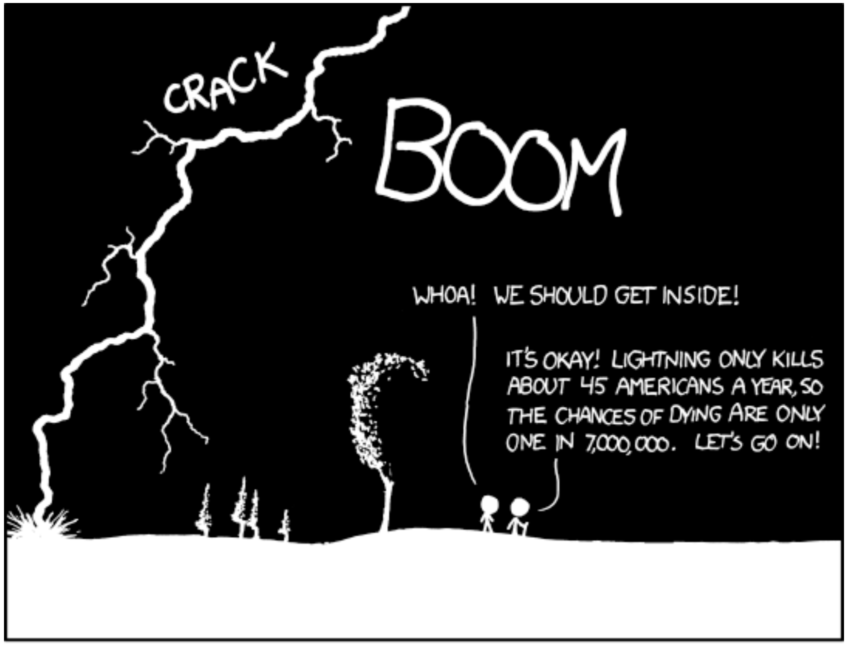
BIOL-UA 45:
BIOSTATISTICS AND HUMAN GENETICS


CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!

THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

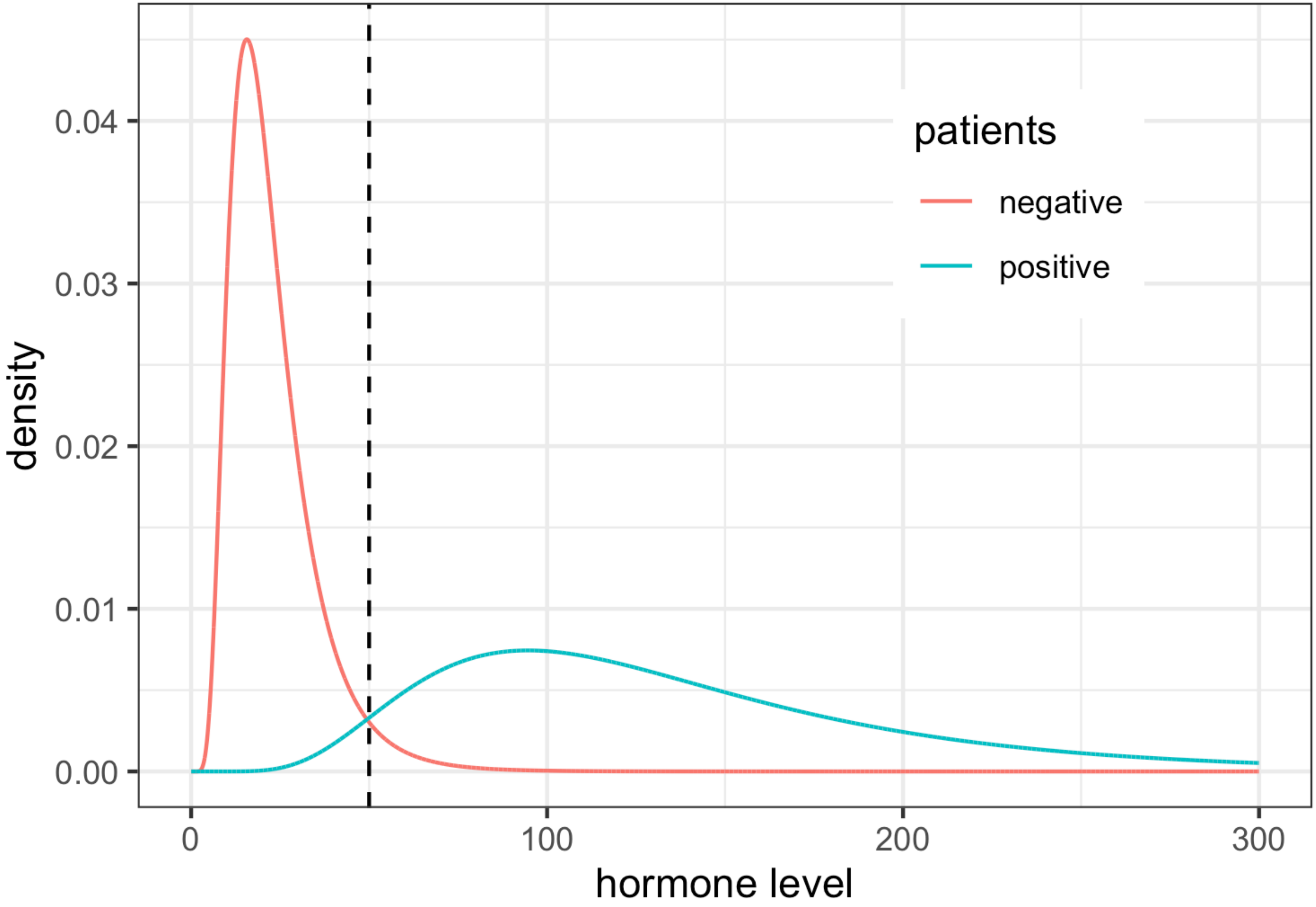
Eugene Plavskin

CLASS 12:
Bayesian inference continued

RECITATION ASSIGNMENTS

- Comment code
- Break up code into lines

TESTING:
BINARIZING CONTINUOUS THINGS



TESTING: BINARIZING CONTINUOUS DISTRIBUTIONS

- In testing, we often have to make a binary decision, e.g.: is a person sick or not?

- But we are basing this decision on data that is continuous, not binary (e.g. hormone levels)

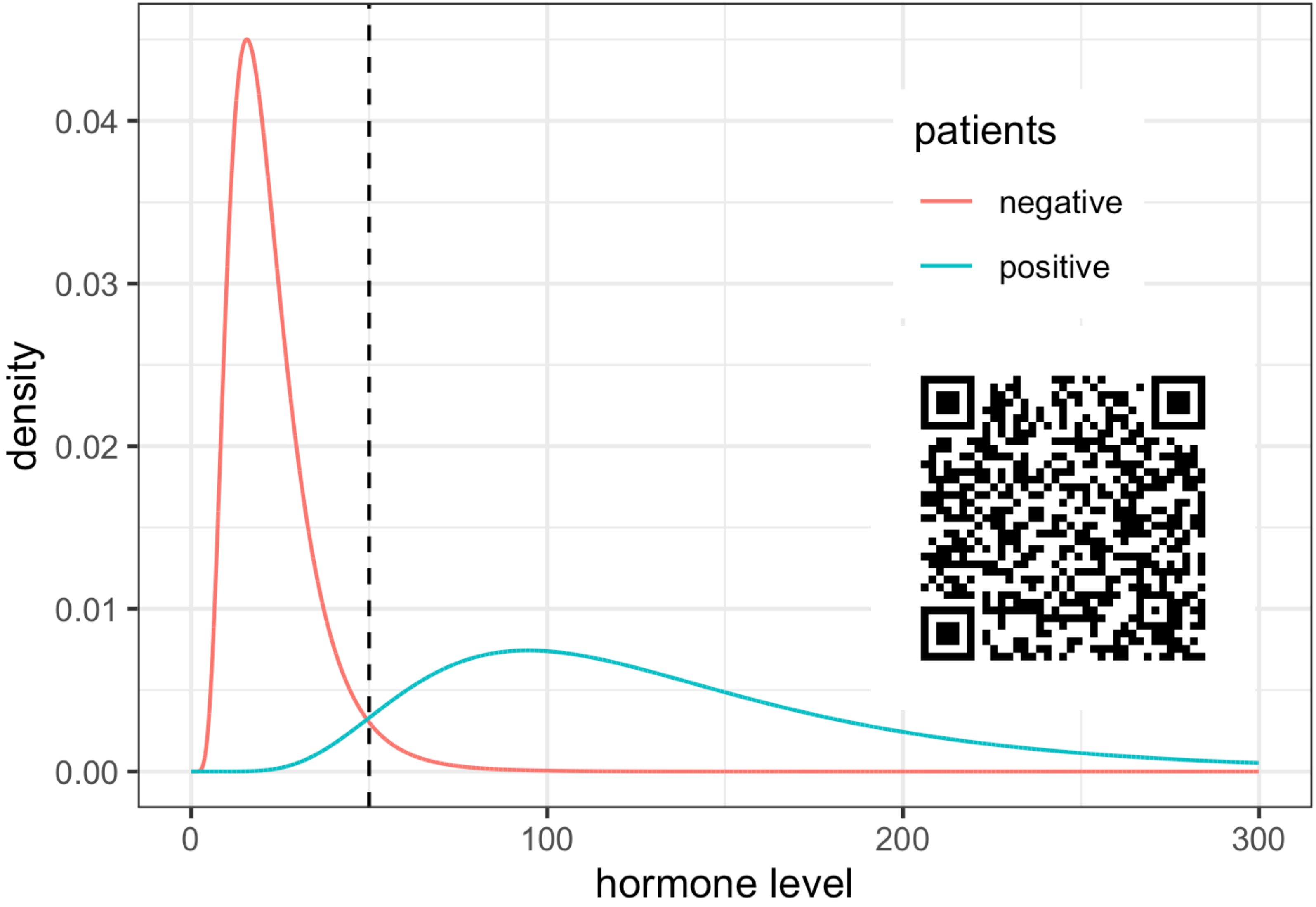
- We solve this by setting a **threshold**

