

A decorative graphic in the top-left corner consisting of two red squares of different sizes and a thin vertical black line extending downwards from the bottom-right corner of the larger square.

Marcello Pagano

[JOTTER 3 PROBABILITY]

Definition of probability and its use in health models, including diagnostic tests and screening.

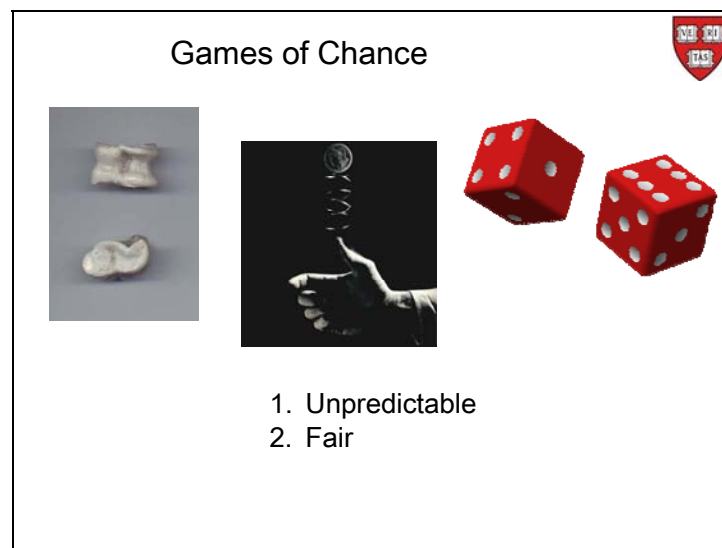
Doubt is not a pleasant condition, but certainty is absurd.

Voltaire 1894—1778

The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account.

— Pierre-Simon Laplace Introduction to *Théorie Analytique des Probabilités*

Probability is a way of quantifying uncertainty. We can use probability to measure risk so as to help us in decision making, and we also use it to assess our inference.



Historically, our main impetus for quantifying uncertainty was through games of chance¹. One of the earliest instruments of chance is the astragalus, a bone which has four faces. You roll them out, and games can be based on which face is up and so on. Unfortunately, the astragalus was not ideal because some faces showed up more frequently than others thus making calculations difficult. So we idealized the situation by introducing dice.

A die has six faces with a one of the numbers, one through six, on each face—note the design for the five, it is called a quincunx. The idealization is that it is just as likely that

¹ It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

— Pierre-Simon Laplace *Théorie Analytique des Probabilités*

any one face will show when the die is rolled as any other face, and a number of games of chance depend on this ideal.

Lastly, to simplify things even more, we do not play dice too often, but we certainly quite often flip a coin to introduce an element of uncertainty into decision making.

There are two aspects of flipping a coin that are important. One is we flip it, for example, to choose who's going to serve in tennis, which side gets the sun, who gets the sun in their eyes, who gets to serve first. We flip because we cannot predict whether a head will come up or whether a tail will come up. (How useful would it be if a head always came up?)

So one, we've got this element of uncertainty. But equally important is this notion of fairness—flipping a coin is a fair way of deciding things.

What does that mean, to be fair? Certainly, if we flip the coin once, it is going to come up a head, or it is going to come up a tail. Excluding cheating, there is no way of telling which comes up.

The idea of fairness requires something more. In repeated trials we can find what we need: If we flip the coin repeatedly, then we expect that roughly half the time we shall get a head, and the other half we shall get a tail.

So this juxtaposition between unpredictability in the short run, but quite predictable in the long run is what makes probability fascinating and useful.

At any rate, I am convinced that He
(God) does not play dice.

Albert Einstein (1879-1955)
Letter to Max Born

God not only plays dice. He also
sometimes throws the dice where
they cannot be seen.

Stephen Williams Hawking (1942-)
Nature 1975, 257.



It is also interesting how the acceptability of the notion of uncertainty has changed through the years. How it has now even become acceptable in the exact sciences, whereas it did not use to be. People much preferred certainty, determinism, instead of uncertainty, but we are learning.



In medicine, it has taken some time to accept the formalism of probability. What makes that ironic is that the first book on probability was actually written by a physician, by the name of Girolamo Cardano.

Cardano, because of his birth situation—he was born illegitimate, something that mattered in those days—was not accepted into the guild to practice medicine in Milano. So he turned to gambling on the side in order to earn a living. What he noticed—he was a brilliant man who wrote about 100 books—was these long run stabilities that he could take advantage of so as to outwit his opponents. The problem is it took more than 100 years for his book to get published, so it did not have the impact it deserved.

Probability



- Element — Event
- Event — set of descriptions
 - proposition
 - everyday sense

An event can *occur* or *not occur*.

Use letters A, B, C, ... to denote events

We talk about the probability of an event happening, and an event is our elemental entity in the theory of probability.

An event is simply a set of descriptions; it is a proposition. We use it here in the everyday sense of the word. So just like when you were introduced to geometry, you started with the elemental point, and built on that to get a line, and then a plane, and so on, so too in probability. We start with an event and build up from there.

Events themselves we denote by capital letters, A, B, C, et cetera, and we focus on whether the event occurred or did not occur (or happened, or did not happen). Or in the future, will it or will it not occur. It makes it easier sometimes to think of the future, because you are then attempting to predict. But you can think retrospectively, too, just gets a little trickier.

In summary, we have that we are concerned about events that are unpredictable in the short run and much more predictable in the long run, and this juxtaposition is the essence of our use of probability.

Operations on events



1° Intersect

The event "A intersect B," denoted $A \cap B$, is the event "both A and B."

A = "A 65 year old woman is alive at 70"

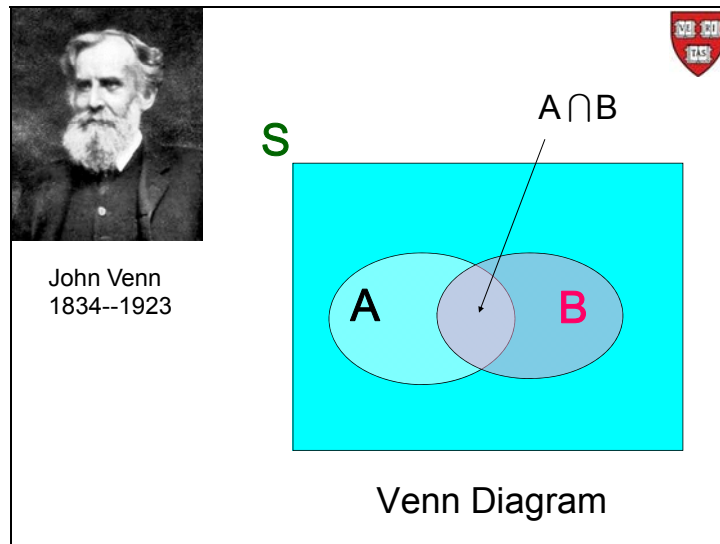
B = "A 65 year old man is dead at 70"

$A \cap B$ = "A 65 year old woman is a widow at 70"

To repeat, probability is a measure defined on events. We have introduced events, so now let us start expanding on the grammar of events by combining events to create new events and thus create a whole library, a whole language with events.

The first operation we investigate is called the intersection: Given two events, say A and B, create a new event that occurs if both A and B occur. The notation we use to denote the new event is A upside down cup B; $A \cap B$.

For example, if the event A is that a 65-year-old woman is alive at 70, and the event B is that a 65-year-old man is dead at 70, and they are married, then the event A *and* B, denoted $A \cap B$, is that she will be a widow at age 70. Because we want both A and B to be true: she is alive at 70, and he is dead at 70.



To help us to think about events, the effect of combining them, and how to determine their probabilities, the philosopher John Venn created a wonderful construct that we now call the Venn diagram. It works as follows: You start off with the whole space, S that represents everything. Let us represent that by a rectangle. So any particular event you can imagine might be represented by a dot in this space S . Now consider the event A ; let us denote it by an ellipse. So if an event occurs that is inside of A , then we say the event A has occurred. For any event outside the ellipse A , we would say that the event A has not occurred. (Note how we use the same language whether we are talking about a single event or a collection of events.)

Now introduce the event B , another ellipse, say, and place it in S .

So we have four areas:

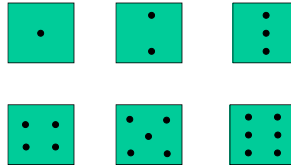
- (i) We start with the area outside of both ellipses. Events out there are neither A nor B .
- (ii) Now, think of the points within A . These can be classified as being in one of two groups: those not in B , and
- (iii) those in B .
- (iv) Finally, we have the events in B that are not in A .

The intersection, $A \cap B$, is represented by the region where the two ellipses A and B (the events in (iii), above) overlap.

2° Union

The event "A union B," denoted $A \cup B$, is "either A or B or both"

e.g.
6 sided
die



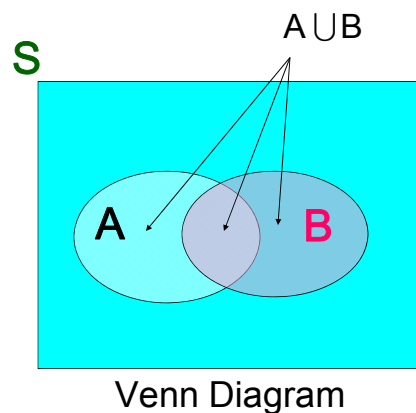
Roll twice:

A = "Roll a 7" B = "Roll an 11"

$A \cup B$ = "Roll a 7 or 11"


The other operation we need is the union. The union of A and B, denoted $A \cup B$, is the event that occurs if either A or B, or both, occur. It is not the exclusionary "or". It is either A, or B, or both.

So when we roll a six sided die, we either get a 1, 2, 3, 4, 5, or a 6. Suppose we roll it twice and sum the results of the two rolls. Define the event A to be we roll a 7, and the event B that we roll an 11. Then $A \cup B$ is that we get a 7 or an 11 when we roll the die twice. So that is the union of two events.



Returning to our Venn diagram, we can see that $A \cup B$ is the total space spanned by the joining of the two ellipses (to make almost a horizontal figure eight!).

Now we have the intersection that says that both have to happen at the same time, and the union that says either one or the other, or both happen at the same time.

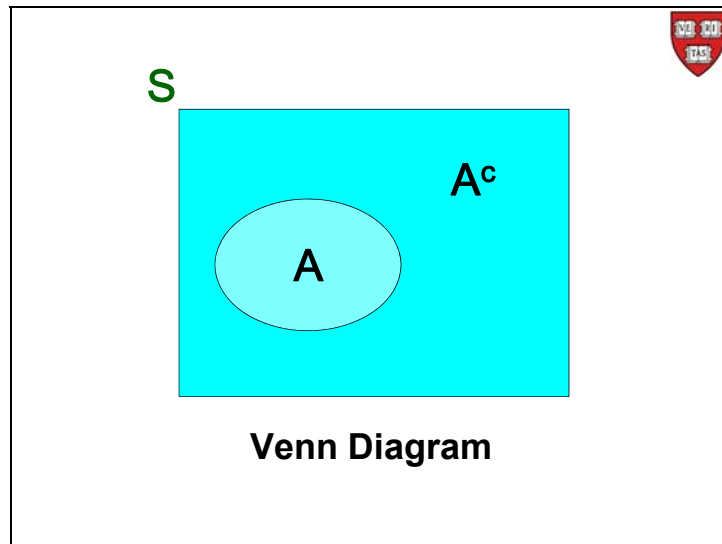


3° Complement

“A complement,” denoted by A^c , is the event “not A.”

