Probability and Mathematical Statistics in Data Science

Lecture 05: Section 2.4: Baye's Rule: Uses and Interpretations

Section 2.5: Independence

Poll

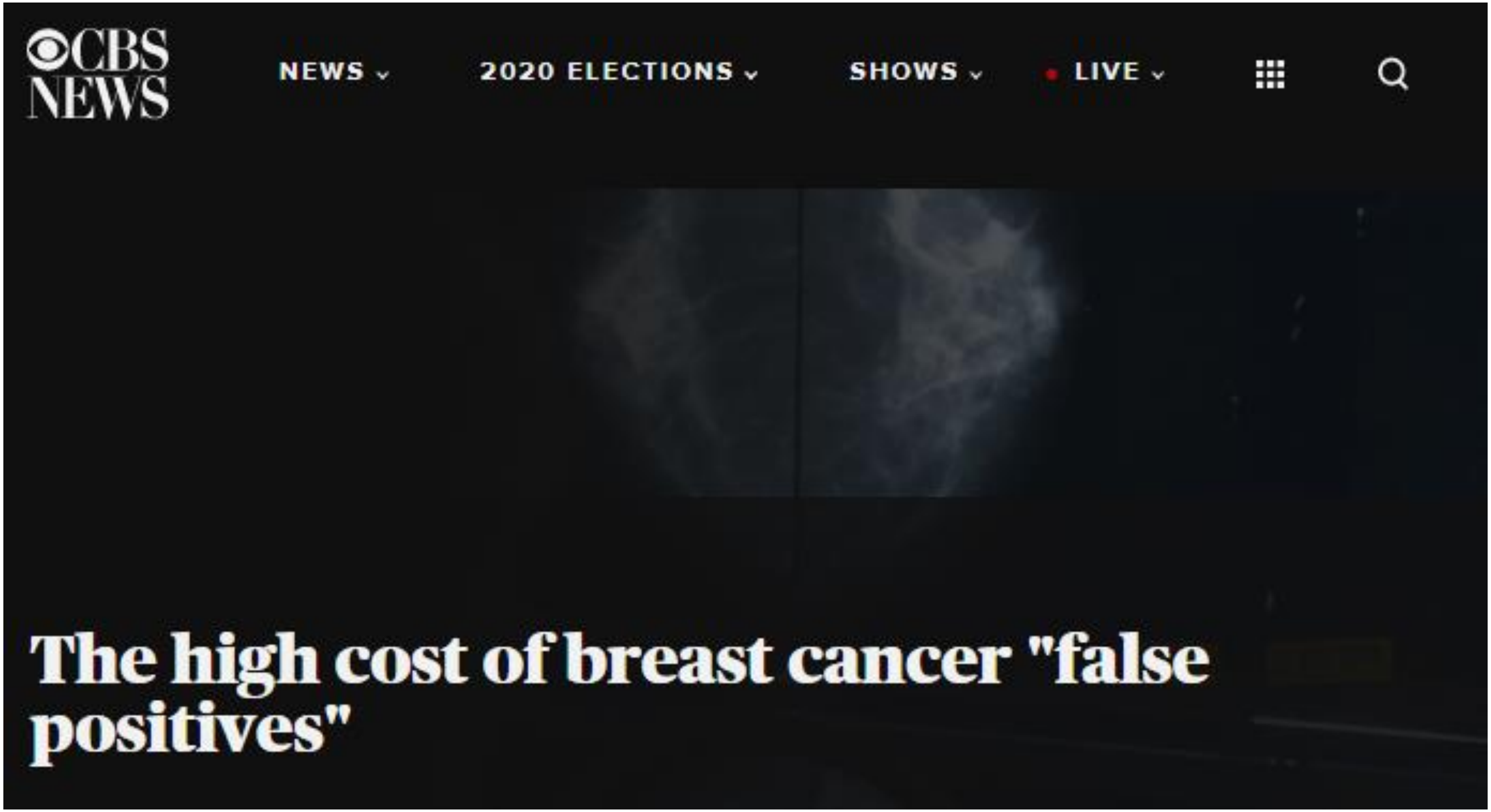
- ***You test positive for rare disease***, your original chances of having disease are 1 in 1000.

