Probability (1)

Random trial

- A random trial is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples: coin tosses, die rolls, iTunes shuffle, etc.
- It can be helpful to model a process as random even if it is not truly random.

Sample space

- Sample space is the collection of all possible outcomes of a trial.
    - A couple has one kid, what is the sample space for the gender of this kid? S = {M, F}
    - A couple has two kids, what is the sample space for the gender of these kids? S = {MM, FF, MF, FM}

sample space vs population?

Event

- An event is a subset of the sample space S
  - $A = \{MF, FM\}$ is an event of $S = \{MM, FF, MF, FM\}$

Event vs sample

Probability

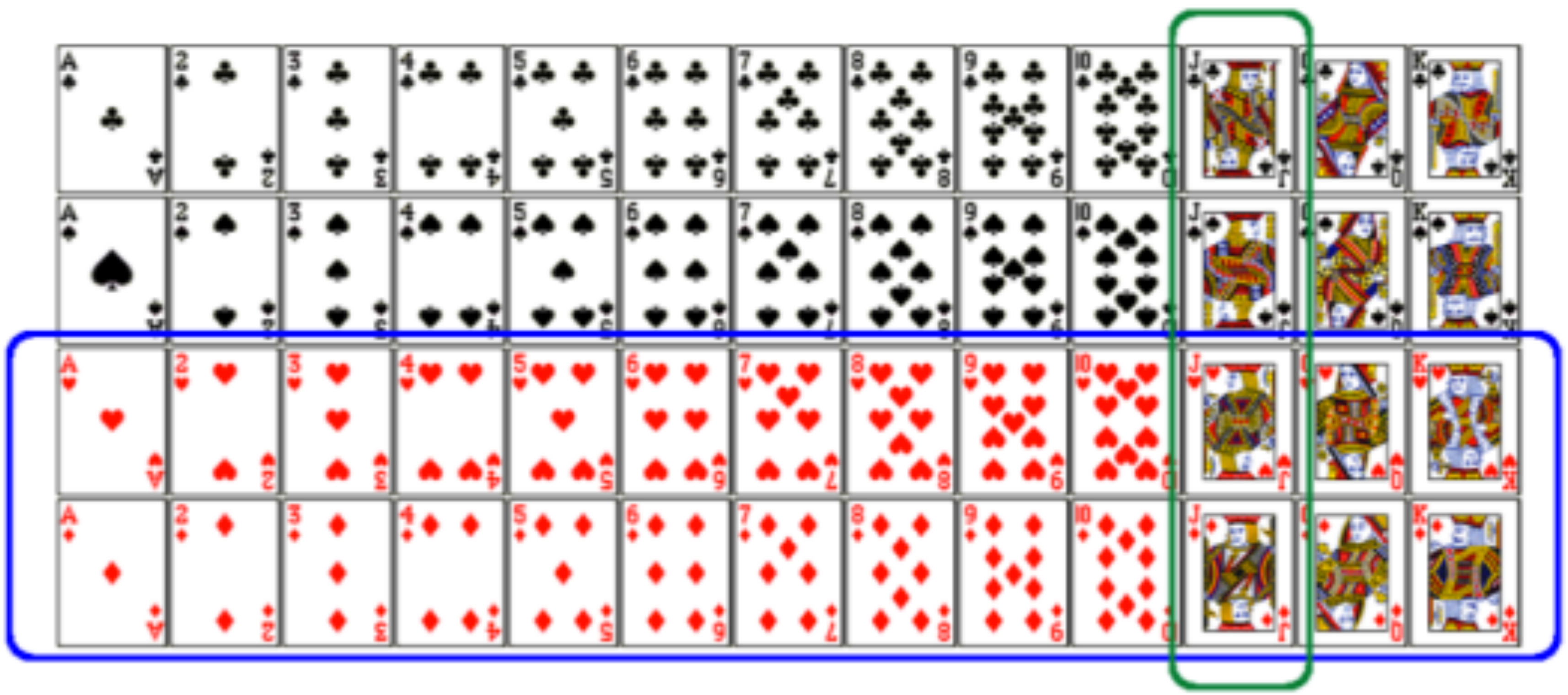
- Frequentist interpretation:
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- Mathematical interpretation:
  - $P(A) =$ probability of event A
  - $0 \leq P(A) \leq 1$
  - $P(S) = 1$

Disjoint

- Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.
    - The outcome of a single coin toss cannot be a head and a tail.
    - A student both cannot fail and pass a class.
    - A single card drawn from a deck cannot be an ace and a queen.
- Non-disjoint outcomes: Can happen at the same time.