A = “live to be 25”
 A^c = “do *not* live to be 25”
= “dead by 25”

Lastly, the third operation we need deals with the complement of an event. It is denoted by a superscript c, A^c , and that denotes the event “not A”. So if the event A is, I live to be 25, then the event A^c is, I do not live to be 25; so dead by 25.



Returning to our Venn diagram, that part of the whole, S , that is not in the ellipse A is the event A^c .

Definitions:

Null event \emptyset

Cannot happen – contradiction

e.g.

$$A \cap A^c = \emptyset$$

It is amazing how much we can do just with those three operations. For example, we can define the null event, usually denoted by \emptyset . It is the event that cannot happen. It is a contradiction. So for example, the event A cannot happen at the same time as A^c . So $A \cap A^c = \emptyset$, the null event. It just cannot happen. You can't have it both ways.

Mutually exclusive events:

Cannot happen together:

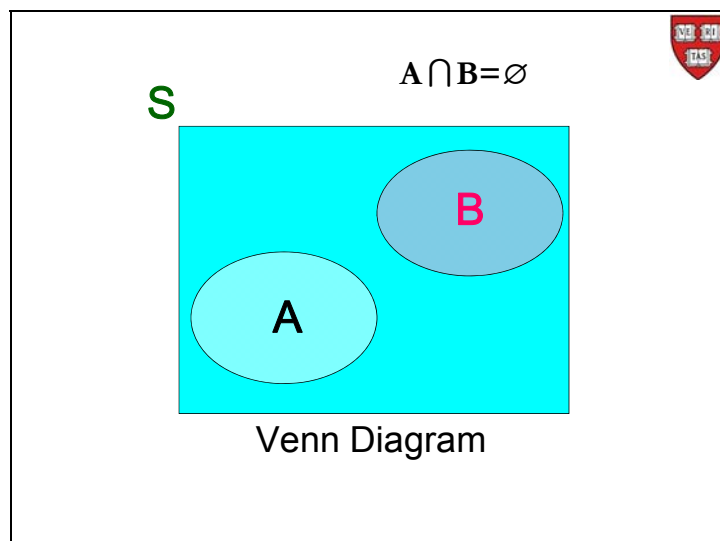
$$A \cap B = \emptyset$$

A = “live to be 25”

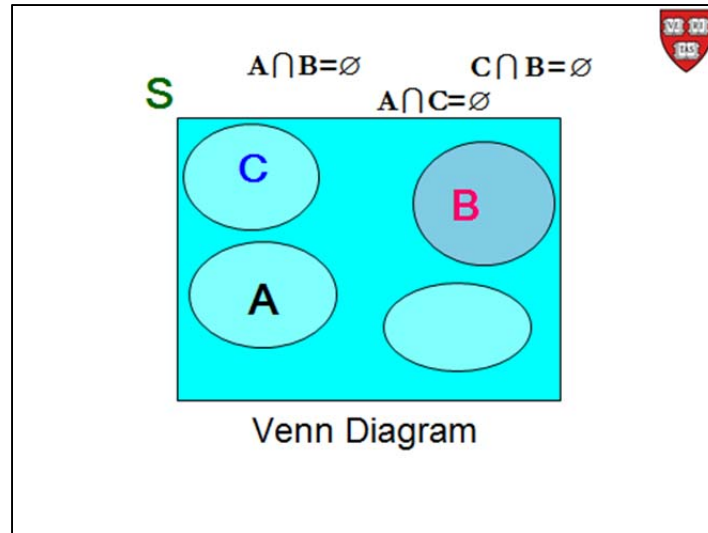
B = “die before 10th birthday”

This leads us to a very important collection of events, and these are mutually exclusive events. The definition of a pair of mutually exclusive events is that they cannot happen simultaneously. So, either one happens, the other, or neither, but you cannot have both of them happening together.

For example, if the event A is to live to be 25, and the event B is die before your 10th birthday. Then you cannot have both, and $A \cap B = \emptyset$. So this can be thought of as an extension of the idea of a complement.



In a Venn diagram, it means that the ellipses are separate.



This concept is extremely useful for attacking problems that are very complex, and involved, and where we have a huge amount of information. Then what we can do is break up the problem into little separate modules, knowing that only one of these modules can happen at a time. That then allows us to solve the problems one aspect at a time until we have the whole solved. This is the genesis of the idea of modularity. It is very important for computer programming, and other tasks, including how we present the material for this course to you! Here we call it mutually exclusive events.

Probability

Probability

If an experiment is repeated n times under essentially identical conditions and the event A occurs m times, then as n gets large the ratio $\frac{m}{n}$ approaches the probability of A .

$$P(A) = \frac{m}{n}$$

Odds of $A = \frac{m}{n - m} = \frac{P(A)}{1 - P(A)}$

Now that we have laid out the groundwork, we are ready to define probability, and we present what is known as the relative frequency definition of probability. There are others, but we do not pursue them.

The definition we are going to use is a practical one. Probability theory itself is a well-formed mathematical theory that is broad and fascinating and can keep us entranced for a lifetime, but we are not going to go down that path, here. We are going to focus on a practical use of the meaning of probability. Returning to our geometry analogy, when you first learned geometry, you learned about straight lines. We know that straight lines that abide according to the geometrical definition only exist in somebody's mind, whereas all around us all sorts of straight lines, that, for example, engineers construct in order to put up buildings, et cetera. Similarly, let us attempt a practical definition of probability.

We focus on the long range stability of events happening that Cardano, von Neumann and Halley took advantage of: Consider an experiment that we repeat n times under essentially identical conditions. For example, flip a coin over and over, under essentially identical condition. Now think of an event A , and suppose this event A occurs m times. Then, as n gets large, the ratio, or proportion, m over n , approaches the probability of A .

We have already noticed this in national mortality tables, except we did not call it probability we called them mortality rates, or proportions.

The definition is deliberately vague in certain spots, because we want a practical definition. So certain phrases such as identical conditions necessarily remain vague. What do we mean by that? Do we need to have the humidity exactly the same? Do we need to have the wind directions exactly the same? And then, what do we mean by n

gets large? Is 3 a large n ? Is 3 million large? And what do we mean by approaches? Is this in a limit, as in mathematics?

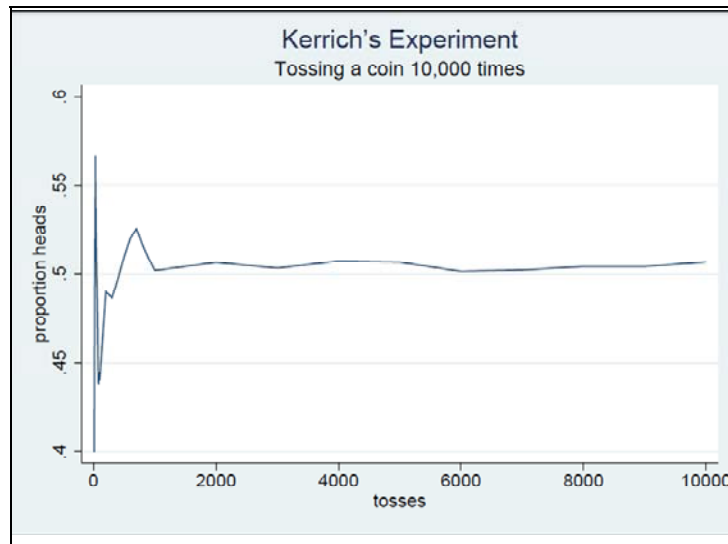
For example, we are going to be applying this definition when talking about people. We need to make statements such as; the probability that this patient will survive the operation is 0.4. Well, why is it 0.4? The answer possibly is, because 40% of previous patients have survived the operation. But were those prior patients identical to this patient? No, they were not, we are all individuals. But that is how we are going to apply this construct. This is the perpetual problem we face whenever we apply an ideal mathematical model to a practical and real situation. We must be prepared to understand when we can apply a model and when not to apply it.

Before leaving this definition, let us point out that there is a mathematically related quantity, and that is the odds of A. So rather than the probability of A, we can also talk about the odds of A.

The only difference is that the odds of A, which is also a ratio, is the ratio of the number of times A occurred to the number of times A did not occur. One should note that Cardano actually wrote his book entirely about odds! Nowadays one mostly hears odds mentioned only at the race track, sporting events, or, in this course, when we get to case control studies, et cetera. By and large, probably because of their training, people tend to favor probability, but the two are clearly related by a mathematical equation, and if you have the one you can uniquely retrieve the other, so their knowledge value is equivalent.

Another difference, which could be of minor importance, is that probability is symmetric around a half. At the edges we are certain—at zero we are certain it will not happen, at one we are certain it will happen. We have maximal uncertainty at the center, when $p=1/2$.

That point of symmetry ($p=1/2$) with odds translates to one; below one we are arguing against the event happening, above one we favor the event happening. But the simple symmetry is no longer there, since, for example, at the left boundary, at zero, the odds are zero, whereas at the right boundary, at one, the odds are infinite. This is indicative of a more complex symmetry, namely the reciprocal symmetry—for example the odds of one third are one over the odds at 3, and thus the symmetry around the point one.



Just before the Second World War, a fellow by the name of John Edmund Kerrich was visiting Copenhagen in Denmark. The Germans invaded Copenhagen and Kerrich was South African, and so he was interred in Denmark for the duration of the war. Being imprisoned with nothing to do, and being a mathematician, he decided to start flipping a coin. He collected pieces of paper so he could keep tabs of his flips, and he had enough paper to note what happened when he flipped the coin 10,000 times.

Above you can see the results of his 10,000 tosses. Specifically, what is plotted is the proportion of heads as the number of flips increased. You can see that initially it is unpredictable (short range), and things go all over the place. But as time progresses (long range), the ratio stabilizes to almost a half (actually 0.5067 after 10,000 tosses of his *real* coin). He wrote a fascinating book about this experiment.²

² John Kerrich, *An experimental introduction to the theory of probability*, E. Munksgaard (Copenhagen) 1946. Also Reprinted in 1950 by BELGISK IMPORT COMPAGNI, LANDEMÆRKET 11, COPENHAGEN

Something must happen:

$$\frac{n}{n} = 1$$

$$\Pr(\text{sure thing}) = 1$$

Impossible:

$$\frac{0}{n} = 0$$

$$\Pr(\text{impossible}) = 0$$

So let us investigate some particular probabilities. First, something must happen. So if your event is that something happens (S) then if you repeat the experiment n times, something happens n times, so its probability is 1 and thus the probability of a sure thing is 1. On the other hand, at the other extreme, if the event is impossible then it never happens, so 0 out of n . So the probability of the impossible event is 0.