- Setting a threshold for continuous traits will result in some incorrect calls

THE PROBLEM WITH BINARIZING CONTINUOUS DISTRIBUTIONS

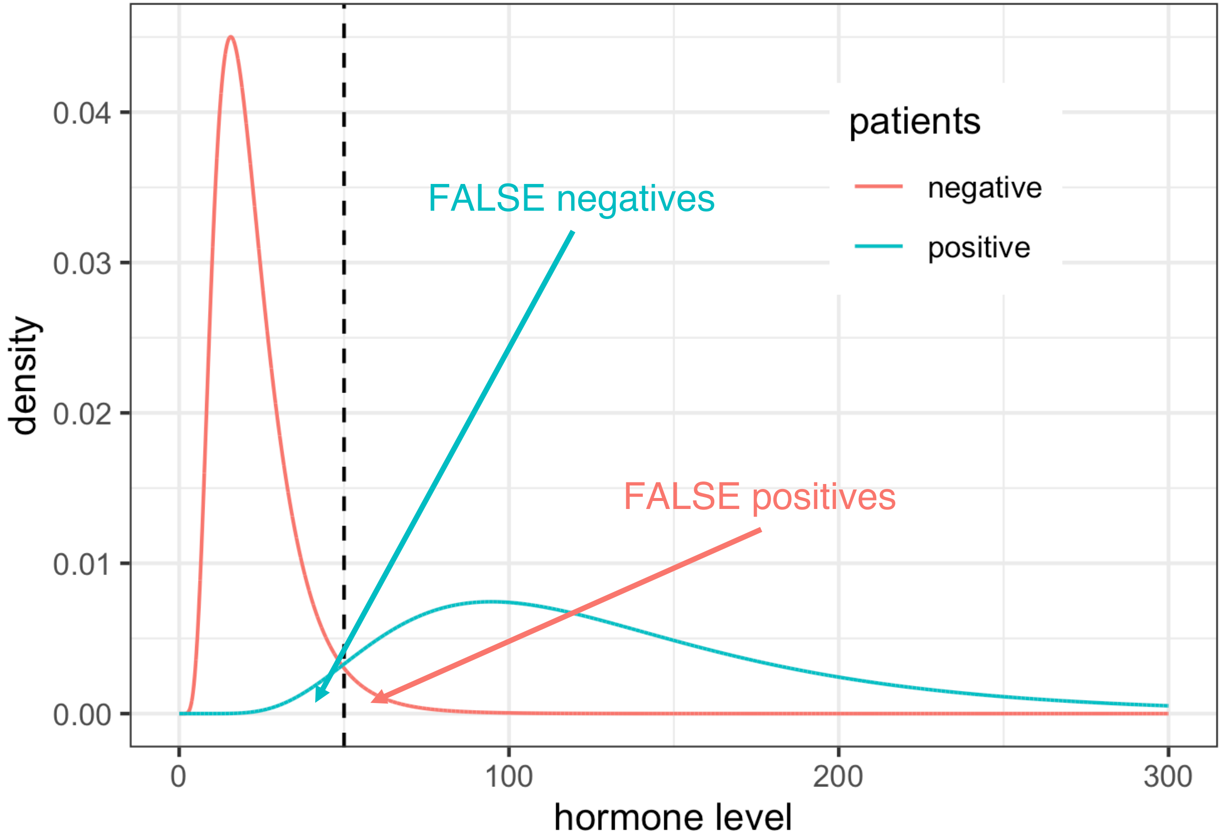
- **False positives**: individuals we think ARE affected, who are NOT actually affected
- **False negatives**: individuals we think are NOT affected, who actually ARE affected

IF WE INCREASE THRESHOLD, WHAT HAPPENS?



THE PROBLEM WITH BINARIZING CONTINUOUS DISTRIBUTIONS

If I increase the threshold, fewer false positives, more false negatives (in this example!)



GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

- Down syndrome has a population frequency of ~1/1000

- Non-invasive prenatal test (NIPT) is a blood test
  - false positive rate of ~5%
  - false negative rate of ~40%

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

I want to know:

In what percept of pregnancies that test positive for Down Syndrome using NIPT does the fetus ACTUALLY have Down Syndrome

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

I want to know:

In what percept of pregnancies that test positive for Down Syndrome using NIPT does the fetus ACTUALLY have Down Syndrome

How do I phrase this in terms of probabilities?

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

I want to know:

In what percept of pregnancies that test positive for Down Syndrome using NIPT does the fetus ACTUALLY have Down Syndrome

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

I want to know:

In what percept of pregnancies that test positive for Down Syndrome using NIPT does the fetus ACTUALLY have Down Syndrome

$P(D|+) = ?$

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

I want to know:

In what percept of pregnancies that test positive for Down Syndrome using NIPT does the fetus ACTUALLY have Down Syndrome

$$P(D|+) = P(+|D)*P(D)/P(+)$$

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

- Down syndrome has a population frequency of ~1/1000
- NIPT: false + ~5%, false - ~40%

- $P(D) =$
- $P(+|D) =$
- $P(+) =$

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

- Down syndrome has a population frequency of ~1/1000
- NIPT: false + ~5%, false - ~40%

- $P(D) = 0.001$
- $P(+|D) = 1-0.4 = 0.6$
- $P(+) = P(+|D)*P(D) + P(+|\text{not } D)*P(\text{not } D) =$

  $.6*.001+.05*.999 = 0.056$

GENETIC TESTING: PRENATAL SCREENING FOR DOWN SYNDROME

- Down syndrome has a population frequency of ~1/1000
- NIPT: false + ~5%, false - ~40%

- $P(D) = 0.001$
- $P(+|D) = 1-0.4 = 0.6$
- $P(+) = 0.056$

- $P(D|+) = P(+|D)*P(D)/P(+) =$ 1.2%

NIPT FOR DOWN SYNDROME: LARGE NUMBERS

- Consider 1,000,000 fetuses
- 1,000 will have DS; 600 of those will test +
- 999,000 will NOT have DS; 49,950 of those will test +
- 49,950+600 = 50,550 + tests, but only 600 of those have DS

WHY DO WE DO THIS!?

