Math 186: Conditional Probability and Bayes' Theorem (2.4) Independence (2.5)

Math 283: Ewens & Grant 1.12.4–5

Prof. Tesler

Math 186 and 283
Fall 2019

Scenario: Flip a fair coin three times

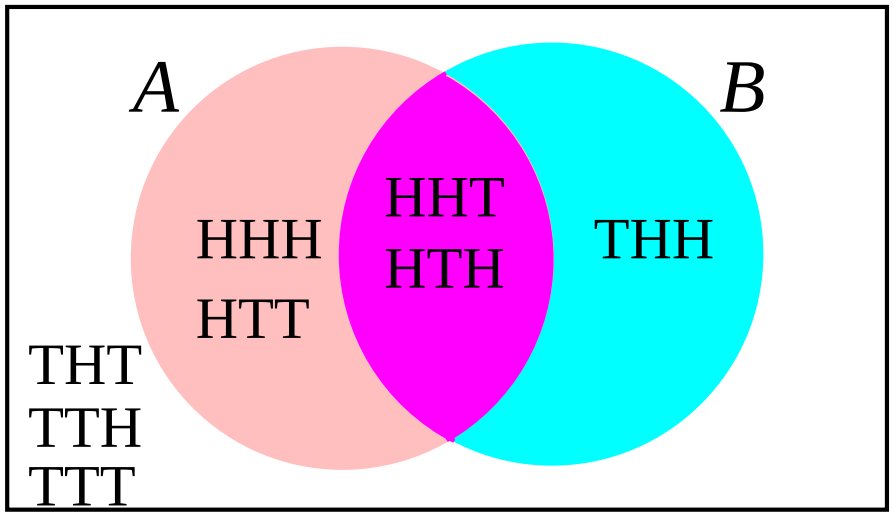
- Flip a coin 3 times. The sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- Define events

$$A = \text{“First flip is heads”} = \{HHH, HHT, HTH, HTT\}$$
$$B = \text{“Two flips are heads”} = \{HHT, HTH, THH\}$$

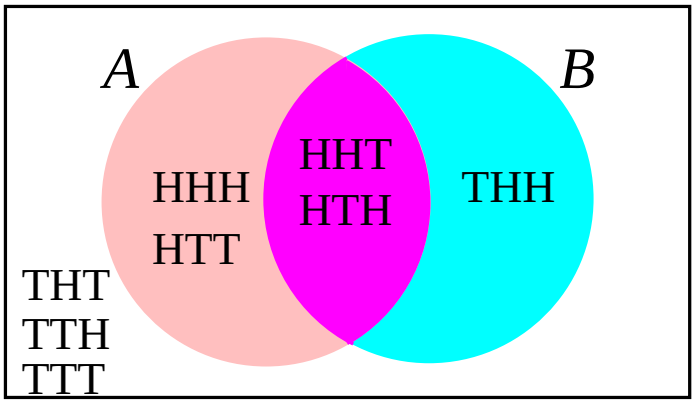
- Venn diagram:



Scenario: Flip a fair coin three times

A = "First flip is heads" $P(A) = \frac{4}{8}$
$= \{HHH, HHT, HTH, HTT\}$

B = "Two flips are heads" $P(B) = \frac{3}{8}$
$= \{HHT, HTH, THH\}$

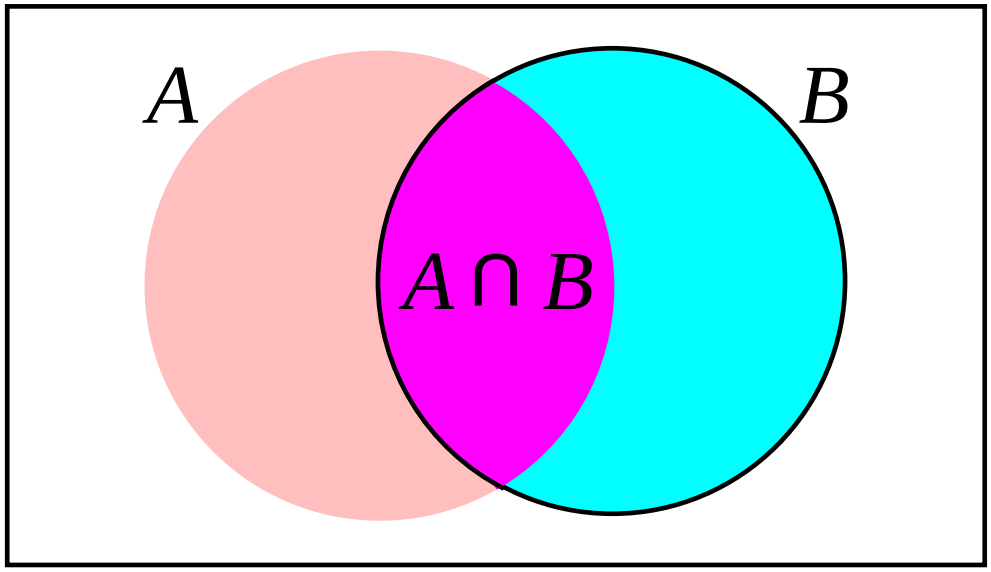


Conditional probability

- Flip a coin 3 times. If there are 2 heads, what's the probability that the first flip is heads?
- *Rephrase:* Assuming B is true, what's the probability of A?
- Since B is true, the coin flips are one of HHT, HTH, or THH.
- Out of those, the outcomes where A is true are HHT and HTH (which is $A \cap B$). So 2 out of the 3 possible outcomes in B give A.
- The probability of A, given that B is true, is

$$\frac{P(\{HHT, HTH\})}{P(\{HHT, HTH, THH\})} = \frac{2/8}{3/8} = \frac{2}{3} \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Conditional probability

