SOME FUN TOPICS IN PROBABILITY

Conditional and joint probabilities

Calculating probabilities from data

Independence and conditional independence

Simpson’s paradox

The rule of total probability

Bayes rule

PROBABILITY

- Our notation: P(some event) = Number between 0 and 1
- 1 means certain, 0 means impossible
- For example:

  P(coin lands heads) = 0.5

  P(flight departs on time) = 0.79

  P(sun in Austin) = 0.85

  P(rain in Dublin) = 0.4

  P(cold day in Hell) = 0.0000000001

  etc.

JOINT PROBABILITY

- A joint probability is the chance that two or more events both happen.
- Our notation: P(A, B) — “P(A and B)” or “the joint probability of A and B”
- For example:

  P(temp < 45F, wind speed > 10mph)

  P(Longhorns score 31 points, Sooners score 28 points)

  P(AAPL up, AMZN down)

CONDITIONAL PROBABILITY

- A conditional probability is the chance that one thing (A) happens, given that some other thing (B) has already happened.
- Our notation: P(A | B) — “Probability of A, given B”
- Conditional probabilities reflect our uncertainty in light of partial knowledge:

  P(rain this afternoon | cloudy this morning)

  P(UT beats OU | UT ahead by a touchdown at halftime)

  P(accepted to Dell Medical School | college GPA > 3.6)

FOR EXAMPLE...

- **Instagram: P(follow @LeoMessi | follow @Cristiano)**
- **Amazon: P(buy organic dog food | buy GPS dog collar)**
- **Netflix: P(watch *Tinker Tailor Soldier Spy* | watch *Sherlock*)**

FOR EXAMPLE...

- Instagram: P(follow @LeoMessi | follow @Cristiano)
- Amazon: P(buy organic dog food | buy GPS dog collar)
- Netflix: P(watch *Tinker Tailor Soldier Spy* | watch *Sherlock*)
- YouTube
- Google
- Spotify
- *The New York Times*
- Twitch
- Facebook
- EBay

KEY FACT:
$P(A \mid B) \neq P(B \mid A)$

This is the single most important fact to remember about conditional probabilities!

EXAMPLE

- Suppose we're looking at these two events:

  A: You can dribble a basketball

  B: You play in the NBA

- What is P(A | B)?
- What is P(B | A)?

P(CAN DRIBBLE BASKETBALL | PLAYS IN NBA) = 1

P(PLAYS IN NBA| CAN DRIBBLE BASKETBALL) ≈ 0

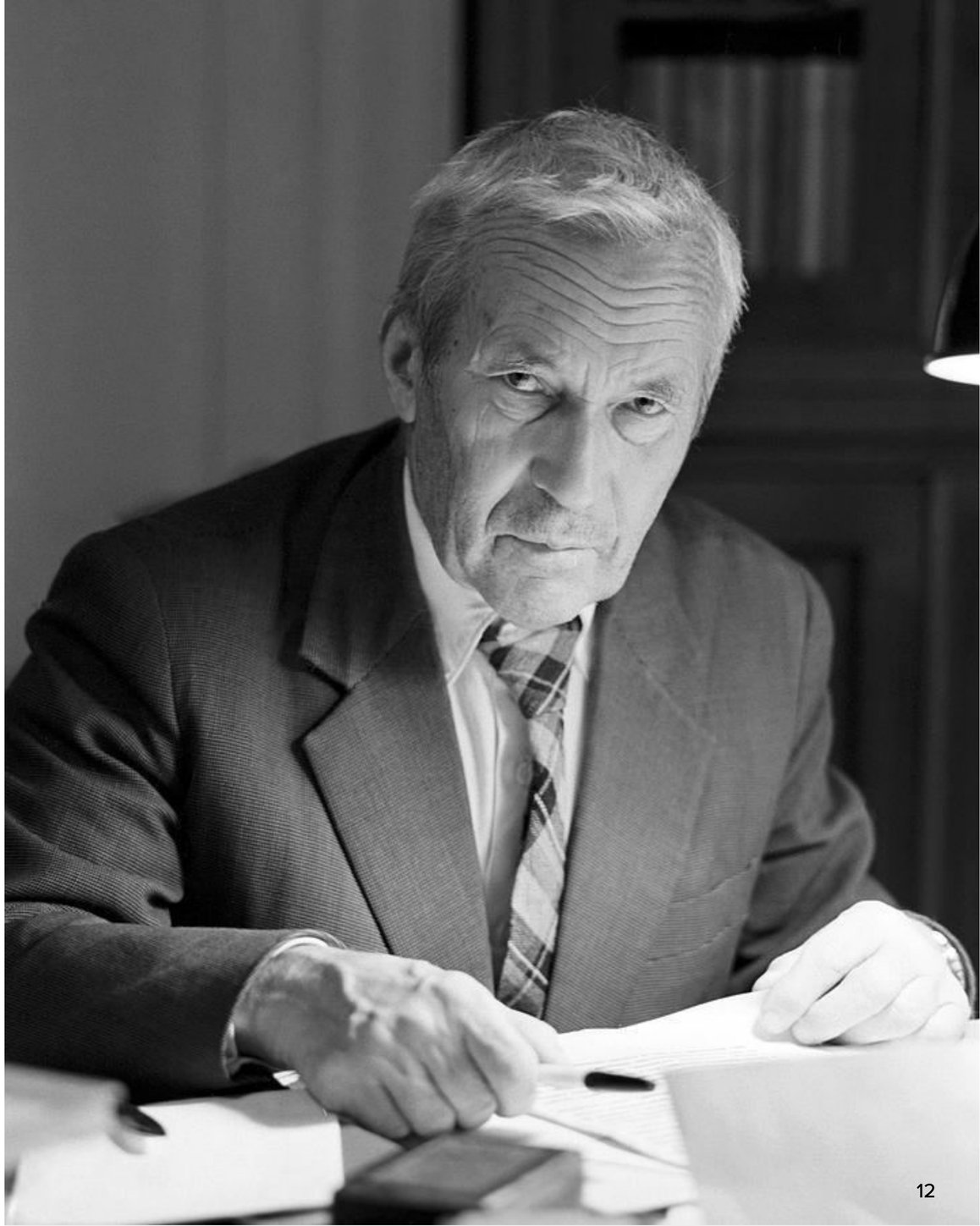
KEY FACT:
$P(A \mid B) \neq P(B \mid A)$

Moral of the story:

Always be specific about what's on the left-hand side, and what's on the right-hand side.

KOLMOGOROV'S AXIOMS (BASIC VERSION)

- Consider an uncertain outcome with sample space Ω. "Probability" P(•) is a set function that maps Ω to the real numbers, such that:
  1. Non-negativity: $P(A) \geq 0$ for all A.
  2. Normalization: $P(\Omega) = 1$ and $P(\varnothing) = 0$.
  3. Additivity: if A and B are disjoint events, i.e. disjoint subsets of the sample space, then $P(A \cup B) = P(A) + P(B)$
- Not that intuitive! No mention of frequencies....



THE CONDITIONAL PROBABILITY RULE

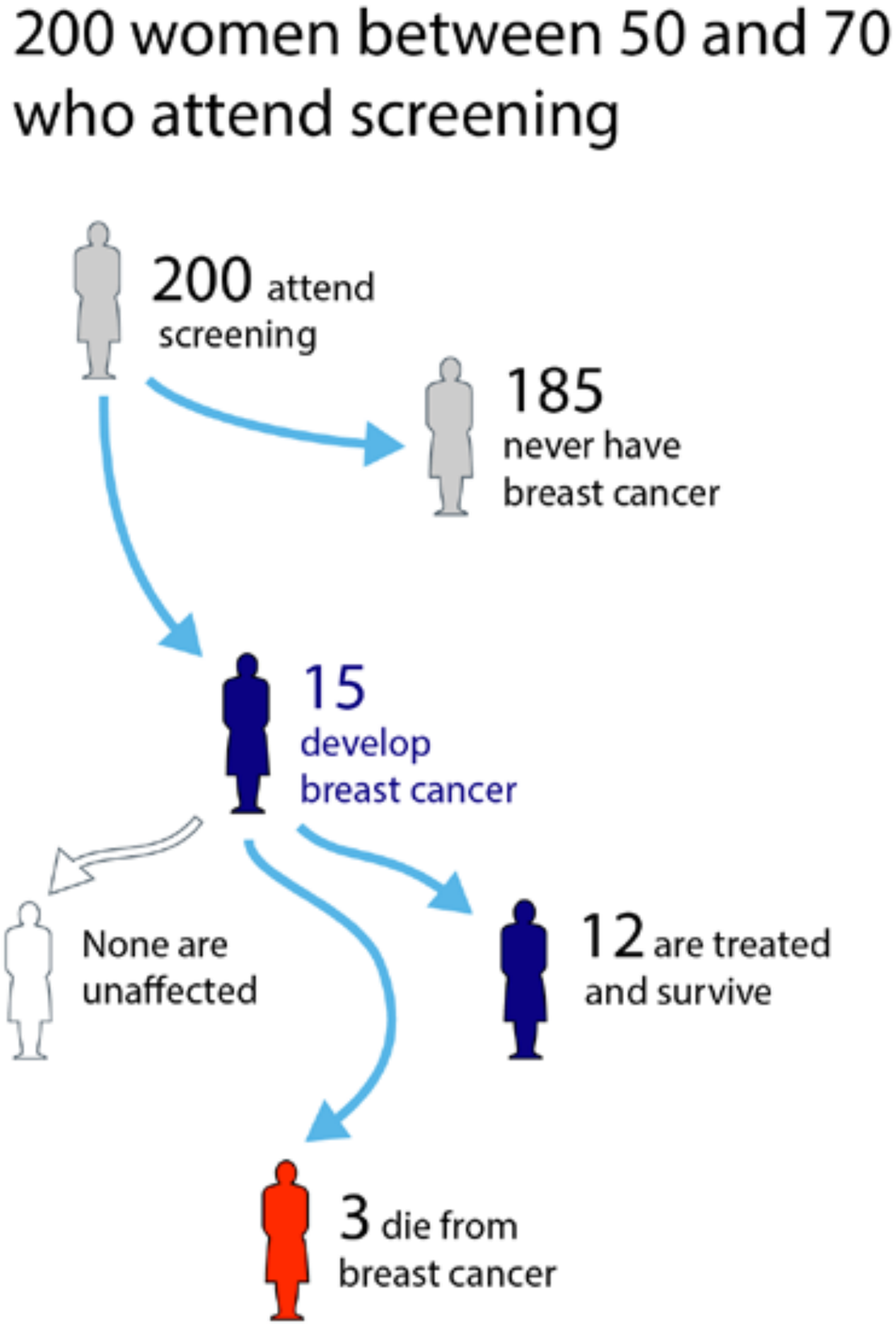
- The following rule establishes a very important relationship between joint and conditional probabilities.

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

- Let’s see an example.

EXAMPLE: MAMMOGRAMS

- The picture at right shows a data visualization from scientists at Cambridge University. It concerns women aged 50-70 who attend regular screening mammograms.
- What is P(die | cancer)?