For *any* event A

$$m \leq n, \text{ so}$$

$$0 \leq \Pr(A) \leq 1$$

Complement

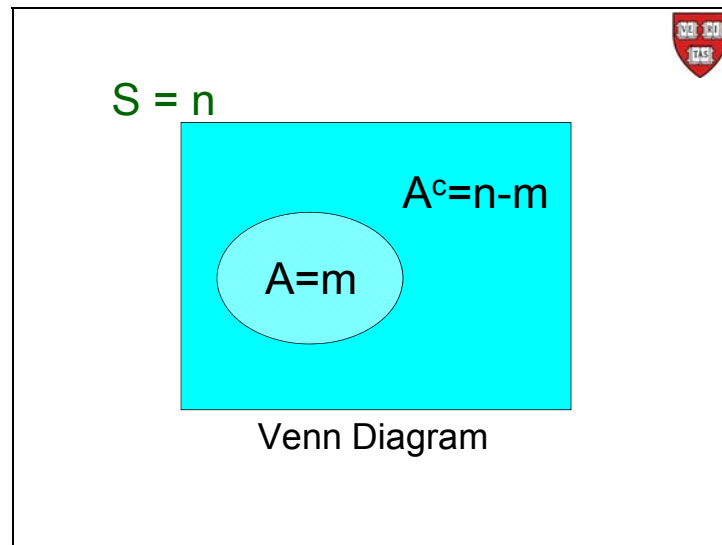
$$P(A) = \frac{m}{n}$$

$$P(A^c) = \frac{n-m}{n} = 1 - P(A)$$

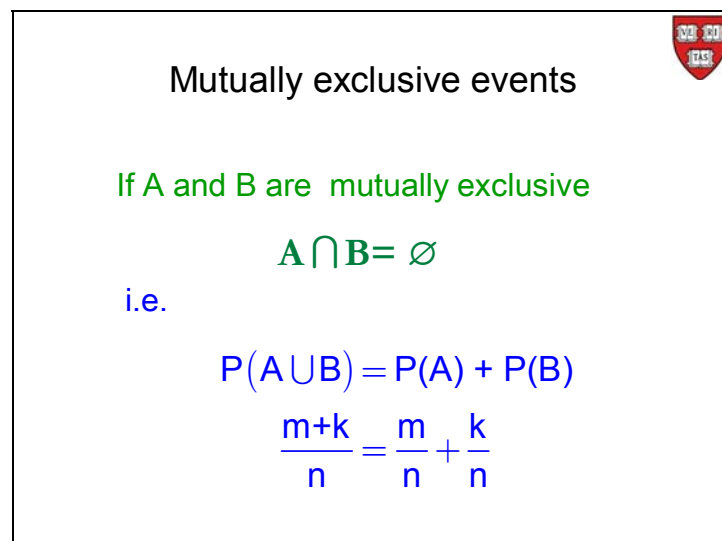
$$P(A) + P(A^c) = 1$$

Indeed, these two events define the limits, namely, the probability of any event is going to be between 0 and 1.

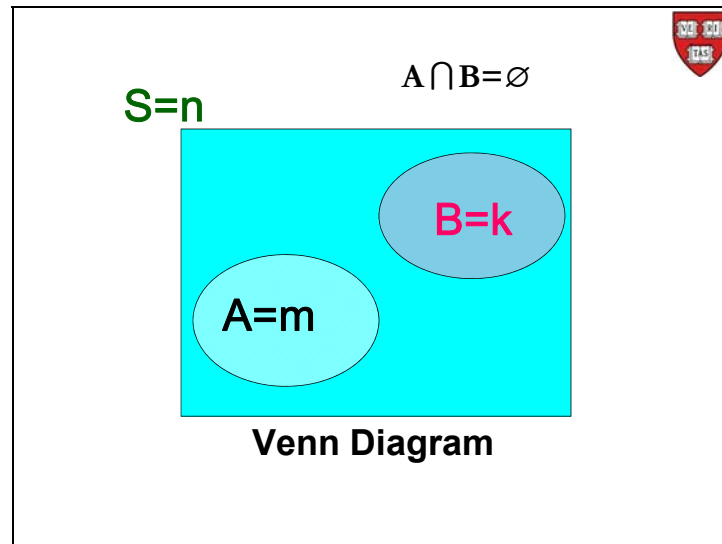
Remember the complement? Well, if A happened m times, that means the other n minus m times A did not happen, or A^c happened. So the complement happens n minus m times, and the probability of the complement is 1 minus the probability of the event. Or another way of saying that is the probability of A plus the probability of A complement is equal to 1.



So here we have it in our Venn diagram.



Now, if A and B are mutually exclusive events, that means they cannot happen at the same time ($A \cap B = \emptyset$). So the probability of A plus the probability of B is the probability of $A \cup B$, because the probability that A happened or B happened is just going to be the probability of A plus the probability of B.



And we see this in our Venn diagram.

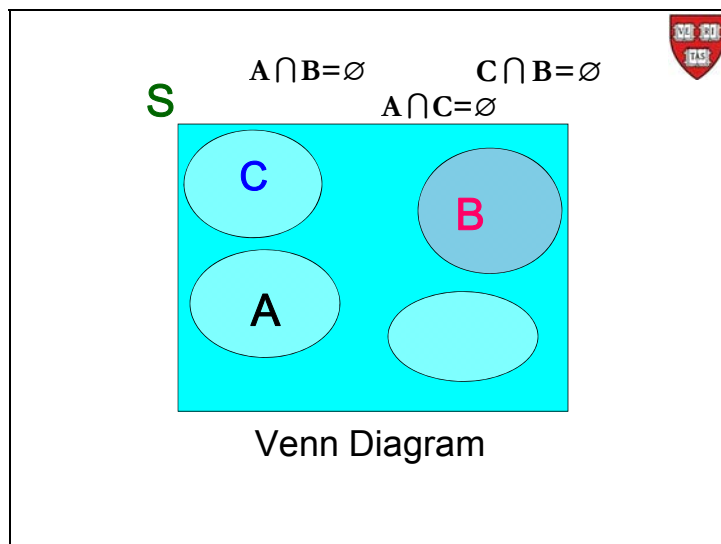
Additive Law

If the events A, B, C, are mutually exclusive – so at most one of them may occur at any one time – then :

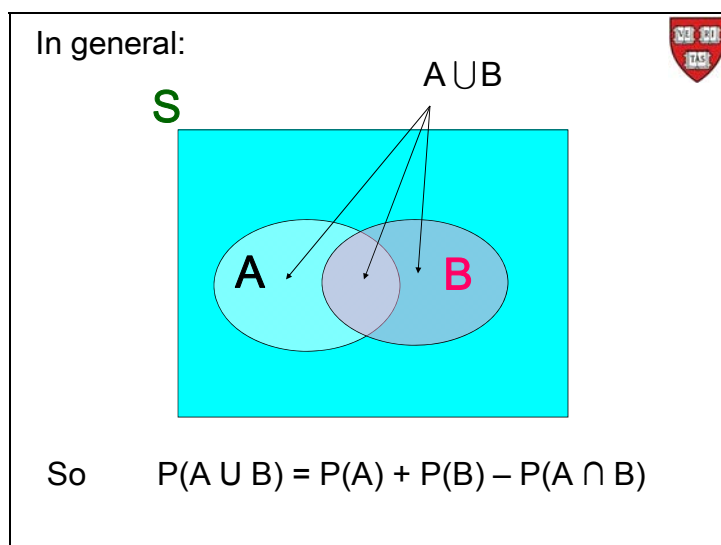
$$P(A \cup B \cup C \dots) = P(A) + P(B) + P(C) \dots$$

We can extend this to more than two mutually exclusive events and we give it a name: it is called the Additive Law. So the Additive Law of Probability tells us that if we have

mutually exclusive events then the probability of their union is the sum of their individual probabilities.




One way to think of probability when looking at Venn diagrams is to think of areas, and you will not go wrong. For example, with mutually exclusive events, as in the diagram above, the $P(A \cup B \cup C \cup \dots) = P(A) + P(B) + P(C) + \dots$.



In general, if events A and B are not mutually exclusive, then there is some overlap in the ellipses. If we now take the area of A and add it to the area of B, and think that that should be the probability of $A \cup B$, we'd be wrong. Why? Because we would be bringing the overlap of A and B, namely $A \cap B$, in twice into our calculation of the area of the union.

So the general formula is the probability of $A \cup B$ is the probability of A plus the probability of B minus the probability of $A \cap B$, and that is the Additive Law.

Conditional probability



Conditional Probability

Notation: $P(B | A)$

is the probability of B *given*, or knowing that the event A has happened.

B="A person in US will live to be 70"

A="A person is alive at age 65"

Then

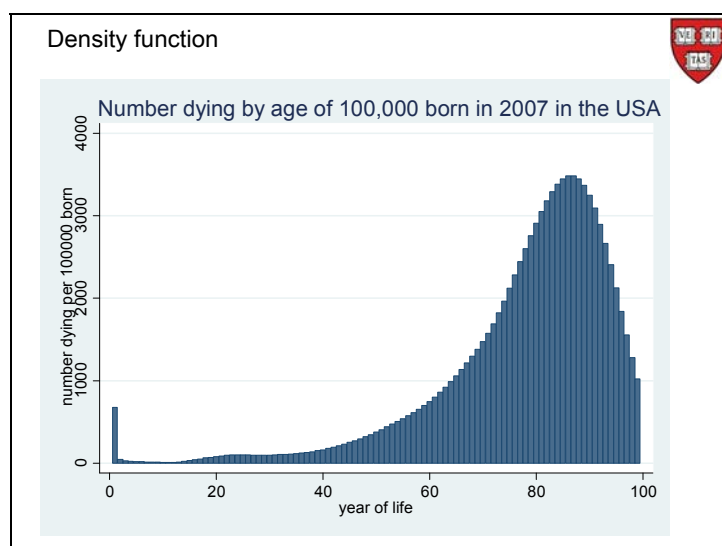
$B | A$ ="A 65 year old person will be alive at 70"

Now that we have probability under our belts, we are almost there. We need to tackle one more concept, called conditional probability. As time evolves, we gather information, information that, if relevant, will possibly change our probabilities. Acknowledging that our probability depends on the information at hand, and how that probability changes, leads us to the concept of conditional probability.

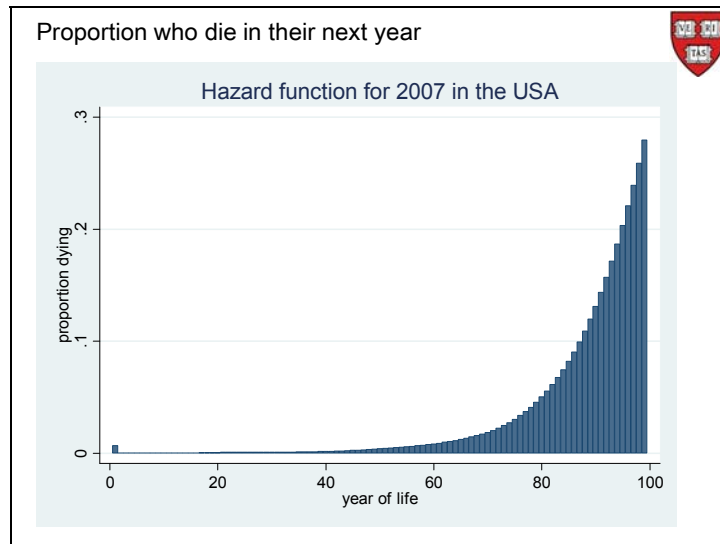
To proceed, we need to expand our notation. Suppose that we retain interest in the probability of the event B, but suppose that now, we also want to bring into consideration the fact that the event A has happened. So what is the probability of B, given that the event A has happened? That is how we approach conditional probability. How do we modify $P(B)$, knowing that A has happened? Intuitively, if A is simpatico to B, then P of B, given A, denoted $P(B|A)$, should be bigger than $P(B)$. If, on the other hand, A is antithetical, or antipatico, to B, then $P(B|A)$ should be smaller than $P(B)$.