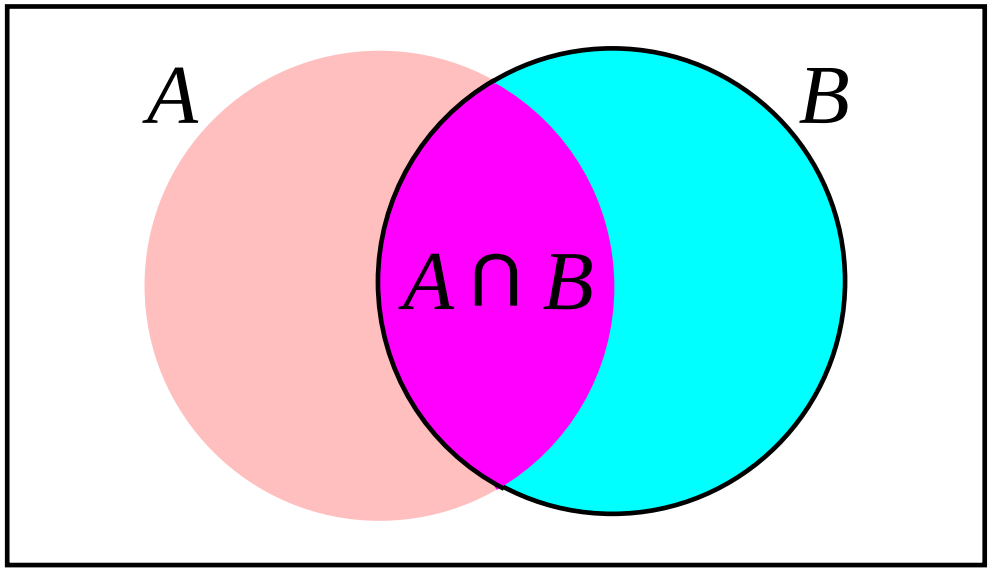


- $P(A)$ = probability of A
  measures A as a fraction of the sample space.

- $P(A \mid B)$ = conditional probability of A, given B
  measures $A \cap B$ as a fraction of B:

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

Conditional probability



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

- We can solve this for:

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

- In the same way,

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

Earwax genetics

$$\cdots \quad T \quad G \quad G \quad C \quad C \quad [C/T] \quad G \quad A \quad G \quad T \quad A \quad \cdots$$

- In humans, a specific position in the DNA sequence of gene ABCC11 can be a C or a T. This is an example of a *Single Nucleotide Polymorphism*, or *SNP* (pronounced like "snip").
- Each cell has two copies of this gene, one inherited from each parent, and the variations have this effect:

Genotype (versions of the gene)	**Phenotype (resulting trait)**
CC	wet earwax, normal underarm odor
CT	wet earwax, less odor
T T	dry earwax, no odor

Earwax in different populations

- The 1000 Genomes Project studies variations like these in thousands of individuals from different ancestral groups. Each participant is considered to be in exactly one of these groups.
- The prevalence of each genotype at this site is approximately*

Population	CC	CT	TT
AFR (African)	98%	2%	0.15%
AMR (Ad-mixed American)	73%	25%	1%
EAS (East Asian)	7%	30%	63%
EUR (European)	75%	22%	2%
SAS (South Asian)	27%	50%	23%

(in some rows, percentages don't total 100% due to rounding)

*1000 Genomes Project Phase 3, Ensembl release 94, Oct. 2018.
On ensembl.org, search for rs17822931, and select population genetics.
For more info, see links on the class website.

Earwax in different populations

Population	CC	CT	T T
AFR (African)	98%	2%	0.15%
AMR (Ad-mixed American)	73%	25%	1%
EAS (East Asian)	7%	30%	63%
EUR (European)	75%	22%	2%
SAS (South Asian)	27%	50%	23%

These are conditional probabilities. For example, the bottom row:

$$P(\text{CC} \mid \text{SAS}) = 0.27$$
$$P(\text{CT} \mid \text{SAS}) = 0.50$$
$$P(\text{T T} \mid \text{SAS}) = 0.23$$

Example: Two groups

Example

- A study sample is 40% AFR and 60% AMR.
- In AFR, the probability of CC is 98%, while in AMR, it's 73%.
- A random individual is chosen from the sample.

Questions

1. What's the probability they're in AFR and have genotype CC?
2. What's the probability their genotype is CC?
3. If the genotype is CC, what's the probability they're in AFR?

