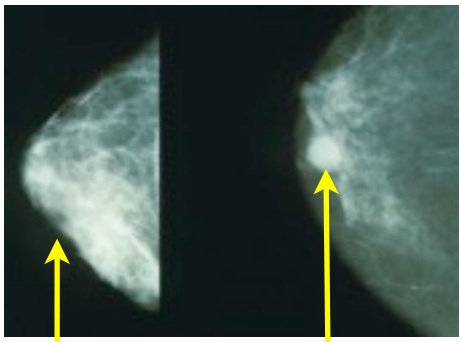
OVERTURNING A TREE

An everyday event: a forty year old woman has a mammography done. There is no evidence of cancer whatsoever but she knows that 1 % of the women her age will develop breast cancer and she wants to make sure. The physician finds some suspicious spots on the X-rays, so the test is "positive" in the medical jargon. But now the key question: what is the probability that this woman has got breast cancer? Has the woman reason to be really worried?
Even physicists have great difficulties in giving the correct figures. A survey of the Max-Planck-Institut für Bildungsforschung of 2001 with physicians shows that you get any answer between 1 and 90 %, the average being 70 %.
What is wrong here? How can people who know the test very well disagree to such an extent?

Mammography is the process of using low-dose X-rays to examine the human breast and is used as a diagnostic as well as a screening tool. The goal of mammography is the early detection of breast cancer, typically through detection of characteristic masses and/or microcalcifications.
In many countries routine mammography of older women is encouraged as a screening method to diagnose early breast cancer. The US Preventive Task Force recommends screening mammography every 1–2 years for women aged 50 and older. Altogether clinical trials have found a relative reduction in breast cancer mortality of 20%, but the two highest-quality trials found no reduction in mortality. Mammograms have been controversial since 2000, when a paper highlighting the results of the two highest-quality studies was published.



No cancer. Cancer, made visible by an increase of the tissue's density.

Mammography is a **medical test** designed to diagnose breast cancer. In this context a "positive" result means that the test detected what it was supposed to detect, a "negative" result means that it could not detect it. Like any other test mammography is not absolutely reliable, as it has two possibilities to fail:
it can be **false negative** or **false positive**.

1) **false negative**
"False negative" means that the test result is negative but wrong. In other words, the woman is told to be free of cancer under the condition that she has got cancer.

1) **false positive**
"False positive" means that the test result is positive but wrong. In other words, the woman is told to have cancer under the condition that she is free of cancer.

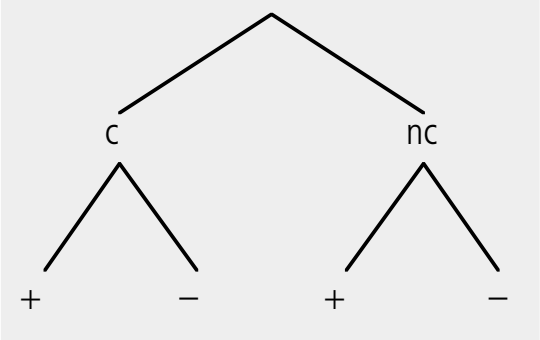
Mammography has a false-negative (missed cancer) rate of around 20 percent. This is partly due to dense tissues obscuring the cancer and the fact that the appearance of cancer on mammograms has a large overlap with the appearance of normal tissues.
For the same reasons mammography has a false-positive rate of around 10 percent.
That means that

if a woman has **got** breast cancer mammography will detect it with a probability of 80 % and miss it with a probability of 20 %.
if a woman has **not** got breast cancer mammography will declare the woman cancer free with a probability of 90 % and declare her suspicious with a probability of 10 %.

Like any other medical test mammography can be treated like a compound probability experiment:

1st level: cancer (c) or no cancer (nc)

2nd level: test positive (+) or negative (-)



All probabilities in the tree are well known:

$p(c) = 1\% = 0.01$

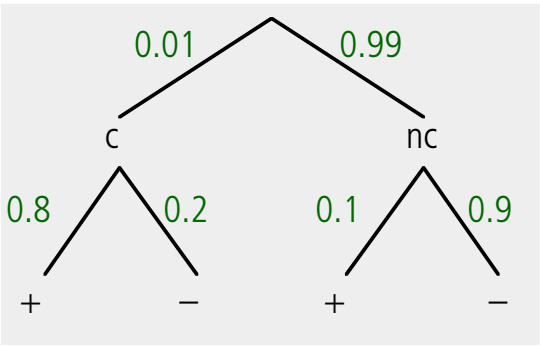
$p(nc) = 1 - 1\% = 0.99$

$p(+|c) = 80\% = 0.8$

$p(-|c) = 20\% = 0.2$

$p(-|nc) = 90\% = 0.9$

$p(+|nc) = 10\% = 0.1$



This is the probability of a positive test result **under the condition that** the patient has got cancer.