The question is, how is the probability of B, given A, changed. For example, return to the USA life table. Suppose that the event B is that a person in the USA will live to be 70, and the event A, is that a person is alive at age 65. Then the event B, given A, is

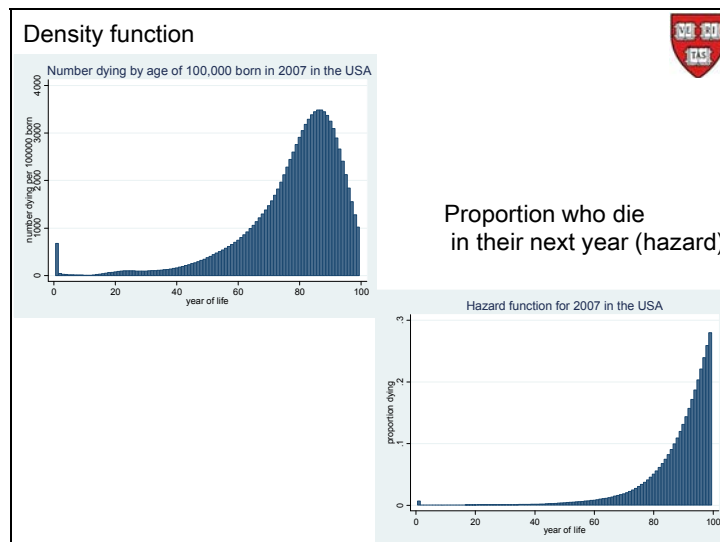
that a 65-year-old person will be alive at 70. That is different from the event B. So the question for the insurance salesman is, if you're selling insurance, would you charge the same thing for a baby to reach age 70 as you would a 65-year-old to reach age 70? Intuitively, the baby has to live another 70 years, including the stretch from 65 to 70, whereas the 65-year-old only has to live another five years—that stretch from 65 to 70. So these two events are not the same, thus their probabilities should be different. The question then is, how do we get from one probability to the other?



To make this pointy more empirically, here is what we saw when we looked at the life table for 2007 in the USA. This is how this fictional cohort of 100,000 people “born” in 2007, are going to die as the years progress. Remember, we said there was a big number around 80 to 95, or so, which is when most of us will die, and then it goes down. And George Burns said, “Very few are going to die beyond a 100,” and that is right, because there are very few left. This is what we call the density function.



We also had, from the life table, the hazard function. The definition of the hazard function was, given that one is alive at a particular age, what is the probability that one dies within the next year? Note that now we do not have the curve thing coming back down beyond 95, because the probability of dying within the next year of life, when we are above 95, say, is very high.



So the density function and the hazard function address two different situations: The probability of dying at any particular age—the density function—versus the probability of dying in the next year, given that you are alive at the beginning of that year—the hazard function.

So there are the two different functions. On the left is the density function giving you the probability of dying at any time after birth, and on the right the hazard function. The

latter is a conditional probability— conditional on being alive at a particular age. Two related but different concepts.

Formula:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$


B = "A person will be alive at 70"

A = "A person will be alive at age 65"

$B | A$ = "A 65 year old person will be alive at 70"

$A \cap B$ = "A person will reach 65 & 70"

= "A person reaches 70"



Here is the formula to get you from one to the other. So the probability of B, given that A has happened, is the probability that both happen divided by the probability of A. I am, of course, assuming that we are not dividing by zero. So I'm assuming that the event A can happen. Returning to our example, if B was that a person will be alive at 70, A, that the person will be alive at 65, and B given A is that a 65-year-old person will be alive at 70, then the event $A \cap B$ is that a person will reach 65, and a person will reach 70. In this example, if a person reaches 70, they will already have reached 65, so $A \cap B = B$, that is it reduces to the event that a person reaches 70.

**Life table (segment) for the total population:
United States, 2007**



Age	Probability of dying between ages x to $x + 1$	Number surviving to age x	Number dying between ages x to $x + 1$
	q_x	l_x	d_x
65-66	0.013600	83,587	1,137
66-67	0.014722	82,451	1,214
67-68	0.015959	81,237	1,296
68-69	0.017288	79,940	1,382
69-70	0.018755	78,558	1,473
70-71	0.020424	77,085	1,574

National Vital Statistics Reports, Vol. 59, No. 9,
September 28, 2011

From the 2007 life table, we see that of the 100,000 who start off, 83,587 reach age 65, and 77,085 reach age 70.

Formula: $P(B | A) = \frac{P(A \cap B)}{P(A)}$



e.g. 2007 lifetable: born: 100,000

65 : 83,587 70 : 77,085

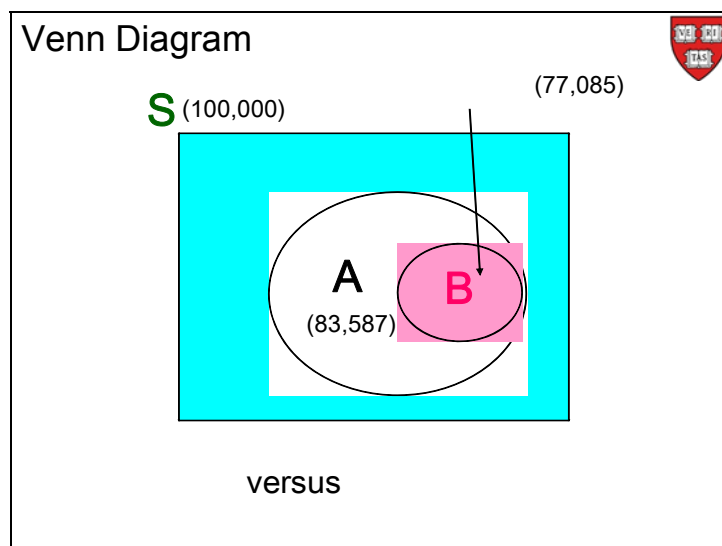
$$P(A \cap B) = \frac{77,085}{100,000} = 0.77$$

$$P(A) = \frac{83,587}{100,000}$$

$$P(B|A) = \frac{77,085/100,000}{83,587/100,000} = \frac{77,085}{83,587} = 0.92$$

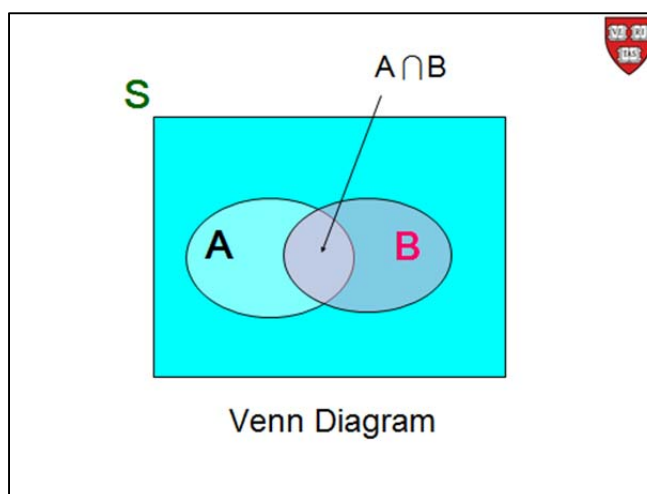
So if we bring all that information together, $A \cap B$ is 77,085 divided by 100,000, so that is 0.77. So 77% of all babies born in this construct, this fictional cohort that we set up in 2007, 77% of those people will reach age 70. Also $P(A)$ is 83,587 divided by 100,000. So dividing those two we get that the probability that a 65 year old reaches 70 is 92%.

So 92% of 65-year-olds will reach age 70. So it's much higher than at birth, as we intuited. That is how much bigger the probability becomes.




In the Venn diagram-- this is a special Venn diagram. Here A is the event you reach 65, and there was 83,587 such people. The event B was that you reach 70, and there were 77,085 such people. But B is also a sub-event of A (one must have reached 65 if one is to reach 70) so the 77,085 are also part of the 83,587— B is completely engulfed by A in the Venn diagram.

So if we calculated the conditional probability directly we would divide 77,085 by 83,587 and get 92%, which agrees with our earlier calculation that used the formula.



In general we can get an intuitive confirmation of the formula for conditional probability from the Venn diagram. If we associate $P(B)$ with the area of B in the diagram, then that works if the area of S is one, otherwise we need to divide the area of B by the area of S . Now if we know that A has happened (so A is the new S) then $P(B|A)$ should be associated with the area of that part of B that contains A , namely $B \cap A$, so we want the area $P(B \cap A)$, but it needs to be normalized by the area of A —the new S — so divide it by $P(A)$, and we get $P(B|A)$.

Multiplicative Law and Independence

From the formula for conditional probability, 
$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

we get the multiplicative law

$$P(A \cap B) = P(A) P(B | A)$$

Note:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

From the formula for conditional probability, multiply both sides of the formula for the conditional probability by $P(A)$ and you have the multiplicative law of probability. Note that even if $P(A)=0$, the law holds since then $P(A \cap B)=0$.

Multiplicative law



$$P(A \cap B) = P(A) P(B|A)$$

Note:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$\begin{aligned} \text{So } P(A \cap B) &= P(A) P(B | A) \\ &= P(B) P(A | B) \end{aligned}$$

A way of talking your way through the multiplicative law is to say, if both A and B are to happen, then if A happens first, then the probability I then want is the probability that B happens given that A has happened. This is just a minor help in remembering the law.

Note, by the way, that there's nothing special about A, nothing special about B. The whole development is symmetric in A and B. So we could either use probability of B given A, or the probability of A given B to obtain the multiplicative law. (Or, I could just as easily have A first or B first.)

Independence



A and B are said to be independent if:

$$P(A \cap B) = P(A) P(B)$$

and since in general

$$P(A \cap B) = P(A) P(B | A)$$

So independence implies

$$P(B | A) = P(B)$$


Similarly

$$P(A | B) = P(A)$$

The multiplicative law leads us to a very fundamental and important concept in probability, statistics, and epidemiology and that is the idea of independence. We say that two events, A and B, are independent if the probability of both A and B happening, $P(A \cap B)$ is $P(A) P(B)$. That means that $P(B|A) = P(B)$. So knowing that A happens does not influence our probability of B happening. Similarly, by symmetry of A and B, we have that $P(A) = P(A|B)$. So knowing that B happens does not influence our probability that A happens. So the label, independent events, is well earned.


We use this condition repeatedly, for example, when we build up our knowledge by taking more and more and more patients. It will be very difficult if after observing the second patient we now have to go back to recalibrate what we learned from the first patient, and then after the third patient go back to the first two etc. So we assume that the patients are independent of each other, so what happened to the first patient is going to be independent of what happens to the second patient, independent of what happens to the third patient, and so on. Of course it must fit the situation or else the model is inappropriate, and we see an example of such a situation shortly. It is probably a little easier to accept with inanimate objects such as a fair roulette wheel. When you spin the ball around the wheel, surely what happened half an hour ago should not have any impact on where the ball lands up now. Surely the wheel does not have a memory! But judging by how people bet, this lack of memory is not believed by all.

The independence assumption needs to be justified, but we make it quite often, and it's extremely useful.