EXAMPLE: MAMMOGRAMS

- The picture at right shows a data visualization from scientists at Cambridge University. It concerns women aged 50-70 who attend regular screening mammograms.
- What is P(die | cancer)?
- The conditional probability rule says that:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

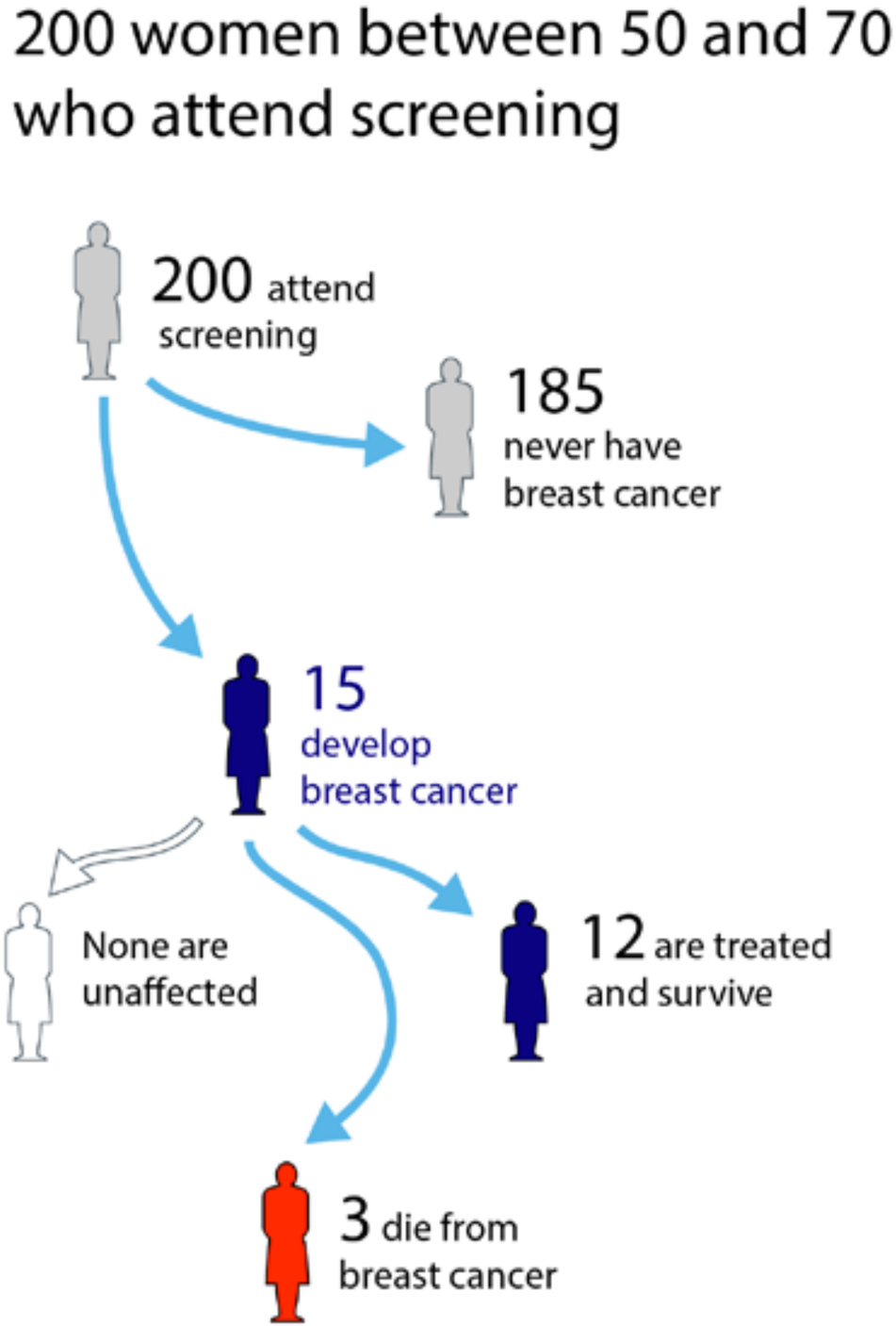
- We know from the diagram that:

  P(cancer) = 15/200

  P(die, cancer) = 3/200

- Therefore

$$P(\text{die} \mid \text{cancer}) = \frac{3/200}{15/200} = 3/15 = 0.2$$



WHERE DO PROBABILITIES COME FROM?

- From careful counting
- From subjective judgment
- From other probabilities
- From data

Basic idea: just use observed proportions/frequencies to approximate probabilities.

PROBABILITIES FROM DATA

- Proportion of U.S. babies that are female:

  100/206 ≈ 0.485

- Proportion of Americans that are in a car accident each year:

  9/1000 = 0.009

- Proportion of Netflix subscribers that like Saving Private Ryan:

  ?

Q: What proportion of Netflix subscribers liked “Saving Private Ryan”?

Subscriber	Liked Saving Private
1. Ben A.	Yes
2. Alejandra C.	Yes
…	…
99. Rashid T.	No
100. Wei Z.	Yes

Raw data

Q: What proportion of Netflix subscribers liked “Saving Private Ryan”?

A: Just count!

Subscriber	Liked Saving Private
1. Ben A.	Yes
2. Alejandra C.	Yes
…	…
99. Rashid T.	No
100. Wei Z.	Yes

Raw data

Count

Yes	No
70	30

Table of counts (“cross tab”)

$$\text{P(Likes SPR)} = \frac{70}{70+30} = 0.7$$

CONDITIONAL PROBABILITIES FROM DATA

- The same idea works for conditional probabilities.

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \approx \frac{\text{Frequency of A and B both happening}}{\text{Frequency of B happening}}$$

- Let's start with an example:

  What is P(likes Saving Private Ryan | likes Band of Brothers)?

  This is the kind of conditional probability that underpins any "recommender system" (Netflix, Amazon, Spotify, etc.)

Subscriber	Liked Saving Private Ryan?	Liked Band of Brothers?
1. Ben Armstrong	Yes	Yes
2. Alejandra Contreras	Yes	Yes
...	...	...
99. Rashid Tannous	No	No
100. Anna Yeo	Yes	No

	Liked *Saving Private Ryan*	Didn’t like it
Liked *Band of Brothers*	56	6
Didn’t like it	14	24

	Liked *Saving Private Ryan*	Didn't like it
Liked *Band of Brothers*	56	6
Didn't like it	14	24

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \approx \frac{\text{Frequency of A and B both happening}}{\text{Frequency of B happening}}$$

	Liked *Saving Private Ryan*	Didn't like it
Liked *Band of Brothers*	56	6
Didn't like it	14	24

$$P(\text{Likes SPR} \mid \text{Likes BB}) = \frac{56/100}{(56+6)/100} = \frac{56}{56+6} \approx 0.9$$

Notice how the total sample size (N=100) cancels in the fraction.

So we can work directly with the counts.

Example 2: ACL Fest

Band	ACL	Bonnaroo	Coachella	Lollapalooza	Outside Lands
Jimmy Cliff	0	1	0	1	0
Pretty Lights	1	1	0	1	0
Lila Downs	0	0	0	0	1
Rebelution	1	1	0	1	0
Black Joe Lewis and the Honeybears	0	0	1	0	0
Explosions In The Sky	0	1	0	1	0
Brand New	0	0	0	1	0
Frank Turner	1	0	1	0	0
Local Natives	0	0	0	1	0
Nas & Damian Marley	0	0	1	0	0

(+ 1,228 more rows)

	Didn’t play Lollapalooza	Played Lollapalooza
Didn’t play ACL	719	361
Played ACL	81	77

P(played ACL | played Lollapalooza) = ?

SUMMARY

- Conditional probabilities depend on two things:
  - What we want to know (whether A will happen)
  - What we know or assume to be true (that B happened)
- Question: what is P(A | B)?
- Answer: go to the data!
  - Form a table of counts of the possible outcomes (a “cross-tab” or “contingency table”).
  - Estimate the conditional probability as:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \approx \frac{\text{Frequency of A and B both happening}}{\text{Frequency of B happening}}$$

  - This is just the “data science” version of the mathematician’s rule for conditional probabilities.

INDEPENDENCE

- Two events A and B are said to be independent if:

  $P(A \mid B) = P(A \mid \text{not } B) = P(A)$

- In words, A and B convey no information about each other:

  $P(\text{coin 2 lands heads} \mid \text{coin 1 lands heads}) = P(\text{coin 2 lands heads})$

  $P(\text{stock market up} \mid \text{bird poops on your car}) = P(\text{stock market up})$

  $P(\text{God exists} \mid \text{Longhorns win title}) = P(\text{God exists})$

INDEPENDENCE

- So if A and B are independent, then by the multiplication rule:

  $P(A, B) = P(A) \cdot P(B \mid A) = P(A) \cdot P(B)$

- Sometimes events are independent:

  $P(\text{flip 1 lands heads, flip 2 lands heads}) = P(\text{flip 1 lands heads}) \cdot P(\text{flip 2 lands heads})$

  $P(\text{AAPL up today, AAPL up tomorrow}) = P(\text{AAPL up today}) \cdot P(\text{AAPL up tomorrow})$

- And sometimes they’re not:

  $P(\text{rain, high winds}) \neq P(\text{rain}) \cdot P(\text{high winds})$

  $P(\text{sib 1 colorblind, sib 2 colorblind}) \neq P(\text{sib 1 colorblind}) \cdot P(\text{sib 2 colorblind})$

CHECKING INDEPENDENCE FROM DATA

- Suppose we have two outcomes A and B and we want to know if they're independent.
- Solution:

  Check whether B happening seems to change the probability of A happening.

  That is, verify using data whether P(A | B) = P(A | not B) = P(A).
- These probabilities won't be *exactly* alike because of statistical fluctuations, especially with small samples.
- But with enough data they should be pretty close if A and B are independent.

EXAMPLE: NBA SHOOTING

- The "hot hand hypothesis" says that if a player makes their previous shot, they're more likely to make their next shot (see: NBA Jam).
- The “independence hypothesis” says that a player’s next shot doesn’t depend on the outcome of the previous shot.

EXAMPLE: NBA SHOOTING

- The "hot hand hypothesis" says that if a player makes their previous shot, they're more likely to make their next shot (see: NBA Jam).
- The “independence hypothesis” says that a player’s next shot doesn’t depend on the outcome of the previous shot.
- Let’s check from data! Compare two probabilities:

  P(makes next | makes previous)

  P(makes next)

- We’ll look at data from Julius Irving’s 1981 Philadelphia 76ers.

Shooting percentage by situation

Player (ordered from most shots to fewest shots)

	3 misses	2 misses	1 miss	overall	1 hit	2 hits	3 hits
Julius Erving	0.52	0.51	0.51	0.52	0.52	0.53	0.48
Caldwell Jones	0.50	0.48	0.47	0.43	0.47	0.45	0.27
Maurice Cheeks	0.77	0.60	0.60	0.54	0.56	0.55	0.59
Daryl Dawkins	0.88	0.73	0.71	0.58	0.62	0.57	0.51
Lionel Hollins	0.50	0.49	0.46	0.46	0.46	0.46	0.32
Bobby Jones	0.61	0.58	0.58	0.47	0.54	0.53	0.53
Andrew Toney	0.52	0.53	0.51	0.40	0.46	0.43	0.34
Clint Richardson	0.50	0.47	0.56	0.50	0.50	0.49	0.48
Steve Mix	0.70	0.56	0.52	0.48	0.52	0.51	0.36

Self-test: which looks more correct to you? The hot-hand hypothesis, or the independence hypothesis?