- The test has a 10% false positive rate and a 10% false negative rate => ***whether you have disease or not, test is 90% likely to give a correct answer***.

- Given you tested positive, what do you think is the ***probability that you actually have disease***?

A. 90% B. 70% C. 35% D. 1%

Mammograms and False Positives



https://www.cbsnews.com/

The high cost of breast cancer "false positives"

Sharpening a medical debate about the costs and benefits of cancer screening, a new report estimates that the U.S. spends $4 billion a year on unnecessary medical costs due to mammograms that generate false alarms, and on treatment of certain breast tumors unlikely to cause problems.

The study published Monday in the journal Health Affairs breaks the cost down as follows: $2.8 billion resulting from false-positive mammograms and another $1.2 billion attributed to breast cancer overdiagnosis. That's the treatment of tumors that grow slowly or not at all, and are unlikely to develop into life-threatening disease during a woman's lifetime.

The cost estimates cover women ages 40-59.

Source: https://www.cbsnews.com

Medical Testing – Some Key Terms

- *To determine probability of a positive test result being accurate, you need:*

- **Sensitivity** of the test – the proportion of people who correctly test positive when they actually have the disease

- **Specificity** of the test – the proportion of people who correctly test negative when they don't have the disease

- **Base rate -** probability that you are likely to have disease, without any knowledge of your test results.

- **Positive Predictive Value (PPV) -** probability of being diseased, given that someone tests positive.

Medical Testing for a Rare Disease

- ***You test positive for rare disease***, your original chances of having disease are 1 in 1000. The base rate is equal to 0.001

- The test has a 10% false positive rate and a 10% false negative rate => ***whether you have disease or not, test is 90% likely to give a correct answer***.

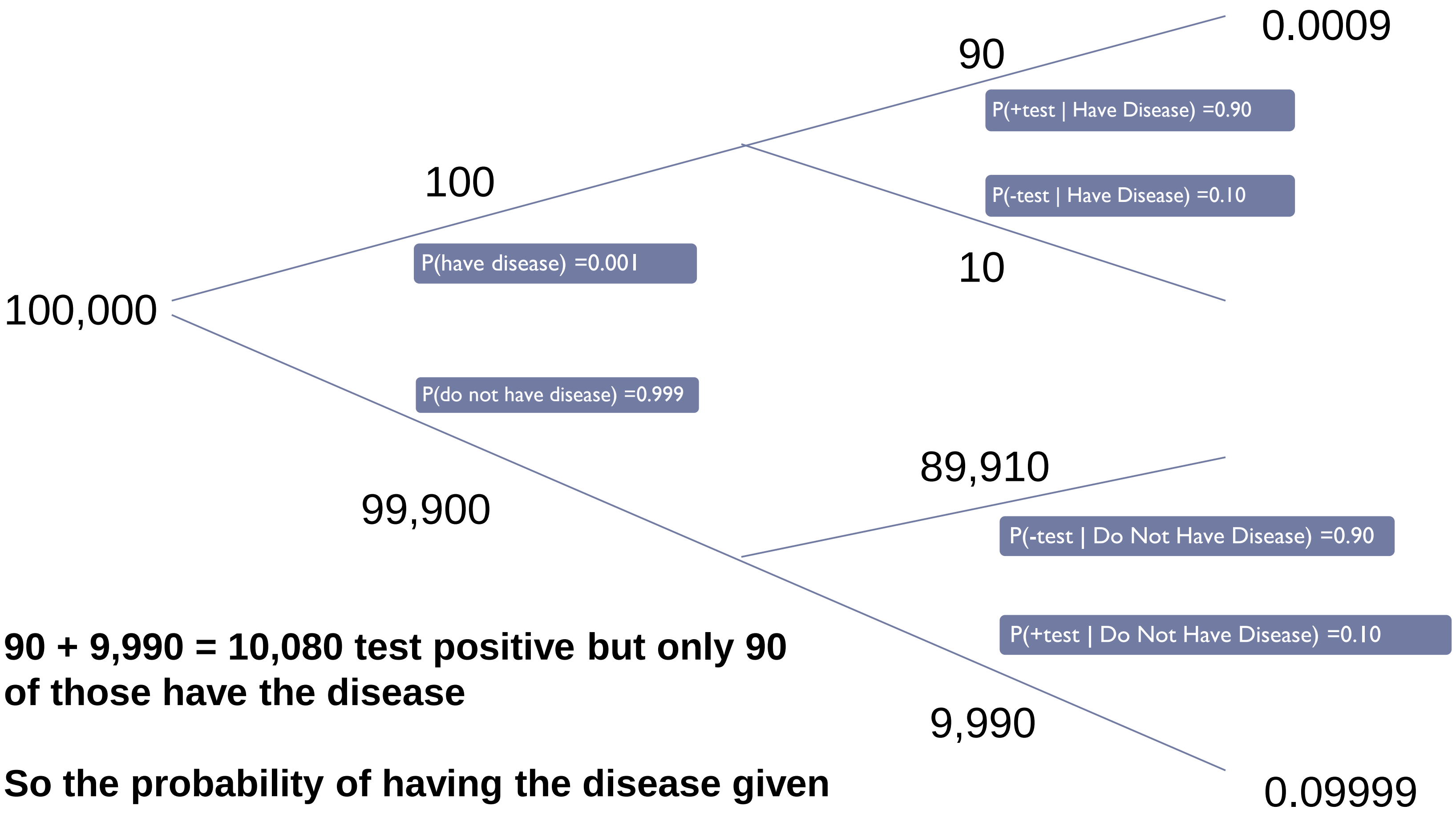
- The **sensitivity and specificity** of the test are both 90%. The proportion of people who correctly test positive or correctly test negative is 90%.