Better testing is RISKY: ~0.5% miscarriage rate

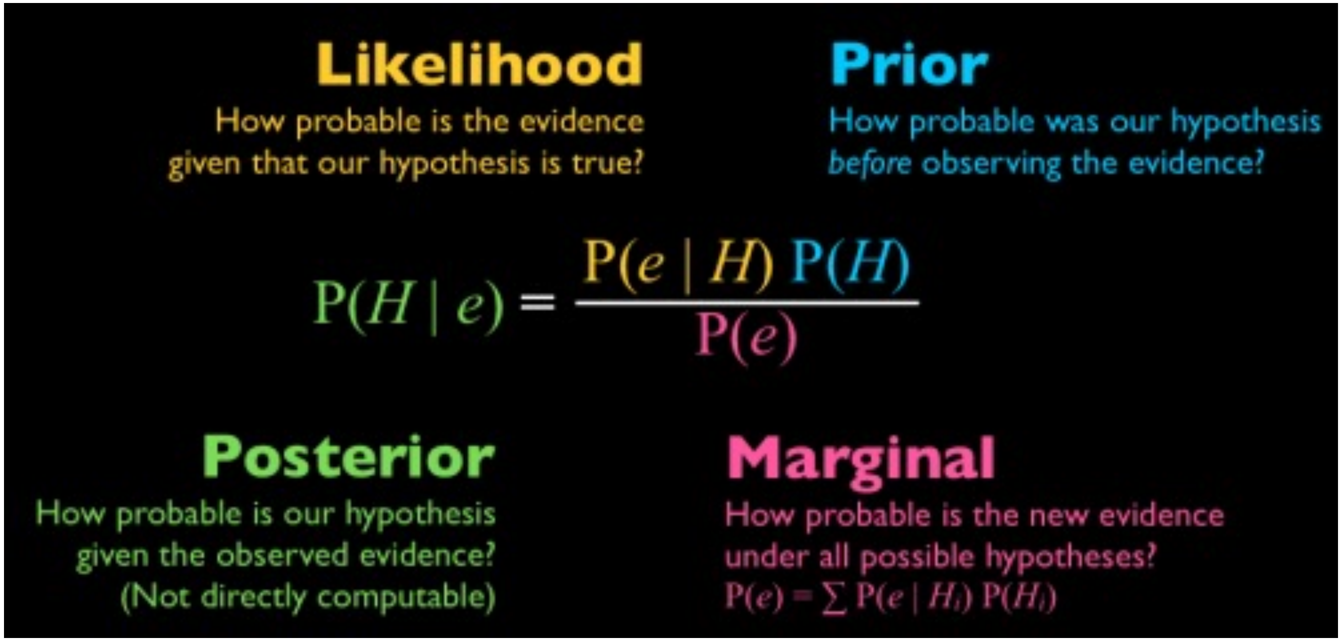
TESTING AND CONDITIONAL PROBABILITY: THE INTUITION

- If we’re testing for a rare condition, our false positives will swamp out any true positives in number
  - We can still get *enrichment* of the individuals we’re testing for

- **LIKELIHOOD THAT TEST RESULT IS CORRECT DEPENDS ON THE FREQUENCY OF THE THING BEING TESTED FOR!**

BAYESIAN INFERENCE AND LIKELIHOODS

Bayesian statistics allows us to convert likelihoods to **actual probabilities**



Likelihood
How probable is the evidence given that our hypothesis is true?

Prior
How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior
How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal
How probable is the new evidence under all possible hypotheses?
$P(e) = \sum P(e \mid H_i)\, P(H_i)$

https://towardsdatascience.com/what-is-bayesian-statistics-used-for-37b91c2c257c

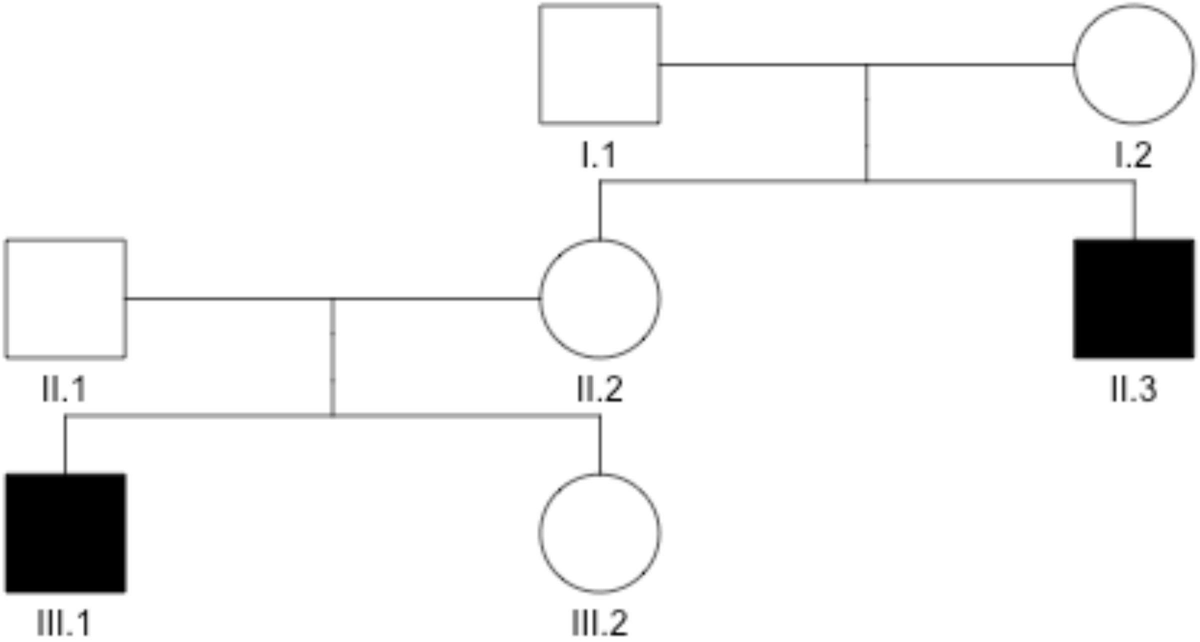
BAYESIAN INFERENCE: PRIORS AND POSTERIORS

- **Prior probability**: probability of a hypothesis BEFORE observing evidence

- **Posterior probability**: probability of a hypothesis AFTER observing evidence

CONDITIONAL PROBABILITIES: INCORPORATING NEW INFORMATION

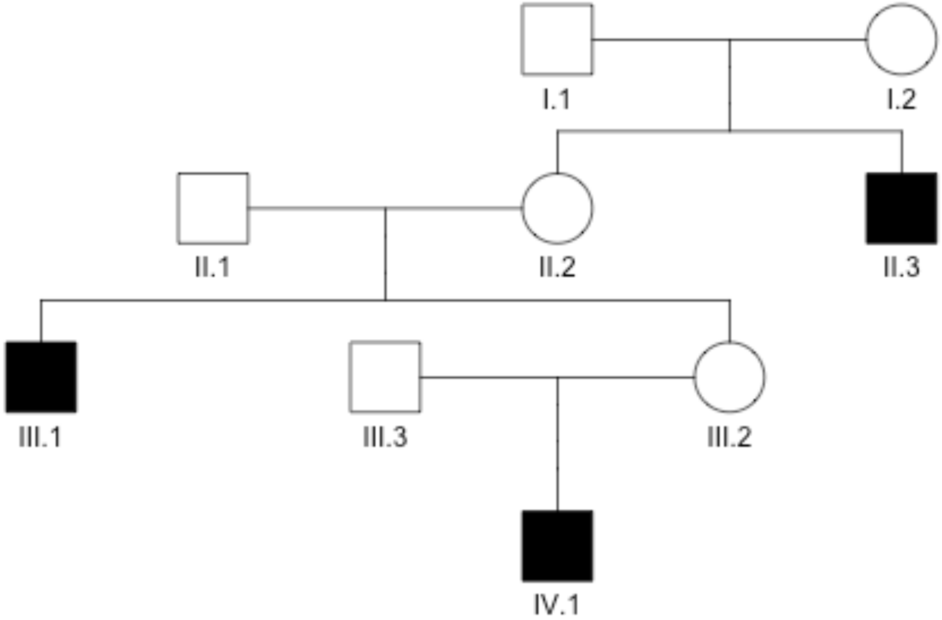
Duchenne Muscular Dystrophy (X-linked recessive)



What is P(III.2 is a carrier)?