CONDITIONAL INDEPENDENCE

- **Two events A and B are conditionally independent, given C, if**

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

- **Said another way, A and B convey no information about each other, once we know C:**

$$P(A \mid B, C) = P(A \mid C)$$

- **Neither independence nor conditional independence implies the other.**

  It is possible for two outcomes to be *dependent* and yet *conditionally independent*.

  Less intuitively, it is possible for two outcomes to be *independent* and yet *conditionally dependent*.

CONDITIONAL INDEPENDENCE

- Let's see an example. Alice and Brianna live next door to each other and both commute to work on the same metro line.

  A = Alice is late for work.

  B = Brianna is late for work.

- A and B are dependent:

  if Brianna is late for work, we might infer that the metro line was delayed or that their neighborhood had bad weather.

  This means Alice is more likely to be late for work: $P(A \mid B) > P(A)$

CONDITIONAL INDEPENDENCE

- **Now let's add some additional information:**

  A = Alice is late for work.

  B = Brianna is late for work.

  C = The metro is running on time and the weather is clear.

- **A and B are conditionally independent, given information C:**

  If Brianna is late for work but we know that the metro is running on time and the weather is clear, then we don't really learn anything about Alice's commute:

  $P(A \mid B) > P(A)$**, but...**

  $P(A \mid B, C) = P(A \mid C)$

CONDITIONAL INDEPENDENCE

- Same characters, different story.

  A = Alice has blue eyes.

  B = Brianna has blue eyes.

- A and B are independent: Brianna’s eye color can't give us information about Alice’s.

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

CONDITIONAL INDEPENDENCE

- **Again, let's add some additional information:**

  A = Alice has blue eyes.

  B = Brianna has blue eyes.

  C = Alice and Brianna are sisters.

- **A and B are *conditionally dependent*, given C:**

  If Brianna has blue eyes, AND we know that Alice is her sister…

  Then we know something about Alice's genes. It is now more likely that she has blue eyes:

  $P(A \mid B) = P(A)$, **but…**

  $P(A \mid B, C) > P(A \mid C)$

A PARADOX

	Low-risk (easier)	High-risk (harder)	Overall
Senior doctor	0.052	0.127	0.076
Junior doctor	0.067	0.155	0.072

Complication rates across 3,690 deliveries at a large maternity hospital in Cambridge, UK

A PARADOX

	Low-risk (easier)	High-risk (harder)	Overall
Senior doctor	0.052	0.127	0.076
Junior doctor	0.067	0.155	0.072

Complication rates across 3,690 deliveries at a large maternity hospital in Cambridge, UK

A PARADOX

- Senior doctors are...

  better at easy cases...

  better at hard cases...

  yet worse overall.

- This is an example of *Simpson's paradox*. How is it possible?

	Low-risk (easier)	High-risk (harder)	Overall
Senior doctor	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctor	0.067 (3169)	0.155 (206)	0.072 (3375)

Complication rates and sample sizes across 3,690 deliveries at a large maternity hospital in Cambridge, UK

RULE OF TOTAL PROBABILITY

- In words: the probability of an event is the sum of the probabilities for all the different ways in which that event can happen:

  P(rain) = P(rain, wind) + P(rain, no wind)

  P(complication) = P(complication, easy delivery) + P(complication, hard delivery)

RULE OF TOTAL PROBABILITY

- In words: the probability of an event is the sum of the probabilities for all the different ways in which that event can happen:

  P(rain) = P(rain, wind) + P(rain, no wind)

  P(complication) = P(complication, easy delivery) + P(complication, hard delivery)

- Suppose that B_1, B_2, ... B_N are mutually exclusive events whose probabilities sum to 1:

$$P(B_i, B_j) = 0 \quad (i \neq j) \quad \text{and} \quad \sum_{i=1}^{N} P(B_i) = 1$$

- Then for any event A:

$$P(A) = \sum_{i=1}^{N} P(A, B_i) = \sum_{i=1}^{N} P(A \mid B_i) \cdot P(B_i)$$

The second part of the equation comes from the multiplication rule: P(A, B) = P(A | B) • P(B)

RULE OF TOTAL PROBABILITY

- So, for example, the overall (total) probability of a complication is:

$$P(\text{comp}) = P(\text{comp, easy}) + P(\text{comp, hard})$$
$$= P(\text{easy}) \cdot P(\text{comp} \mid \text{easy}) + P(\text{hard}) \cdot P(\text{comp} \mid \text{hard})$$

RULE OF TOTAL PROBABILITY

	Low-risk (easier)	High-risk (harder)	Overall
Senior doctor	0.052 (213)	0.127 (102)	0.076 (315)
Junior doctor	0.067 (3169)	0.155 (206)	0.072 (3375)

- So, for example, the overall (total) probability of a complication is:

$$P(\text{comp}) = P(\text{comp, easy}) + P(\text{comp, hard})$$
$$= P(\text{easy}) \cdot P(\text{comp} \mid \text{easy}) + P(\text{hard}) \cdot P(\text{comp} \mid \text{hard})$$

- For senior doctors:

$$P(\text{comp}) = \left(\frac{213}{213+102}\right) \cdot 0.052 + \left(\frac{102}{213+102}\right) \cdot 0.127 = 0.076$$

- For junior doctors:

$$P(\text{comp}) = \left(\frac{3169}{3169+206}\right) \cdot 0.067 + \left(\frac{206}{3169+206}\right) \cdot 0.155 = 0.072$$

PARADOX 1 RESOLVED

- Senior doctors are…

  better at easy cases and better at hard cases…

  yet worse overall.

- This is an example of *Simpson's paradox*. Here's how it's possible:

  P(comp | easy) and P(comp | hard) are both lower for senior doctors…

  yet senior doctors work fewer easy cases: P(easy) is lower in the mixture!

- Moral of the story:

  Make sure you're asking the right question!

  Always be sensitive to whether probabilities are conditional or overall/total, and which type of probability makes more sense for your situation.

	Stones < 2cm	Stones ≥ 2cm	Overall
Open surgery	0.93	0.73	0
Percutaneous nephrolithotomy	0.83	0.69	0

Success rates in a medical study of two modes of treatment for kidney stones (from Julious and Mullee, BMJ 1994)

	Stones < 2cm	Stones ≥ 2cm	Overall
Open surgery	0.93	0.73	0
Percutaneous nephrolithotomy	0.83	0.69	0

Success rates in a medical study of two modes of treatment for kidney stones (from Julious and Mullee, BMJ 1994)

Which procedure has the higher *overall* success rate?

- **Open surgery**
- **Percutaneous nephrolithotomy**
- **It is impossible to tell.**
- **It is possible to tell, but we must know the sample sizes by procedure and stone type.**

	Stones < 2cm	Stones ≥ 2cm	Overall
Open surgery	0.93	0.73	0
Percutaneous nephrolithotomy	0.83	0.69	0

Success rates in a medical study of two modes of treatment for kidney stones (from Julious and Mullee, BMJ 1994)

Which procedure has the higher *overall* success rate?

- **Open surgery**
- **Percutaneous nephrolithotomy**
- **It is impossible to tell.**
- **It is possible to tell, but we must know the sample sizes by procedure and stone type.**

	Stones < 2cm	Stones ≥ 2cm	Overall
Open surgery	0.93 (87)	0.73 (263)	0.78 (350)
Percutaneous nephrolithotomy	0.83 (270)	0.69 (80)	0.80 (350)

Success rates in a medical study of two modes of treatment for kidney stones (from Julious and Mullee, BMJ 1994)

Which probabilities would you care about as a patient?

abc NEWS

Just Say Yes? Teens Not Always Honest About Drug Use

Teens don't always tell the truth about illegal drug use, a study says.

By **KIM CAROLLO, ABC News Medical Unit**

October 22, 2010, 11:15 AM • 5 min read

...and parents everywhere are shocked.

THE RESEARCH

- Virginia Delaney-Black and her colleagues at Wayne State University gave an anonymous survey to 432 teenagers, asking whether they had used various drugs.
- Of these 432 teens, 211 agreed to give a hair sample.
- Hair samples were analyzed in the aggregate: no hair sample could be traced back to an individual survey or teen.

V. Delaney-Black et. al. "Just Say I Don't: Lack of Concordance Between Teen Report and Biological Measures of Drug Use." *Pediatrics* 165:5, pp. 887-93 (2010)

THE RESULTS

- The two sets of results were strikingly different.

  Of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine.

  When the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

V. Delaney-Black et. al. "Just Say I Don't: Lack of Concordance Between Teen Report and Biological Measures of Drug Use." *Pediatrics* 165:5, pp. 887-93 (2010)

THE RESULTS

- The two sets of results were strikingly different.

  Of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine.

  When the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

- And the parents were asked about cocaine use as well:

  Only 6.1% said yes.

  But 28.3% of the hair samples came back positive.

V. Delaney-Black et. al. "Just Say I Don't: Lack of Concordance Between Teen Report and Biological Measures of Drug Use." *Pediatrics* 165:5, pp. 887-93 (2010)

WHY THIS IS A PROBLEM

- There were folks who:
  - were guaranteed anonymity.
  - wouldn’t be arrested or fired for saying yes,
  - willingly agreed to provide a hair sample.
- Yet a big fraction lied about their drug use anyway. Yikes!
  - Drug abuse is a huge social problem.
  - Pediatricians, schools, and governments all rely on self-reported measures of drug use to guide their thinking on this issue.
  - Can we trust the numbers?

OTHER RESEARCH SHOWS...

- **People lie about plenty of things in surveys:**

  Churchgoers overstate the amount of money they give when the hat gets passed around during the service.

  Gang members embellish the number of violent encounters they have been in.

  Men exaggerate their salary.

  Ravers will “confess” to having gotten high on drugs that do not actually exist.

OTHER RESEARCH SHOWS...

- People lie about plenty of things in surveys:

  Churchgoers overstate the amount of money they give when the hat gets passed around during the service.

  Gang members embellish the number of violent encounters they have been in.

  Men exaggerate their salary.

  Ravers will “confess” to having gotten high on drugs that do not actually exist.

- Paradoxically, this is good news!

  When people lie, it’s for predictable reasons and in predictable ways.

  “Predictable” means we can use probability to get at the truth.

OUR SURVEY

- Flip a coin and keep the result private.

  If you don’t have a coin handy, Google “flip a coin” and use that one.

- If heads, answer Q1: Is the last digit of your tax ID number (SSN) odd?
- If tails, answer Q2: Have you ever smoked marijuana?
- Only you will ever know which question you were answering.

REMEMBER THE RULE OF TOTAL PROBABILITY

- Let's use the following notation:

  Y: a respondent answers yes

  Q1: the respondent was answering question 1 (about their tax ID number)

  Q2: the respondent was answering question 2 (about marijuana use)

- The key equation is:

$$P(Y) = P(Y, Q1) + P(Y, Q2)$$
$$= P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2)$$

- Let's calculate!

BAYES' RULE

- Bayes' rule is an equation that tells us how to learn:

  We start with some prior belief.

  We observe some new data.

  How do we combine these two sources of information to update our prior belief into a new and improved belief?

- For example:

  Suppose we know that 10% of all professional cyclists take EPO (which is banned).

  Now some specific cyclist tests positive for EPO—but the test isn't perfect!

  Before: P(guilty) = 0.1

  After: P(guilty | positive test) = ? (It's not 100%, since no drug test is perfect!)