1. Probability they're in AFR and have genotype CC

- A study sample is 40% AFR and 60% AMR.
- In AFR, the probability of CC is 98%, while in AMR, it's 73%.
- A random individual is chosen from the sample.

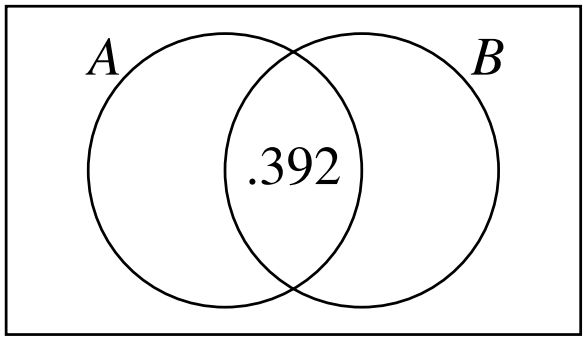
Express the data using event notation

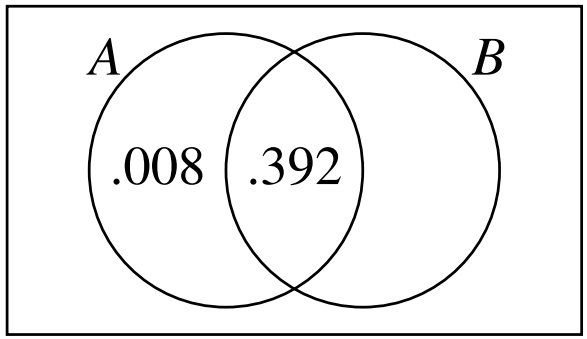
- Event $A =$ individual is in AFR, $\quad A^c =$ individual is in AMR
$$P(A) = .40 \qquad P(A^c) = .60$$
- Event $B =$ genotype CC
$$P(B|A) = .98 \qquad P(B|A^c) = .73$$

1. Probability they're in AFR and have genotype CC

- Events: $A =$ individual is in AFR, $B =$ genotype CC.
- A study sample is 40% AFR and 60% AMR:
$$P(A) = 0.40, \qquad P(A^c) = 0.60.$$
- In AFR, the probability of CC is 98%, while in AMR, it's 73%:
$$P(B|A) = 0.98, \qquad P(B|A^c) = 0.73.$$

- Express the question using event notation: $P(A \cap B) = ?$
- We showed $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.
We have the info for the second of these.
- So $P(A \cap B) = P(B|A)P(A) = (.98)(.40) = \boxed{\mathbf{.392}} = \boxed{\mathbf{39.2\%}}$.

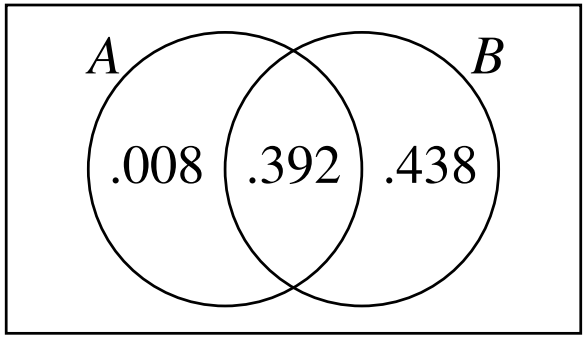


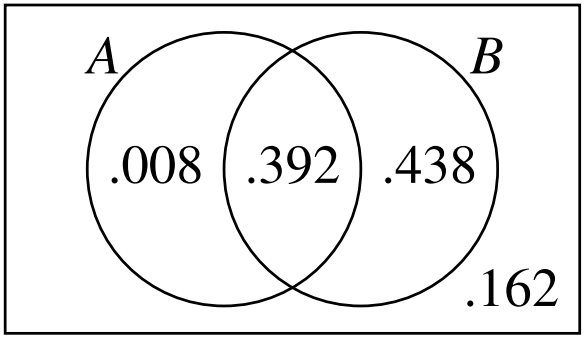


2. Probability genotype is CC

- Events: $A =$ individual is in AFR, $B =$ genotype CC.
- A study sample is 40% AFR and 60% AMR:
$$P(A) = 0.40, \qquad P(A^c) = 0.60.$$
- In AFR, the probability of CC is 98%, while in AMR, it's 73%:
$$P(B|A) = 0.98, \qquad P(B|A^c) = 0.73.$$

- Express the question using event notation: $P(B) = ?$
- $$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A^c) \\ &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= (.98)(.40) + (.73)(.60) = .392 + .438 = \boxed{\mathbf{.830}} = \boxed{\mathbf{83.0\%}} \end{aligned}$$





3. If the genotype is CC, what's the probability they're in AFR?

- Events: $A =$ individual is in AFR, $B =$ genotype CC.
- A study sample is 40% AFR and 60% AMR:
$$P(A) = 0.40, \qquad P(A^c) = 0.60.$$
- In AFR, the probability of CC is 98%, while in AMR, it's 73%:
$$P(B|A) = 0.98, \qquad P(B|A^c) = 0.73.$$

- Express the question using event notation: $P(A|B) = ?$
- $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B|A)P(A)}{P(B)} = \dfrac{(.98)(.40)}{.830}$ $\boxed{\approx \mathbf{.472} \approx \mathbf{47.2\%}}$

Bayes' Theorem (simple version)

Theorem (Bayes' Theorem)

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

This lets us express the probability of A given B, in terms of the probability of B given A.