DEFINITION
If A and B are two events, then the **conditional probability of B given event A** is

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

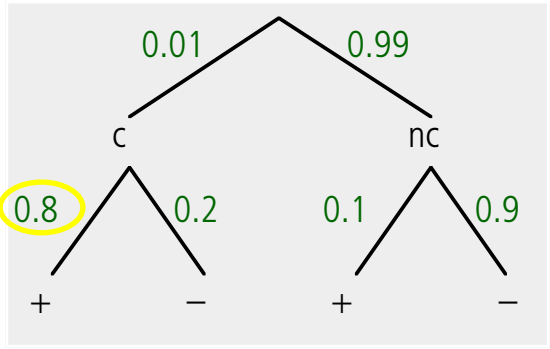
NB

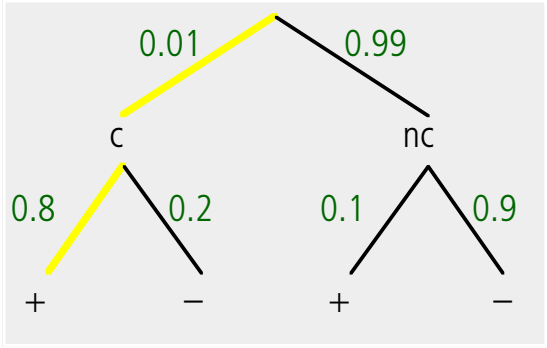
1) We are used to conditional probabilities: all probabilities in a tree but the ones on the first level are conditional probabilities, e.g. $p(+|c)$.

2) This definition is the deeper reason for the multiplication rule (see workshop 6), which states that the probability of a whole branch from top to bottom is calculated by multiplying all green probabilities along the branch.
The multiplication rule can be derived directly from the definition, e.g.

$$p(+|c) = \frac{p(c \cap +)}{p(c)}$$

$$\Leftrightarrow \quad p(c \cap +) = p(c) \cdot p(+|c)$$





1 What is the percentage of all women who underwent a mammography and
a) have cancer and are tested positive?
b) have cancer and are tested negative?
c) are tested positive and have no cancer?
d) are tested negative and have no cancer?

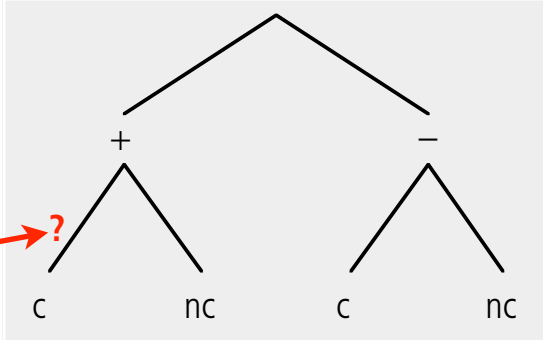
The positive tested woman is interested in the answer to a completely different question:

What is the probability that a woman who is tested positive really has cancer?

$$p(c|+)=?$$

The problem is that the order of the events is overturned, + is now the condition under which the probability of c is searched for. That means that in the tree the first level now is + or -, the second level c or nc. The tree is overturned.

What are the probabilities along the branches, especially this one?

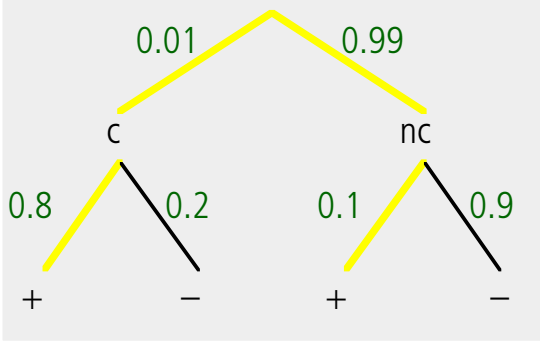


The old tree provides all the necessary information:

$$p(+)=p(c)\cdot p(+|c)+p(nc)\cdot p(+|nc)$$

$$=0.01\cdot 0.8+0.99\cdot 0.1=0.107$$

$$p(c\cap +)=p(c)\cdot p(+|c)=0.01\cdot 0.8=0.008$$



The definition of the conditional probability states $p(B|A)=\frac{p(A\cap B)}{p(A)}$.

