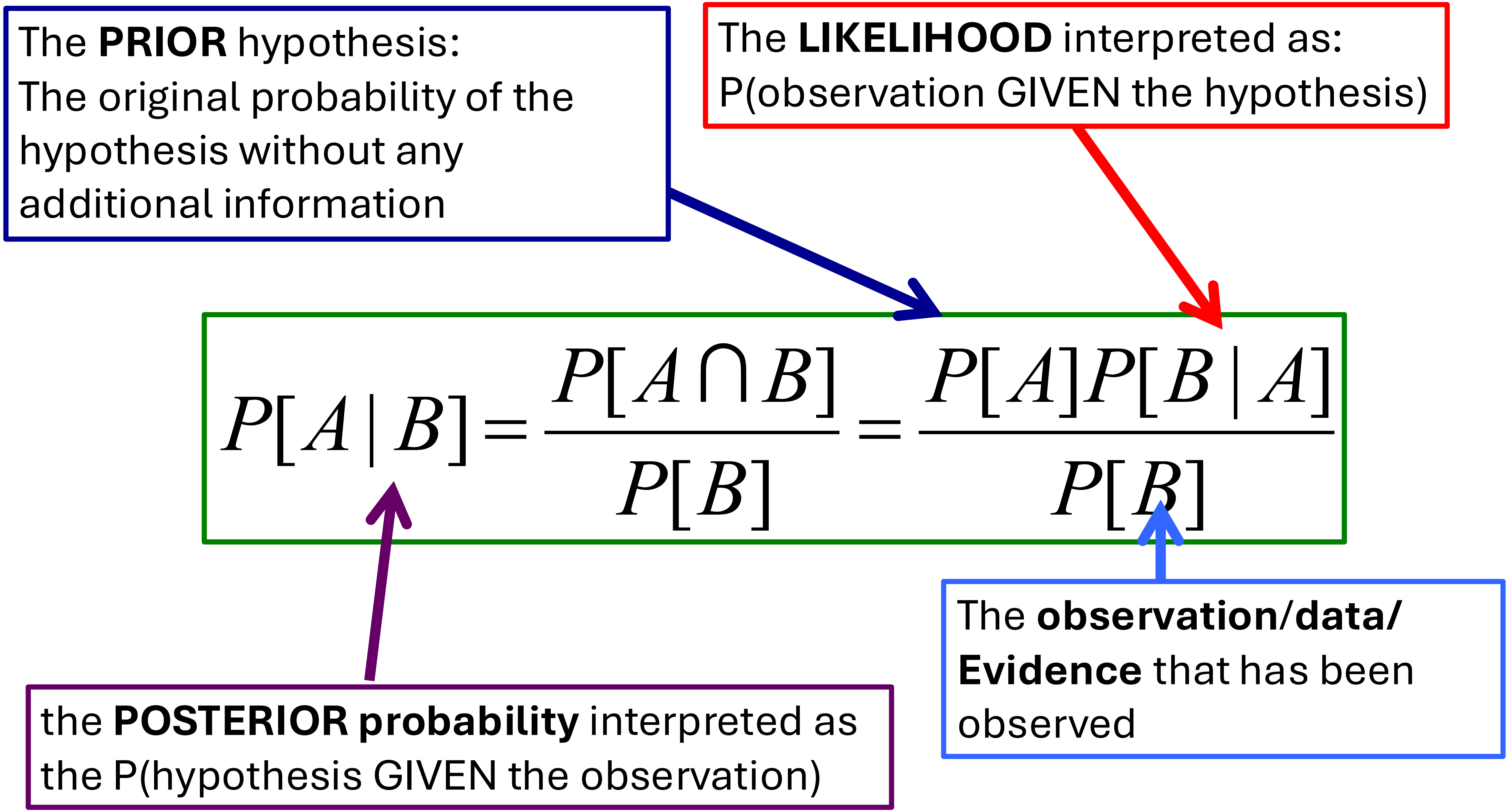
Module 2 : Probability

Frequentist and Bayesian building blocks

Agenda:

- Bayesian Probability
  - Structure of Bayes' Theorem: $P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B|A]P[A]}{P[B]}$
  - The Monty Hall Problem: illustrating the philosophical difference with Frequentist camp - ability to update probability with new information
  - Examples:
    - Pedigree Analysis



The **PRIOR** hypothesis:
The original probability of the hypothesis without any additional information

The **LIKELIHOOD** interpreted as:
P(observation GIVEN the hypothesis)

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B \mid A]}{P[B]}$$

the **POSTERIOR probability** interpreted as the P(hypothesis GIVEN the observation)

The **observation/data/ Evidence** that has been observed

Example: Suppose we want to calculate the probability that someone will die of lung cancer **given** that they smoke. We study a cohort of individuals, determining who smoke and which ones do not and track them until they died. Then we could calculate the number of smokers who died of lung cancer.

There is an easier way, however....

USE BAYES

$$\boxed{P[A \mid B] = \frac{P[A]P[B \mid A]}{P[B]}}$$

Specify the question: What is event 'A' and what is event 'B'? A=lung cancer death; B=smoker

Probabilities	Where/how do we get them?
P[Death due to lung cancer | Smoker]	
P[Death due to lung cancer]	estimated from death records
P[Smoker]	Polling appropriate population
P[Smoker | Death due to lung cancer]	estimated from death records

Example:

$$\boxed{P[A \mid B] = \frac{P[A]P[B \mid A]}{P[B]}}$$

Specify the question: What is event 'A' and what is event 'B'?

Probabilities	Where/how do we get them?
P[Death due to lung cancer | Smoker]	
P[Death due to lung cancer]	estimated from death records
P[Smoker]	Polling appropriate population
P[Smoker | Death due to lung cancer]	estimated from death records

P[Smoker] = 0.5

P[Smoker | Death due to lung cancer] = 0.9

P[Death due to lung cancer] = 0.3

P(Smoker|Death by lung cancer) + P(Nonsmoker|Death by lung cancer) = 1

P[Death due to lung cancer | Smoker] = $\frac{0.9 \times 0.3}{0.5} = 0.54$

Note: Using Bayes, also gives us a bonus calculation: **P[Non-Smoker | Death due to lung cancer] = 0.1**

P[Death due to lung cancer | Non-smoker] = $\frac{0.1 \times 0.3}{0.5} = 0.06$

Jim was bitten by a mosquito during his trip to South Sudan. He gets tested for Malaria. What is the probability that Jim has Malaria given a positive test result, considering the following facts: Malaria occurs in 1 in 1,000 people in South Sudan, the test for Malaria has an 85% probability of detecting Malaria, but there is also a 10% false positive rate as well.

Let's gather our information.

P[malaria]=1/1000 = 0.001
P[no malaria]=999/1000 =0.999
P[pos test|malaria] = 0.85
P[pos test|no malaria] =0.10

$$\textbf{P[malaria| positive test]} = \frac{\textbf{P[pos test|malaria]P[malaria]}}{\textbf{P[positive test]}} = \frac{\mathbf{0.85*0.001}}{\mathbf{(0.85*0.001+0.10*0.999)}} = \mathbf{0.00844}$$

There are two ways to have a positive test: because you have malaria or because the test is inaccurate.

P[positive test] = P[Positive|malaria]*P[malaria]+P[positive|no malaria]*P[no malaria]
= 0.85*0.001+0.10*0.999

1/1000 →8.4/1000

There are a handful of other probabilities and terms that are often given:

Specificity (a rate) **= P[negative test | don’t have condition]/P[don’t have disease]**

From this, you can get:

False alarm = 1-Specificity = 1-P[positive test | don’t have condition]

We have:

Sensitivity (like **Likelihood** but a rate**)**

= P[positive test | have condition]/P[have the disease]

