

Tree diagrams, absolute frequencies, and diagnostic testing

Corinne Riddell

September 28, 2020

Today's agenda

- Use absolute frequencies to calculate probabilities
- Use tree diagrams to calculate probabilities
- Apply these skills to diagnostic testing
 - Sensitivity, specificity, positive predictive value, negative predictive value, true positives, false positives, true negatives, and false negatives
- Learn Bayes' theorem

Unintended pregnancies

- Approximately 9% of all births in the US are to teen mothers (aged 15-19), 24% to young-adult mothers (ages 20-24) and the remaining 67% to adult mothers (aged 25-44).
- A survey found that only 23% of births to teen mothers are intended. Among births to young adult women, 50% are intended, and among women aged 25-44 75% are intended

Define events using probability notation

Express all the percents on the previous slide using probability notation.

- Let M denote the age of the mother and B denote whether the birth was intended. Then we can define the events on the previous slides as:
 - $P(M = \text{teen}) = 0.09$
 - $P(M = \text{young adult}) = 0.24$
 - $P(M = \text{older adult}) = 0.67$
 - $P(B = \text{intended} | M = \text{teen}) = 0.23$
 - $P(B = \text{intended} | M = \text{young adult}) = 0.5$
 - $P(B = \text{intended} | M = \text{older adult}) = 0.75$

Question to answer

- What is the probability that any given live birth in the U.S. is unintended?
 - Rewrite this question as a probability statement
- We will cover two ways to answer this question:
 - a) Using absolute frequencies (not covered in the book)
 - b) Using tree diagrams

Method a: Absolute Frequencies

- Pretend there are 1000 women. Given that 9%, 24%, and 67% of the mothers are teens, younger, and older mothers (respectively) this means that out of the 1000:
 - 90 are teens
 - 240 are younger mothers
 - 670 are older mothers

Method a: Absolute Frequencies

- Now, conditional on being a teen, 23% of the pregnancies are intended.
- This means that $90 \times 23\% = 20.7$ teen mothers had intended pregnancies.
- We can calculate these joint probabilities for each age group:
 - 90 are teens, $90 \times 23\% = 20.7$ teens with intended pregnancies (and 69.3 teens with unintended pregnancies).
 - 240 are younger mothers, $240 \times 50\% = 120$ younger mothers with intended pregnancies (and 120 younger mothers with unintended pregnancies).
 - 670 are older mothers, $670 \times 75\% = 502.5$ older mothers with intended pregnancies (and 167.5 with unintended pregnancies).