CONDITIONAL PROBABILITIES: INCORPORATING NEW INFORMATION

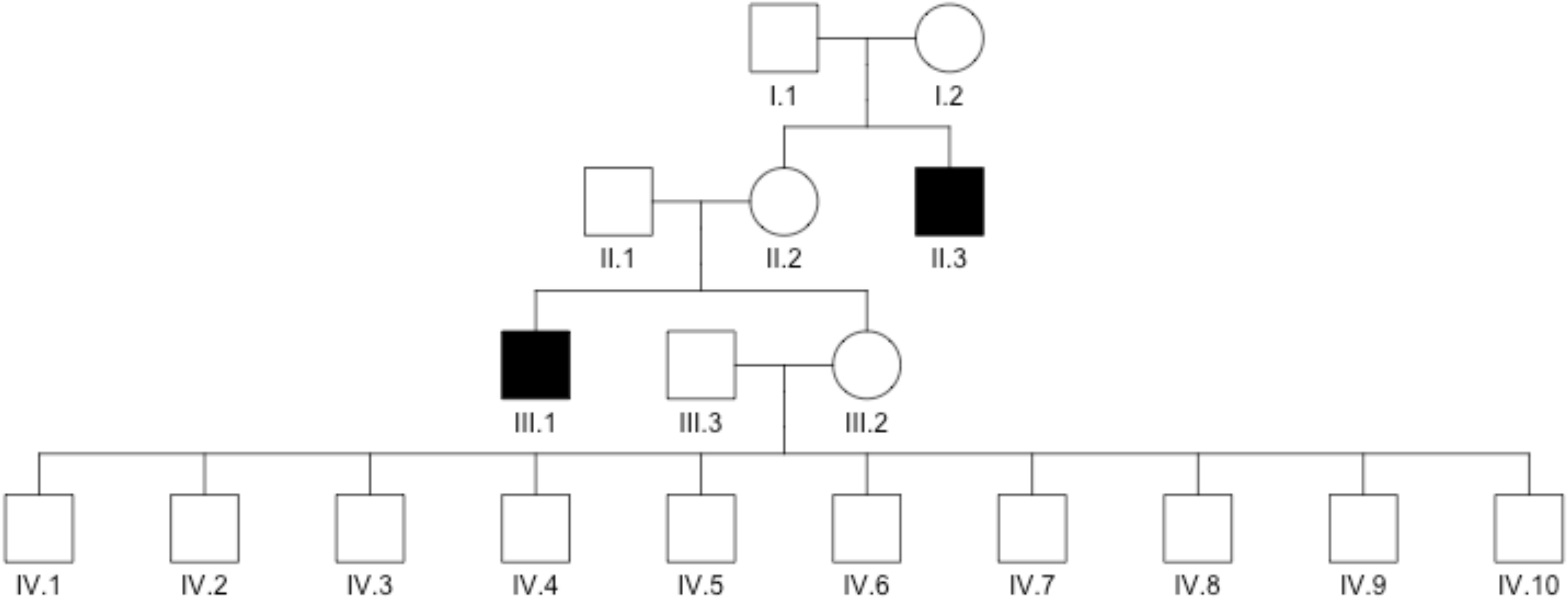
Duchenne Muscular Dystrophy (X-linked recessive)



What is P(III.2 is a carrier)?

CONDITIONAL PROBABILITIES: INCORPORATING NEW INFORMATION

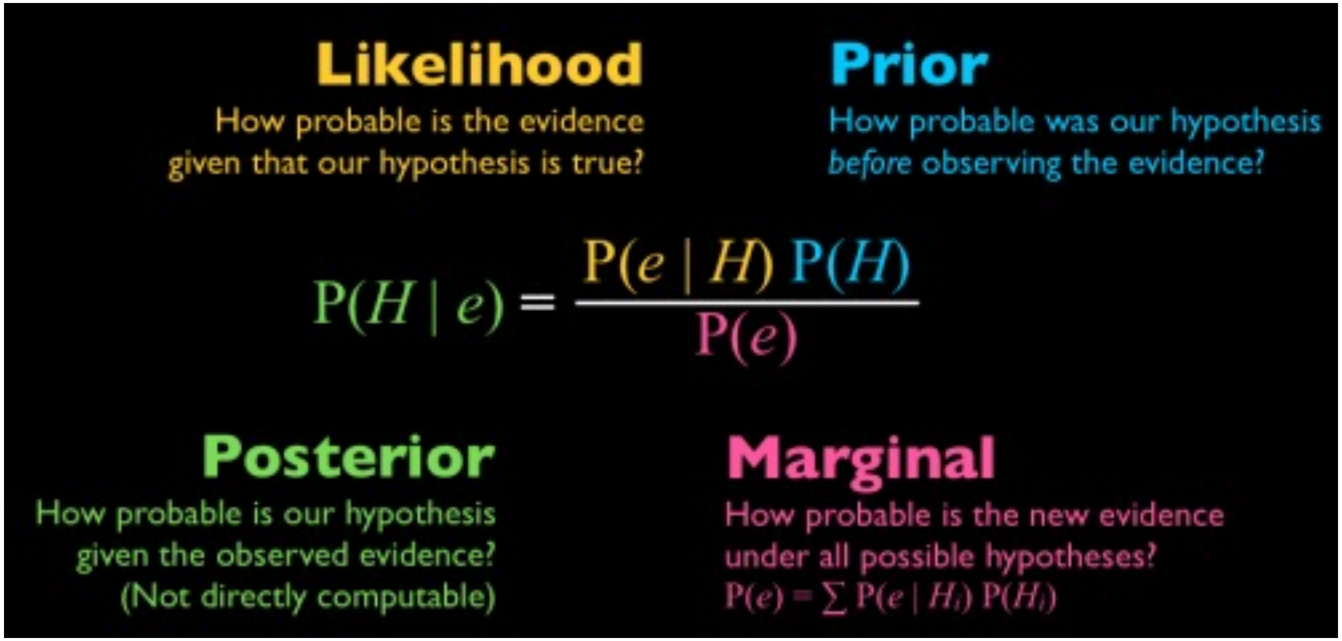
Duchenne Muscular Dystrophy (X-linked recessive)



What is P(III.2 is a carrier)?

BAYESIAN INFERENCE AND LIKELIHOODS

Bayesian statistics allows us to convert likelihoods to **actual probabilities**



Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

Posterior

How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

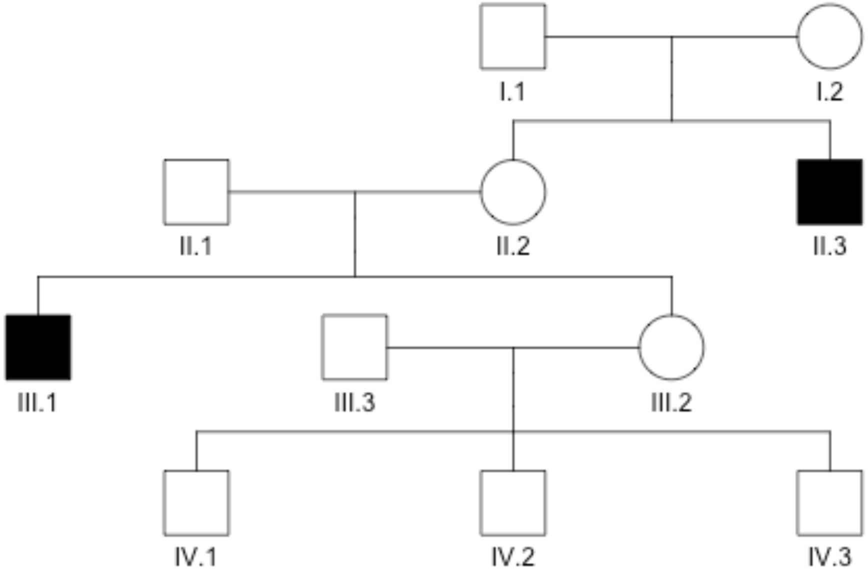
How probable is the new evidence under all possible hypotheses?

$\mathrm{P}(e) = \sum \mathrm{P}(e \mid H_i)\,\mathrm{P}(H_i)$

https://towardsdatascience.com/what-is-bayesian-statistics-used-for-37b91c2c257c

CONDITIONAL PROBABILITIES: INCORPORATING NEW INFORMATION

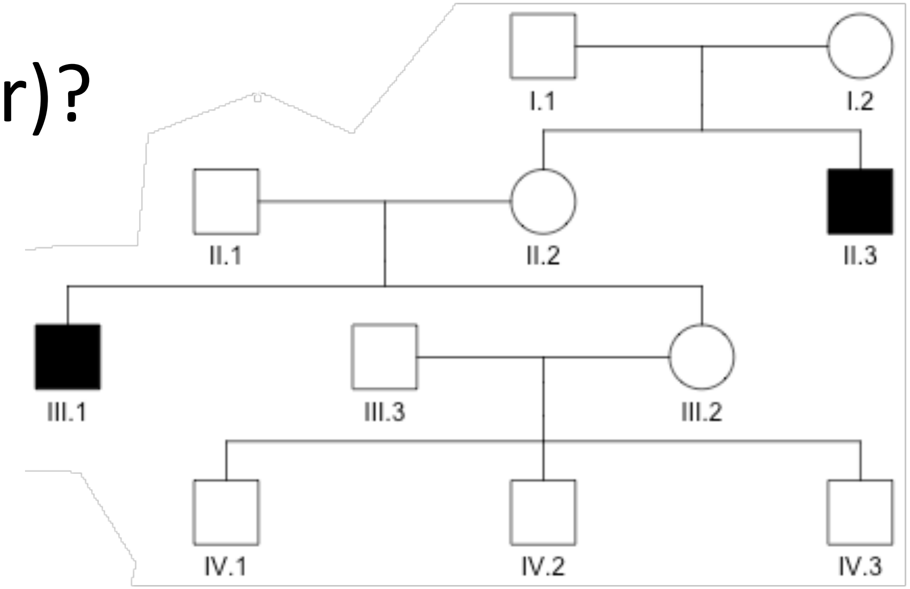
Duchenne Muscular Dystrophy (X-linked recessive)



What is P(III.2 is a carrier)?