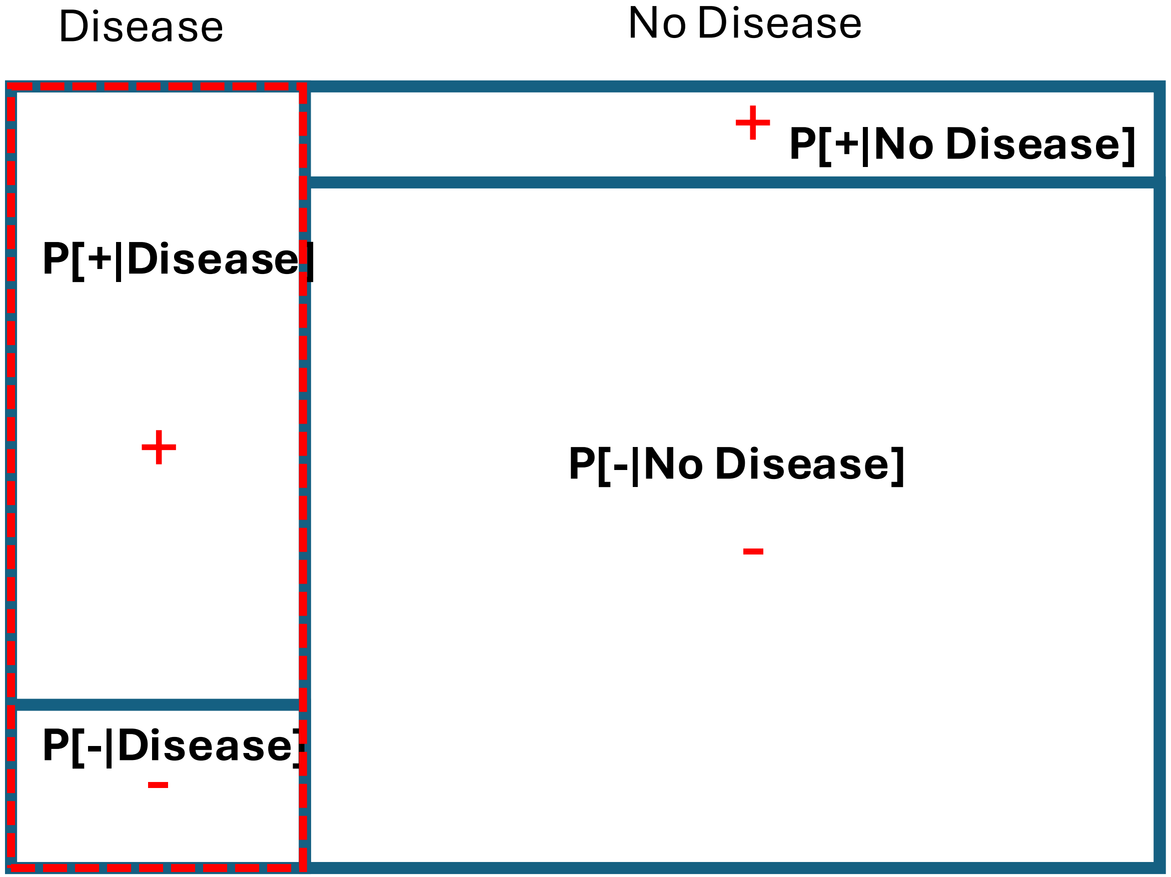
Prevalence = Prior = P[condition in general pop.]

* I updated the parts in purple to be more precise than I was in the video. This aligns with the information on the next page

Here is a three-page worksheet that contains definitions and a clear worked example for each of these terms:
https://statstutor.ac.uk/resources/uploaded/stcp-rothwell-diagnostictests.pdf

Prevalence= P[Disease]