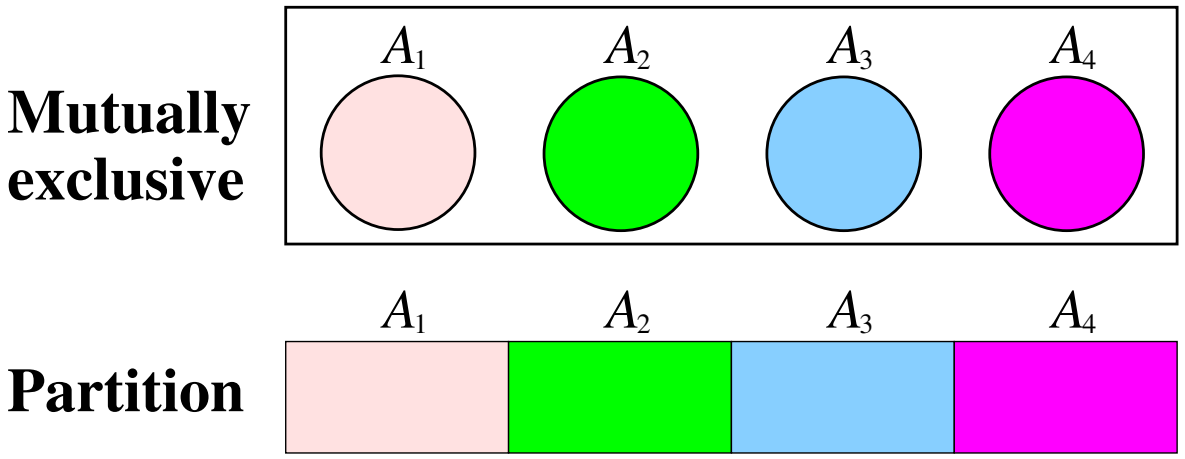
Alternate formulation of Bayes' Theorem

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

where we used

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

Partition of a sample space



Definition (Partition of S)

Events $A_1, \ldots, A_n$ *partition* the sample space S when

- $P(A_i) > 0$ for all i.
- $A_i \cap A_j = \emptyset$ for $i \neq j$. *(pairwise mutually exclusive)*
- $S = A_1 \cup \cdots \cup A_n$.

In a partition, every element of the sample space is in exactly one of the parts $A_1, \ldots, A_n$. Vs. for mutually exclusive, there could be elements in S outside of those parts.

Example: Multiple groups

Data

- A study sample is 10% AFR, 20% AMR, 30% EAS, and 40% SAS. It's designed so that every individual is in exactly one group.
- The probability of genotype CC in each group is
  AFR: 98% AMR: 73% EAS: 7% SAS: 27%
- A random individual is chosen from the sample.

Event notation

- There are four groups:
  $A_1 =$ AFR $A_2 =$ AMR $A_3 =$ EAS $A_4 =$ SAS
- The sample space is $S = A_1 \cup A_2 \cup A_3 \cup A_4$.
- Since the groups don't overlap, $A_i \cap A_j = \emptyset$ when $i \neq j$.
- $B =$ genotype CC.

Example: Multiple groups

Sample space, events, and probabilities

$$\begin{array}{cccc} A_1 = \text{AFR} & A_2 = \text{AMR} & A_3 = \text{EAS} & A_4 = \text{SAS} \\ P(A_1) = 10\% & P(A_2) = 20\% & P(A_3) = 30\% & P(A_4) = 40\% \\ P(B|A_1) = 98\% & P(B|A_2) = 73\% & P(B|A_3) = 7\% & P(B|A_4) = 27\% \end{array}$$

where sample space $S = A_1 \cup \cdots \cup A_4$ and $B =$ genotype CC.

Breaking down the probabilities of events

Sample space, events, and probabilities

$$\begin{array}{cccc} A_1 = \text{AFR} & A_2 = \text{AMR} & A_3 = \text{EAS} & A_4 = \text{SAS} \\ P(A_1) = 10\% & P(A_2) = 20\% & P(A_3) = 30\% & P(A_4) = 40\% \\ P(B|A_1) = 98\% & P(B|A_2) = 73\% & P(B|A_3) = 7\% & P(B|A_4) = 27\% \end{array}$$

where sample space $S = A_1 \cup \cdots \cup A_4$ and $B =$ genotype CC.