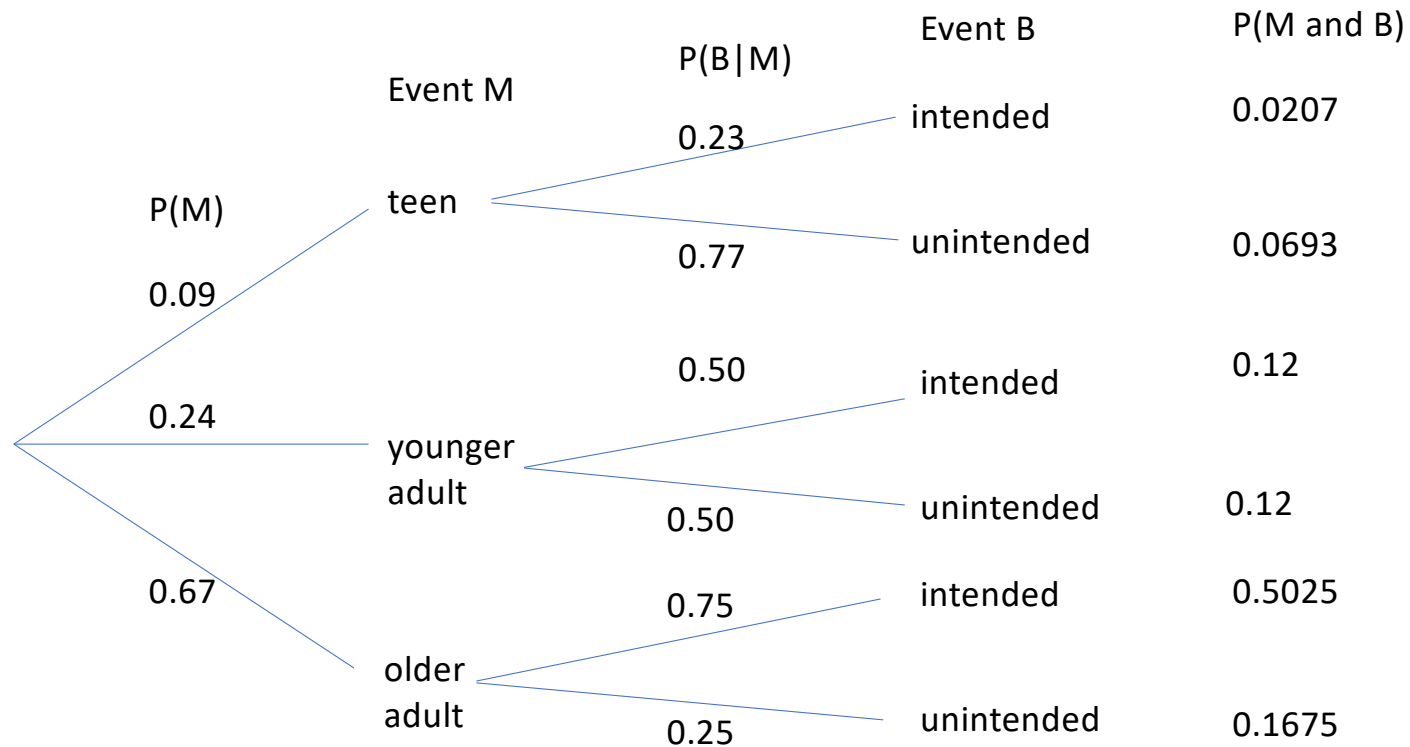
Method a: Absolute Frequencies

- Then, we can add on the number of unintended pregnancies across all the mothers:
 - $69.3 + 120 + 167.5 = 356.8$
- The last step is to convert this back to a probability.
- To do that, remember that there were 1000 women in the population. So $356.8/1000 = 35.7\%$
- Conclusion: The chance that a live birth in the US is unintended is 35.7%.

