Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

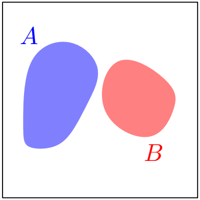
Today: Chapter 2 (continued)

- Practice the generalized "and" rule (non-independent events)
- Tree diagrams
- Practice joint/marginal/conditional probabilities on trees
- Bayes rule
- Dependence/independence in small/large populations

Recap of Last Lecture: "or" Rule

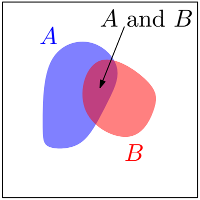
- If two events A and B are <u>disjoint</u>,

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



- In general,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$



Remark: The second formula becomes the first when A and B are disjoint, since $P(A \text{ and } B) = 0$.

Recap of Last Lecture: "and" rule

- If two events A and B are <u>independent</u>,

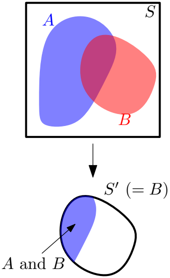
$$P(A \text{ and } B) = P(A) \times P(B).$$

- In general,

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

Can also write

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



Remark: The second formula becomes the first when A and B are independent, since $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Dependent VS Independent Events

A box contains 2 red balls and 3 green ones.
You pick two balls **with replacement** (= you pick a ball, note its color, put it back in the box, and then pick a second one).
What is the probability that both balls are red? Write

$$A = \text{“the first ball is red”}$$
$$B = \text{“the second ball is red”}$$

<u>As you replace the ball after the first draw</u>, B is unaffected by A.

By independence, $P(A \text{ and } B) = P(A) \times P(B) = \dfrac{2}{5} \times \dfrac{2}{5} = \dfrac{4}{25}$.

Dependent VS Independent Events

A box contains 2 red balls and 3 green ones.
You pick two balls **without replacement** (= you pick a ball, note its color, do not put it back in the box, and then pick a second one).
What is the probability that both balls are red? Write

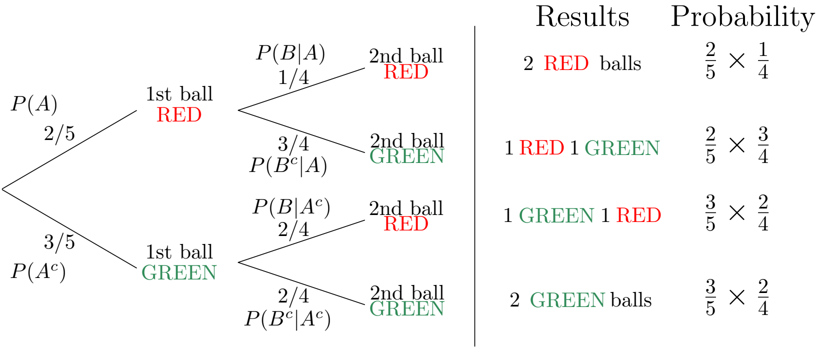
$$A = \text{“the first ball is red”}$$
$$B = \text{“the second ball is red”}$$

B is affected by A, since if A occurs, there is only 1 red ball left among the 4 balls remaining, so B is less likely to happen.

$$P(A \text{ and } B) = P(B|A) \times P(A) = \frac{1}{4} \times \frac{2}{5} = \frac{1}{10}.$$

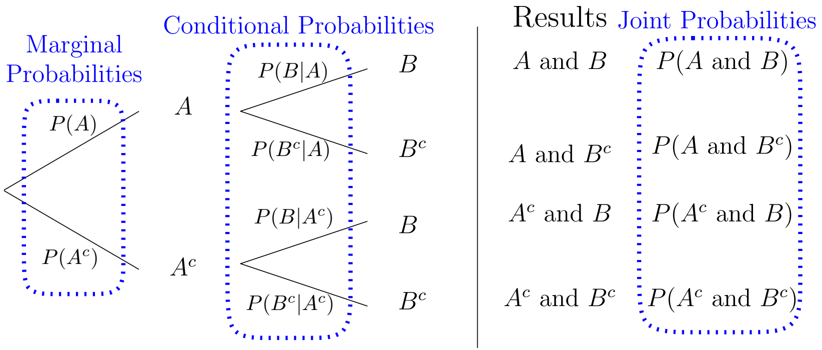
Tree Diagram

A box contains 2 red balls and 3 green ones.
You pick two balls **without replacement**.