Mendelian Segregation

Gregor Mendel
1823-1884

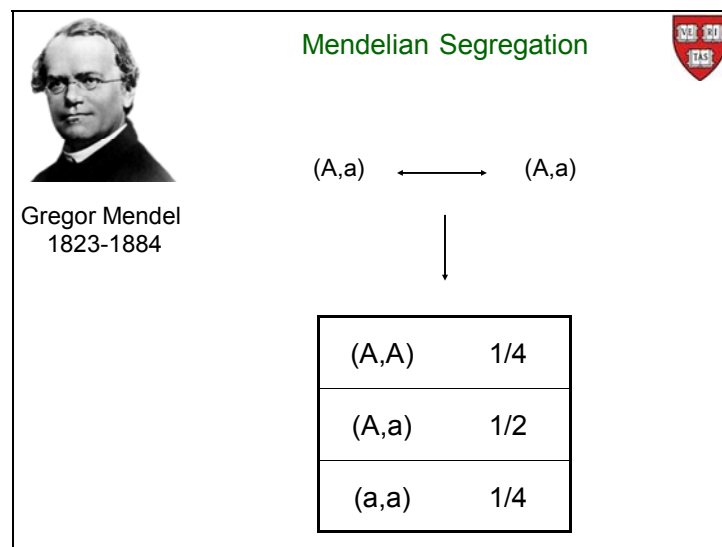


	Sperm	
	A	a
	1/2	1/2
Egg	A	Aa
	1/2	1/4
	a	aa
	1/2	1/4

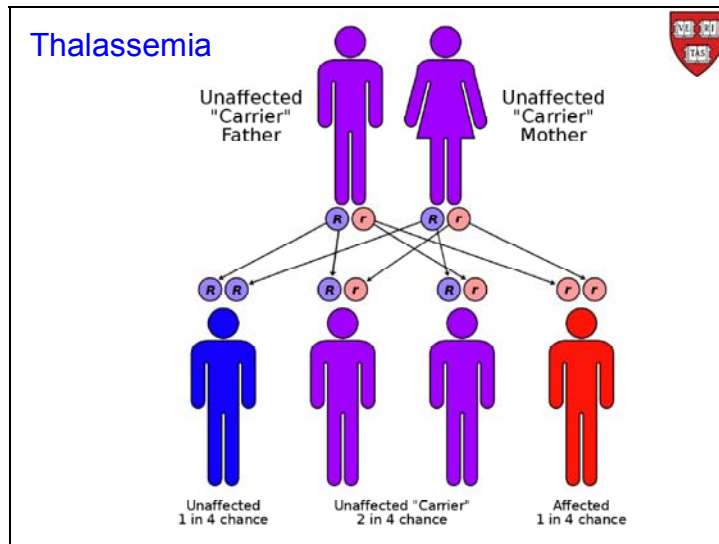
Probabilities

Here is one example of independence, and it has to do with Mendelian Segregation. If we have a sperm from a heterozygous male fertilizing an egg from a heterozygous female, then the first part of Mendelian Segregation tells us that it is equally likely that the sperm carries the dominant (so $\frac{1}{2}$ the time) as the recessive allele ($\frac{1}{2}$ the time), and it's equally likely to mate with the egg that carries the dominant ($\frac{1}{2}$) or the recessive ($\frac{1}{2}$) allele.

Step two says that you have independence. So what mates with what is independent of the allele in either the sperm or the egg. So the result is that you would get AA a quarter of the time (from the independence assumption $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$) so too with Aa, aA, and aa. Lastly, there is no difference between Aa and aA.



So finally, combining these four terms together, from a mating of heterozygous pairs, we should get dominant-dominant a quarter of the time, recessive-recessive a quarter of the time, and mixed half the time.



Indeed, this is what we see with thalassemia—or as it used to be called, Mediterranean anemia. This is what happens when you have an unaffected but carrier father, with an unaffected but carrier mother. It's a little deceptive because all the offspring are male in this diagram, but they don't have to be, of course. You can have any sex mixture of the offspring you want in there.

Sally Clark and Roy Meadow

Sally Clark was a British solicitor
 Had a son in September of 1996.
 He died in December of 1996.
 Had a son in November of 1997.
 He died in February of 1998.
 She is accused and tried for murder.
 Found guilty.

"Expert" witness, a pediatrician, Roy Meadow claims that the chance of two SIDS deaths in a family is "one in 73 million", and that carried the day.

Independent events are extremely important to us, and they play an integral role in a large number of models, but it is not always true that we have independent events. And if we falsely assume that there are independent events, then we can have an error that can have horrendous effects, such as what happened to Sally Clark.

Sally Clark was a British solicitor who had a son at the end of 1996. He died within a few months, in December of 1996.

She recovered from that event and had another son in November of 1997. This one died in February of 1998. She was accused of and tried for murder of both kids, and she was found guilty.

The defense argued that possibly these were cases of Sudden Infant Death Syndrome (SIDS) and so the prosecution called an expert witness—a pediatrician—one by the name of Roy Meadow, who went off and did some probability calculations, even though his expertise was not in probability but rather, in pediatrics. He guessed the chance that a particular family would have two SIDS deaths to be 1 in 73 million. He quite graphically explained that 1 in 73 million is like betting on 80 to 1 horses and having them win 4 races in a row³—in other words, 1 in 73 million is virtually impossible. This argument carried the day. She was found guilty, and she was imprisoned.

Meadow's Law:

one cot death is a tragedy, two cot deaths is suspicious and, until the contrary is proved, three cot deaths is murder.

The CESDI study looked at 472,823 live births.
363 deaths were identified as SIDS.

$$P(\text{SIDS}) = 363/472,823 = 1/1300$$

Meadow testimony

(i) $P(\text{SIDS}) = 1/8543$

(ii) $P(2 \text{ SIDS}) = (1/8543)^2 = 1/73 \text{ million.}$

Fleming P, Bacon C, Blair P, Berry PJ (eds). *Sudden Unexpected Deaths in Infancy, The CESDI Studies 1993-1996*. London: The Stationery Office, 2000.



³ <http://news.bbc.co.uk/2/hi/health/4432273.stm>

Meadow's law⁴ was that one cot death is a tragedy. Two cot deaths-- that's what the British call SIDS—two cot deaths is suspicious, and until the contrary is proved, three cot deaths is murder. But the question still remained how did he get the 1 in 73 million?

It is of note that there is the CESDI report that studied 472,823 live births at roughly the same period in England, between 1993 and 1996, and they found 363 deaths that were identified as SIDS, or cot deaths⁵. So using their empirical evidence, the chance of a cot death should be 1 per 1,300 births.

Meadow argued that this family was a middle class family—both husband and wife were solicitors—so he used the divisor of 8,543, instead of 1,300. It is difficult to determine the basis for this creation.

But then he made the independence assumption. It then followed that, because of independence, the probability of two SIDS is $1/8543$ squared, and that is how he got his 1 in 73 million.

In the CESDI report, they carried out a case-control study:



Among the 323 SIDS families studied, there were 5 previous SIDS,

$$P(\text{prev. SIDS in 323}) = 5/323 \approx 0.0155$$

Among the 1288 control families, there were 2 previous SIDS.

$$P(\text{prev. SIDS in 1288}) = 2/1288 \approx 0.00156$$

It is dangerous to be right when the government is wrong.
Voltaire 1694 - 1778

Ray Hill, Multiple sudden infant deaths – coincidence or beyond coincidence?
Paediatric and Perinatal Epidemiology 2004, 18, 320–326

If we return to the CESDI study, they have evidence of more than one SIDS death in a family. They used a case-control methodology to compare two groups, and what they found was that in the group with 323 SIDS deaths there were 5 previous SIDS, which


⁴ Ray Hill, *Paediatric and Perinatal Epidemiology* 2004, 18, 320–326

⁵ Fleming P, Bacon C, Blair P, Berry PJ (eds). *Sudden Unexpected Deaths in Infancy, The CESDI Studies 1993-1996*. London: The Stationery Office, 2000.

comes out to be 0.0155 and already raises some doubt about the 1 in 73 million figure. Then they compared this number to a control group of 1,288 families and found that the probability amongst that group of a previous SIDS was 0.00156.

So if there is one SIDS death already in the family, it is 10 times as likely that there will be another one. So this casts a doubt on Meadow's assumption of independence.

Voltaire had a saying for this one, too, "It is dangerous to be right when the government is wrong," and poor Sally Clark tragically suffered the consequences. She died shortly after they let her out of prison, which they did once the proper information was made available to the appeals court.



Clarification aid:

IF A and B are **mutually exclusive** then
(**Additive Law**)

$$P(A \cup B) = P(A) + P(B)$$

IF A and B are **independent** then
(**Multiplicative Law**)

$$P(A \cap B) = P(A) \times P(B)$$

Here is just a small clarification aid. We have introduced two situations where we have looked at properties of two events: one of them is when the two events are mutually exclusive, and then the additive law says that the probability of the union is the sum of the probabilities. The other dealt with independent events, in which case the multiplicative law says that the probability of the intersection is the product of the probabilities. Clearly, if two events are mutually exclusive they cannot be independent.

Students sometimes get a little bit confused about these two states. You can use as an aid to help you remember the distinction, union is like the addition of probability measures. Intersection is like the product of probability measures. So union is additive. You add things up and make them bigger when you unite them. Whereas, when you take an intersection it is like multiplying things and making them smaller. Just a little note to remind you, to make life a little bit easier for you.

Bayes' Theorem

Return to the multiplication rule and now look at $P(A | B)$:

$$\begin{aligned} P(A | B) &= \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(A) P(B | A)}{P(B)}, \end{aligned}$$

assuming $P(B) > 0$.

This is known as Bayes' Theorem

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A^c) \\ &= P(A) P(B | A) + P(A^c) P(B | A^c) \end{aligned}$$


Returning to the multiplication rule, let us look at the conditional probability, $P(A|B)$. It says that this probability is $P(B \cap A)$ divided by $P(B)$. Once again, I am assuming that $P(B) > 0$. Returning to the multiplicative rule we have that $P(B \cap A)$ can also be written as $P(A) P(B|A)$.

So after some straightforward algebra, we have on the left hand side a probability conditional on B, and on the right hand side the reverse: here we have that A is the conditioning event. This we are going to find very useful when we get to diagnostic testing.

This formula is called Bayes' Theorem. It is a beautiful theorem that is used repeatedly.

Sometimes it's not expressed this way, but rather the $P(B)$ in the denominator is expanded, as shown. The first line in the expansion shows the use of the additive law, and the second equality shows the use of the multiplicative law.

Diagnostic Tests



Diagnostic tests

$A = D = \text{"have disease"}$


$A^c = D^c = \text{"do not have disease"}$

$B = T^+ = \text{"positive screening result"}$

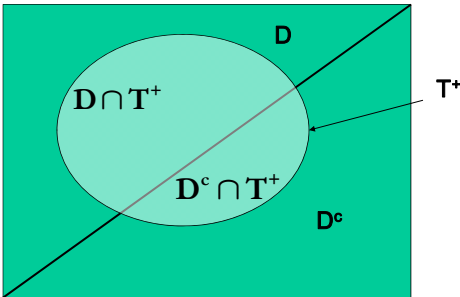
Find $P(D | T^+)$

In diagnostic tests we associate the event A with the event D , the event having the disease in question. The event B is replaced with the letter T to denote testing, and T^+ will denote testing positive for the disease.

Ideally if the test were perfect, anyone with the disease would test positive, and only people with the disease test positive. Unfortunately, there is little perfection in the world and that is certainly true of biological tests. We are thus interested how good a test is. Specifically, in the field we are interested in the probability that someone who tests positive for a condition or disease actually has that condition or disease.



Venn diagram of Bayes' Theorem


$$P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{P(D \cap T^+)}{P(D \cap T^+) + P(D^c \cap T^+)}$$

With a perfect test, of course, the probability we seek is one. But to see what happens with an imperfect test let us go back to our Venn diagram. For convenience, label the upper triangle, D , so it denotes the people with the disease. The lower triangle shows the people without the disease.

Now place an ellipse on the space to show the people who test positive. The diagonal line also intersects the ellipse, with the people in the ellipse above the line being those with the disease who test positive ($D \cap T^+$), and those in the ellipse but below the diagonal the ones who do not have the disease but test positive nonetheless ($D^c \cap T^+$).

If we had a perfect test, there would be nobody without the disease who tests positive, and there would be nobody with the disease who does not test positive.

The people who test positive who are not diseased ($D^c \cap T^+$) are called false positives. They shouldn't be testing positive. They should be testing negative. Those who have the disease but are not testing positive are called the false negatives.

Bayes' Theorem is displayed at the bottom of the diagram.