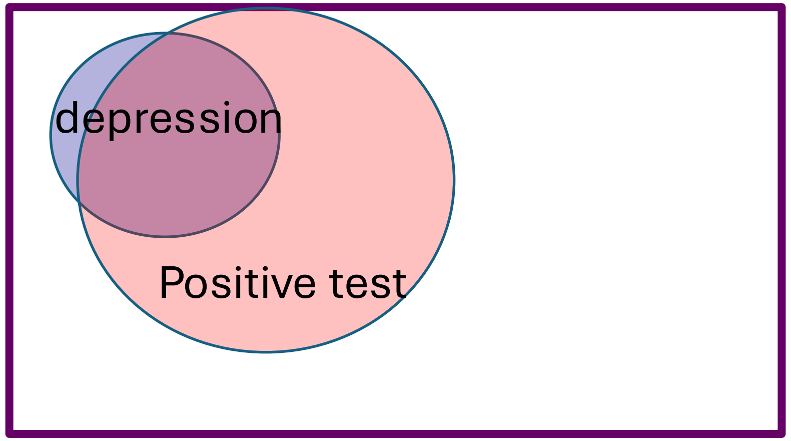
$$\textbf{Specificity}=\frac{P[-|\text{No Disease}]}{P[+|\text{No Disease}]+P[-|\text{No Disease}]}$$



$$\textbf{Sensitivity}=\frac{P[+|\text{Disease}]}{P[+|\text{Disease}]+P[-|\text{Disease}]}$$

$$P[\text{Disease}|+]=\frac{P[+|\text{Disease}]P[\text{Disease}]}{P[+]}$$

MOST POSITIVES ARE FALSE POSITIVES



Blue = depression proportion
Red= positive test for depression

P[depression]=0.1

P[negative Test|No Depression]=0.8

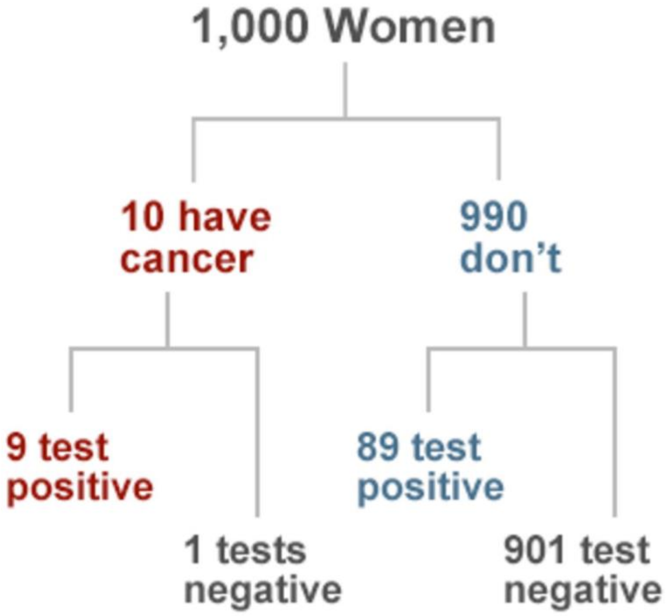
P[positive test | depression] = 0.9

$$\textbf{P[depression | positive test] =} \frac{0.9*0.1}{(0.9*0.1 + 0.8*0.9)} = 0.33$$

Don’t mix up P[A|B] with P[B|A] … that’s a different mistake called “The Prosecutor’s Fallacy”

Diagnostic Example

- We don’t know the truth about the disease state, we only have access to tests (data)
- There is no such thing as a perfect test: every test has a trade off between sensitivity and specificity
- **For a test to be useful,** it is not necessary for both sensitivity and specificity to be high, **but it IS** necessary for the user (health care provider) to interpret positives/negatives correctly.