P(III.2 IS CARRIER): CALCULATING THE "PRIOR"

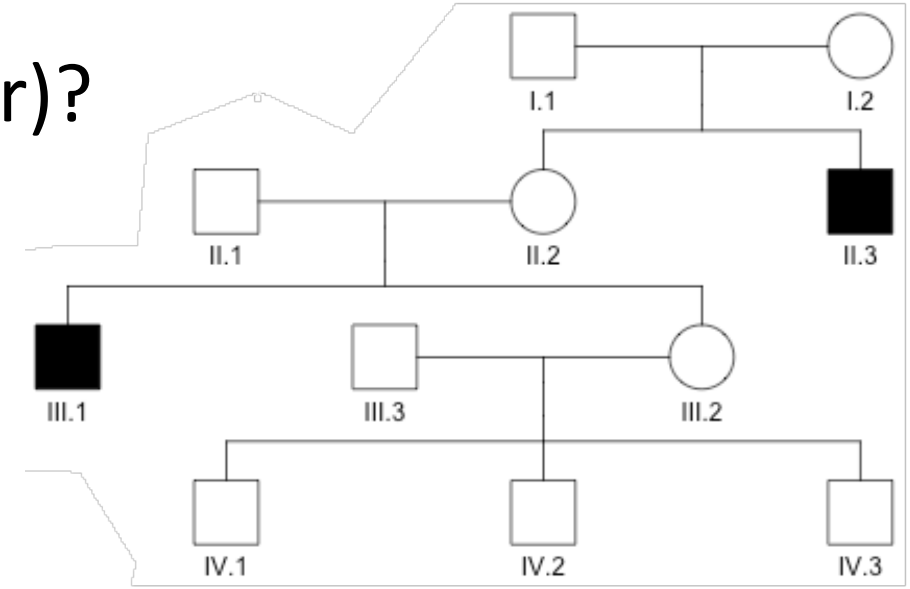
- What is P(III.2 = carrier)?



$$P(H \mid e) = \frac{P(e \mid H)\,P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE “PRIOR”

- What is P(III.2 = carrier)?



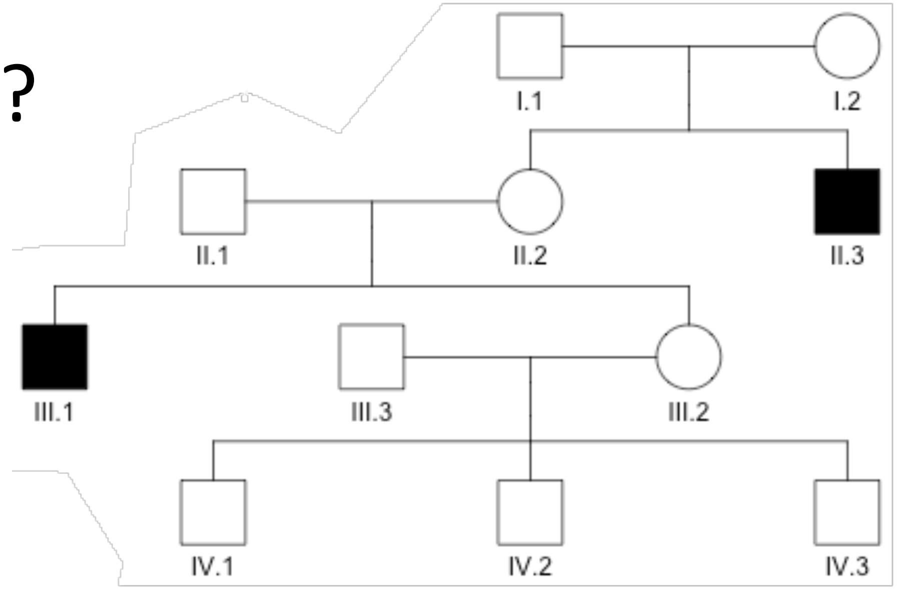
- P(III.2 = carrier) = P(H) = 0.5

$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

P(III.2 IS CARRIER):
CALCULATING THE "LIKELIHOOD"

- What is our 'evidence'?

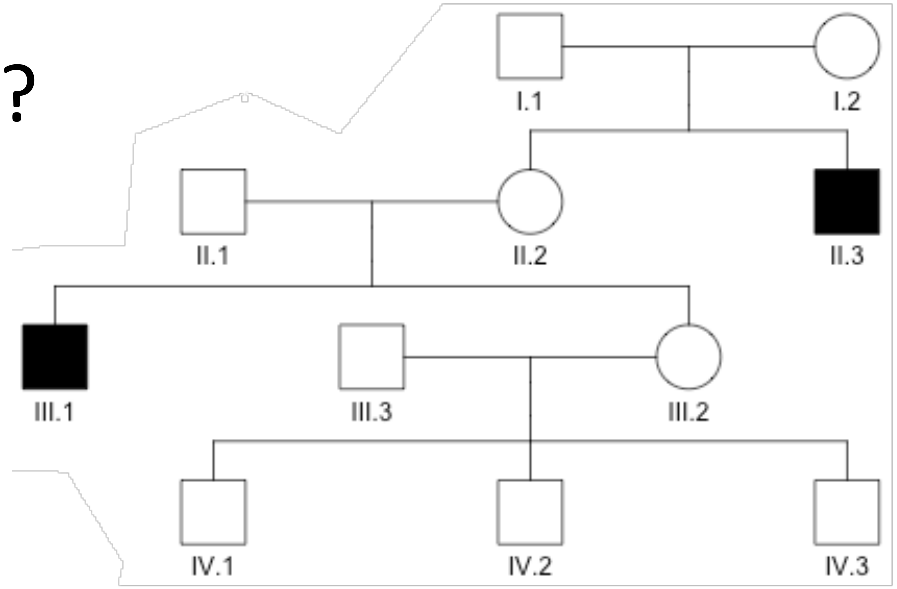
- What is P(e|H)?



$$P(H \mid e) = \frac{P(e \mid H)\,P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE "LIKELIHOOD"

- What is our 'evidence'?



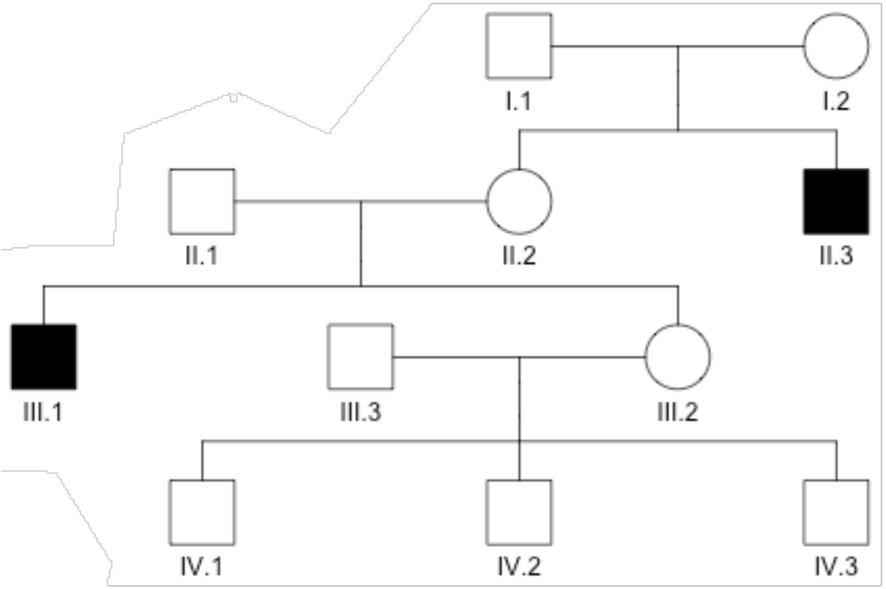
- What is P(e|H)?

- P(e|H) =
  P(3 successes in 3 trials | P(success) = 0.5) =
  0.125

$$P(H \mid e) = \frac{P(e \mid H)\,P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE “LIKELIHOOD”

- What is P(e)?