Venn diagram

	A_1	A_2	A_3	A_4
B	$B \cap A_1$	$B \cap A_2$	$B \cap A_3$	$B \cap A_4$
B^c	$B^c \cap A_1$	$B^c \cap A_2$	$B^c \cap A_3$	$B^c \cap A_4$

Breaking down the probabilities of events

Sample space, events, and probabilities

$A_1 = \text{AFR}$	$A_2 = \text{AMR}$	$A_3 = \text{EAS}$	$A_4 = \text{SAS}$
$P(A_1) = 10\%$	$P(A_2) = 20\%$	$P(A_3) = 30\%$	$P(A_4) = 40\%$
$P(B\|A_1) = 98\%$	$P(B\|A_2) = 73\%$	$P(B\|A_3) = 7\%$	$P(B\|A_4) = 27\%$

where sample space $S = A_1 \cup \cdots \cup A_4$ and $B = \text{genotype CC}$.

Venn diagram with probabilities

	A_1	A_2	A_3	A_4	Total
B	$P(B \cap A_1)$	$P(B \cap A_2)$	$P(B \cap A_3)$	$P(B \cap A_4)$	$P(B)$
B^c	$P(B^c \cap A_1)$	$P(B^c \cap A_2)$	$P(B^c \cap A_3)$	$P(B^c \cap A_4)$	$P(B^c)$
Total	$P(A_1)$	$P(A_2)$	$P(A_3)$	$P(A_4)$	1

Breaking down the probabilities of events

Sample space, events, and probabilities

$$\begin{array}{llll} A_1 = \text{AFR} & A_2 = \text{AMR} & A_3 = \text{EAS} & A_4 = \text{SAS} \\ P(A_1) = 10\% & P(A_2) = 20\% & P(A_3) = 30\% & P(A_4) = 40\% \\ P(B|A_1) = 98\% & P(B|A_2) = 73\% & P(B|A_3) = 7\% & P(B|A_4) = 27\% \end{array}$$

where sample space $S = A_1 \cup \cdots \cup A_4$ and $B =$ genotype CC.

Venn diagram with probabilities

Fill in top row with $P(B \cap A_i) = P(B|A_i)P(A_i)$,
and fill in column totals $P(A_i)$.

	A_1	A_2	A_3	A_4	Total
B	$(.98)(.1) = .098$	$(.73)(.2) = .146$	$(.07)(.3) = .021$	$(.27)(.4) = .108$	$P(B)$
B^c	$P(B^c \cap A_1)$	$P(B^c \cap A_2)$	$P(B^c \cap A_3)$	$P(B^c \cap A_4)$	$P(B^c)$
Total	.1	.2	.3	.4	1

Breaking down the probabilities of events

Sample space, events, and probabilities

$$\begin{array}{cccc} A_1 = \text{AFR} & A_2 = \text{AMR} & A_3 = \text{EAS} & A_4 = \text{SAS} \\ P(A_1) = 10\% & P(A_2) = 20\% & P(A_3) = 30\% & P(A_4) = 40\% \\ P(B|A_1) = 98\% & P(B|A_2) = 73\% & P(B|A_3) = 7\% & P(B|A_4) = 27\% \end{array}$$

where sample space $S = A_1 \cup \cdots \cup A_4$ and $B =$ genotype CC.

Venn diagram with probabilities

Fill in rest of table to complete column totals. Then compute row totals.

	A_1	A_2	A_3	A_4	Total
B	.098	.146	.021	.108	.373
B^c	.002	.054	.279	.292	.627
Total	.1	.2	.3	.4	1

$$\begin{aligned} P(B) &= P(B \cap A_1) + \cdots + P(B \cap A_4) \\ &= P(B|A_1)P(A_1) + \cdots + P(B|A_4)P(A_4) \end{aligned}$$

Questions

Events and probabilities

$P(A_1) = .1 \quad P(B|A_1) = .98$

$P(A_2) = .2 \quad P(B|A_2) = .73$

$P(A_3) = .3 \quad P(B|A_3) = .07$

$P(A_4) = .4 \quad P(B|A_4) = .27$

	A_1	A_2	A_3	A_4	Total
B	.098	.146	.021	.108	.373
B^c	.002	.054	.279	.292	.627
Total	.1	.2	.3	.4	1

- What is the total probability of CC? $P(B) = \boxed{\mathbf{.373}} = \boxed{\mathbf{37.3\%}}$
- If the sample size is 10000, approximately how many individuals have genotype CC? $(10000)(.373) = \boxed{\mathbf{3730}}$
- If a random individual has genotype CC, what's the probability they're from the i^{th} group?
  - AFR: $P(A_1|B) = \dfrac{P(B|A_1)P(A_1)}{P(B)} = \dfrac{(.98)(.10)}{.373} \approx \boxed{\mathbf{.263}}$
  - AMR: $P(A_2|B) = \dfrac{(.73)(.20)}{.373} \approx \boxed{\mathbf{.391}}$
  - EAS: $P(A_3|B) = \dfrac{(.07)(.30)}{.373} \approx \boxed{\mathbf{.056}}$
  - SAS: $P(A_4|B) = \dfrac{(.27)(.40)}{.373} \approx \boxed{\mathbf{.290}}$

Full version of Bayes' Theorem

Let $A_1, \ldots, A_n$ be mutually exclusive events that partition sample space S, and B be any event on S. Then

- $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$
- If $P(B) > 0$ then for each $j = 1, \ldots, n$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Events can be named based on the problem instead of A, B