BAYES' RULE IS EVERYWHERE

- Search engines
- Recommender systems
- Medical testing
- Doping control
- ancestry.com and similar sites
- The search for extrasolar planets
- Satellite tracking
- Asset-pricing models
- And many more!

EXAMPLE 1

- A 45-year old woman goes to the doctor for a routine screening mammogram. (No family history or clinical symptoms.)
- The mammogram comes back positive.
- What is P(cancer | positive test)?

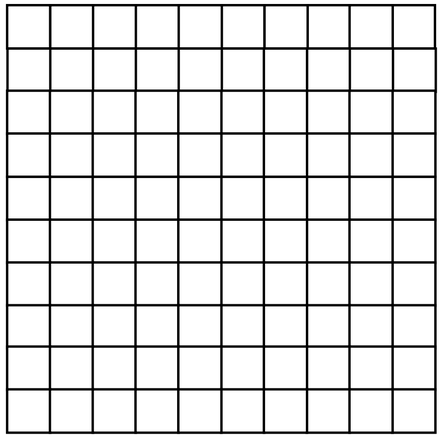
SOME FACTS

- For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer, and 990 don’t.

- Out of 10 cancer cases, we would expect a screening mammogram to correctly detect about 8 of them, on average.

- If a woman does not have breast cancer, the mammogram has a small chance of resulting in a positive test anyway. Out of 100 such cases, it will wrongly flag about 10 of them, on average.

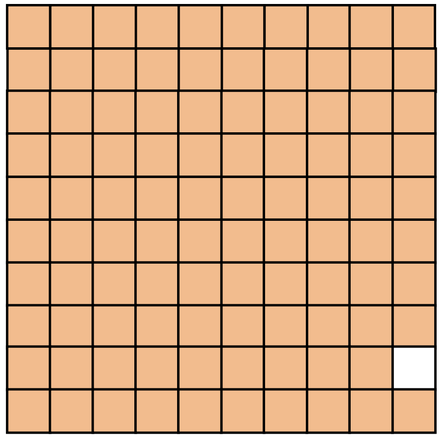
1,000 women go in for screening



= 10

Each square represents 10 women

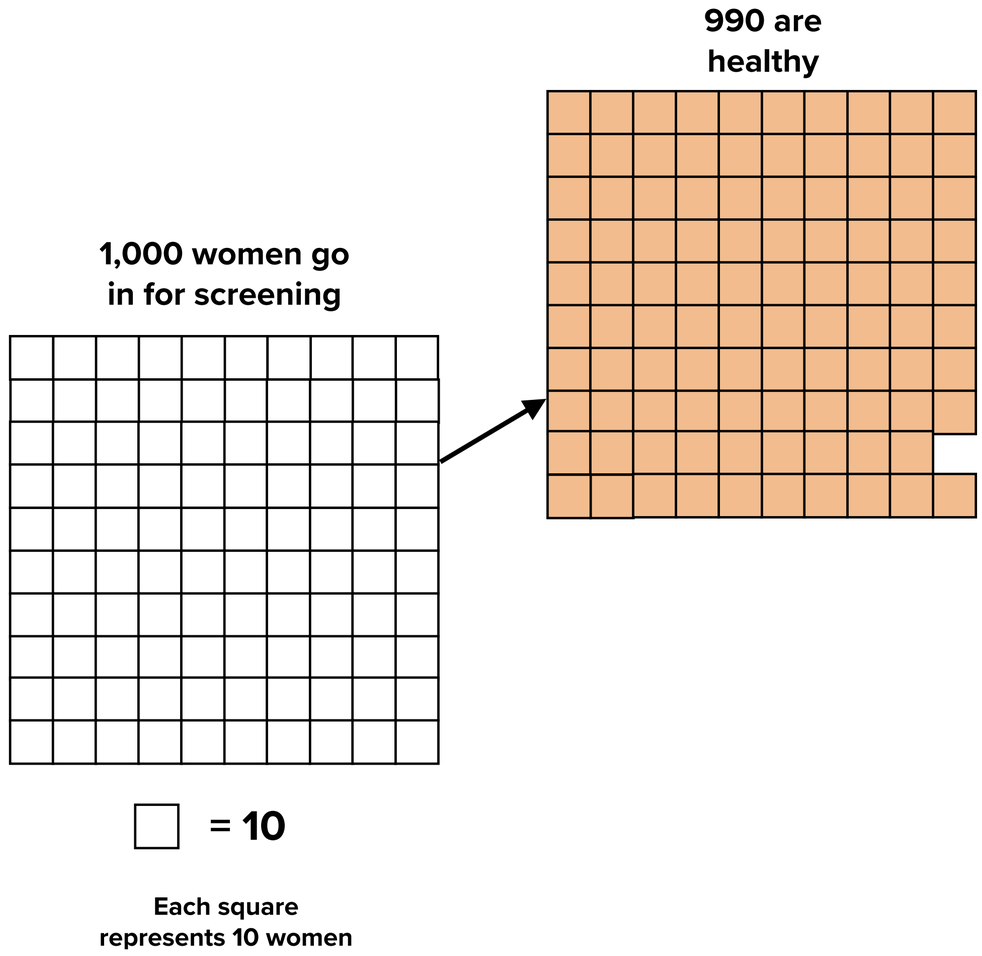
1,000 women go
in for screening



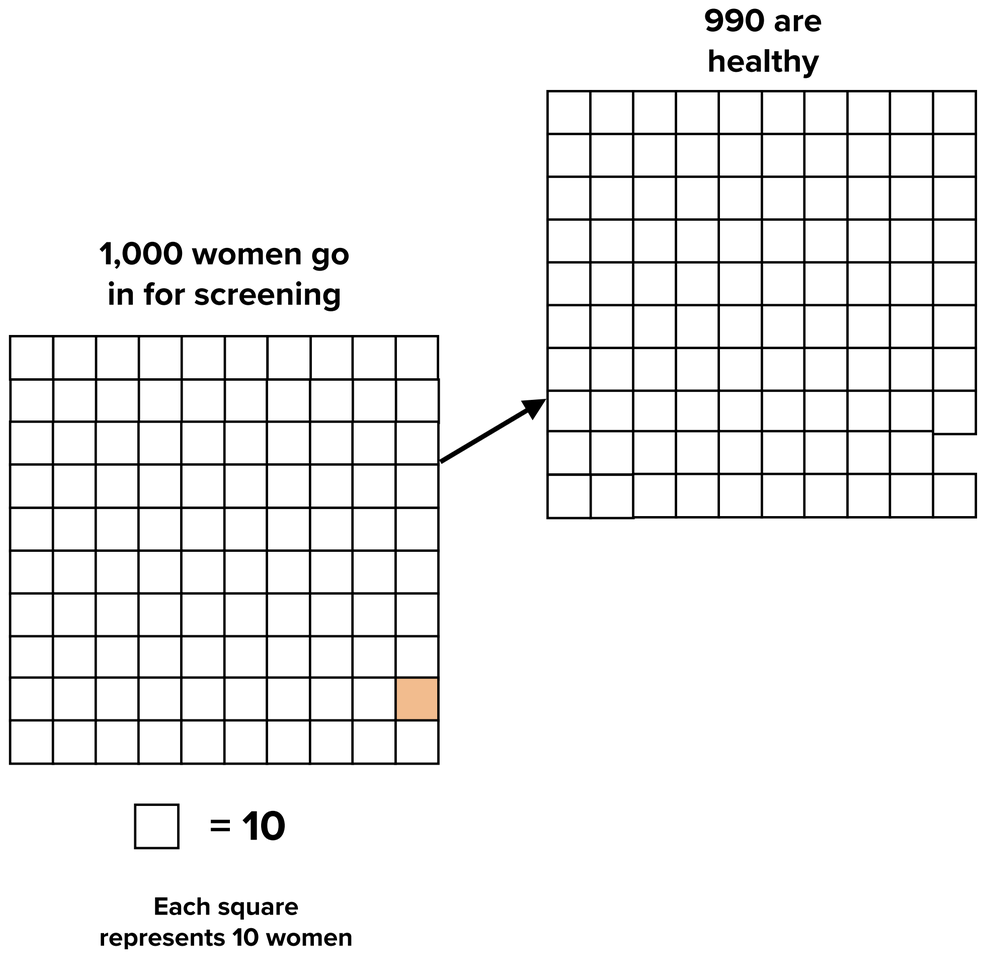
= 10

Each square
represents 10 women

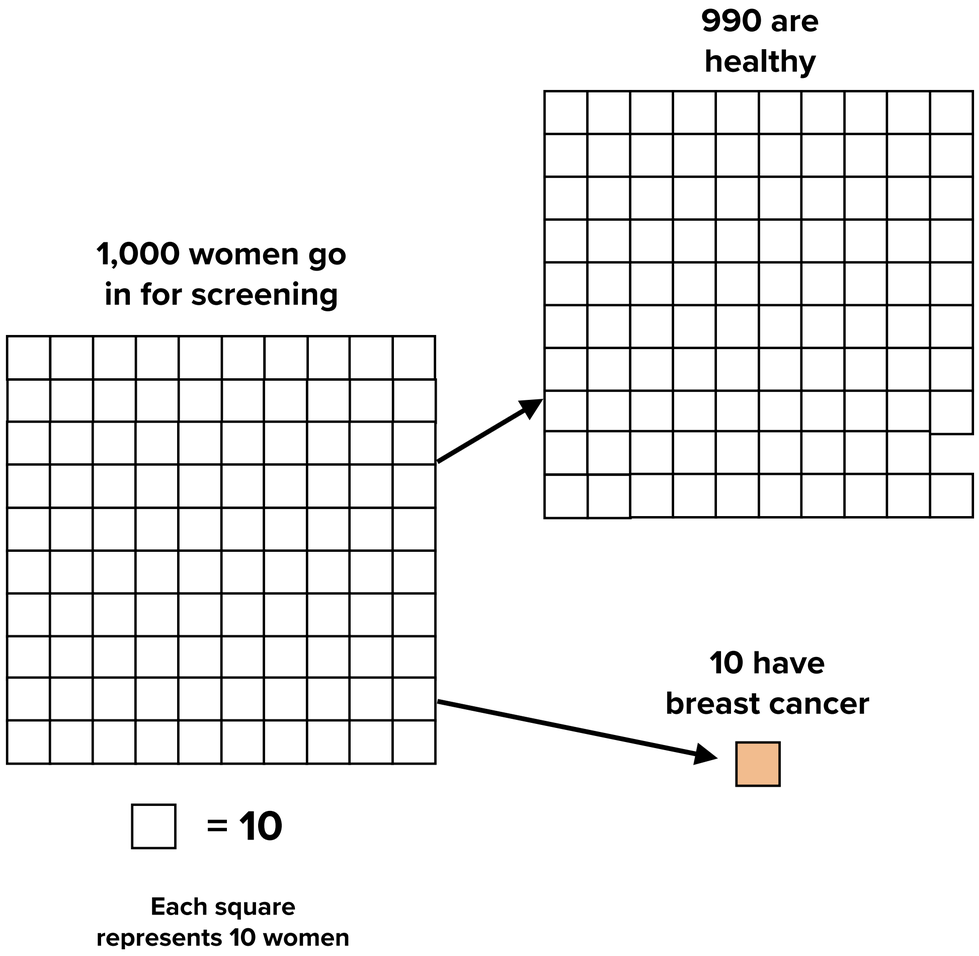
“For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer.”