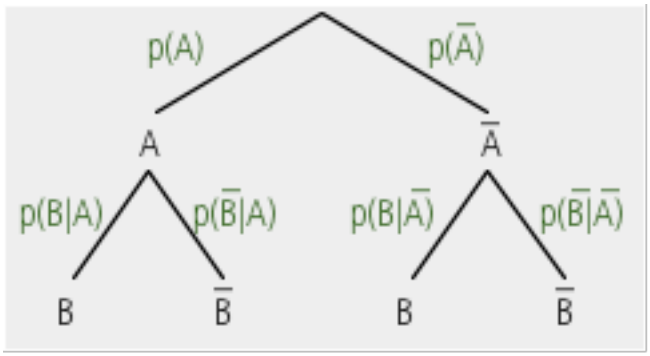
Our patient is interested in $p(c|+)$. With A = + and B = c one gets $p(c|+)=\frac{p(+\cap c)}{p(+)}$.

Because $p(+\cap c)=p(c\cap +)$ one gets $p(c|+)=\frac{p(c\cap +)}{p(+)}=\frac{0.008}{0.107}=\mathbf{0.075}$

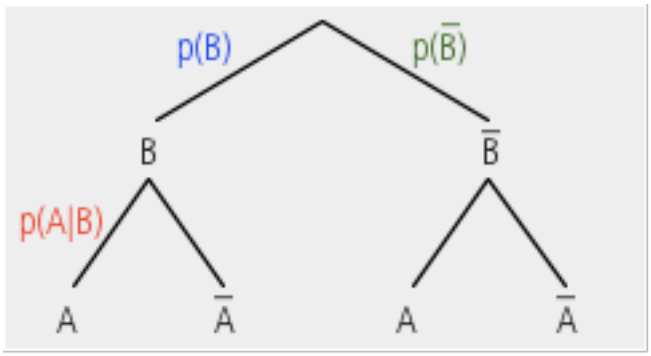
That means that out of 1000 women who are tested positive only **75** really have got breast cancer! In other words: a positive tested woman has a chance of a mere **7.5 %** to have cancer.

The example above can be generalised.

Starting point is a compound experiment with two levels:



Goal is a tree with inverted levels, i.e. with inverted conditions:



All the probabilities in this tree have to be calculated, especially $p(A|B)$.

With the first tree $p(A \cap B) = p(A) \cdot p(B|A)$

$$p(B) = p(A) \cdot p(B|A) + p(\bar{A}) \cdot p(B|\bar{A})$$

With the second tree $p(B \cap A) = p(A \cap B) = p(B) \cdot p(A|B)$

$$\Leftrightarrow \quad p(A|B) = \frac{p(A \cap B)}{p(B)}$$

BAYES' THEOREM

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

2 What is the probability of a woman who underwent a mammography and was tested negative to have breast cancer?

NB 1

There is a completely different approach to inverting the conditions. Instead of a tree and Bayes' Theorem a table for all possibilities with the events cancer/no cancer and test result +/– is used:

	c	nc	total
+			
–			
total			

Then a figure, e.g. 1000 is filled into the last cell. Such we are looking at a sample of 1000 women undergoing a mammography:

	c	nc	total
+			
–			
total			1000

Now with the known probabilities all the figures are calculated and filled into the corresponding cells:

	c	nc	total
+			
–			
total	10	990	1000

	c	nc	total
+	8	99	
–	2	891	
total	10	990	1000

	c	nc	total
+	8	99	107
–	2	891	893
total	10	990	1000

Number of women who are tested positive = 107
Number of women who are tested positive and have cancer = 8

$$\Rightarrow \quad p(c|+) = \frac{8}{107} \approx \mathbf{0.075}$$

	c	nc	total
+	8	99	107
–	2	891	893
total	10	990	1000

3 What is the probability of a woman who underwent a mammography and was tested negative to have breast cancer, calculated with tables?