- We could have called the events

$$\begin{aligned} \text{AFR, AMR, EAS, SAS} \quad &\text{instead of} \quad A_1, \ldots, A_4, \\ \text{and CC} \quad &\text{instead of} \quad B. \end{aligned}$$

- In this notation, the initial data is

$$\begin{aligned} P(\text{AFR}) &= 10\% & P(\text{CC}|\text{AFR}) &= 98\% \\ P(\text{AMR}) &= 20\% & P(\text{CC}|\text{AMR}) &= 73\% \\ P(\text{EAS}) &= 30\% & P(\text{CC}|\text{EAS}) &= 7\% \\ P(\text{SAS}) &= 40\% & P(\text{CC}|\text{SAS}) &= 27\% \end{aligned}$$

- The total probability of CC is

$$\begin{aligned} P(\text{CC}) = P(\text{CC}|\text{AFR})P(\text{AFR}) + P(\text{CC}|\text{AMR})P(\text{AMR}) \\ + P(\text{CC}|\text{EAS})P(\text{EAS}) + P(\text{CC}|\text{SAS})P(\text{SAS}) \end{aligned}$$

- If a random individual has genotype CC, the probability they're from each group is

$$P(\text{AFR}|\text{CC}) = \frac{P(\text{CC}|\text{AFR})P(\text{AFR})}{P(\text{CC})}, \text{ etc.}$$

Independence (2.5)

Independence

Independence

Events A and B are independent when

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

Derivation from conditional probability

A and B are independent when knowledge of one event doesn't affect the probability of the other event:

$$P(A|B) = P(A) \quad \Leftrightarrow \quad \frac{P(A \cap B)}{P(B)} = P(A) \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B)$$

Independence examples

Rolling two dice (red and green)

- $P(\text{red} = 1) = 1/6$
- $P(\text{green} = 2) = 1/6$
- $P(\text{red} = 1 \text{ and green} = 2) = (1/6)(1/6) = 1/36$
- The two rolls are independent.

Dealing cards

- Draw two cards X, Y from a standard 52 card deck.
- Separately, without knowledge of the other card:
$$P(X \text{ is red}) = 1/2 \qquad \text{and} \qquad P(Y \text{ is red}) = 1/2$$
- Recall that $P(A \cap B) = P(A|B)P(B)$:
$$P(X \text{ is red and } Y \text{ is red}) =$$
$$P(X \text{ is red} \mid Y \text{ is red})P(Y \text{ is red}) = (25/51)(1/2) = \tfrac{25}{102}$$
- This doesn't equal $(1/2)(1/2) = 1/4$, so the cards are dependent.

Independence for multiple events

Rolling two dice (red and green)

$A =$ "red is even" $\quad P(A) = 1/2$

$B =$ "green is even" $\quad P(B) = 1/2$

$C =$ "red+green is even" $\quad P(C) = 1/2$

- **Question:** If red is even and red+green is odd, what's the parity of green? odd
- Any two of the above imply the third, so they are not independent.
- We need a way to check this.

Independence for multiple events

Rolling two dice (red and green)

- $A =$ "red is even", $B =$ "green is even", $C =$ "red+green is even"
- $S = \{(r,g) : r = 1,\ldots,6 \text{ and } g = 1,\ldots,6\}$
- $A \cap B = \{(r,g) : r = 2,4,6 \text{ and } g = 2,4,6\}$
- $P(A \cap B) = 3^2/6^2 = 9/36 = 1/4$
  $P(A)P(B) = (1/2)(1/2) = 1/4$ so A and B are independent.
- $A \cap B = A \cap C = B \cap C = \{(r,g) : r = 2,4,6 \text{ and } g = 2,4,6\}$
  Likewise, A and C are independent, and B and C are independent.
- **Three-way intersection:**

$$A \cap B \cap C = \{(r,g) : r = 2,4,6 \text{ and } g = 2,4,6\}$$

$$P(A \cap B \cap C) = 1/4 \quad \neq \quad P(A)P(B)P(C) = (1/2)(1/2)(1/2) = 1/8$$

A, B, C are dependent.

Independence for multiple events

Events $A_1, A_2, \ldots, A_n$ are independent if *all combinations* of them have multiplicative probabilities:

All pairs:	$P(A_i \cap A_j) = P(A_i)P(A_j)$	i, j distinct
All triples:	$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$	i, j, k distinct