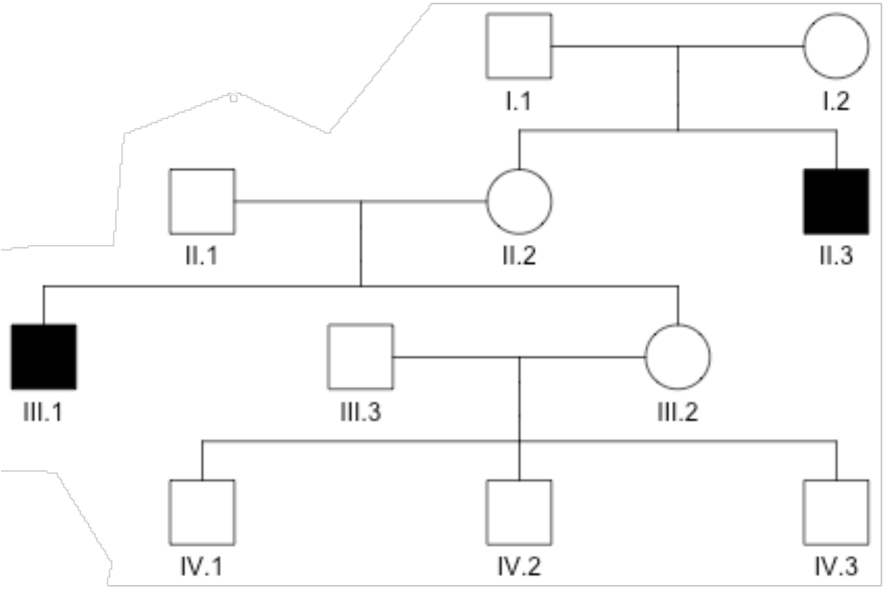
$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

P(III.2 IS CARRIER): CALCULATING THE "MARGINAL"

- What is P(e)?

- **P(e) is ALL THE WAYS 'Evidence' can happen, i.e. with our hypothesis being true OR with our hypothesis NOT being true!**

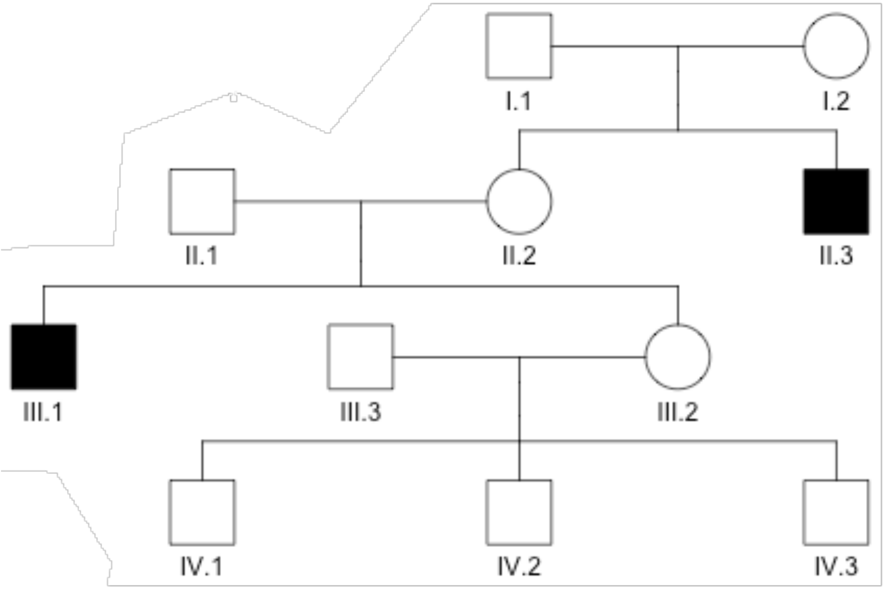
- **LAW OF TOTAL PROBABILITY**



$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

P(III.2 IS CARRIER): CALCULATING THE "MARGINAL"

- What is P(e)?

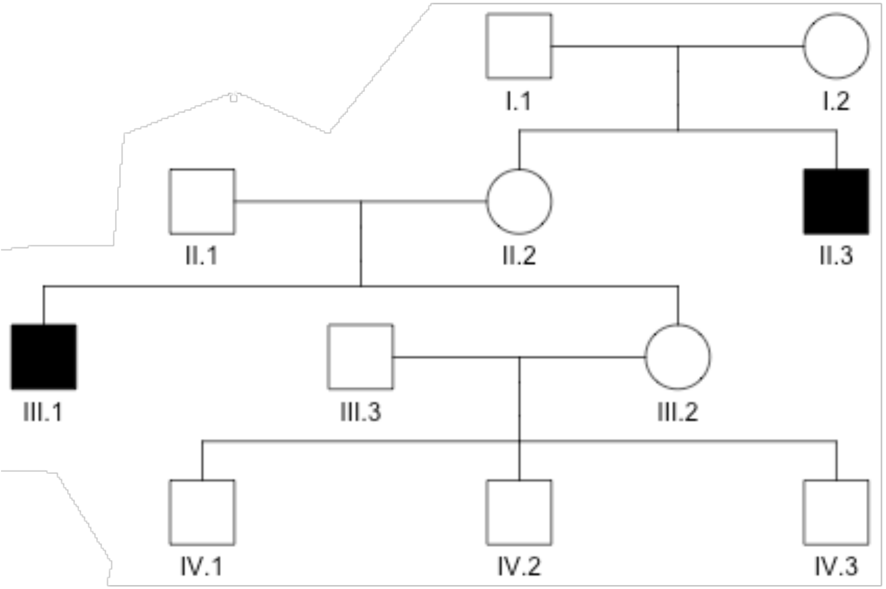


- P(e) = P(e ∩ H) + P(e ∩ !H)

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE “MARGINAL”

- What is P(e)?

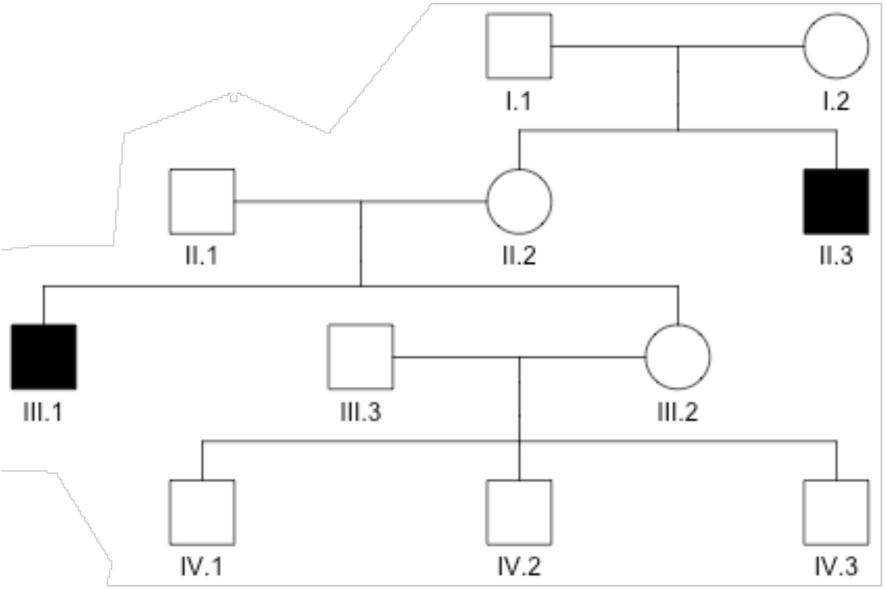


- P(e) = P(e ∩ H) + P(e ∩ !H) = P(e | H)P(H) + P(e | H)P(!H)

$$P(H \mid e) = \frac{P(e \mid H)\,P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE "MARGINAL"

- What is P(e)?



