you would still observe a difference of this magnitude?

One way to *start* trying to answer this question is to consider the following thought experiment: imagine that, among all people in the world, there is NOT a higher incidence of cancer among smokers (as compared to non-smokers). If that were the case, you would expect to see

$$P(\text{Cancer}|\text{Smoke}) = P(\text{Cancer}|\text{Not Smoke}).$$

Or, written slightly differently:

$$P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke}) = 0.$$

In comparison, you observed the following in your initial sample:

$$P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke}) = 0.184 - 0.124 = 0.06.$$

So, you could pose the following question: what is the probability that, among the whole population, smokers do not have higher risk of cancer; but, among the random sample of 1000 people that you observed, there is a 6% (or greater) increased incidence of cancer among smokers as compared to non-smokers? This type of question is the basis for **hypothesis testing**. Often, in hypothesis testing, we form a **null hypothesis** (in this case, the null hypothesis might be that smokers and non-smokers have equal cancer incidence among the full population) and **alternative hypothesis** (in this case, the alternative hypothesis might be that smokers have at least 6% higher risk of cancer than non-smokers). If you are interested in learning more about how statisticians use probability distributions to answer these types of questions, there are plenty of resources online. In this book, however, we will attempt to answer this question using a simulation.

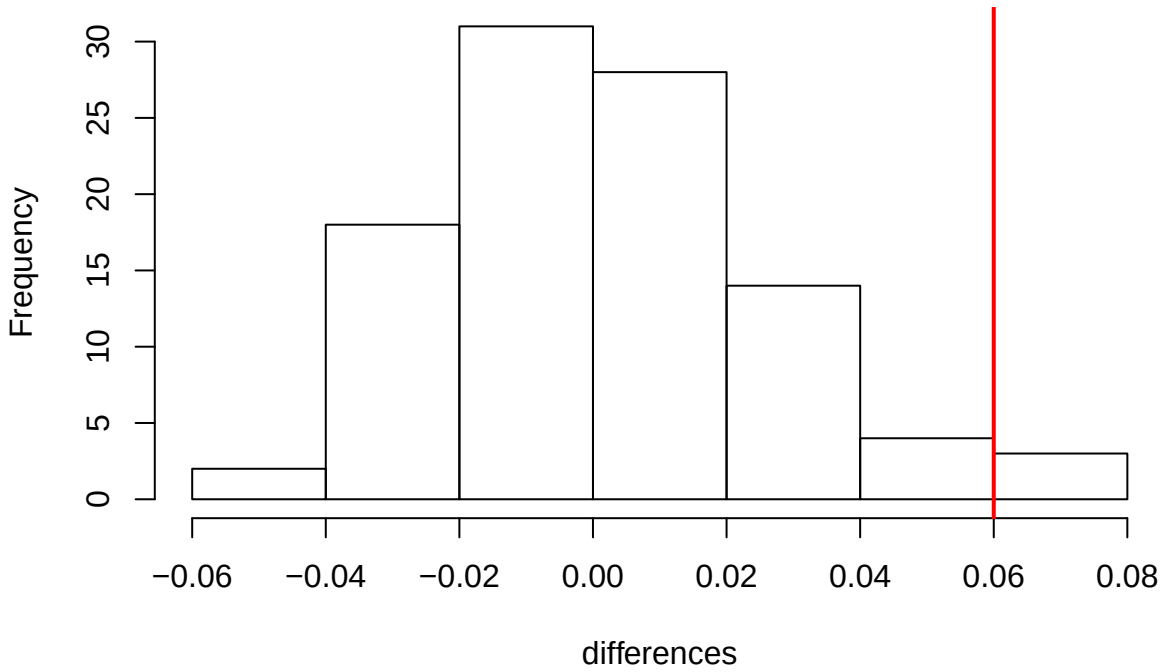
Based on the sample you observed, you could estimate that approximately $\frac{46+204}{1000} * 100 = 25$ percent of the population smokes and approximately $\frac{46+93}{1000} * 100 = 13.9$ percent of the population has been diagnosed with cancer. If there is no real difference in cancer incidence among smokers and non-smokers, then these two variables are independent: 25% of your sample randomly decided to smoke, and 13.9% were randomly diagnosed with cancer. It turns out that it's very easy to simulate datasets under this assumption. All we have to do is, in two completely separate steps, randomly assign 25% of people to be smokers and randomly assign 13.9% of people to get cancer. Then, for each of these simulated datasets (of 1000 people each), we can calculate $P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke})$ and observe what types of differential proportions could be observed by random chance. Then we can calculate the proportion of these differences that are greater than or equal to 0.06 in order to understand the chances that we observe a difference of that size when it doesn't actually exist:

```
set.seed(513)
```

```
nIter = 100 #set some number of iterations
differences = vector(length = nIter) #create vector to save differences in proportions

for(i in 1:nIter){ #repeat the following process nIter times

  #create some fake data and save it as "fakedata"
  fakedata = data.frame(Smoke = sample(c("Y", "N"), size=1000, prob=c(.25, .75), replace = T),
                        Cancer = sample(c("Y", "N"), size=1000, prob=c(.139, .861), replace = T))

  #use the fake data to calculate P(cancer|smoke)
  CgivenS = table(fakedata)[2,2]/sum(table(fakedata)[2,])

  #use the fake data to calculate P(cancer|not smoke)
  CgivenNS = table(fakedata)[1,2]/sum(table(fakedata)[1,])

  #save P(cancer|smoke) - P(cancer|not smoke) in the ith location of differences
  differences[i] <- CgivenS - CgivenNS

}

#calculate proportion of differences greater than or equal to .06
sum(differences >= .06)/nIter
```

```
## [1] 0.03
```

```
#plot a histogram of the differences with a red vertical line at .06
hist(differences, main="Histogram of P(cancer|smoke) - P(cancer|not smoke)")
abline(v=.06, lwd=2, col=2)
```



As you might expect, the histogram of simulated differences ($P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke})$) is centered around zero. If there's no real difference, then you should expect to observe (close to) zero differences among any random sample of 1000 people. That said, you'll see from the histogram that it is still possible, by random chance, to observe a difference as large as 8%.

But let's get back to our original question: You might notice that a very small proportion of the simulated values of were greater than or equal to .06. (about .03 or 3%). This might help convince you that there was a relatively low probability of observing $P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke}) \geq .06$ among your sample of 1000 people if the reality was that $P(\text{Cancer}|\text{Smoke}) - P(\text{Cancer}|\text{Not Smoke}) = 0$ among the full population.

We've made some good progress, but you might still have some concerns: 1) there is always some risk that there is no real difference in cancer incidence among these two groups, but you observed a large difference in your sample nonetheless and 2) even if there is a real difference, there are a lot of possible reasons that this difference could exist without a direct causal relationship. That said, if you can convince yourself that there is a very small probability that the association you observed could have occurred randomly, then you might want to move on to a new question: can you explain why is there an association?

1.8.1 Things that are not causation

You may have heard the expression, "correlation is not causation."