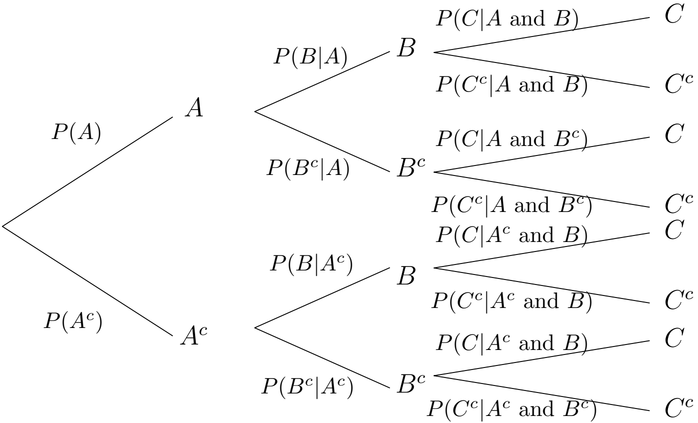
$$P(2 \text{ red}) = \frac{2}{20}, \quad P(1 \text{ red and } 1 \text{ green}) = \frac{12}{20}, \quad P(2 \text{ green}) = \frac{6}{20}.$$

Tree Diagram: General Case



Bigger Tree Diagrams?

If you deal with tree events A, B and C, you can still make a tree.

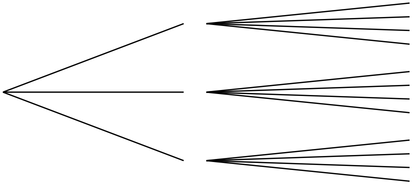


Examples:

- Draw 3 balls/cards without replacement
- Studying the weather of 3 consecutive days
- ...

Richer Tree Diagrams?

If you deal with variables having more than 2 outcomes, you can branch more broadly.

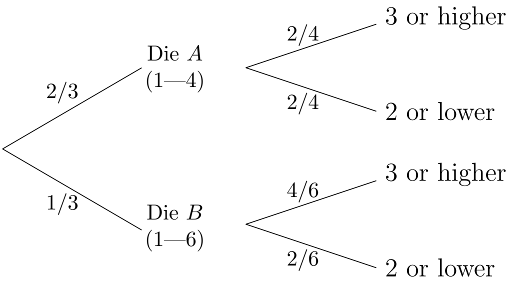


Examples:

- Outcome of a die roll (6 branches)
- Color of a ball in a box containing red,green and blue ones (3 branches)
- ...

Exercise

Consider a board game in which you need to kill an enemy that has 3 health points. You begin by spinning a dial that comes up A with probability 2/3, and B with probability 1/3. In the event of A, you roll a four-sided die (1—4); In the event of B, you roll a six-sided die (1—6). Your roll is the amount of damage you deal the opponent. What is the probability the enemy dies?



$$\frac{2}{3} \times \frac{2}{4} = \frac{4}{12}$$

$$\frac{1}{3} \times \frac{4}{6} = \frac{4}{18}$$

$$P(\text{enemy dies}) = \frac{4}{12} + \frac{4}{18} = \frac{5}{9}.$$

Bayes' Rule

Remember that

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

Switching the roles of A and B

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

Since $P(B \text{ and } A) = P(A \text{ and } B)$, we get

$$P(A|B) \times P(B) = P(B|A) \times P(A).$$

Dividing by $P(B)$ gives **Bayes' rule**:

$$\boxed{P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}.}$$

Bayes' Rule: Example

In a Math 183 class, 30% of students got an A on Midterm I. 40% of students studied a long time. Suppose the probability that you got an A given that you studied a long time is 70%.
What is the probability that you studied a long time if you got an A?

$$\begin{aligned} A &= \text{“earning an A grade”} \\ S &= \text{“studying a long time”} \end{aligned}$$