- P(e) = P(e ∩ H) + P(e ∩ !H) = P(e | H)P(H) + P(e | !H)P(!H) = 0.125*.5 + 1*.5 = 0.5625

$$P(H \mid e) = \frac{P(e \mid H)\,P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE “POSTERIOR”

- Hypothesis: III.2 is carrier
- Evidence: 3 non-affected male offspring

- P(III.2 = carrier GIVEN the evidence) = P(3 non-affected offspring if III.2 is carrier) * P(III.2 is carrier before evidence) / P(all the ways evidence can happen)

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

P(III.2 IS CARRIER): CALCULATING THE “POSTERIOR”

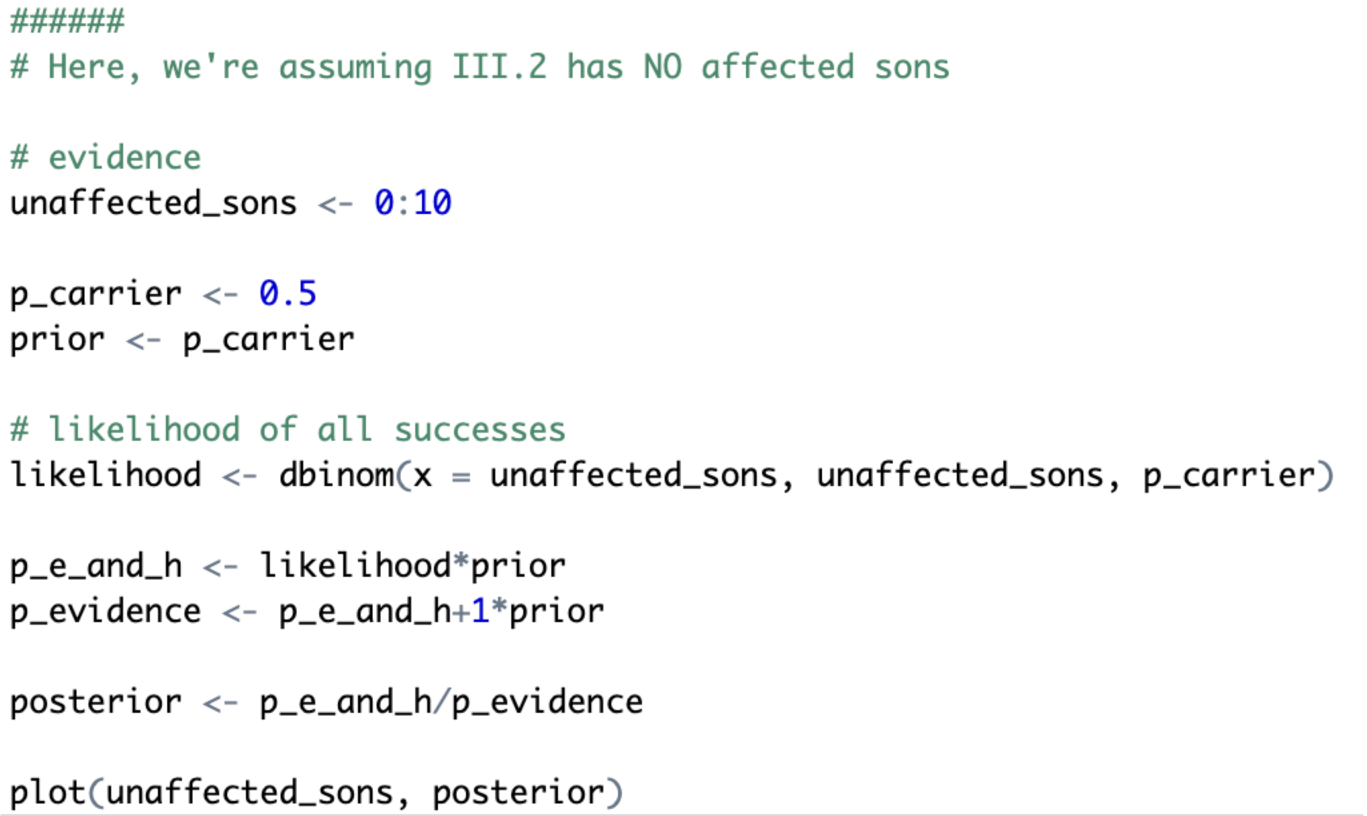
- Hypothesis: III.2 is carrier
- Evidence: 3 non-affected male offspring

- P(III.2 = carrier GIVEN the evidence) =
  P(3 non-affected offspring if III.2 is carrier) *
  P(III.2 is carrier before evidence) /
  P(all the ways evidence can happen) =
  0.125 * 0.5 / 0.5625 = 0.11

$$\mathrm{P}(H \mid e) = \frac{\mathrm{P}(e \mid H)\,\mathrm{P}(H)}{\mathrm{P}(e)}$$

WE CAN GET POSTERIOR FOR DIFFERENT NUMBERS OF UNAFFECTED SONS!

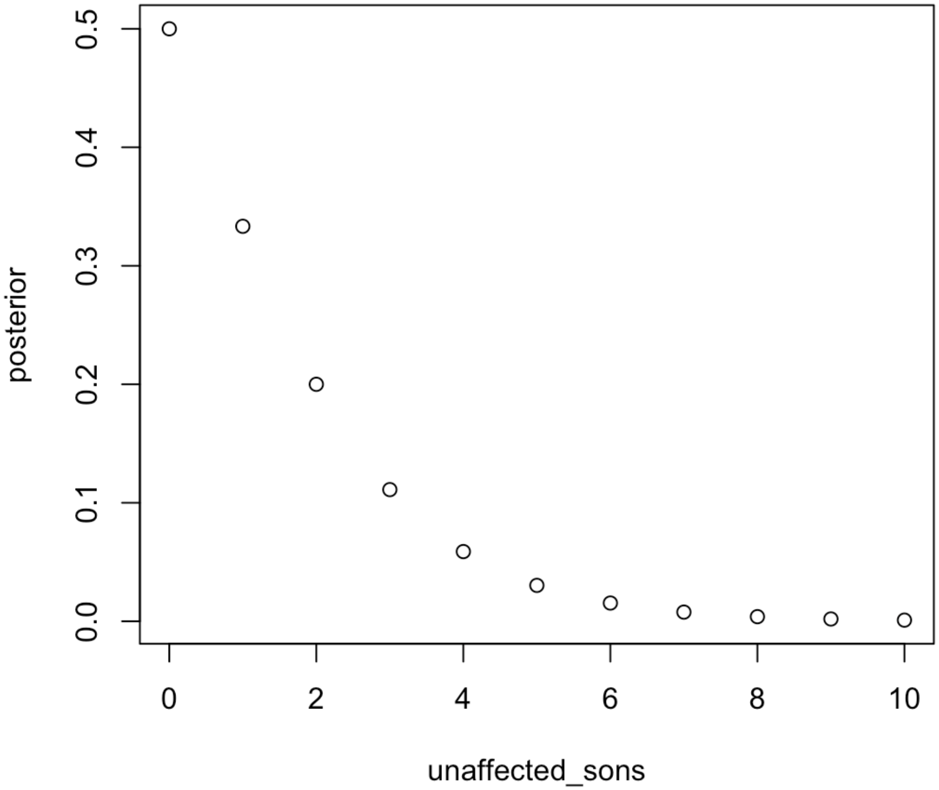
Let’s see how our belief that III.2 is a carrier changes for different numbers of UNAFFECTED sons that she has (assuming no affected sons)



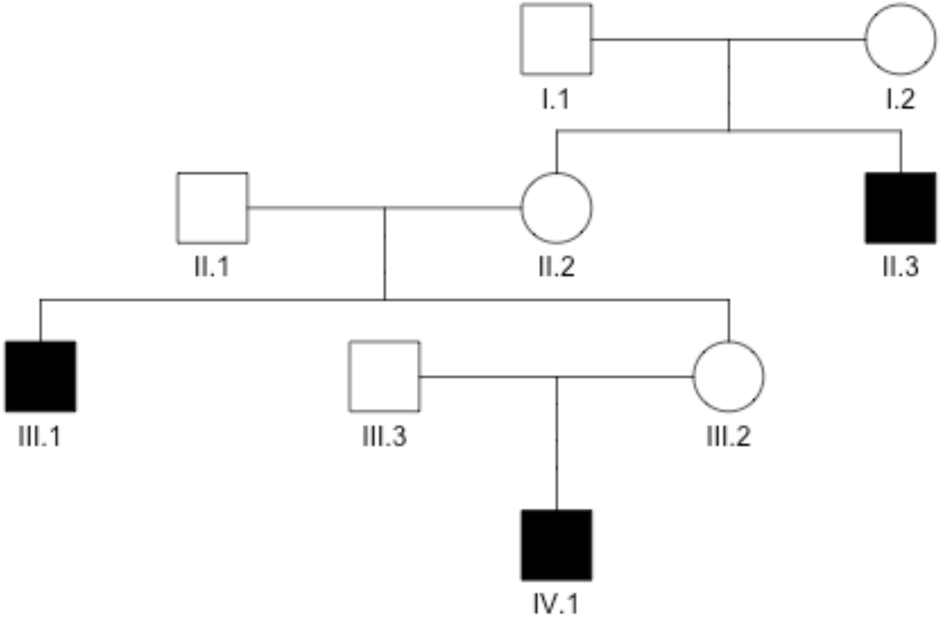
```r
######
# Here, we're assuming III.2 has NO affected sons

# evidence
unaffected_sons <- 0:10

p_carrier <- 0.5
prior <- p_carrier

# likelihood of all successes
likelihood <- dbinom(x = unaffected_sons, unaffected_sons, p_carrier)

p_e_and_h <- likelihood*prior
p_evidence <- p_e_and_h+1*prior

posterior <- p_e_and_h/p_evidence

plot(unaffected_sons, posterior)
```

WE CAN GET POSTERIOR FOR DIFFERENT NUMBERS OF UNAFFECTED SONS!