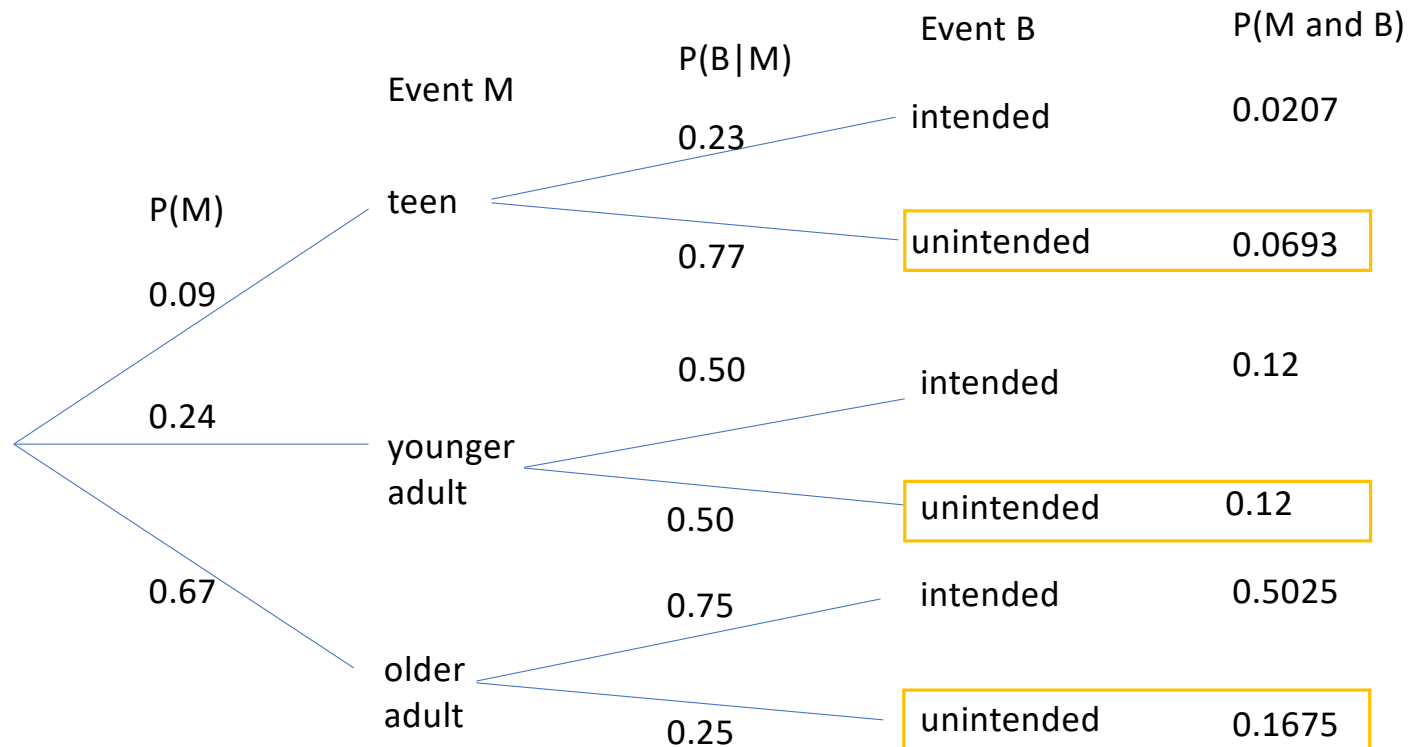
Method b: Tree diagram

- Rather than using absolute frequencies, you might prefer to draw this information using a tree diagram
- These diagrams are helpful when you know information about conditional probabilities and when the events of interest have more than two states (which is when Venn diagrams are used)

Method b: Tree diagram



Method b: Tree diagram



$$P(B=\text{unintended}) = 0.0693 + 0.12 + 0.1675 = 35.7\%$$

Diagnostic Testing

Recall the question I asked a few days ago...

- Suppose that there is test for a specific type of cancer that has a 90% chance of testing positive for cancer if the individual truly has cancer and a 90% chance of testing negative for cancer when the individual does not have it.
- 1% of patients in the population have the cancer being tested for.
- What is the chance that a patient has cancer given that they test positive?
 - a) Between 0% - 24.9%
 - b) Between 25.0% - 49.9%
 - c) Between 50.0% - 74.9%
 - d) Between 75.0% - 100%

Rewrite this information using prob. notation

- Let C be the true cancer status. $C = \text{cancer}$ for individuals who truly have cancer and $C = \text{no cancer}$ for individuals who truly do not have cancer.
- Let T be the test result. $T = +$ for individuals who test positively for cancer and $T = -$ for individuals who test negative for cancer. Then:
 - $P(C=\text{cancer})=0.01$
 - $P(\text{Test} = \text{positive} | C=\text{cancer}) = 0.90$
 - $P(\text{Test} = \text{negative} | C=\text{no cancer}) = 0.90$
- The question is “What is the chance that a patient has cancer given that they test positive”. Rewrite the question using this probability notation.

Diagnostic testing definitions

- **Sensitivity:** The test's ability to appropriately give a positive result when a person tested has the disease, or **$P(T = + | C = \text{cancer})$**
- **Specificity:** The test's ability to appropriately give a negative result when a person tested does not have the disease, or **$P(\text{test} = - | C = \text{no cancer})$**

Diagnostic testing definitions

- **Positive predictive value:** The chance that a person truly has cancer, given that the test is positive, or $P(C=\text{cancer} | T=+)$
- **Negative predictive value:** The chance that a person truly does not have cancer, given that the test is negative, or $P(C=\text{no cancer} | T=-)$