Prior to testing				
	Has Disease D	Disease Free D^c		
Test Positive T^+	$P(T^+ D)$ sensitivity	$P(T^+ D^c)$		$P(T^+)$
Test Negative T^-	$P(T^- D)$	$P(T^- D^c)$ specificity		$P(T^-)$
	$P(D)$ prevalence	$P(D^c)$ $= 1 - P(D)$		

Prior to using the test in the field we can quantify the properties displayed above. Create a so-called two-by-two table by cross-classifying individuals according to two binary variables: disease status and result of the test. Now fill the four cells of the table with the probability of being in that cell. Complete the table by filling in the probabilities of the margins. In the top left hand corner we have $P(T^+ | D)$, and we call that the sensitivity of

the test—how sensitive it is at detecting the disease. In the bottom right hand corner we have the $P(T|D^C)$, and we call this the specificity—how specific is the test to the condition we are investigating, in other words will it just go positive for any condition, or just specifically for the one for which it was designed to test.

The column sums of the two cell probabilities are both one.

As often happens, the cells we name, are the ones where we want to achieve high values. So we would like tests with as high as possible sensitivities and specificities.

The people with the disease are sometimes called prevalent cases, and the probability of being in that state is called the prevalence.

The sensitivity and specificity of a test can be established before the test is used in the field, and guidance is available from the Food and Drug Administration⁶. Local conditions will determine the prevalence. All three of these quantities are of importance when testing an individual.

Post testing				
	Has Disease D	Disease Free D^C		
Test Positive T^+	$P(D T^+)$ PPV	$P(D^C T^+)$		$P(T^+)$
Test Negative T^-	$P(D T^-)$	$P(D^C T^-)$ NPV		$P(T^-)$ $=1 - P(T^+)$
	$P(D)$ prevalence	$P(D^C)$ $=1 - P(D)$		


PPV=positive predictive value NPV=negative predictive value

Post testing we can replace the probabilities in the two-by-two table. The margins remain the same. Prior to testing the column classification determines the conditioning event in the cell probabilities, whereas post-testing, the row classifier determines the conditioning. In the top left-hand corner we have the probability that someone who tests positive actually has the disease, and this is called the positive predictive value. In the

⁶ <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm>

bottom right-hand corner we have the probability that someone who tests negative does not have the disease. This is called the negative predictive value.

For this table the row sums are one.



From Bayes' theorem:

Positive predictive value:

$$P(D|T^+) = \frac{P(D) P(T^+ | D)}{P(D)P(T^+ | D) + P(D^c)P(T^+ | D^c)}$$
$$= \frac{\text{prevalence} \times \text{sensitivity}}{\text{prev} \times \text{sens} + (1 - \text{prev}) \times (1 - \text{specificity})}$$

As often happens, the diagonal entries are important. But the point to remember is that prior to testing, we have one set of measures; sensitivity, specificity. Post testing, we have another set of measures; the positive predictive value and the negative predictive value. What ties these quantities together is Bayes' theorem.

Sensitivity and Specificity

Example: X-ray screening for tuberculosis

X-ray	Tuberculosis		Total
	Yes	No	
Positive	22	51	73
Negative	8	1739	1747
Total	30	1790	1820

Sensitivity = $\frac{22}{30} = .7333$

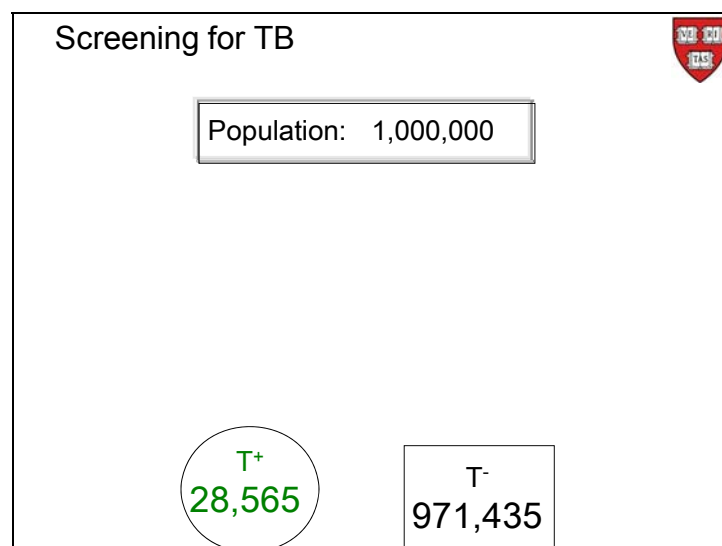
Specificity = $\frac{1739}{1790} = .9715$

As an example of imperfect testing, suppose we want to use X-rays to screen for tuberculosis. In this study 30 people who had tuberculosis, had their X-rays read and 22 tested positive. Unfortunately 8 tested negative—the false negatives.

So the proportion of those with TB who tested correctly is 0.733 and that we can use as an estimate of the sensitivity.

For the 1790 without TB, 1739 correctly tested negative, but unfortunately 51 tested positive—the false positives. So we estimate the specificity to be .9715.

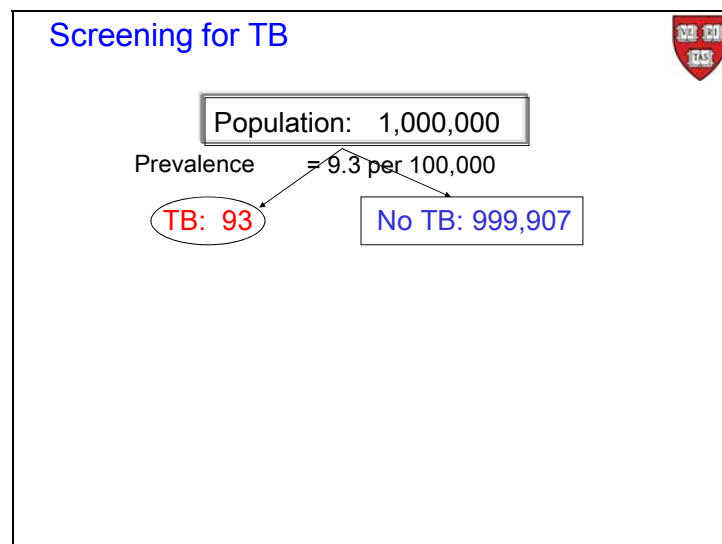
So this study establishes the properties of the testing procedure.



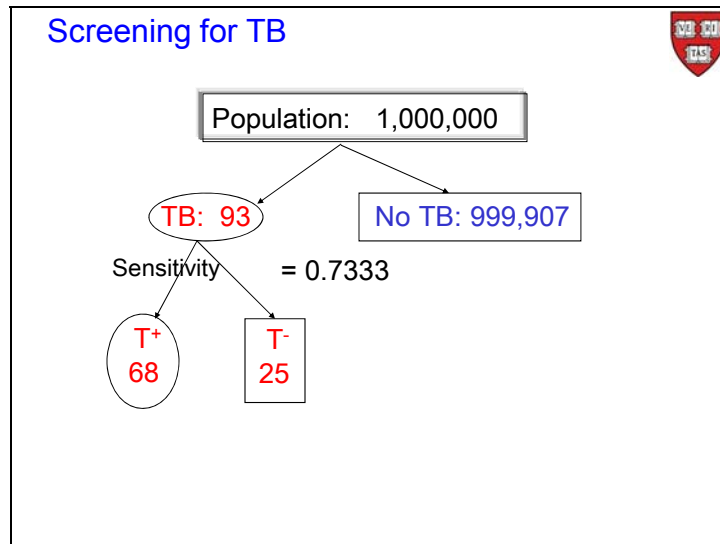
Having established the sensitivity and specificity, let us see what happens when the procedure is used in the field.

Suppose we have a population of a million people to be screened for TB and we use this test. When we implement a screening we find that 28,565 of this population actually tested positive, and the other 971,435 tested negative.

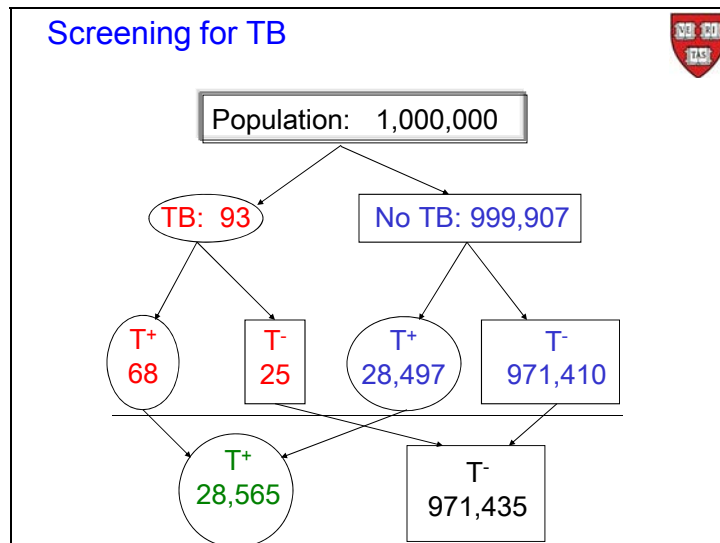
Before understanding the procedure, there is a third quantity we need, and that is the prevalence. How many of these one million people actually have TB? This thinking may seem circular, since determining how many people have TB is the point of the screening. But what we hope to do is look behind the scenes to see how these numbers are generated.



Let us suppose, for argument's sake, that the prevalence is 9.3 per 100,000. So that means 93 per million. We know that the test acts differently on those with TB and those without TB.

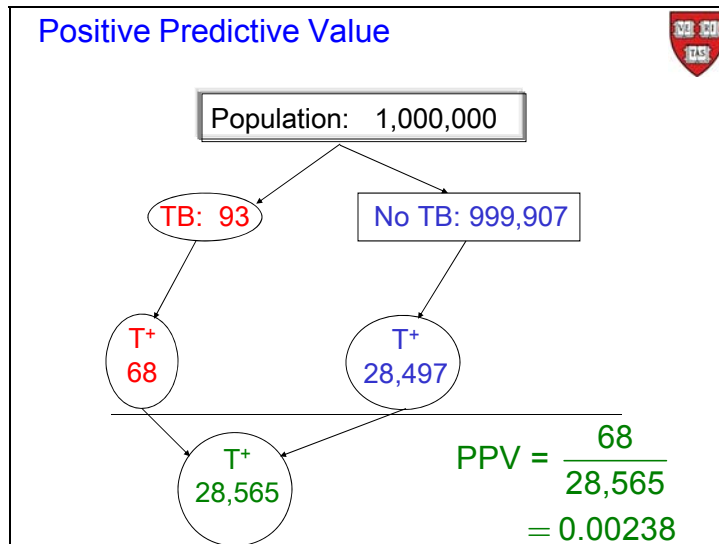


What would happen to these 93 when tested. The sensitivity is 0.7333 so 88 (=93x0.7333) would test positive and the other 25 would test negative.



For those without TB, we need to look to the specificity and see that of the 999,907 without TB, 971,410 (=999,907x0.9715) would test negative. That means the other 28,497 test positive. This is a large number of false positives.

We see the impact of this when we draw the line and see that beneath the line, the 28,565 who test positive are mostly (28,497) false positives. The total number of negatives is 971,435.



To determine the positive predictive value, we look at what proportion of those testing positive are true positives. In this case it is 68 of the 28,565. So the positive predictive value is 0.00238. That means that about 2 per thousand of those who test positive have TB.