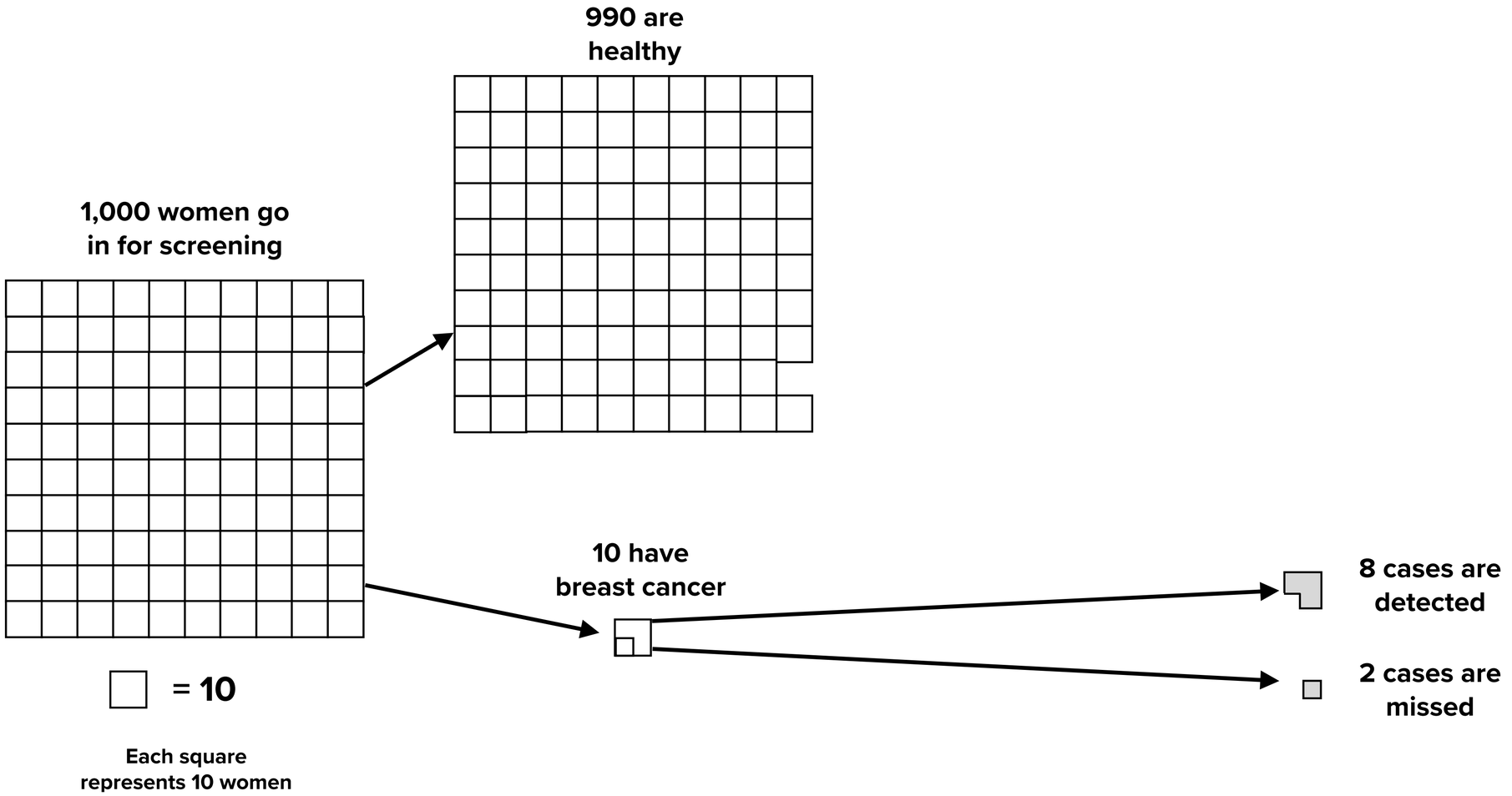
“For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer.”



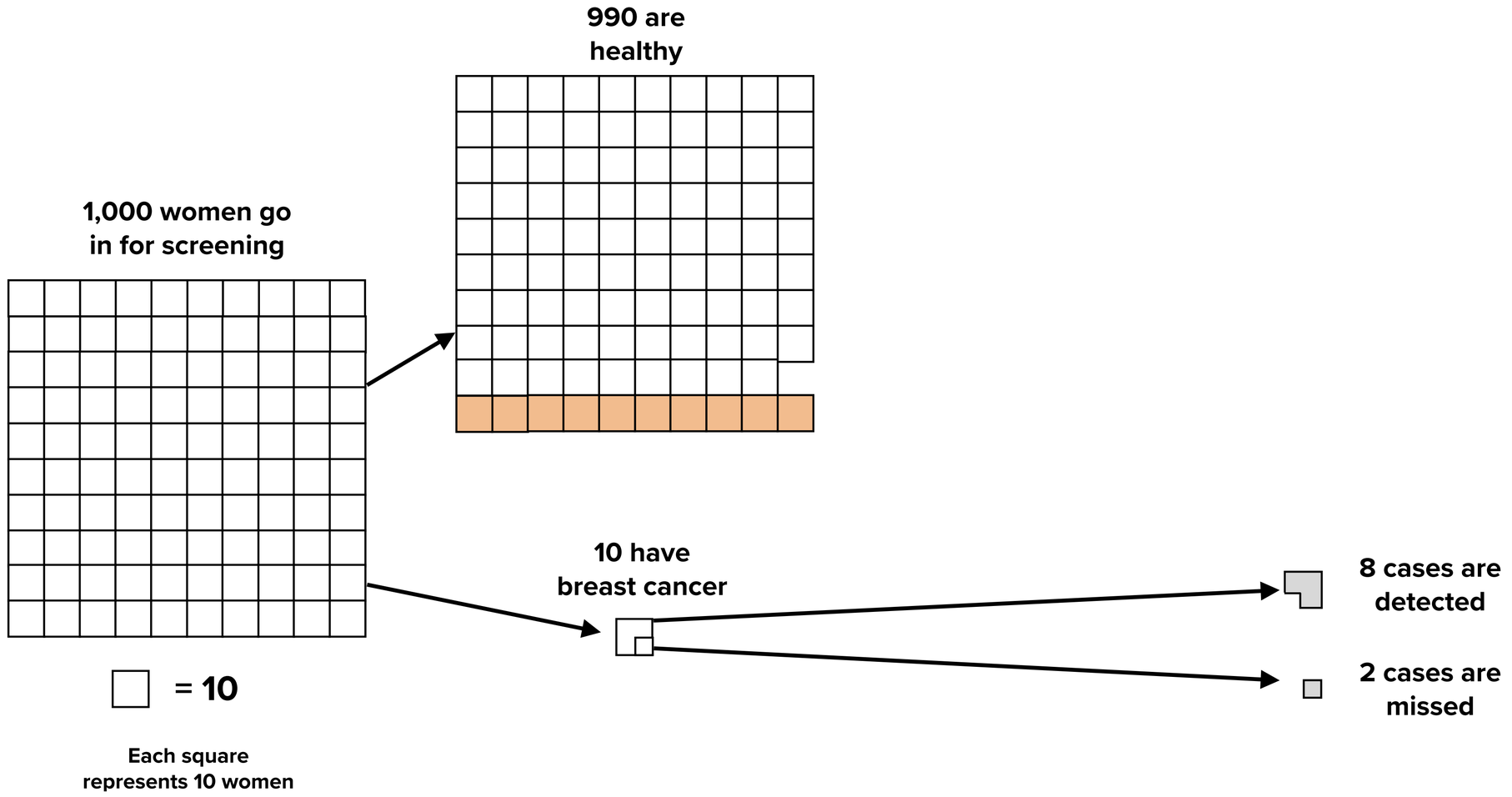
“For every 1,000 45-year-old women who participate in a routine screening mammogram, about 10 of them actually have breast cancer.”



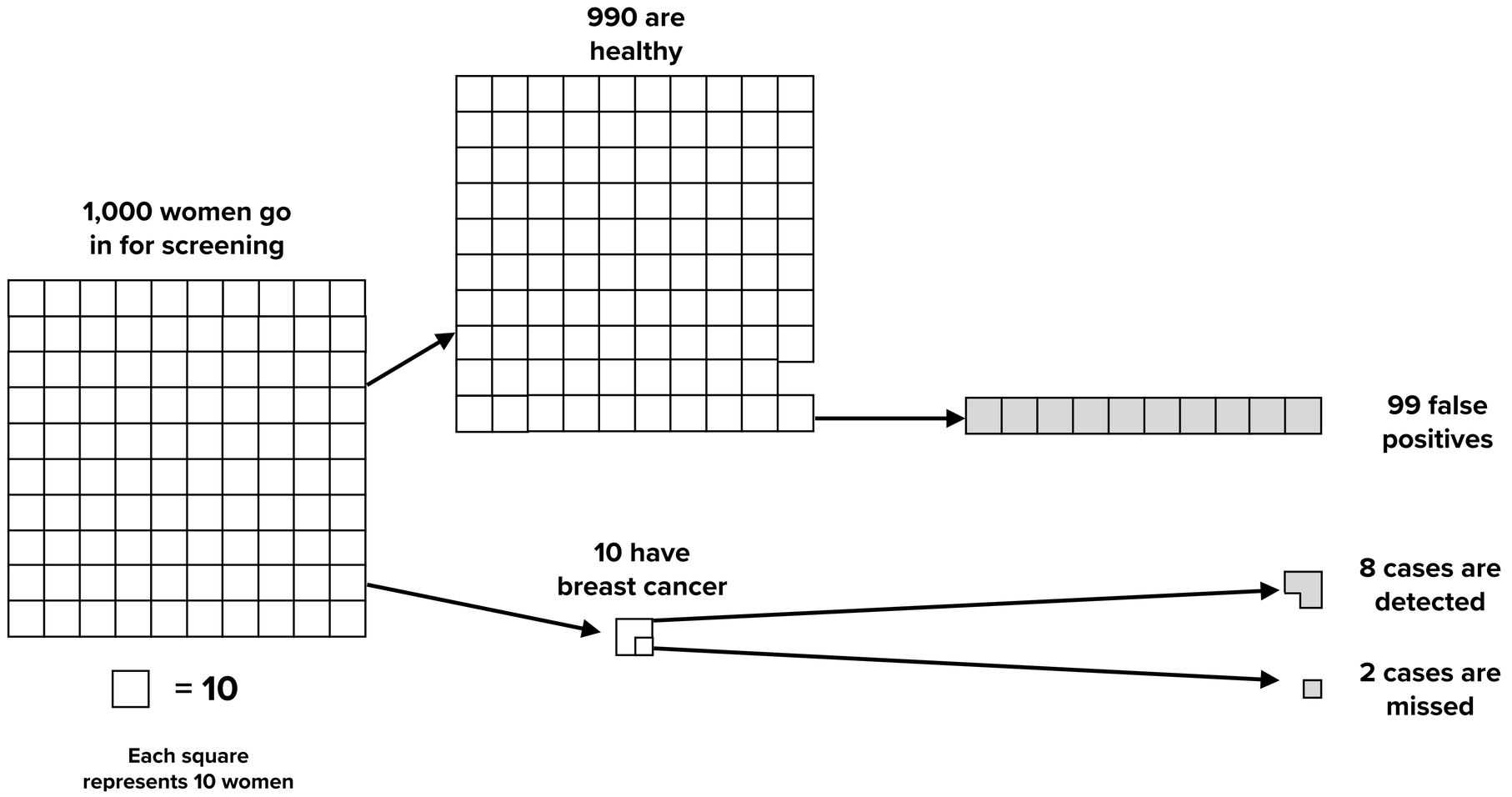
“Out of 10 cancer cases, we would expect a screening mammogram to correctly detect about 8 of them, on average.”