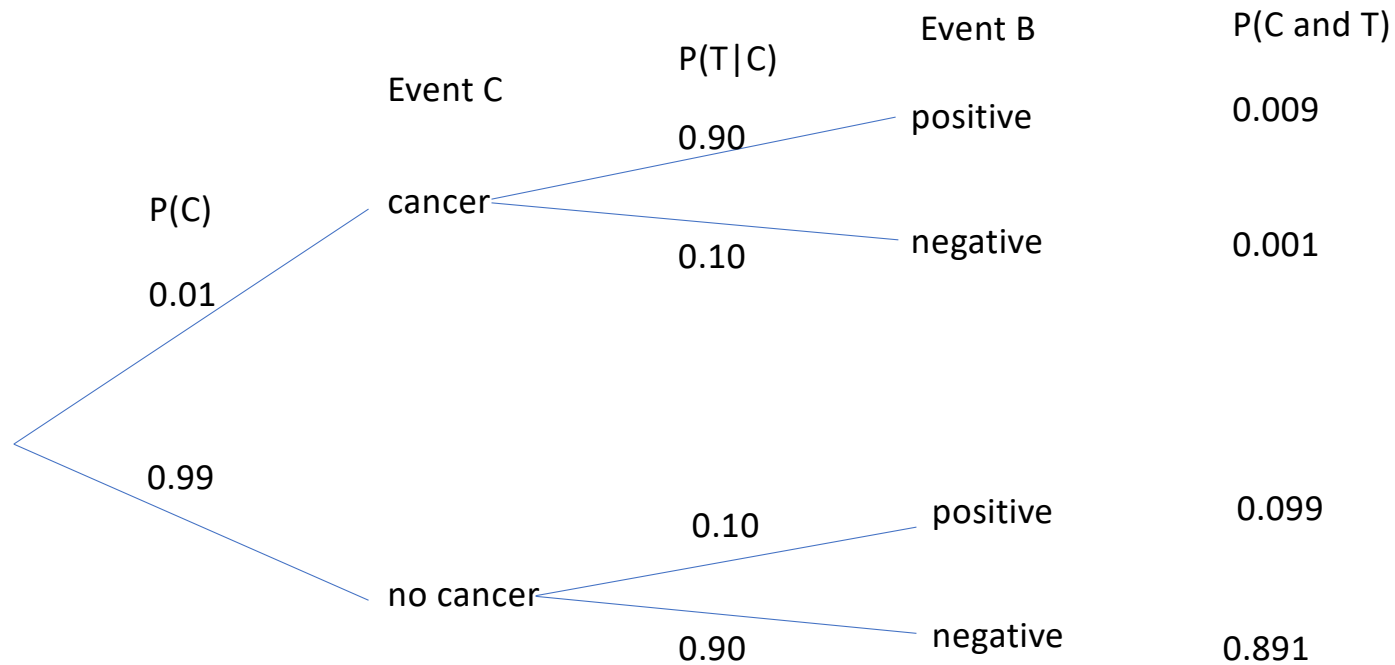
Back to the question

- Going back to the question... The question provided us information on the test's **sensitivity** and **specificity** as well as the **prevalence** of cancer in the underlying population
- The question asks us for the test's **positive predictive value**.
- We can use absolute frequencies or a tree diagram to answer the question.