1,000 Women

10 have cancer

990 don’t

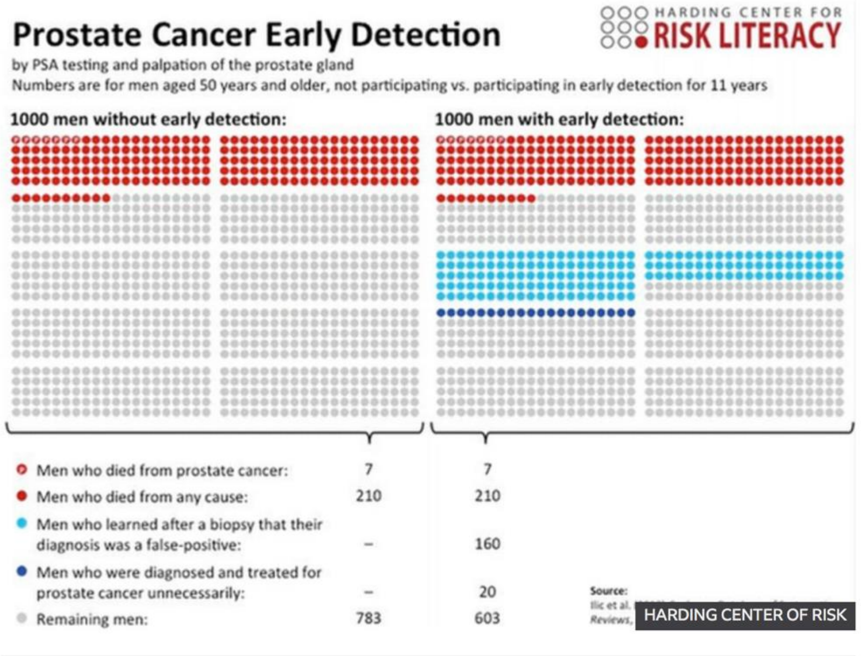
9 test positive

1 tests negative

89 test positive

901 test negative

https://www.bbc.com/news/magazine-28166019



Prostate Cancer Early Detection

by PSA testing and palpation of the prostate gland

Numbers are for men aged 50 years and older, not participating vs. participating in early detection for 11 years

	1000 men without early detection:	1000 men with early detection:
○ Men who died from prostate cancer:	7	7
● Men who died from any cause:	210	210
● Men who learned after a biopsy that their diagnosis was a false-positive:	–	160
● Men who were diagnosed and treated for prostate cancer unnecessarily:	–	20
● Remaining men:	783	603

Source:
Ilic et al.
Reviews,

HARDING CENTER OF RISK

Major Caveat

- Prior probabilities and prevalence are established with certain ancestries in mind

- For genetic information, >85% of our major databases are European ancestry ← this is a **massive problem**

- It is challenging to get prevalence rates in other ancestries.

What do I mean when I say, “Bayes allows us to easily update our information?”

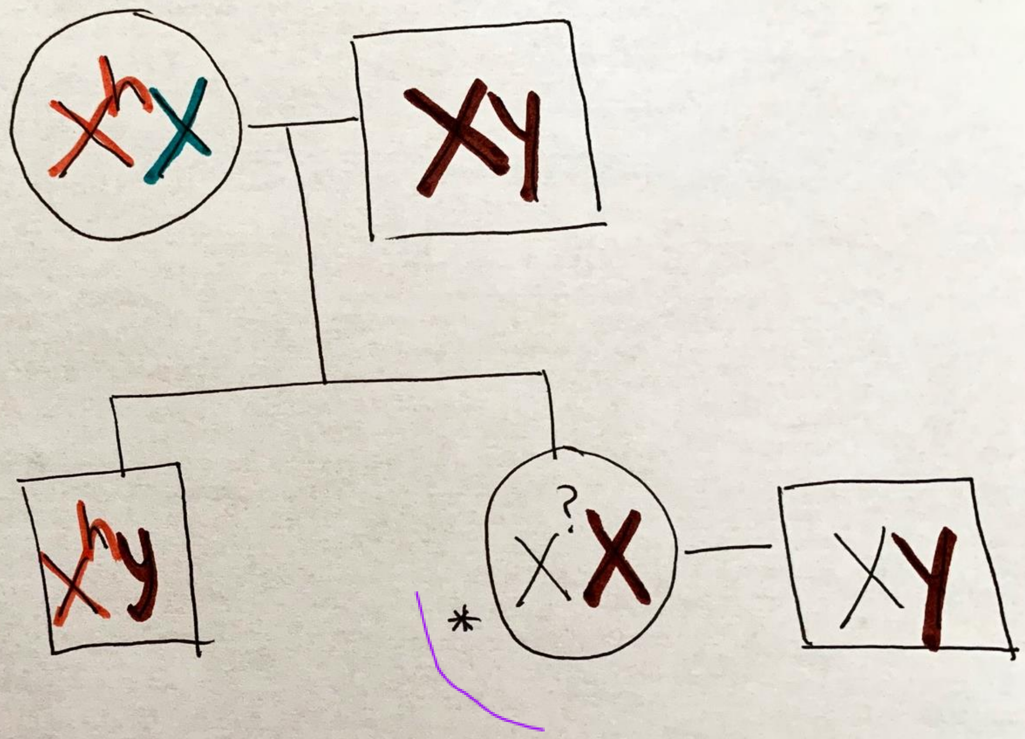
* Before offspring:

Probability of being carrier:

$P(\theta=1) = \frac{1}{2}$

Probability of not being carrier:

$P(\theta=0) = \frac{1}{2}$



X-linked condition

- Woman is unaffected by Hemophilia, but she has a brother who is affected by Hemophilia. She refuses to get genetically tested.
- Hemophilia allele is located on the X chromosome
- Hemophilia is a recessive trait
- Their father is unaffected, and their mother is phenotypically unaffected (but she must be a carrier)

Since the woman has a brother with the disease, she can be a carrier for the recessive allele, or she may have inherited a typical X (doesn’t carry the recessive allele) from her mother (we know that she inherited the typical X from her father because he is unaffected and therefore must have a typical X)

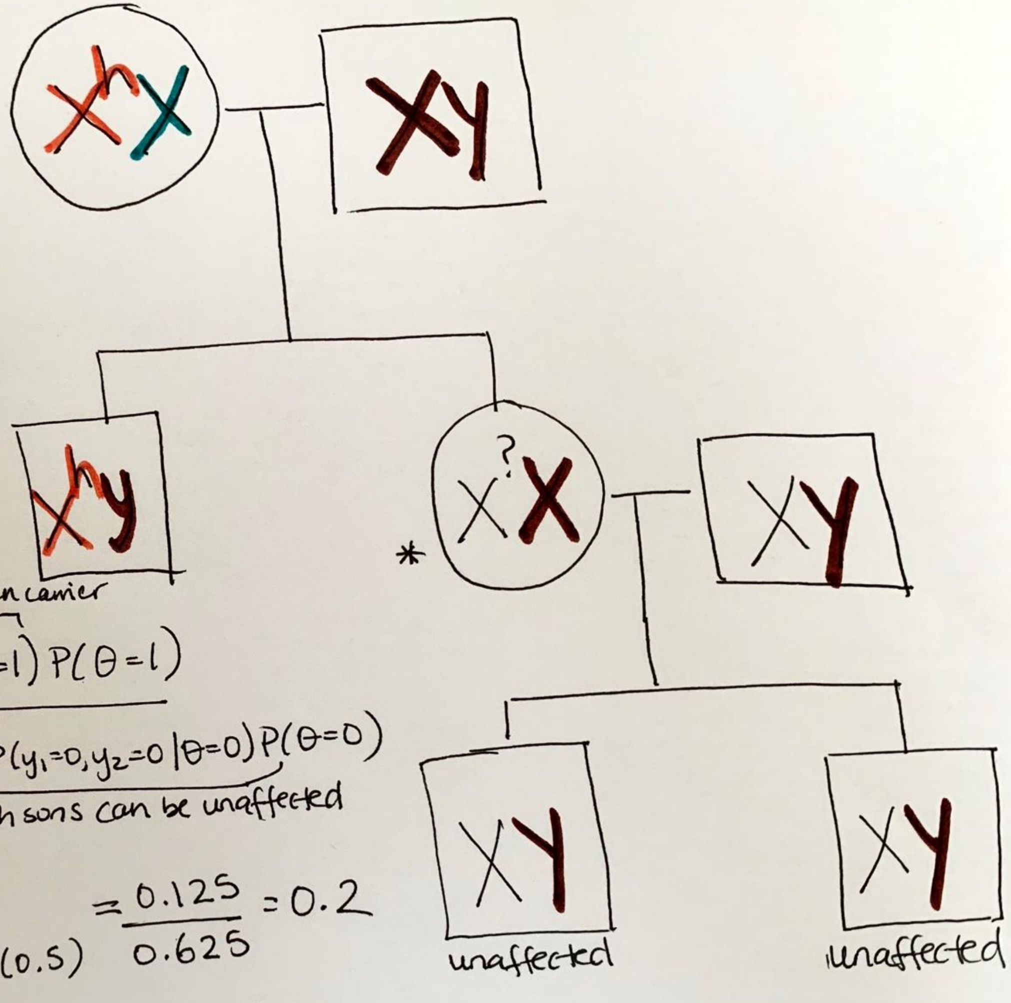
* Before offspring:

Probability of being carrier:

$$P(\theta=1) = \frac{1}{2}$$

Probability of not being carrier:

$$P(\theta=0) = \frac{1}{2}$$



After having two unaffected sons:

(Bayes)

$$P(\theta=1 \mid y_1=0, y_2=0) = \frac{\overbrace{P(y_1=0, y_2=0 \mid \theta=1)}^{\text{Two unaffected when carrier}} P(\theta=1)}{\underbrace{P(y_1=0, y_2=0 \mid \theta=1) P(\theta=1) + P(y_1=0, y_2=0 \mid \theta=0) P(\theta=0)}_{\text{all ways both sons can be unaffected}}}$$

$$= \frac{(0.25)(0.5)}{(0.25)(0.5) + 1\cdot(0.5)} = \frac{0.125}{0.625} = 0.2$$

- WITH EXTRA INFORMATION (2 unaffected sons), the probability that mom is a carrier has gone from 0.5 → 0.2
- NOW what happens if mom has one more unaffected son?

Based on the pedigree and the known inheritance mechanism
P(woman being carrier) =P(Θ=1)= 0.5
P(woman not being carrier) =P(Θ=0)= 0.5

P(woman being carrier | two sons are unaffected)
= P(two unaffected sons and carrier)/P(all ways unaffected sons)

$$P[\Theta=1 \mid y_1=0, y_2=0] = \frac{P(y_1=0, y_2=0 \mid \Theta=1) * P(\Theta=1)}{P(y_1=0, y_2=0 \mid \Theta=1) * P(\Theta=1) + P(y_1=0, y_2=0 \mid \Theta=0) * P(\Theta=0)}$$

$$P[\Theta=1 \mid y_1=0, y_2=0] = \frac{0.25*0.5}{0.25*0.5+1*0.5} = \frac{0.125}{0.625} = 0.2$$

$$P[\Theta=1 \mid y_1=0, y_2=0] = \frac{0.25*0.5}{0.25*0.5+1*0.5} = \frac{0.125}{0.625} = 0.2$$

Based on the pedigree and the known inheritance mechanism
P(woman being carrier) =P(Θ=1)= 0.5
P(woman not being carrier) =P(Θ=0)= 0.5

$$P[\Theta=1\mid y_1=0, y_2=0]=\frac{0.25*0.5}{0.25*0.5+1*0.5}=\frac{0.125}{0.625}=0.2$$

We have now <u>updated our prior probability</u> of the woman being a carrier from a starting prior of <u>0.5</u> to <u>0.2</u>!

If she went on to have a **third unaffected son**, this would provide additional evidence and would continue to change the woman's probability of being a carrier:

We have now updated our prior probability of the woman being a carrier from a starting prior of 0.5 to 0.2!

If she went on to have a **third unaffected son**, this would provide additional evidence and would continue to change the woman's probability of being a carrier. Note: we would use our new updated prior probability of 0.2 (instead of the original prior of 0.5):

$$P[\Theta=1 \mid y_1=0, y_2=0, y_3=0]=\frac{P(y_1=0, y_2=0, y_3=0 \mid \Theta=1) * P(\Theta=1)}{P(y_1=0, y_2=0, y_3=0 \mid \Theta=1) * P(\Theta=1)+P(y_1=0, y_2=0, y_3=0 \mid \Theta=0) * P(\Theta=0)}$$

$$P[\Theta=1 \mid y_1=0, y_2=0, y_3=0]=\frac{0.5 * 0.2}{0.5 * 0.2+1 * 0.8}=0.111$$