Union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck?



$$P(\textit{jack or red}) = P(\textit{jack}) + P(\textit{red}) - P(\textit{jack and red})$$

$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

Addition rule

- General addition rule: Let A and B be two events,

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

Probability distribution

- A probability distribution lists all possible events and the probabilities with which they occur.
    1. The events listed must be disjoint
    2. Each probability must be between 0 and 1
    3. The probabilities must total 1

- The probability distribution for the genders of two kids:

Event	MM	FF	MF	FM
Probability	0.25	0.25	0.25	0.25

Probability of two events

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below:

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?

	relapse	no relapse	total
desipramine	*10*	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	*72*

P(relapsed and desipramine) = 10/72

Marginal probability

What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	*48*	24	*72*

P(relapsed) = 48/72

Conditional probability

If we know that a patient relapsed, what is the probability that they received the antidepressant (desipramine)?

	relapse	no relapse	total
desipramine	*10*	14	24
lithium	18	6	24
placebo	20	4	24
total	*48*	24	72

P(desipramine | relapsed) = 10/48

Conditional probability

- The conditional probability of the outcome of interest A given condition B is calculated as

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

Conditional probability

	relapse	no relapse	total
desipramine	*10*	14	24
lithium	18	6	24
placebo	20	4	24
total	*48*	24	72

$$P(\text{desipramine} \mid \text{relapse}) = \frac{P(\text{desipramine} \cap \text{relapse})}{P(\text{relapse})}$$

$$= \frac{10/72}{48/72} = \frac{10}{48}$$

Conditional probability

	relapse	no relapse	total
desipramine	*10*	14	*24*
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{relapse} \mid \text{desipramine}) = \frac{P(\text{relapse} \cap \text{desipramine})}{P(\text{desipramine})}$$

$$= \frac{10/72}{24/72} = \frac{10}{24}$$

Multiplication rule

- If A and B represent two outcomes or events, then

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

- Note that this formula is simply the conditional probability formula, rearranged.

Independence

- Two trials are independent if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the first card drawn from a deck is an ace may (if without replacement) provide useful information for determining the probability of drawing an ace in the second draw.

Dependence

- Knowing that the first card drawn from a deck is an ace may (if without replacement) provide useful information for determining the probability of drawing an ace in the second draw.
    - Outcomes of two draws from a deck of cards (without replacement) are dependent.

Mathematical definition

- A and B are independent iff

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

- Or equivalently,

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

Independent and identically distributed

- A sequence of random trials is independent and identically distributed (i.i.d.) if each trial has the same probability distribution as the others and all are mutually independent.
- Drawing by sampling with replaces is i.i.d.
- Preferred by most of the statistical and machine learning methods due to its simplicity.

Homework (簡單隨機抽樣)

1. 考慮我們從N個母體中按照取出不放回(sampling without replacement)的方式隨機抽出n個樣本(假設母體中每件東西被抽中的機會均等)。試論證:

   (a) 此n個樣本並非i.i.d.樣本

   (b) 當$n/N \to 0$時，此n個樣本的行為會接近i.i.d.