Medical Testing for a Rare Disease

Breakdown of Actual Status versus Test Status for a Rare Disease			
	Test shows positive	Test shows negative	Total
Actually sick	90	10	100
Actually healthy	9,990	89,910	99,900
Total	10,080	89,920	100,000

- The sensitivity is the P(test + | disease) = 90/100 = 0.90
- The specificity is the P(test - | no disease) = 89,910/99,900 = 0.90
- 90 + 9,990 = 10,080 test positive but only 90 of those have the disease
- So the probability of having the disease given you test positive is 90/10,080 = 0.009 or 1%

Medical Testing for a Rare Disease



90 + 9,990 = 10,080 test positive but only 90 of those have the disease

So the probability of having the disease given you test positive is 90/10,080 = 0.009 or 1%

This is called the Positive Predictive Value

Medical Testing for a Rare Disease

- P(Disease - D) = 0.001 P(No Disease - ND) = 0.9999
- P(Test + | D) = 0.90) P(Test - | ND) = 0.90)
- P(D and Test +) = P(D)P(Test + | D) = (0.001)(0.90) = 0.0009
- P(ND and Test +) = P(ND)P(Test + | ND) = (0.999)(0.10) = 0.09999

P(D | Test +) = 0.0009 / (0.0009 + 0.09999)