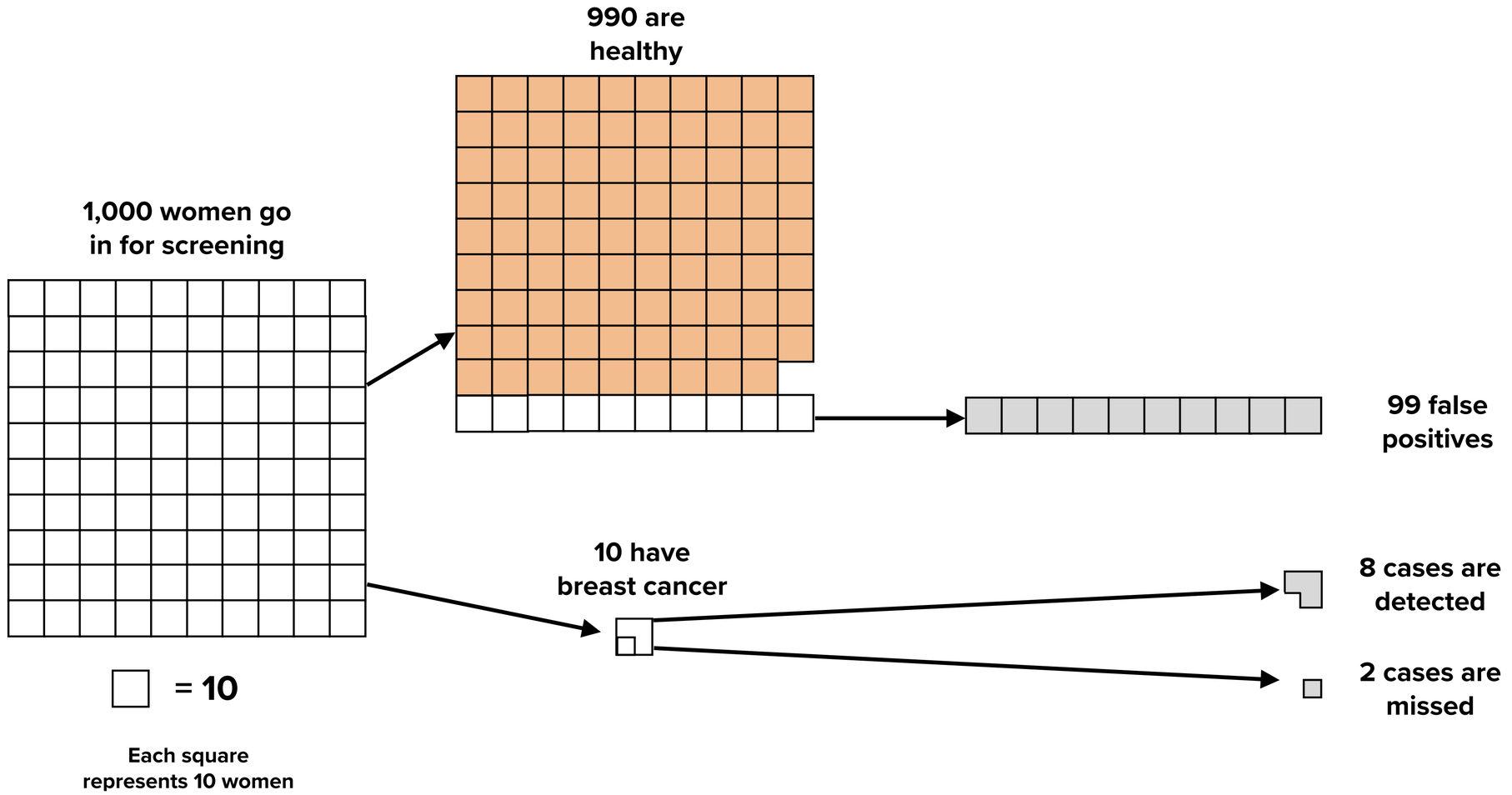
"Out of 10 cancer cases, we would expect a screening mammogram to correctly detect about 8 of them, on average."



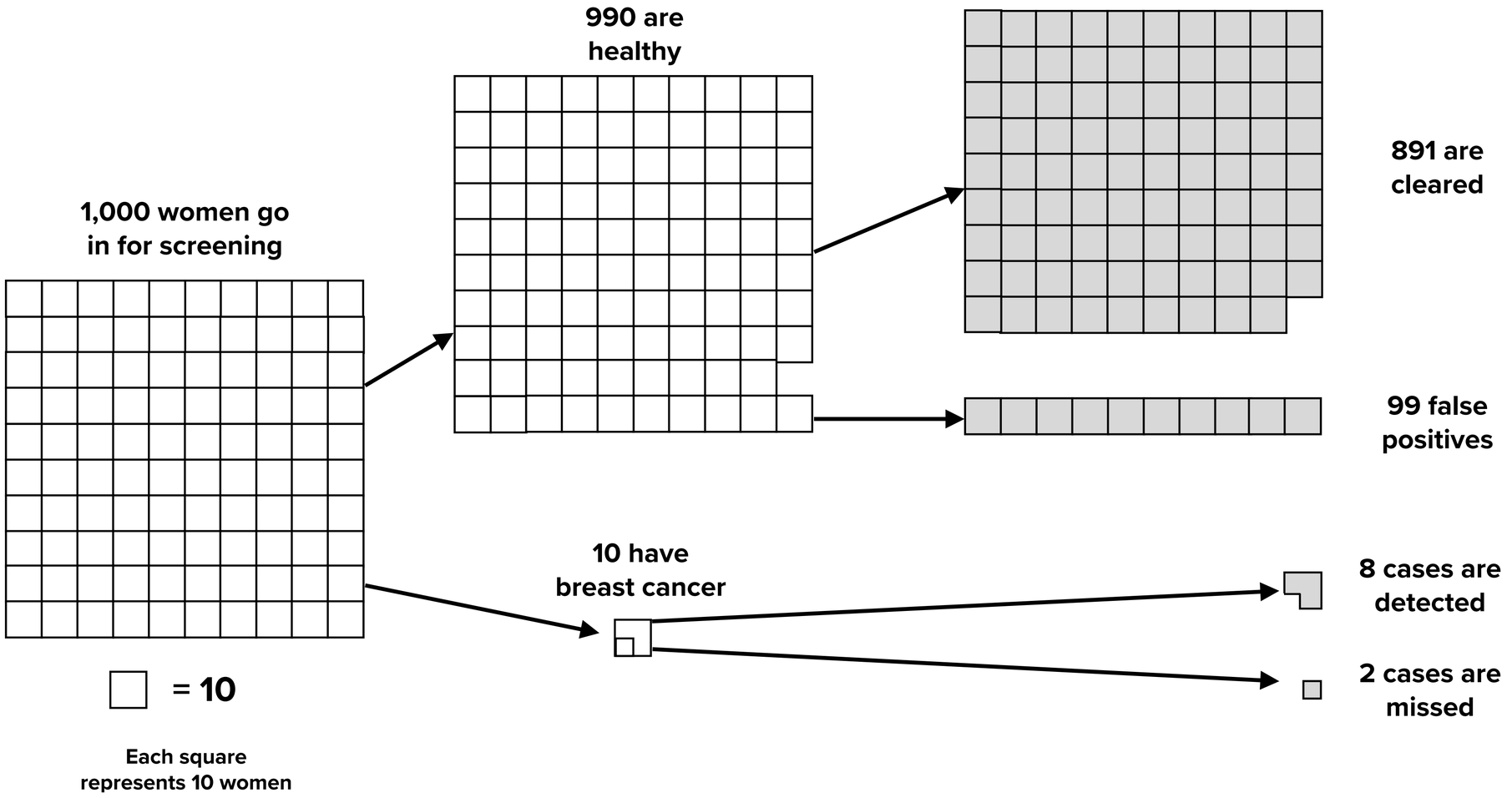
“If a woman does not have breast cancer, the mammogram can still result in a positive test. Out of 100 such cases, it will wrongly flag about 10 of them, on average.”

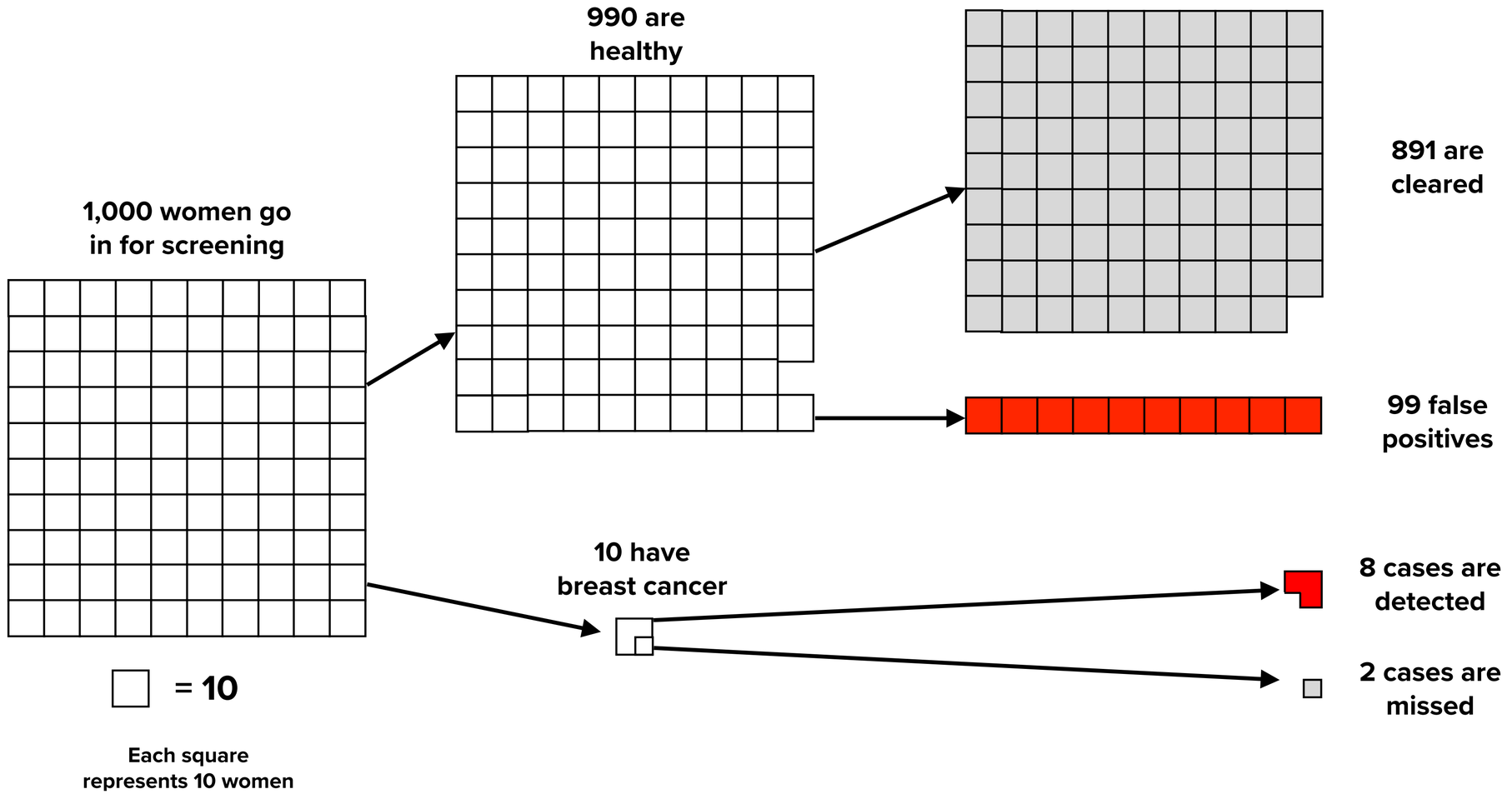


“If a woman does not have breast cancer, the mammogram can still result in a positive test. Out of 100 such cases, it will wrongly flag about 10 of them, on average.”