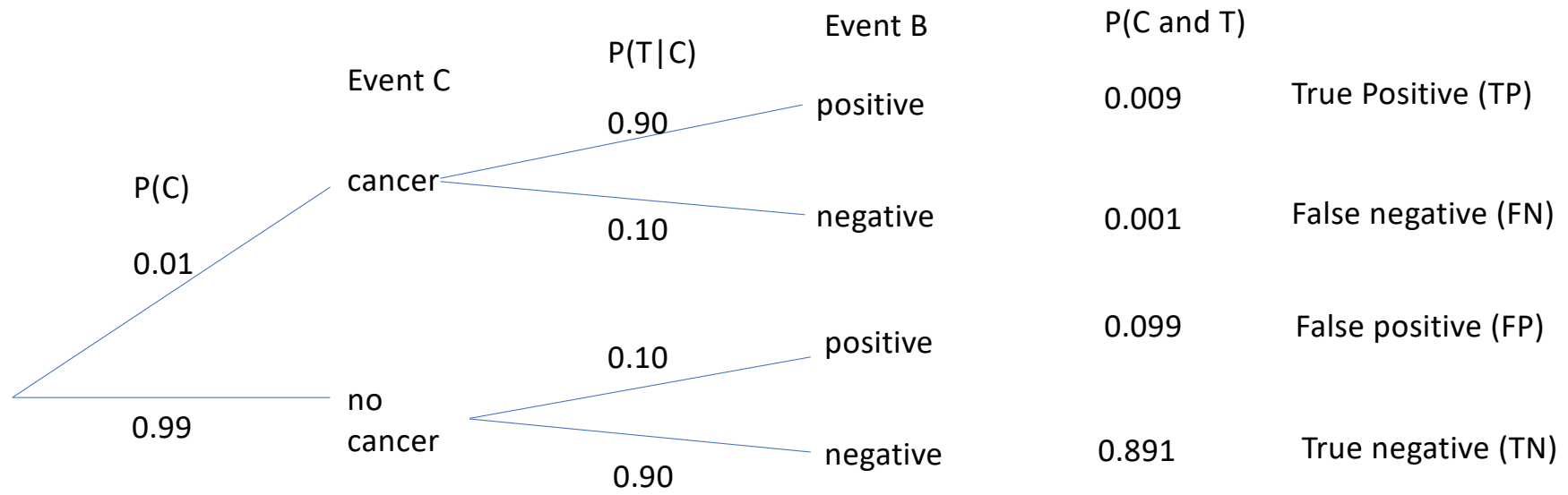
Absolute frequency approach

- Suppose that there are 1000 women in the population
- Translate the probabilities provided into absolute frequencies:
 - 1% truly have cancer \rightarrow 10 women truly have cancer, 990 women do not.
 - 90% sensitivity \rightarrow Among the 10 who truly have cancer, 9 women will test positive and 1 will test negative.
 - 90% specificity \rightarrow Among the 990 who do not have cancer, 891 will test negative, and 99 will test positive.
 - So, we have $9 + 99 = 108$ women detected with cancer
 - Of these 108 women, only 9 truly have cancer. Thus, $9/108 = 8.3\%$ of those detected for cancer actually have it.

Method b: Tree diagram



Method b: Tree diagram



$$\begin{aligned}
 P(C=\text{cancer} | T=+) &= P(\text{cancer \& test positive}) / P(\text{test positive}) \\
 &= P(\text{cancer \& test positive}) / [P(\text{test positive \& cancer}) + P(\text{test positive \& no cancer})] \\
 &= P(\text{true positive}) / [P(\text{true positive}) + P(\text{false positive})] \\
 &= 0.009 / (0.009 + 0.099) = 8.3\%
 \end{aligned}$$

Bayes' Theorem

- To answer this question, we started with information on $P(T|C)$ and $P(C)$ and used it to calculate $P(C|T)$.
- We can generalize how we did this using a rule known as Bayes' Theorem.
- To begin, recall the formula for conditional probability from last class:

$$P(A|B) = \frac{P(A \& B)}{P(B)}$$

Bayes' Theorem

- To begin, recall the formula for conditional probability from last class:

$$P(A|B) = \frac{P(A \& B)}{P(B)} \text{ [Formula 1]}$$

- This formula also implies:

$$P(B|A) = \frac{P(A \& B)}{P(A)}$$

which can be rearranged as: $P(B|A) \times P(A) = P(A \& B)$ [Formula 2]

Bayes' Theorem

- Plug Formula 2 into Formula 1:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \text{ [Formula 3]}$$

- If A only has two states, either A occurs or it does not (A' occurs), then P(B) can be partitioned into two pieces:

$$P(B) = P(B \& A) + P(B \& A') = P(B|A)P(A) + P(B|A')P(A')$$

- Then we can plug in this result into Formula 3:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

- This is Bayes' Theorem
- It allows to calculate a conditional probability (here, $P(A|B)$), when we only have information on the reverse condition ($P(B|A)$), as well as information on the overall probability of A ($P(A)$)
- This is how we went from calculating the positive predictive value, $P(C=\text{cancer} | T=+)$, when we only knew the Sensitivity ($P(T=+ | C=\text{cancer})$), Specificity ($P(T=- | C=\text{no cancer})$), and Prevalence of cancer ($P(C=\text{cancer})$)

Bayes' Theorem, Generalized

- Rather than only having A and A', suppose that A could take the values 1, 2, 3, and so on through A=k, where each of these states are disjoint and their probabilities are non-zero and add to 1.
- Then for B whose probability is not 0 or 1,

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + \dots + P(B|A_k) \times P(A_k)}$$

- Don't worry too much about understanding this formula
- Rather, focus on practicing the calculations for diagnostic testing like the one shown on the previous slide.
- You can watch [this video](#) (6 mins) to see how Bayes' Theorem is used in AI today.

Recap

- Absolute frequencies or tree diagrams
 - Use the method you like best to solve for probabilities
 - Or, use a Venn diagram. Apply the method that makes the most sense to you and suits the question.
- Diagnostic testing
 - Key lesson: Just because Sensitivity and Specificity are high, this does not imply that the Positive predictive value is also high. In lab, you will explore why this is the case
- Bayes' Theorem
 - We used it without even knowing it!
 - Don't worry about the formula, just know how to solve for probabilities using the method that you understand best.