Let’s see how our belief that III.2 is a carrier changes for different numbers of UNAFFECTED sons that she has (assuming no affected sons)

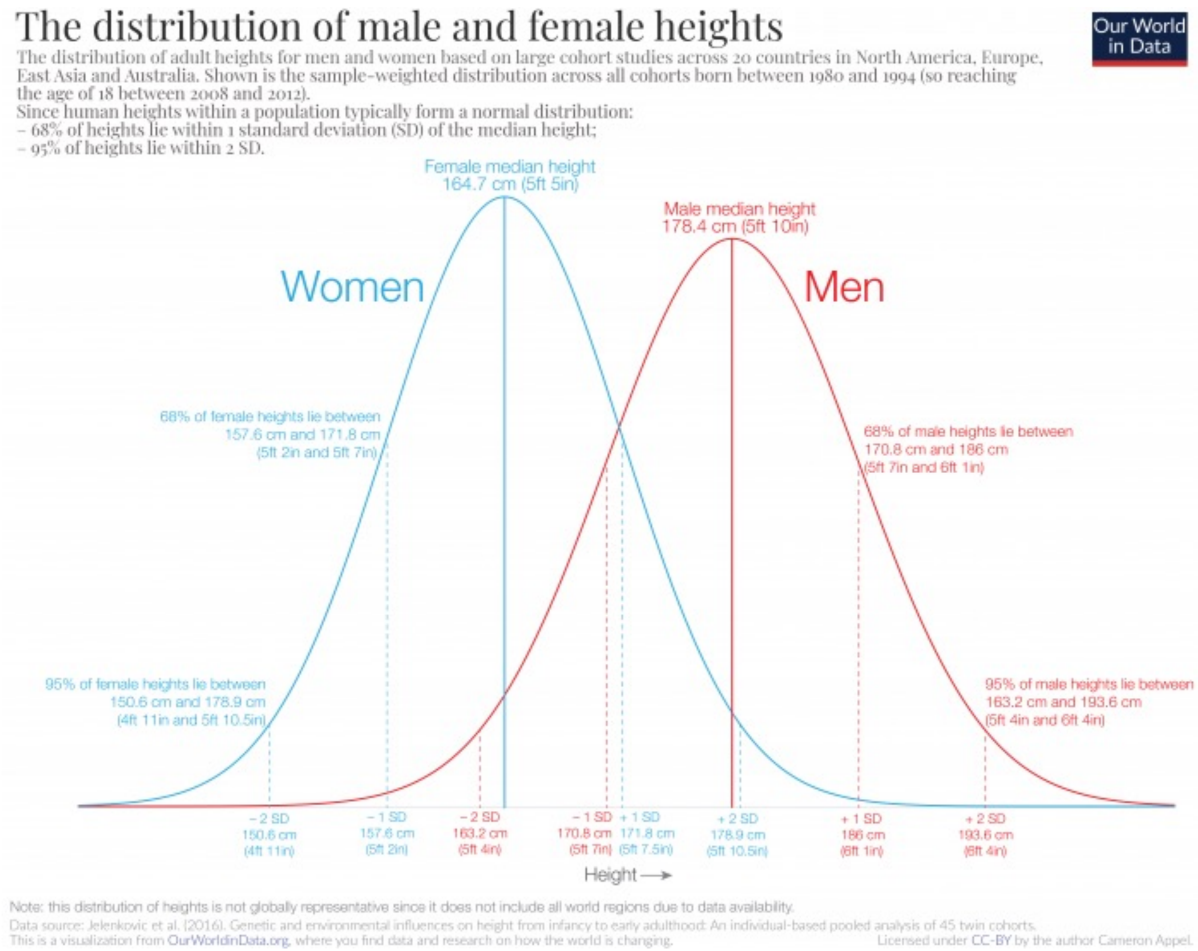


PRACTICE:
CALCULATE POSTERIOR IF 1 SON AFFECTED



$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

DISTRIBUTION OF HUMAN HEIGHT



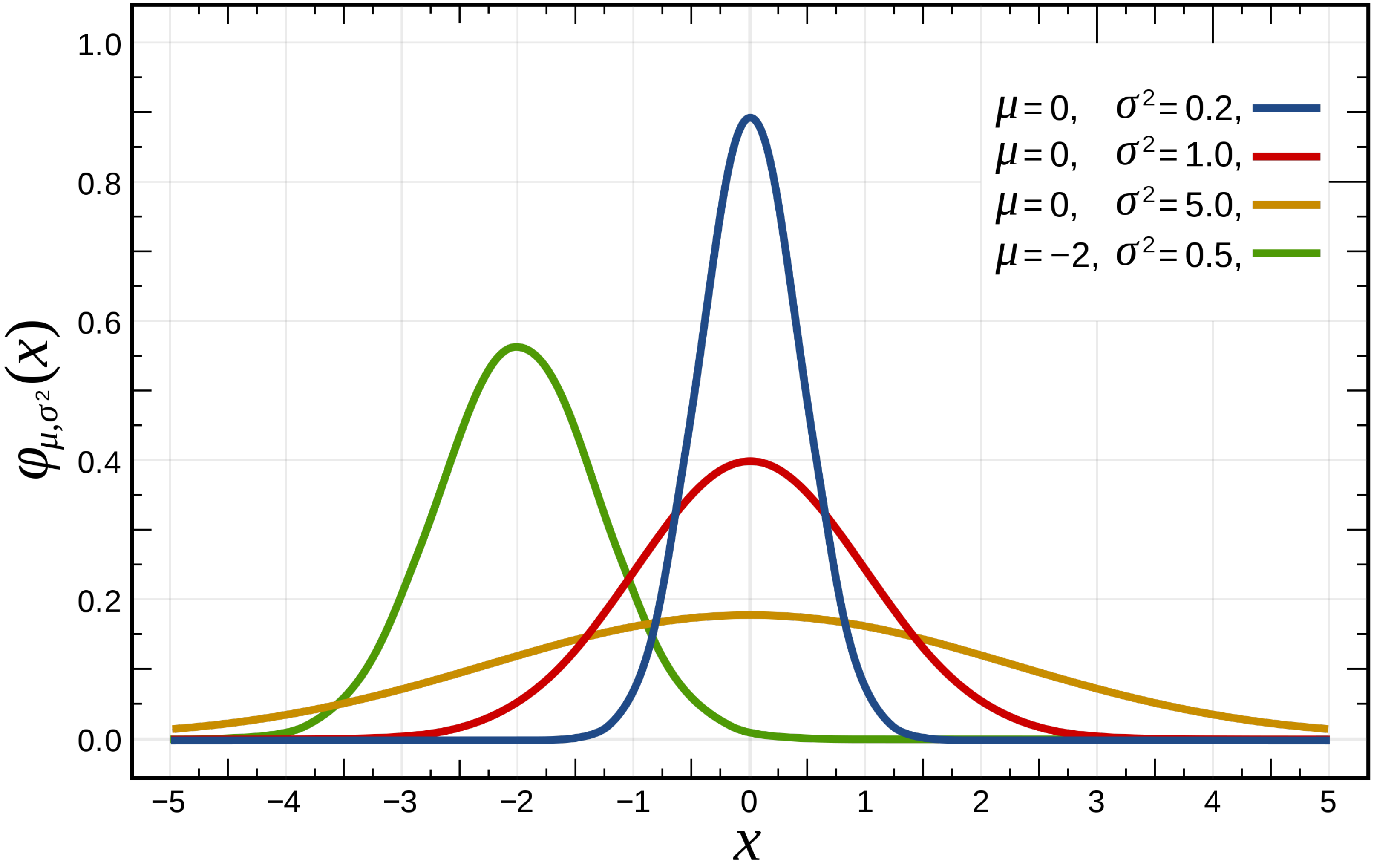
The distribution of male and female heights

The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.

Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.
Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.
This is a visualization from OurWorldinData.org, where you find data and research on how the world is changing. Licensed under CC-BY by the author Cameron Appel.

https://ourworldindata.org/human-height

NORMAL DISTRIBUTIONS



By Inductiveload - self-made, Mathematica, Inkscape, Public Domain,
https://commons.wikimedia.org/w/index.php?curid=3817954