= 0.0009 / 0.10080 = 0.009 or **0.01** to two decimal places

Bayes's Rule

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^C)P(B^C)}$$

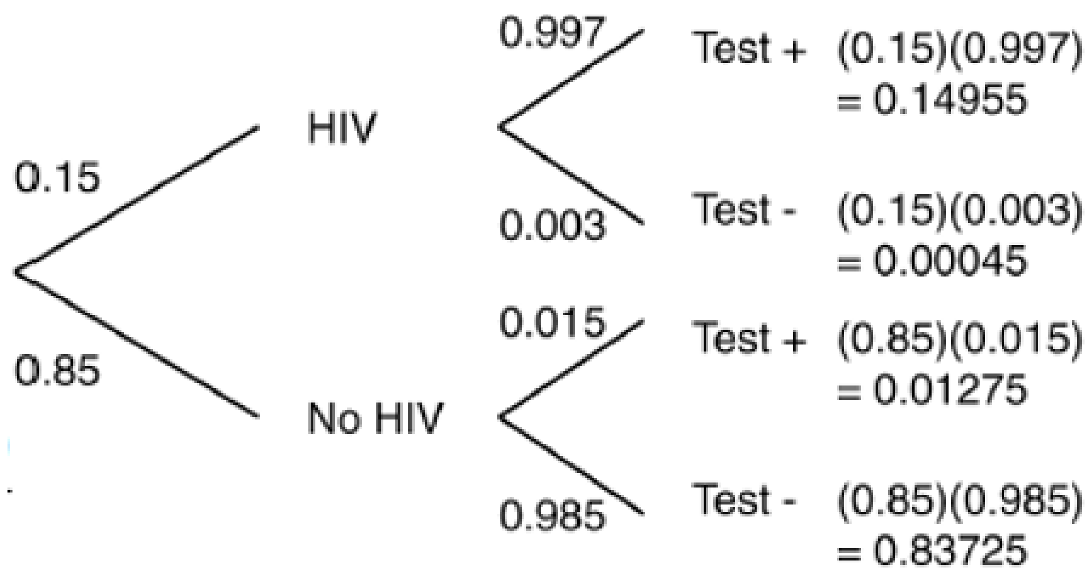
$$PV+ = \frac{P(test+ \mid disease) * P(disease)}{P(test+ \mid disease) * P(disease) + P(test+ \mid nodisease) * P(nodisease)}$$

- Let A – Test Positive B – Have Disease
- P(B|A) = (0.90)(0.001) / (0.90)(0.001) + (.10)(.999)
- P(B|A) = 0.0009 / 0.0009 + 0.0999
- = 0.0009 / 0.1008
- = 0.009 or 1%

Question

- **HIV testing** In July 2005, the journal *Annals of Internal Medicine* published a report on the reliability of HIV testing. Results of a large study suggested that among people with HIV, 99.7% of tests conducted were (correctly) positive, while for people without HIV 98.5% of the tests were (correctly) negative. A clinic serving an at-risk population offers free HIV testing, believing that 15% of the patients may actually carry HIV. What's the probability that a patient testing negative is truly free of HIV?

Tree Diagram



Q. What is the P(No HIV| Test -)?

Q. What is the P(HIV|Test +)?

Review: Independence of Events

- Two events, A and B, are **independent** if the fact that A occurs does not affect the probability of B occurring. This can be written as follows:

$$P(\text{A and B}) = P(\text{A}) \times P(\text{B})$$

- From our understanding of the conditional probability rule, we can also say that two events, A and B, are independent when:

$$P(\text{B}|\text{A}) = P(\text{B}) \text{ or } P(\text{A}|\text{B}) = P(\text{A})$$

Multiple Events

- Events $A_1, \ldots, A_n$ are mutually independent if for every k ($k = 2, 3, \ldots, n$) and every subset of indices $i_1, i_2, \ldots, i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})\, P(A_{i_2}) \ldots P(A_{i_k}).$$

- Independence is very very important!

Poll: Hospital Example

$P(S)=0.14$ $P(O)=0.08$

$P(S \text{ and } O)=0.02$ $P(S|O) = 0.25$

Q1. Are surgery and obstetrics independent events?

Q2. Are they disjoint events?

A. Yes, Yes
B. No, No
C. Yes, No
D. No, Yes

Misleading statistics were presented as facts in Sally Clark trial

The shadow of Sally Clark's wrongful conviction for murdering two of her babies, and the use of misleading statistics by expert witnesses, will have hung heavy over the jury.

The evidence in the 1999 trial was so badly flawed that the Royal Statistical Society of London wrote to the Lord Chancellor in concern and suggested that courts use expert statistical witnesses to avoid a similar debacle. The case also raised concerns about the way that statistics are handled in court. The Sally Clark jury was told by Prof Sir Roy Meadow that the chances of a mother losing two babies to sudden infant death syndrome were **"one in 73 million"**.

Source: https://www.telegraph.co.uk

The Tragic Case of Sally Clark

Professor Sir Roy Meadow, a pediatrician testified that the chance of two children from an affluent family suffering sudden infant death syndrome was 1 in 73 million,

He relied on a study which showed that for non-smoking professional families, the probability of cot death is 1/8500.

$$1/8500 \times 1/8500 = 1/73 \text{ million}$$

He assumed independence of the two events.

Do you think that was a fair assumption to make? Why?