Prior to the test we have:

$$P(D) = \frac{93}{1,000,000} = 0.000093$$

Post(erior) to the test we have:

$$P(D | T^+) = \frac{68}{28,565} = 0.00238$$

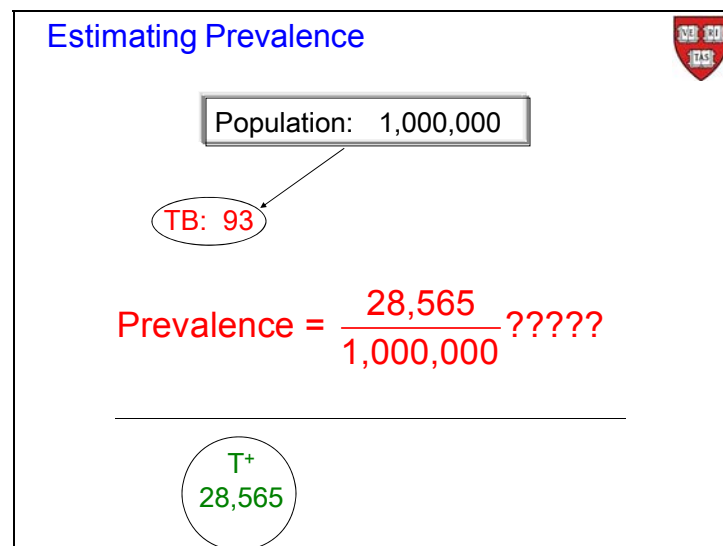
Ratio:

$$\frac{0.00238}{0.000093} = 25.6$$

The problem here is that we are looking at a very rare disease. So if the specificity is not one we stand to observe a relatively large number of false positives, irrespective of how good the sensitivity is. A fairer way to appraise this number is to compare our ability to detect a case of TB before the screening (9.3 per 100,000, or about 1 in 10,000) with our ability after the test (approximately 2.4 per 1,000). So our detection capability has been increased by a factor of 25.6 (taking the ratio of pre to post probabilities of detection.)

If now we have a second test we can apply to those who screened positive on the X-ray test, then we can increase our detection probability even further since now we start with a prevalence of 2.38 per 1,000 as opposed to 9.3 per 100,000.

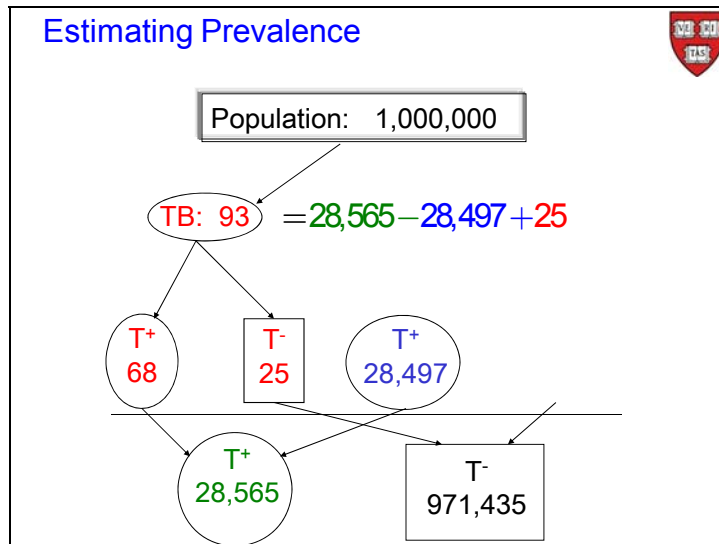
Estimating prevalence



Often, as a result of screening we would like to estimate the prevalence of the disease. Indeed, what public health measures one takes might be dependent upon the prevalence. For example, if we are concerned with controlling schistosomiasis, then what the WHO prescribes as a public health measure very much depends on the prevalence of the disease.⁷

Given the results in our screening, how would we estimate the prevalence? If we divide all who tested positive by how many were tested, so $28,565/1,000,000$ we would estimate the prevalence to be 28,565 per million and be off by a factor of 275 (= $28,565/93$) because we do not have a perfect test.

⁷ Page 42, Table A2.2 in http://whqlibdoc.who.int/publications/2006/9241547103_eng.pdf




To see how to obtain a good estimate of the prevalence, let us backtrack over the way the 28,565 were generated. From these 28,565 we would need to subtract the false positives (28,497) and add the false negatives (25) we lost. If we did that we would reduce the 28,565 down to 93.

$$\begin{aligned}
 93 &= 28,565 - 28,497 + 25 \\
 &= 28,565 - \\
 &\quad \{(1 - \text{prev}) \times (1 - \text{spec})\} 1,000,000 + 25
 \end{aligned}$$

We obtain the false positives from the $1,000,000 \times (1 - \text{prevalence})$ who did not have the disease, and then multiply those by $(1 - \text{specificity})$ to see how many tested positive.

Formula for estimating prevalence




$$\begin{aligned}
 93 &= 28,565 - 28,497 + 25 \\
 &= 28,565 - \{(1 - \text{prev}) \times (1 - \text{spec}) - \\
 &\quad \text{prev} \times (1 - \text{sens})\} 1,000,000 \\
 \text{prev} &= \frac{\frac{28,565}{1,000,000} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}
 \end{aligned}$$

$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$

The 25 false negatives, on the other hand we got from the $1,000,000 \times \text{prevalence}$ who tested negative, and then tested negative, so multiply by $(1 - \text{sensitivity})$. Dividing through by 1,000,000 we get the formula above.

Detection Limit

Conditions for formula to make sense



$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$

$$\begin{aligned}
 \text{"prop +ve"} - (1 - \text{spec}) &\geq 0 \\
 \text{"prop +ve"} &\geq (1 - \text{spec})
 \end{aligned}$$

$$\begin{aligned}
 \text{sens} - (1 - \text{spec}) &\geq 0 \\
 \text{or} \\
 \text{sens} + \text{spec} &\geq 1
 \end{aligned}$$

We can think of our testing as utilizing an instrument, which like any other measurement instrument has its precision. Just like a cheap ruler to measure the length of an object


might only measure to the closest inch. On the other hand, if you had an electron microscope to measure the length of your object, you would have a much more precise measurement.

So what is the detection limit of our testing device? We can look at the prevalence formula and we know that prevalence has to be non-negative. That means that the numerator and denominator must both be of the same sign. Consider the case when they are both positive, leaving the other case for you to ponder.

Consider first the numerator. The proportion positive must thus be greater than one minus the specificity.

For the denominator to be positive we need that the sensitivity plus the specificity of the test must sum to more than one. Since we can get a sum of one by spinning a coin, this is not asking too much of our test. So let us assume that this is always the case. Indeed, with the X-ray we have a sensitivity of 0.733 and a specificity of 0.97, so their sum is indeed greater than one.

So the final constraint to ensure that we get a positive estimate of prevalence requires that the proportion positive be greater than one minus the specificity, and that is the lower bound to our detectability.


HIV newborn screening New York 11/87—3/90 			
Region	Positive	Tested	Percent
NYS not NYC	601	346,522	0.17
NYC Suburban	329	120,422	0.27
Mid-Hudson	71	29,450	0.24
Upstate Urban	119	88,088	0.14
Upstate Rural	82	108,562	0.08
New York City	3650	294,062	1.24
Manhattan	799	50,364	1.59
Bronx	998	58,003	1.72
Brooklyn	1352	104,613	1.29
Queens	424	67,474	0.63
Staten Island	77	13,608	0.57

Here is an example of an HIV screening carried out amongst newborns in New York State, in the period November '87 through March '90. Now this is old data because they used to test all babies at birth, but then some politicians got involved and, as a result, this screening is no longer carried out and we do not have this monitoring information.

Over the period when we had the data, they reported separately for two regions: New York State (NYS), not New York City(NYC); and NYC. In NYS not NYC they tested 601 positive babies of the 346,522 tested. So the percent positive was 0.17%.

In NYC itself, there were 3,650 babies tested positive of the 294,062 births. So there, the rate was 1.24%.

Detection limit of instrument


$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$
$$\text{"prop +ve"} - (1 - \text{spec}) \geq 0$$
$$\text{"prop +ve"} \geq (1 - \text{spec})$$

So, for Upstate Urban NY where 119 tested positive out of 98,088 = 0.14% we need a specificity of better than

$$1 - 0.0014 = 0.9986 \text{ or } 99.86\%$$

Drilling down on these numbers, we can ask the question of whether some of these numbers truly represent infected babies. Our detection limits are such that the observed ratio must be greater than one minus the specificity. So for Upstate Urban NY where 119 tested positive of the 98,088 babies (=0.14%), that means that for those to be true positives would require that we have a specificity better than 99.86% (1-0.14%). This is an unrealistically large number for the specificity. So we don't expect that, indeed, these 119 were truly positive.

ROC

Cotinine Level (ng/ml)	Smokers
0--13	78
14--49	133
50--99	142
100--149	206
150--199	197
200--249	220
250--299	151
300+	412
Total	1539

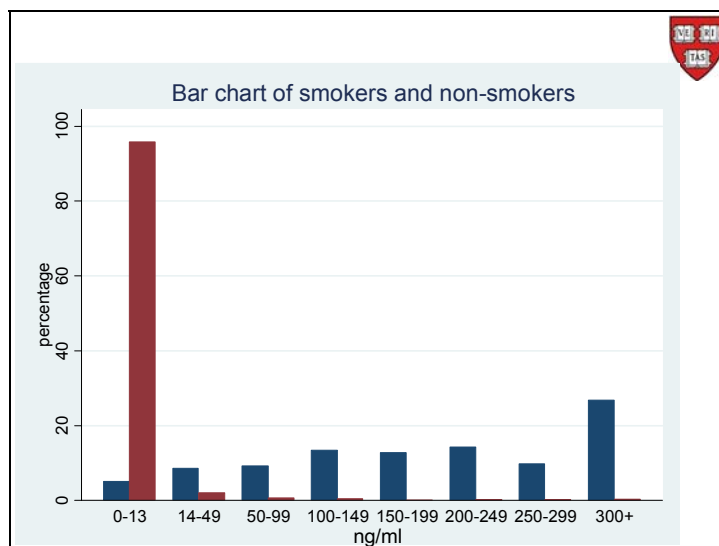
Continuing our study of imperfect tests, let us look at the rather common situation where a test is designed by measuring a certain biological quantity. If the test is applied to a sample that measures above a certain threshold then the test is declared positive, if below, then the result is negative. The determination of the threshold requires judgment. If set too high it would result in a number of false negatives, whereas if it is set too low it would result in a large number of false positives.

Here is a case in point. Cotinine is a metabolite for nicotine. Something like 40% of the nicotine metabolizes to cotinine. So measuring the cotinine level, in an individual's blood is a more reliable method than self-identification as a way to classify an individual as a smoker or non-smoker, so the theory goes.

Here are cotinine levels in this study of 1,539 identified as smokers. And 78 of them had very low cotinine levels, something less than 14 nanograms per milliliter (ng/ml), 133 of them had 14ng/ml to 49ng/ml. 142 of them have 50ng/ml to 99ng/ml, et cetera. So this is the distribution of cotinine level in smokers.

Cotinine Level (ng/ml)	Smokers	Non-smokers
0--13	78	3300
14--49	133	72
50--99	142	23
100--149	206	15
150--199	197	7
200--249	220	8
250--299	151	9
300+	412	11
Total	1539	3445

They also studied 3,445 non-smokers and found that 3,300 had less than 14 ng/ml in their system. They also found 72 with levels between 14ng/ml and 49 ng/ml, etc. as displayed above.



So if we draw the bar graph, the red bars are the non-smokers and the blue bars the smokers. To differentiate between the two on the basis of cotinine measure, it seems sensible to have a cutoff and if a person has cotinine below the cutoff, then classify that person as a non-smoker, and if above the cutoff, then classify that person as a smoker. Moving the cutoff right should mean more smokers will be falsely classified as non-smokers, and moving the cutoff to the left should mean that non-smokers will be falsely classified as smokers. A judgment has to be made about the judicious placement of the cutoff. Usually what enter into consideration for making this judgment are the consequences of potential errors.