Homework (分層抽樣)

2. 考慮母體可被分為m個互斥的群體，每個群體各自有N_i個成員，其中$N_1 + N_2 + \cdots + N_m = N$。若我們從每個群體按照$N_i$的比例各自以取出放回的方式抽出$n_i$個樣本，其中$\sum_{i=1}^{m} n_i = n$及$\frac{n_i}{n} \approx \frac{N_i}{N}$。試論證以此方式抽出的$n$個樣本是否為i.i.d.樣本。

Bayes rule

- Bayes' theorem is an important tool in data science:

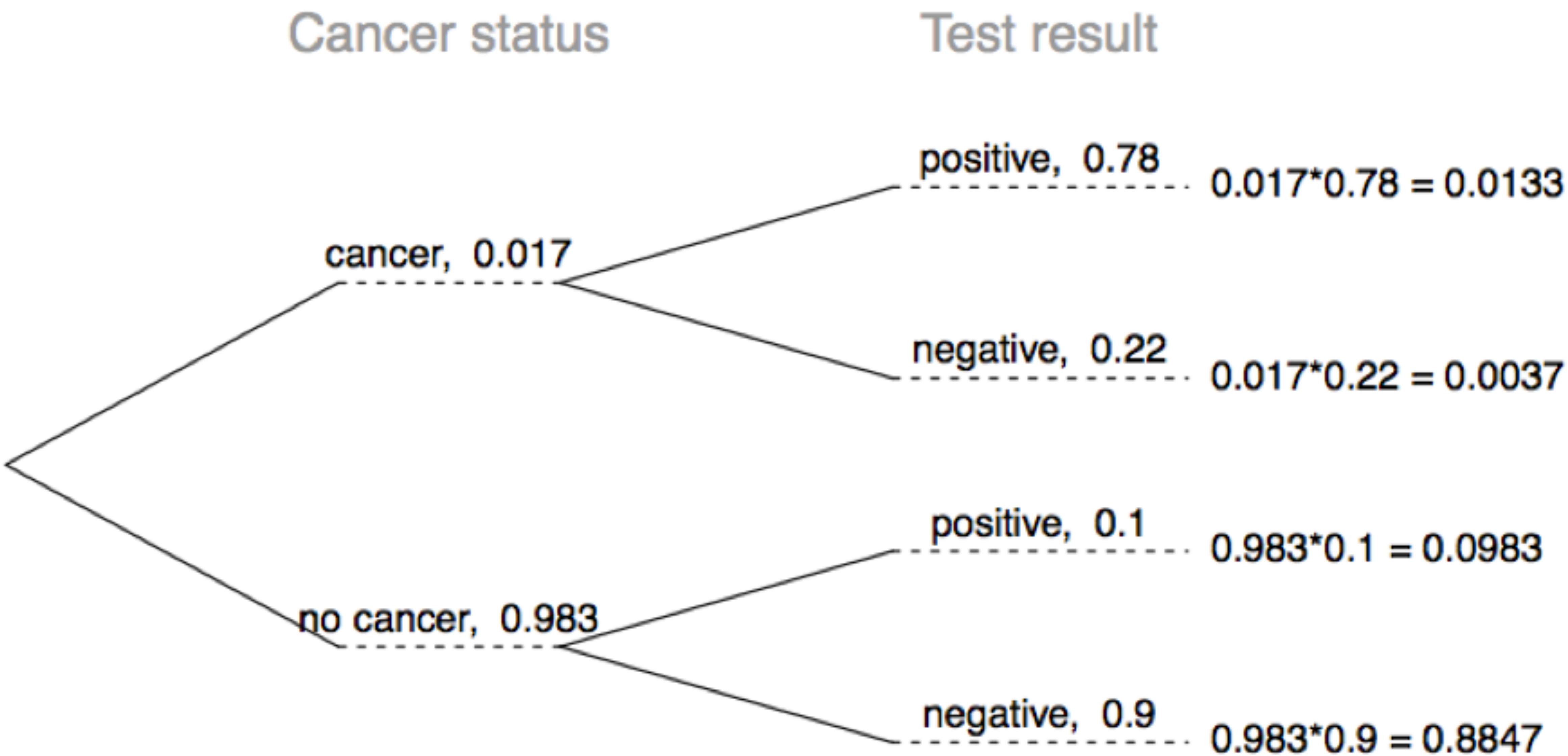
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

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

Cancer screening

- When a patient goes through breast cancer screening there are two competing claims: patient had cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient actually has cancer?

Cancer screening



$$P(C \mid +) = \frac{P(C \cap +)}{P(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$