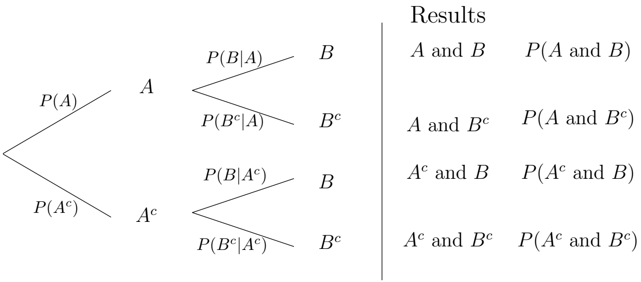
$P(A) = 0.3$ $\qquad$ $P(S) = 0.4$ $\qquad$ $P(A|S) = 0.7$

$$P(S|A) = \frac{P(A|S) \times P(S)}{P(A)} = \frac{0.7 \times 0.4}{0.3} \simeq 0.93.$$

Bayes' Rule: Level 2

One cannot use only Bayes' rule if $P(B)$ is unknown:

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



Looking at the tree, we see that:

$$P(B) = P(B|A) \times P(A) + P(B|A^c) \times P(A^c).$$

(Exercise for CS/Math majors: prove this rigorously!)

Bayes' Rule: Level 2

We just derived the **advanced Bayes' rule**:

$$\boxed{P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|A^c) \times P(A^c)}.}$$

Another Example

Doctors wanted to study the effectiveness of a new HIV test. When they used the test on people known to have HIV, it gave a positive result 99.9% of the time. On those known to not have HIV, it gave negative result 99% of the time. Supposing someone from San Diego (HIV rate: 0.0172%) tests positive on this new test, what is the probability the person actually has HIV?

Name your events:

$+$	"the test gives a positive result"
-	"the test gives a negative result"
yes	"the person does have HIV"
no	"the person does not have HIV"

The questions amounts to find $P(yes|+)$.

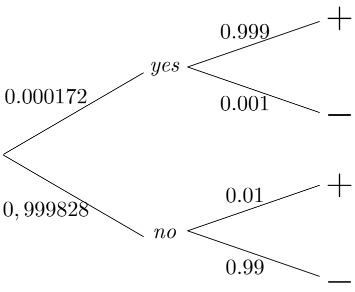
Another Example

Doctors wanted to study the effectiveness of a new HIV test. When they used the test on people known to have HIV, it gave a positive result 99.9% of the time. On those known to not have HIV, it gave negative result 99% of the time. Supposing someone from San Diego (HIV rate: 0.0172%) tests positive on this new test, what is the probability the person actually has HIV?

$P(+|yes) = 0.999$ $\qquad$ $P(\text{-}|no) = 0.99$ $\qquad$ $P(yes) = 0.000172$

$$P(yes|+) = \frac{P(+|yes) \times P(yes)}{P(+)} = \frac{0.999 \times 0.000172}{??}$$

Another Example



$$0.000172 \times 0.999 = 0,000171828$$

$$0.999828 \times 0.01 = 0.00999828$$

$$P(yes|+) = \frac{P(+|yes) \times P(yes)}{P(+)} = \frac{0.999 \times 0.000172}{0.000171828 + 0.00999828} \simeq 1.7\%.$$

What's Happening There?

How is it that a test so accurate could give such a terrible result?

Here, we have a great mismatch between

- The accuracy of the test (only 99.9% and 99%)
- The extreme rareness of the disease (0.0172%) in the population.

For positive test results to be useful (that is, for $P(\text{yes}|+)$ to be high), you need the orders of magnitude of "test accuracy" and "disease prevalence" to be better matched.

Dependence/Independence in Small/Large Populations

You pick 5 balls in a box containing green and red balls balls.
Study the event $A =$ "all 5 balls are green".

- 3 red 7 green (10 total)
  - **With replacement:** $P(A) = \left(\frac{7}{10}\right)^5 \simeq 0.168.$
  - **Without replacement:** $P(A) = \frac{7}{10}\frac{6}{9}\frac{5}{8}\frac{4}{7}\frac{3}{6} \simeq 0.083.$

- 30 red 70 green (100 total)
  - **With replacement:** $P(A) = \left(\frac{70}{100}\right)^5 \simeq 0.168.$
  - **Without replacement:** $P(A) = \frac{70}{100}\frac{69}{99}\frac{68}{98}\frac{67}{97}\frac{66}{96} \simeq 0.161.$

When you draw a sample without replacement, your choices are always dependent.
Although dependence is still present, it tends to disappear when population is large.

Lesson About Sampling

Moral (to be used later in the course):

If your sample size is less than 10% of the population you are sampling from, you may assume the choices are independent (even though they aren't).

Said differently: for samples $< 10\%$ population size, the distinction between "with replacement" and "without replacement" is unnecessary.