All 4-way, All 5-way, . . . , All n-way

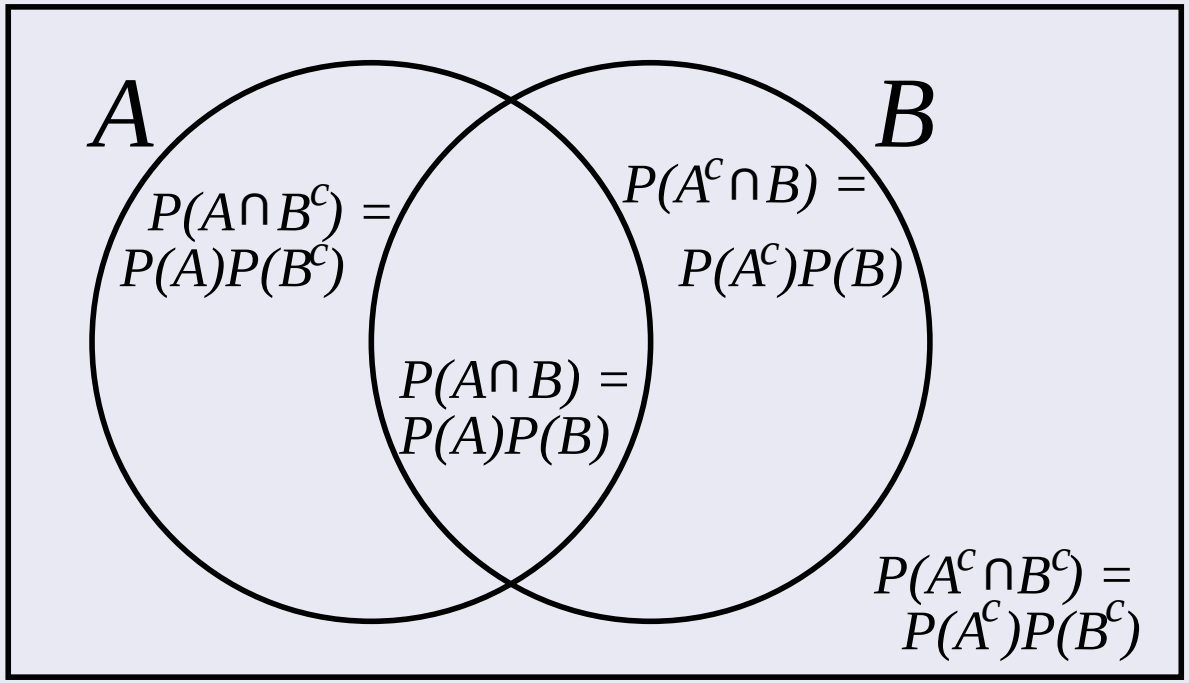
If any of the above equations fail to hold, then $A_1, A_2, \ldots, A_n$ are **dependent**.

Venn diagram of independence

- Event A is split into $A \cap B$ and $A \cap B^c$.
- If A and B are independent, then

$$P(A \cap B^c) = P(A) - P(A \cap B)$$
$$= P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c)$$

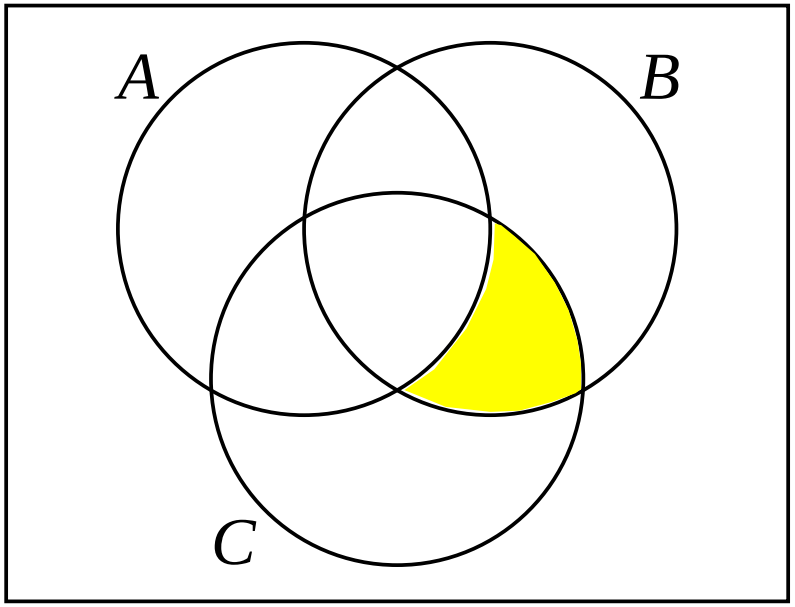
A and B are independent iff all regions of the Venn diagram have multiplicative probabilities



Venn diagram of independence for multiple events

- A, B, C are independent iff all 8 regions follow the multiplication rule; e.g., for the region indicated,

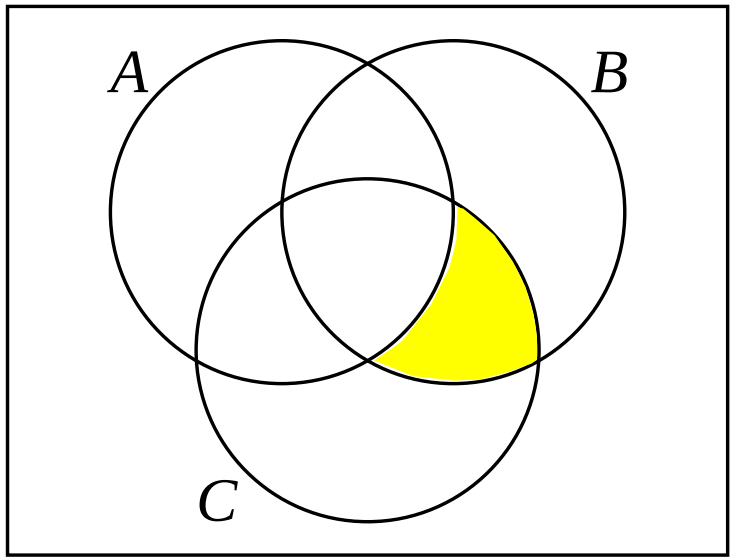
$$P(A^c \cap B \cap C) = P(A^c)P(B)P(C)$$