The impact of the placement very much depends on the context. For example, in donated blood, then each unit of blood typically might impact some eight individuals because of the use of blood products. Thus when testing the donated blood, a false negative test for HIV or Hepatitis B, for example, might result in eight different people being infected as a result. The consequences of a false negative are thus dire.

On the other hand, a donation falsely labeled positive will result in the loss of one unit of blood. Of course, one must not ignore the false classification of the blood donor, but that can be rectified with further testing. The consequences of a unit of blood being falsely being labeled positive, do not seem to be as serious as a false negative.

In these examples, false negatives are highly consequential, whereas false positives might not be, but this is not always the direction of the imbalance. For example, with antenatal tests, one must be extremely careful about false positives⁸.

Cotinine Level (ng/ml)	Smokers	Non-smokers
0--13	78	3300
14--49	133	72
50--99	142	23
100--149	206	15
150--199	197	7
200--249	220	8
250--299	151	9
300+	412	11
Total	1539	3445

Returning to the cotinine numbers, consider the uncertainty in labeling by calculating the false positives and false negatives. Suppose we draw the cutoff line between 13 and 14, and call everybody below 14 a non-smoker and above 13 a smoker.

⁸ <http://www.nlm.nih.gov/medlineplus/prenataltesting.html>

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	
50--99	142	23	
100--149	206	15	
150--199	197	7	
200--249	220	8	
250--299	151	9	
300+	412	11	
Total	1539	3445	

In that case, the sensitivity of this procedure would be, as measured by these data, that we would lose 78 of the 1,539 smokers. That leaves 1,461 which over 1,539 gives us the sensitivity associated with this cutoff.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	1328/1539
50--99	142	23	
100--149	206	15	
150--199	197	7	
200--249	220	8	
250--299	151	9	
300+	412	11	
Total	1539	3445	

What if we move the line down and use between 49 and 50 as our cutoff? Less than 50, would be called a non-smoker, and more than 49, would be called a smoker. Then we

will lose a further 133 smokers, and the sensitivity would now be 1,328 divided by 1,539.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	1328/1539
50--99	142	23	1186/1539
100--149	206	15	980/1539
150--199	197	7	783/1539
200--249	220	8	563/1539
250--299	151	9	412/1539
300+	412	11	
Total	1539	3445	

We can continue this logic line by line all the way down the table to obtain this table.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	0.95
14--49	133	72	0.86
50--99	142	23	0.77
100--149	206	15	0.64
150--199	197	7	0.51
200--249	220	8	0.37
250--299	151	9	0.27
300+	412	11	
Total	1539	3445	

Expressing these fractions as decimals, we see that the sensitivity goes from 0.95 to 0.86, to 0.77, to 0.64, and so on. So as we move the line down, our sensitivity decreases. But what you gain in the roundabout, you should be losing on the swings or vice versa, so let us look at what happens to the non-smokers—let us calculate the specificities.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	
50--99	142	23	0.77	
100--149	206	15	0.64	
150--199	197	7	0.51	
200--249	220	8	0.37	
250--299	151	9	0.27	
300+	412	11		
Total	1539	3445		

In anticipation of the next graph, let us not calculate the specificities but rather calculate one minus the specificity—namely, look at the proportions of the non-smokers who get incorrectly classified. So, starting at the top again, if the cutoff is between 13 and 14, then all but 3,300 of the 3,445 non-smokers would be correctly classified, or 145/3,445 would be the proportion incorrectly classified.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	73/3445
50--99	142	23	0.77	
100--149	206	15	0.64	
150--199	197	7	0.51	
200--249	220	8	0.37	
250--299	151	9	0.27	
300+	412	11		
Total	1539	3445		

Moving the cutoff down one line to between 49 and 50, then the proportion of non-smokers who would be misclassified would be 73/3,445.

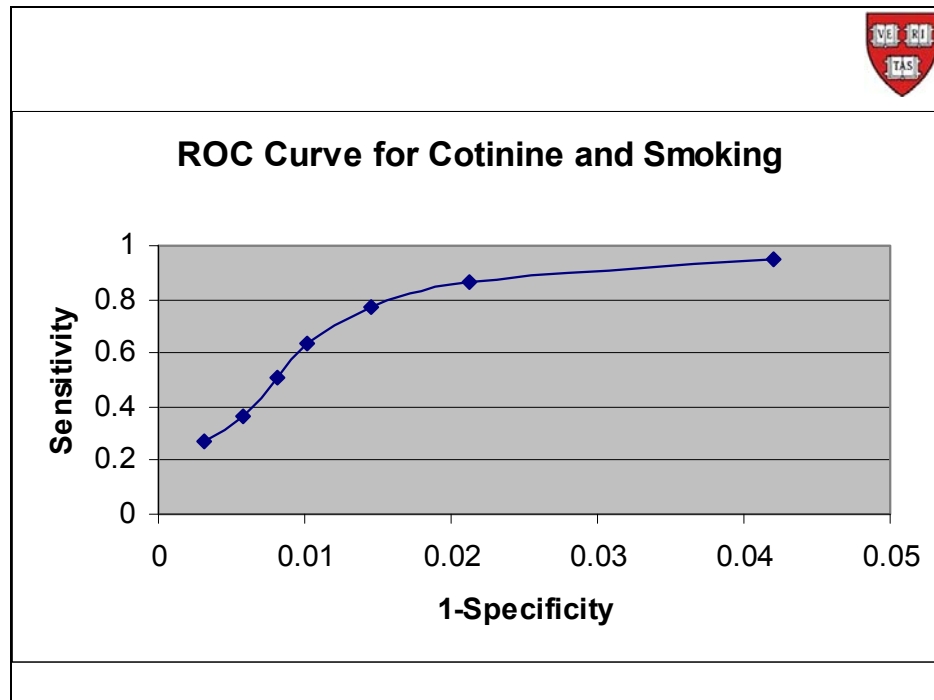
Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	73/3445
50--99	142	23	0.77	50/3445
100--149	206	15	0.64	35/3445
150--199	197	7	0.51	28/3445
200--249	220	8	0.37	20/3445
250--299	151	9	0.27	11/3445
300+	412	11		
Total	1539	3445		

Moving down the table one line at the time we finally get this table.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	0.04
14--49	133	72	0.86	0.02
50--99	142	23	0.77	0.01
100--149	206	15	0.64	0.01
150--199	197	7	0.51	0.01
200--249	220	8	0.37	0.006
250--299	151	9	0.27	0.003
300+	412	11		
Total	1539	3445		

Finally, replacing all the fractions with their decimal equivalents we get this table. We see that as the cutoff is made higher (going down the table) the specificity goes up (one minus the specificity goes down) as, as we observed previously, the sensitivity goes down, as our intuition told us should happen.

We can plot the last two columns of the table against each other:

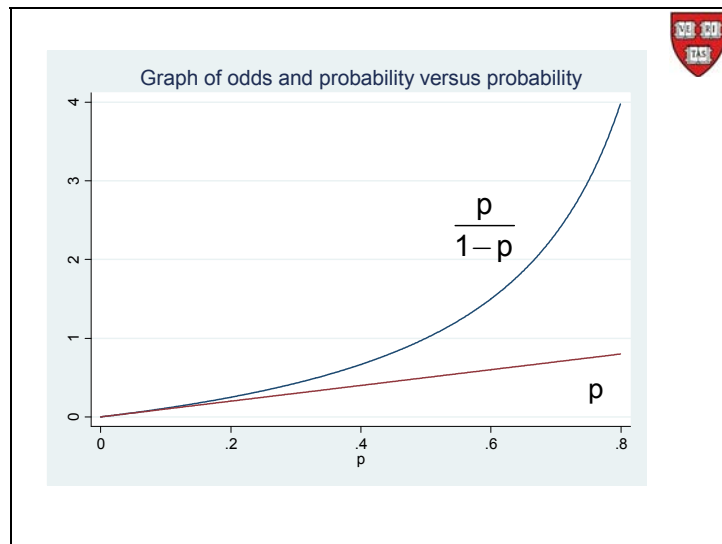


This plot is called the ROC, or Receiver Operator Characteristic, curve. This label comes from World War II signal detection.

This particular example is stopped at 0.05 on the horizontal axis, but it can continue all the way up to 1.

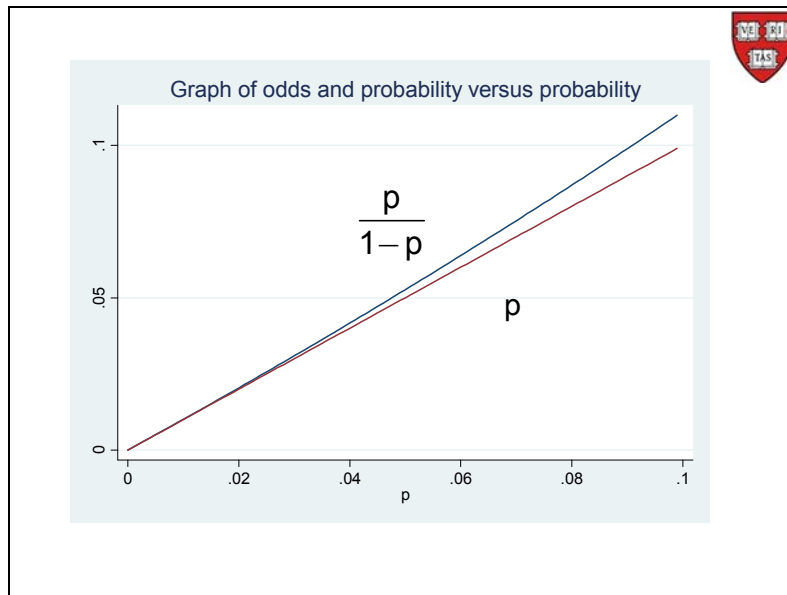
The ideal ROC would be zero at the origin, and one elsewhere. This one is not ideal, of course, few if any of any interest are, but it is quite good. To evaluate a test, or two compare two testing procedures, it is common practice to look at the area under the ROC curve (AUC), with the curve extending all the way to one on the horizontal. Presumably, when comparing two tests, the test with higher AUC is to be preferred.

Probability and Odds



Let us briefly contrast probability with odds. If the probability of an event is p , then we said the odds of the event would be p over 1 minus p .

Mathematically, what happens to p over 1 minus p , as p gets close to 1 is that it goes to infinity, so above it is only drawn to $p=0.8$. In general, as p gets large the two, the probability and the odds, diverge. What is interesting is to look at them for small p when the two are close in value to each other.



Amplifying the left side of the last graph, we have this graph that only extends to when p equals 0.1. We see that they do start separating a little bit, but that for small p they are close to each other.

p	$p/(1-p)$ = odds	odds- p = Δ	Δ / p %	Δ / odds %
0.02	0.02	0.00	2.04	2.00
0.03	0.03	0.00	3.09	3.00
0.04	0.04	0.00	4.17	4.00
0.05	0.05	0.00	5.26	5.00
0.06	0.06	0.00	6.38	6.00
0.07	0.08	0.01	7.53	7.00
0.08	0.09	0.01	8.70	8.00
0.09	0.10	0.01	9.89	9.00
0.10	0.11	0.01	11.1	10.00

Here, when in tabular form, we see that when p is less than 0.07, the difference between the two is less than 0.005. For p less than or equal to 0.10, then even the relative difference is about 11% or less. The two are not the same, but when talking about rare events ($p < 0.1$), there is not much difference between the two.

Venn Diagram Tattoos