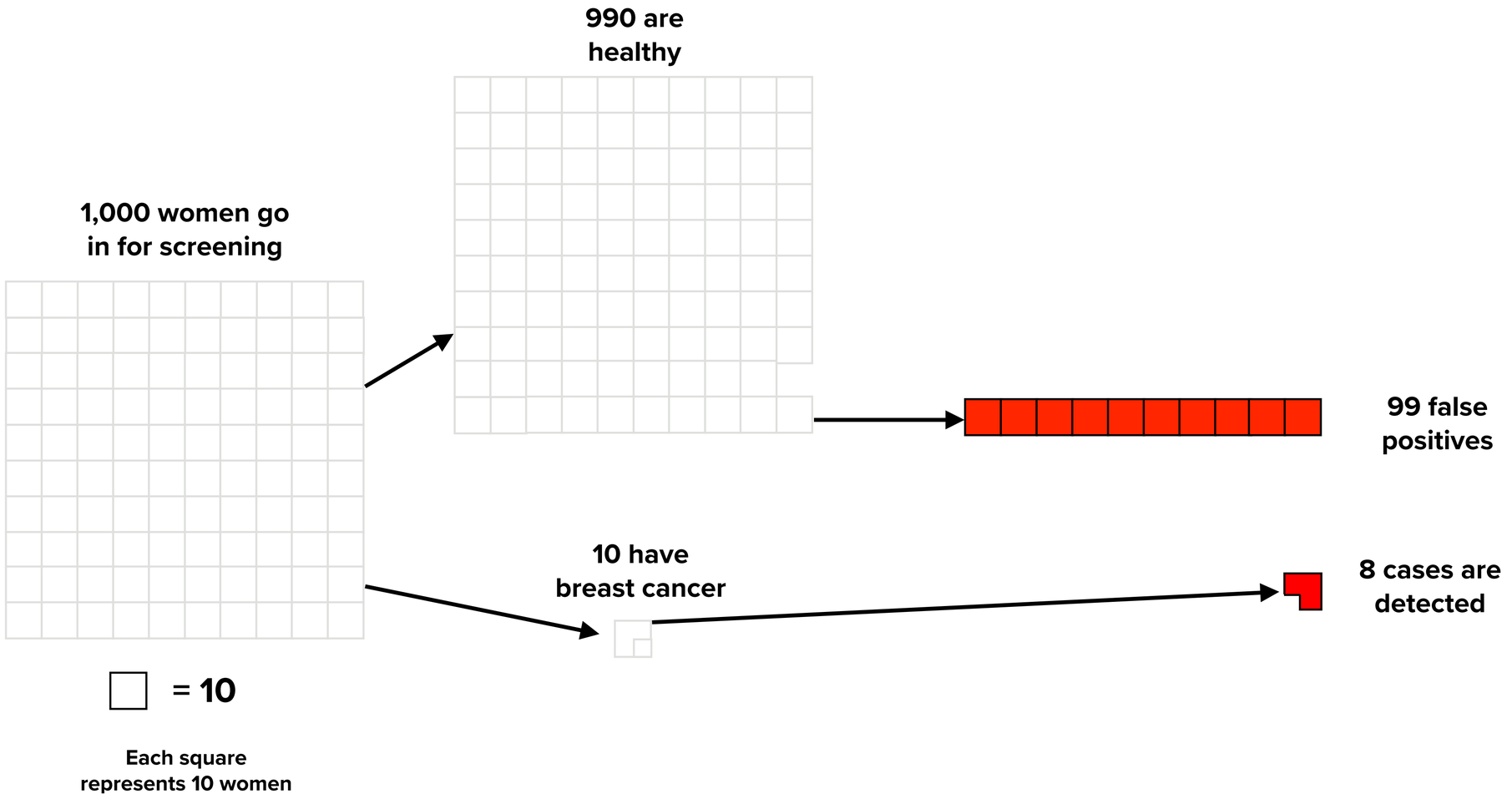


“If a woman does not have breast cancer, the mammogram can still result in a positive test. Out of 100 such cases, it will wrongly flag about 10 of them, on average.”

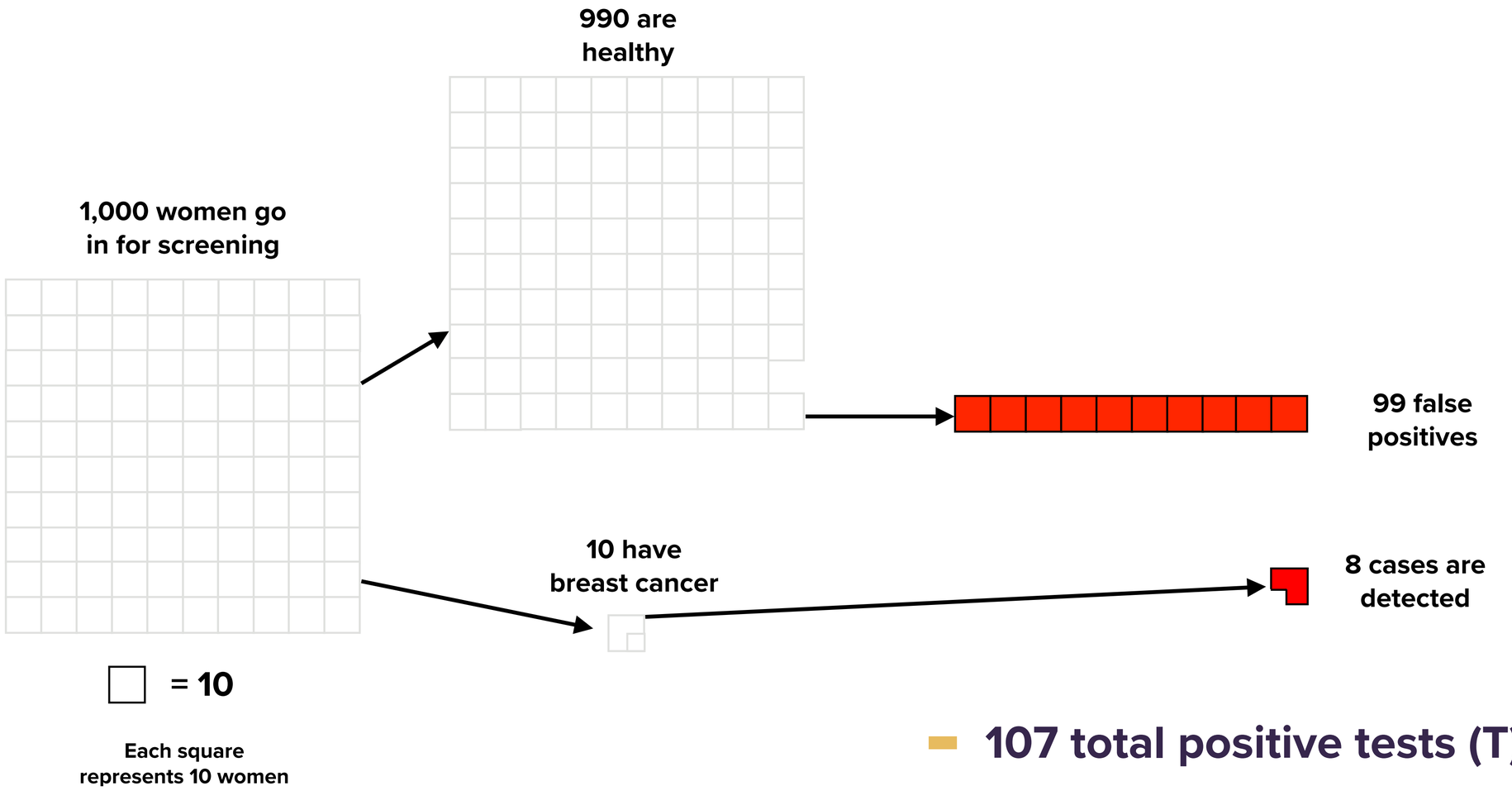

990 are
healthy
1,000 women go
in for screening
891 are
cleared
99 false
positives
10 have
breast cancer
8 cases are
detected
2 cases are
missed
= 10
Each square
represents 10 women



“The mammogram comes back positive. What is P(cancer | positive test)?”



“The mammogram comes back positive. What is P(cancer | positive test)?”



- 107 total positive tests (T)
- 8 are cancer cases (C)
- 99 are false positives
- So P(C | T) = 8/107 ≈ 0.075

FROM PRIOR TO POSTERIOR

- A 45-year old woman goes to the doctor for a routine screening mammogram. (No family history or clinical symptoms.)
- The mammogram comes back positive.
- What is P(cancer | positive test)? Let's abbreviate this as P(C | T).

  $P(C \mid T) = 8/107 \approx 0.075$

- Compare this with our prior knowledge, before we saw the outcome of the test:

  $P(C) = 10/1000 = 0.01$

FROM PRIOR TO POSTERIOR

- A 45-year old woman goes to the doctor for a routine screening mammogram. (No family history or clinical symptoms.)
- The mammogram comes back positive.
- What is P(cancer | positive test)? Let's abbreviate this as P(C | T).

  P(C | T) = 8/107 ≈ 0.075 (posterior probability, i.e. after the data)

- Compare this with our prior knowledge, before we saw the outcome of the test:

  P(C) = 10/1000 = 0.01 (prior probability, i.e. before the data)

- This is Bayes' rule in action: we updated our prior probability into a posterior probability, using the data from the test.
- Let's review how we got there, so we can see the general rule.

BAYES' RULE AS AN EQUATION

- **Bayes' rule is traditionally expressed as an equation:**

$$P(H \mid D) = \frac{P(H) \cdot P(D \mid H)}{P(D)}$$

- **This equation follows from using the multiplication rule once...**

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

BAYES' RULE AS AN EQUATION

- **Bayes' rule is traditionally expressed as an equation:**

$$P(H \mid D) = \frac{P(H) \cdot P(D \mid H)}{P(D)}$$

- **This equation follows from using the multiplication rule once and then twice:**

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

Notice we've just taken the numerator here....

$$= \frac{P(H) \cdot P(D \mid H)}{P(D)}$$

And re-written it using the multiplication rule a second time.

BAYES' RULE AS AN EQUATION

- **Bayes' rule is traditionally expressed as an equation:**

$$P(H \mid D) = \frac{P(H) \cdot P(D \mid H)}{P(D)}$$

- **Each term in Bayes' rule has a name:**

  P(H): the prior probability, or what you believe before seeing the data.