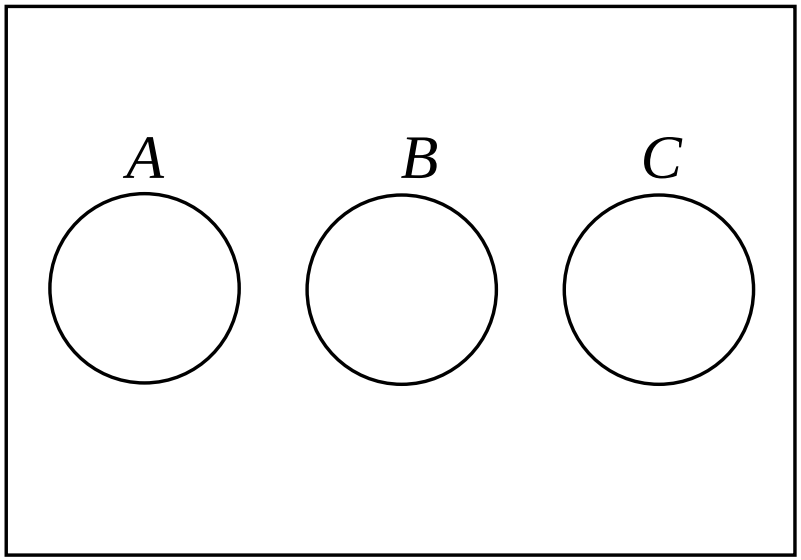
- For a Venn diagram on n sets, the sets are independent iff all 2^n regions obey the multiplication rule.

Independent vs. mutually exclusive

Independent



Mutually exclusive



- A, B, C **independent:**
  - Full Venn diagram with intersecting sets.
  - Intersections and Venn diagram regions have probabilities given by the multiplication formulas.
- A, B, C **mutually exclusive:** No overlaps between sets.

Repeated independent trials

Repeat an experiment over and over, with all trials independent.

Roll a die over and over

The probabilities of the values of the rolls are not influenced by previous rolls, so they are independent.

Draw cards from a deck without replacement

The card values are influenced by previous draws, so they are not independent.

Roll a die 10 times

Probability of at least one 3

The rolls are $R_1, R_2, \ldots, R_{10}$.

$$P(\text{rolling at least one 3}) = 1 - P(\text{no 3's})$$

$$P(\text{no 3's}) = P(R_1 \neq 3)P(R_2 \neq 3)\cdots P(R_{10} \neq 3) = (5/6)^{10}$$

$$P(\text{rolling at least one 3}) = \boxed{\mathbf{1 - (5/6)^{10}}}$$

Probability of rolling exactly one 3

$$P(\text{roll exactly one 3}) = \sum_{i=1}^{10} P(R_i = 3, \text{others} \neq 3)$$

$$= \sum_{i=1}^{10} (1/6)(5/6)^9 = \boxed{\mathbf{10(1/6)(5/6)^9}}$$

Review of geometric series

Geometric series

$$a + ar + ar^2 + ar^3 + \cdots = \sum_{i=0}^{\infty} ar^i = \frac{a}{1-r}$$

where a is the initial term
and r is the ratio, with $|r| < 1$.

A solitaire game

The Rules

Roll a die repeatedly.

- Win if it shows 3.
- Lose if it shows 4.
- Try again otherwise.

Events

- What are the probabilities of winning; losing; and playing forever without winning or losing?
- Events: $A =$ "win", $B =$ "lose", $C =$ "play forever".
- A, B, C are mutually exclusive and $C = (A \cup B)^c$.

Probability of winning

- $A =$ "win" $= A_1 \cup A_2 \cup A_3 \cup \cdots = \bigcup_{k=1}^{\infty} A_k$
  where A_k is the event that you win on the kth roll.
- To win on the kth roll,
  - each of the first $k-1$ rolls must be one of 1, 2, 5, or 6,
  - and the kth roll must be 3.
- $P(A_k) = (4/6)^{k-1}(1/6)$
- $P(A) = \sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} (4/6)^{k-1}(1/6)$
- **Geometric series:**

  First term (plug in $k = 1$): $(4/6)^0(1/6) = 1/6$

  Ratio: $4/6$

  Sum: $P(A) = \frac{1/6}{1-(4/6)} = \frac{1/6}{2/6} = \frac{1}{2}$

- Probability of losing is similarly computed as $P(B) = 1/2$.
- Probability of never winning or losing:

  $C = (A \cup B)^c$

  $P(C) = 1 - P(A) - P(B) = 1 - (1/2) - (1/2) = 0.$