NB 2

All medical tests are not completely reliable, the results are prone to be false negative or false positive. A good test has a low false negative **and** a low false positive rate. Unfortunately decreasing the false negative rate increases the false positive rate and vice versa.
A test, for example, that declares **all patients positive** (i.e. to have the disease) has a false negative rate of 0 but of course a false positive rate of 1!
A test that declares **all patients negative**, i.e. to be healthy, has a false positive rate of 0 but a false negative rate of 1!

4 Prove the statement above!
Breast cancer attacks 1 % of all women between 40 and 60.
a) A fictional test declares all patients positive. Calculate its false positive rate.
b) A fictional test declares all patients negative. Calculate its false negative rate.

5 Statistics shows that a jewellery shop in a certain district of a town has a chance of 0.1 % to be robbed every night.
The owner decides to have an alarm system installed. If a burglary takes place the system alerts the police in 99 out of 100 cases. Unfortunately the system is so sensitive that it goes off without burglary too, but only once every 200 nights.
Evaluate the probability for one night that the system alerts the police.

6 Much more interesting for the police is the question:
The alarm goes off. What is the probability that a burglary is taking place?

With rare events most alarms are false alarms!

7 A factory produces semiconductors. Experience shows that 1 out of 500 produced semiconductors does not work properly. To maintain the company's good reputation only "good" semiconductors should leave the factory. Therefore each semiconductor has to pass a final check. The "bad" ones are thrown into a bin, the "good" ones are packed to be sent to the customers.
Nevertheless 2 % of the "bad" products pass the check and 5 % of the "good" ones end up in the bin.

a) What is the probability that a semiconductor that has not passed the check is "good"?
In other words: What is the rate of the "good" ones in the bin?
b) What is the probability that a semiconductor that has passed the check is "bad"?
In other words: What is the rate of the "bad" ones sent to customers?

8

In 1995 the former American football star Orenthal James Simpson was charged by he state of California with the murder of his ex-wife and her friend. The O. J. Simpson murder case was the longest and the most publicised jury trial in history. Over nine months the trial was lengthily broadcast by television. Simpsons lawyers could finally persuade the jurors that there was reasonable doubt. So it ended with a "not guilty", a verdict not all spectators agreed with.
During the trial Harvard-professor Alan Dershowitz, consultant of the defence, explained that only 0.1 % of all men who beat their wife, which Simpson did, finally kill her. This statement should make the jurors believe that the probability that a beating husband kills his wife is 0.001.
The statistician J. J. Good reacted immediately, stating that the crucial question is a completely different one: how likely is it that a beating husband has killed his wife **under the condition that his wife is dead**!
He calculated a probability of 50 % and suggested to teach probability calculus at school to help Harvard-professors avoid mistakes like that.

Orenthal James "O. J." Simpson (born July 9, 1947), nicknamed "The Juice", is a retired American football player, actor and convicted felon.
He was a famous running back with the Buffalo Bills (in Buffalo, New York) and the first one to pass 2,000 yards in a season (1973). He was elected to the Pro Football Hall of Fame in 1985. His amiable personality helped him to start an acting career (e.g. The Naked Gun Trilogy) which was interrupted by his arrest for murder.
During all his life he had faced many charges, from domestic violence and tax fraud to murder. In 2007 he was arrested for armed robbing and kidnapping and sentenced to 9 years of prison. He is serving his sentence in Nevada and will not be released before 2017.

To check on Good's result a model has to be created. Assume the following:

1) 100 people, her husband included, had access to the victim, and all are equally likely to have been present at the exact time of her death.
2) The probability that a beating husband kills his wife is 0.1 % (as indicated by Dershowitz), that of the average person to kill a wife is 0.0005 %.

Calculate the probability that O. J. Simpson has killed his wife under the condition that his wife is dead.

9

There are three identical boxes with cookies stored in the basement. Box A is filled with brownies only, box B with equal amounts of brownies and butter fingers, all mixed, box C with equal amounts of brownies, butter fingers and ginger breads, all mixed again.
One evening in the dark you sneak down to the basement, you choose a box and you grab the first cookie you feel, all completely at random. Up in your room you realise that you have picked a brownie. What is the probability that you opened box A?