  P(H | D): the posterior probability, or what you believe after seeing the data.

  P(D | H): the likelihood, or the conditional probability of observing data D, given hypothesis H. "How likely would this data D be if H were true?"

  P(D): the total (or overall) probability of the data. This is always calculated using the rule of total probability. It's the hardest part of Bayes' rule.

EXAMPLE 1, REVISITED

- Let's recall our basic facts from the mammogram example. Here's what we know:

  P(C) = 0.01 (1% have cancer)

  P(not C) = 0.99 (99% don't have cancer)

  P(T | C) = 0.8 (80% chance of detecting a true cancer case)

  P(T | not C) = 0.1 (10% chance of returning a false positive for a non-cancer case)

- Goal:

  Find P(C | T), the posterior probability of cancer, given a positive test

EXAMPLE 1, REVISITED

- **Bayes' rule tells us how:**

$$P(C \mid T) = \frac{P(C) \cdot P(T \mid C)}{P(T)}$$

- **We know both terms upstairs:**

$$P(C) = 0.01$$
$$P(T \mid C) = 0.8$$

- **What takes some work is P(T), the term downstairs.**

  Key idea: use the rule of total probability.

  Let's do this calculation...

THE FULL BAYES RULE CALCULATION

$$P(C \mid T) = \frac{P(C) \cdot P(T \mid C)}{P(T)}$$

Use the rule of total probability

$$= \frac{0.01 \cdot 0.8}{0.01 \cdot 0.8 + 0.99 \cdot 0.1}$$

$$= \frac{0.008}{0.107}$$

$$\approx 0.075$$

Non-experts generally find it easier to understand the tree picture.

Use this fact to your advantage when communicating.

Example 2: Robert Julian-Borchak Williams


SHINOLA
DETROIT
441
SHINOLA
DETROIT
SHINOLA
DETROIT
GRAND CHEROKEE


Robert Williams

“THEY REALLY NEED TO GET THEIR EVIDENCE RIGHT... SOMEONE COULD BE LOCKED UP OR HURT OR KILLED...

ROBIN TAYLOR (ROBERT’S COUSIN)

THE EVIDENCE

- Let’s use this shorthand for our two events:

  G: Robert Williams is guilty

  M: the face-recognition algorithm matches Mr. Williams’ face to the security footage

- We care about P(G | M): the posterior probability of guilt, given the match.

THE EVIDENCE

- Let's use this shorthand for our two events:

  G: Robert Williams is guilty

  M: the face-recognition algorithm matches Mr. Williams' face to the security footage

- We care about P(G | M): the posterior probability of guilt, given the match.
- Let's use Bayes' rule:

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(M)}$$

$$= \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$

(note how we've expanded the term downstairs using the rule of total probability)

THE EVIDENCE

- We can make a decent guess for P(G).

  There are 4.3 million people in the greater Detroit Metropolitan area.

  Only 1 of them committed the robbery.

  Without other evidence, the prior probability that *any specific person* is guilty is:

  $P(G) = 1/4.3 \text{ million} \approx 2.33 \times 10^{-7} = 0.000000233$

  $P(\text{not } G) = 1 - 0.000000233 = 0.999999767$

THE EVIDENCE

- We can make a decent guess for P(G).

  There are 4.3 million people in the greater Detroit Metropolitan area.

  Only 1 of them committed the robbery.

  Without other evidence, the prior probability that *any specific person* is guilty is:

  $P(G) = 1/4.3 \text{ million} \approx 2.33 \times 10^{-7} = 0.000000233$

  $P(\text{not } G) = 1 - 0.000000233 = 0.999999767$

- So this leaves us here:

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$
$$= \frac{0.000000233 \cdot P(M \mid G)}{0.000000233 \cdot P(M \mid G) + 0.999999767 \cdot P(M \mid \text{not } G)}$$

THE EVIDENCE

- **Now we need some assumptions about the accuracy of face recognition software.**

  Let's be *very generous* and assume that P(M | G) is 1: if someone is actually present on the security footage, the software is guaranteed to produce a match.

  So what about P(M | not G), the probability of a false positive match?

THE EVIDENCE

- **Now we need some assumptions about the accuracy of face recognition software.**

  Let's be *very generous* and assume that P(M | G) is 1: if someone is actually present on the security footage, the software is guaranteed to produce a match.

  So what about P(M | not G), the probability of a false positive match?

- **What if P(M | not G) is 0.1? (One false match in 10.) Then:**

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$

$$= \frac{0.000000233 \cdot 1}{0.000000233 \cdot 1 + 0.999999767 \cdot 0.1}$$

$$= \frac{0.000000233}{0.1000002} \approx 2.3 \times 10^{-6}$$ (about 2 in a million)

THE EVIDENCE

- **Now we need some assumptions about the accuracy of face recognition software.**

  Let's be *very generous* and assume that P(M | G) is 1: if someone is actually present on the security footage, the software is guaranteed to produce a match.

  So what about P(M | not G), the probability of a false positive match?

- **What if P(M | not G) is only 0.01? (One false match in 100.) Then:**

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$

$$= \frac{0.000000233 \cdot 1}{0.000000233 \cdot 1 + 0.999999767 \cdot 0.01}$$

$$= \frac{0.000000233}{0.01000002} \approx 2.3 \times 10^{-5}$$ (about 2 in a 100,000)

THE EVIDENCE

- **Now we need some assumptions about the accuracy of face recognition software.**

  Let's be *very generous* and assume that P(M | G) is 1: if someone is actually present on the security footage, the software is guaranteed to produce a match.

  So what about P(M | not G), the probability of a false positive match?

- **What if P(M | not G) is only 0.00001? (One false match per 100,000.) Then:**

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$

$$= \frac{0.000000233 \cdot 1}{0.000000233 \cdot 1 + 0.999999767 \cdot 0.00001}$$

$$= \frac{0.000000233}{0.000010233} \approx 0.023$$ (about 2 in 100)

THE EVIDENCE

- **Now we need some assumptions about the accuracy of face recognition software.**

  Let's be *very generous* and assume that P(M | G) is 1: if someone is actually present on the security footage, the software is guaranteed to produce a match.

  So what about P(M | not G), the probability of a false positive match?

- **What if P(M | not G) is only 0.0000002? (Two false matches per 10 million.) Then:**

$$P(G \mid M) = \frac{P(G) \cdot P(M \mid G)}{P(G) \cdot P(M \mid G) + P(\text{not } G) \cdot P(M \mid \text{not } G)}$$

$$= \frac{0.000000233 \cdot 1}{0.000000233 \cdot 1 + 0.999999767 \cdot 0.0000002}$$

$$= \frac{0.000000233}{0.000000434} \approx 0.538$$ (about 1 in 2)

THE EVIDENCE

- **Conclusion: in order for police to even have a 50% posterior probability that Robert Williams was guilty, their facial recognition software had to be very, very accurate.**

  False negative are *actually* impossible: P(M | G) = 1

  False positives are *almost* impossible: P(M | not G) = 2 out of 10 million

- **And that only gets them to P(G | M) ≈ 0.53.**

FACE RECOGNITION SOFTWARE

- So the natural question is: how accurate is face-recognition software?
- It depends!

> "In ideal conditions, facial recognition systems can have near-perfect accuracy. Verification algorithms used to match subjects to clear reference images (like a passport photo or mugshot) can achieve accuracy scores as high as 99.97% on standard assessments like NIST's Facial Recognition Vendor Test (FRVT). This is comparable to the best results of iris scanners. This kind of face verification has become so reliable that even banks feel comfortable relying on it to log users into their accounts."

"How Accurate are Facial Recognition Systems – and Why Does It Matter?"
Center for Strategic and International Studies (CSIS) white paper, April 2020.

FACE RECOGNITION SOFTWARE

- So the natural question is: how accurate is face-recognition software?
- It depends!

  > "In ideal conditions, facial recognition systems can have near-perfect accuracy. Verification algorithms used to match subjects to clear reference images (like a passport photo or mugshot) can achieve accuracy scores as high as 99.97% on standard assessments like NIST's Facial Recognition Vendor Test (FRVT). This is comparable to the best results of iris scanners. This kind of face verification has become so reliable that even banks feel comfortable relying on it to log users into their accounts."

- Note: even 99.97% (0.9997) means an error rate of 0.0003. That's 1500 times higher than the 0.000002 error rate we need to get us to $P(G \mid M) \approx 0.5$.

"How Accurate are Facial Recognition Systems – and Why Does It Matter?"
Center for Strategic and International Studies (CSIS) white paper, April 2020.

FACE RECOGNITION SOFTWARE

- So the natural question is: how accurate is face-recognition software?
- It depends!

> "However, this degree of accuracy is only possible in ideal conditions where there is consistency in lighting and positioning, and where the facial features of the subjects are clear and unobscured. In real world deployments, accuracy rates tend to be far lower. For example, the FRVT found that the error rate for one leading algorithm climbed from 0.1% when matching against high-quality mugshots to 9.3% when matching instead to pictures of individuals captured "in the wild," where the subject may not be looking directly at the camera or may be obscured by objects or shadows."

"How Accurate are Facial Recognition Systems – and Why Does It Matter?"
Center for Strategic and International Studies (CSIS) white paper, April 2020.

FACE RECOGNITION SOFTWARE

- So the natural question is: how accurate is face-recognition software?
- It depends!

  > "However, this degree of accuracy is only possible in ideal conditions where there is consistency in lighting and positioning, and where the facial features of the subjects are clear and unobscured. In real world deployments, accuracy rates tend to be far lower. For example, the FRVT found that the error rate for one leading algorithm climbed from 0.1% when matching against high-quality mugshots to 9.3% when matching instead to pictures of individuals captured "in the wild," where the subject may not be looking directly at the camera or may be obscured by objects or shadows."

- A 9.3% error rate (0.093) is 46,500 times higher than the 0.000002 error rate we need to get us to $P(G \mid M) \approx 0.5$.

"How Accurate are Facial Recognition Systems – and Why Does It Matter?"
Center for Strategic and International Studies (CSIS) white paper, April 2020.

This is what “ideal conditions” look like: some actual images used to train face-recognition software.

This is what the security footage looked like.

Base-Rate Fallacy

IGNORING THE BASE RATE INFORMATION (PRIOR PROBABILITY) IN FAVOR OF THE CASE-SPECIFIC INFORMATION, RATHER THAN CORRECTLY INTEGRATING THE TWO

LESSONS

- Think about the purpose of using facial recognition in police work:

  With a small number of suspects, humans can do it as well or better than algorithms.

  The only purpose for using the algorithm is to screen faces at scale.

- But “screening at scale” means trawling for matches through databases of photos.

  Lots of photos in the database, only one guilty person.

  So P(guilty) for any randomly sampled photo in the database is very small.

- Therefore, even if P(match | guilty) is pretty high, P(guilty | match) will still be low.

LESSONS

- Never let *anyone* try to distract you with P(match | guilty).

  What matters is P(guilty | match).

  You need P(match | guilty) to calculate P(guilty | match)…

  But they’re not the same thing!

  Ignoring the difference is called the “Prosecutor’s Fallacy” or the “Base-rate Fallacy.”

- You need to use Bayes’ rule to integrate two sources of information:

  P(G), the prior or “base rate.”

  P(M | G), the case-specific evidence.