YANDEX SCHOOL OF DATA SCIENCE

Probability and Statistics for Data Science

Lecture 1 – Sample space and probability

Course logistics

Each week

- Lecture slides
- Paper-and-Pencil (P&P) assignments
- Programming assignments (Python)

2 weeks to complete

Platforms

- Google classroom
- Discussions in Slack

Course logistics

Study groups

- Pairs: same for programming and P&P assignments
- By 27.10 - form study groups (or be assigned randomly)
- By 31.10 - try the assigned groups

(tell us if the randomly assigned group does not work for you)

- Do homework together and alone - find the right balance for you
- Discuss with your classmates
- Have fun! You'll miss theory when things get messy with real data

About me

Rachel Buchuk

- Obtained a BA and MA in Statistics from the Hebrew University
- My Master’s thesis dealt with estimation methods for hospital-acquired infections
- I used probabilistic models to describe the behavior of patients in hospitals
- On the theoretical side: I was the TA in all the probability courses that the department of statistics offers.

Our Course

Probability

- Define relationships between random events
- Build formal models for uncertainty situations

Statistics

- Use probabilistic models to explain the real world
- Estimate parameters that define these models using real data
- Test whether reality support our assumptions

Probability and statistics are tools for solving problems with uncertainty

Today

- Random experiment
- Sample space
- Set theory
- The probability function
- Conditional probability
- Independence
- Bayes theorem
- Naïve Bayes classifier
- Combinatorics
- Bernoulli trials

Random experiment

- **Def:** A random experiment is an experiment in which the outcome cannot be predicted.

In order to define a random experiment, we need to:

1. Define a set with all possible outcomes (sample space)
2. Define different subsets of outcomes (random events)

This means that we need to know how to deal with **sets** and with **counting**.

Examples: Rolling a die; tossing a coin 2 times; picking a random phrase from a book; shooting into a target; generating random characters until a period character is sampled; opening 3 envelopes in a random order; …

Sample space

Def: A sample space is the set of all possible outcomes in a random experiment and is usually denoted by Ω.

Let $\Omega = \{\omega_1, \omega_2, \dots\}$ be a sample space. We need Ω to satisfy the following conditions:

1. The outcomes must be mutually exclusive, i.e. if ω_i occurs, then no other ω_j will take place $\forall i \neq j$.
2. The outcomes must be collectively exhaustive, i.e. on every experiment there will always take place some outcome $\omega_j \in \Omega$.
3. Irrelevant information must be removed from the sample space and the right abstraction must be chosen.

Random event

A random event is a subset of possible outcomes.

- Events that consist of a single outcome are called elementary events
- For any event, we should be able to tell if it happens or not
- For any countable number of events, we should be able to tell whether at least one of them happens

Remark: These conditions determine that the set of all events is a σ-algebra. This is a space on which probability can be properly defined.

Set theory

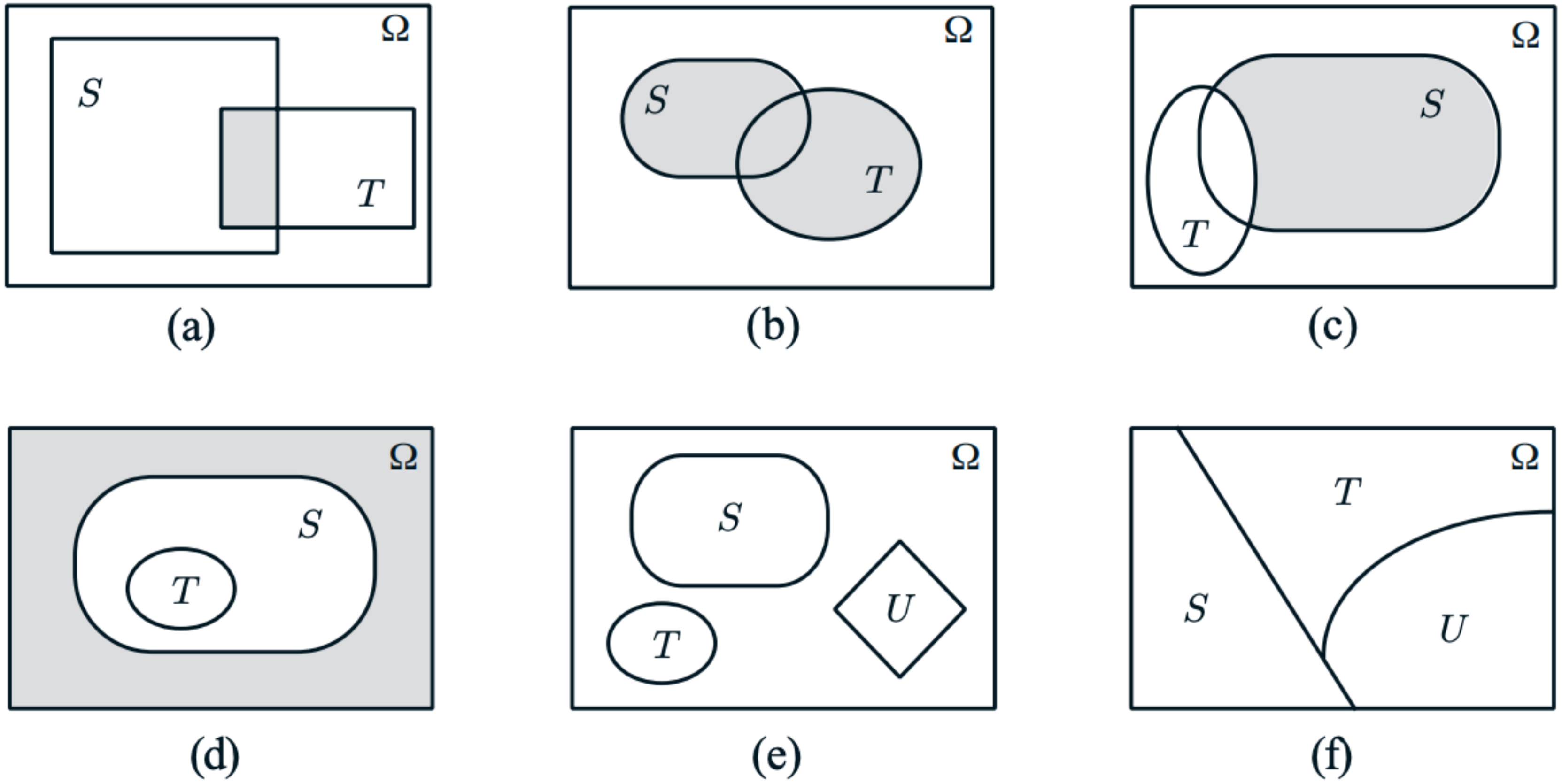
Let A, B be two subsets of Ω. We say that:

1. $x \in A \cap B$ if $x \in A$ and $x \in B$
2. $x \in A \cup B$ if $x \in A$ or $x \in B$ (or in both of them)
3. $\mathrm{B} \setminus \mathrm{A}$ are all the elements in B but not in A
4. $x \in A^c$ if $x \notin A$ ($A^c = \Omega \setminus A$)

Example: $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$, $A = \{\omega_1, \omega_4, \omega_6\}$, $B = \{\omega_2, \omega_4\}$,
$\mathrm{C} = \{\omega_5, \omega_6\}$.
$A \cap B$? $A \cup B$? A^c? $B \setminus A$? $B \cap C$?

Venn diagram

Venn diagrams help us to visualize sets and set operations

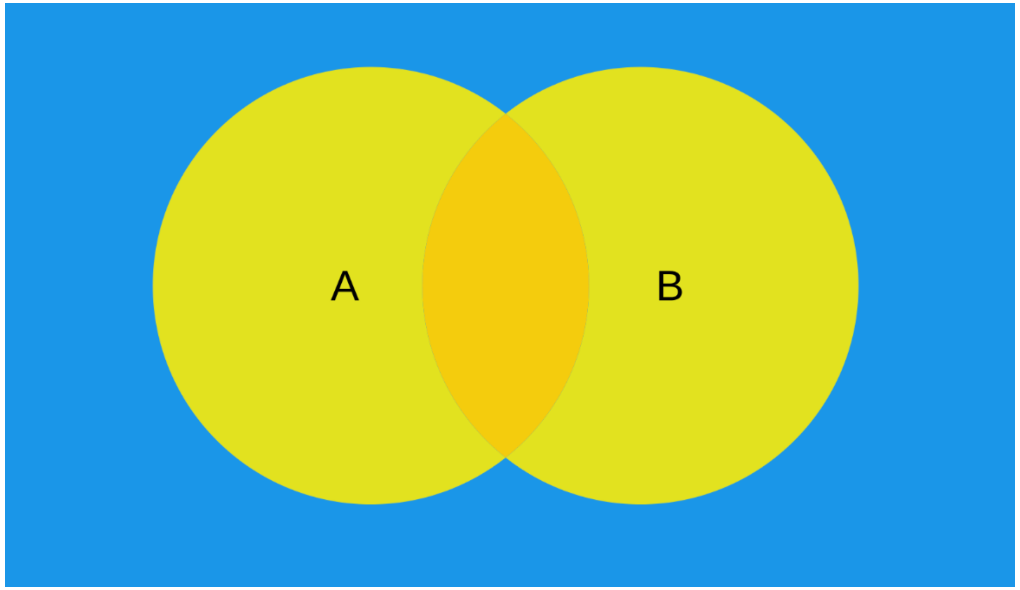


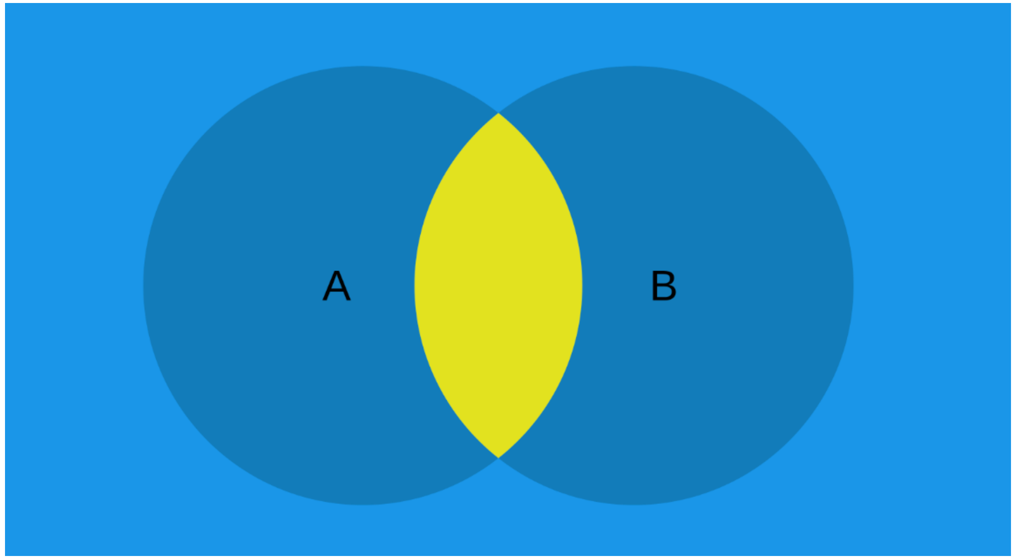
De-Morgan's Laws

Let A and B be two events. Then:

1. $(A \cup B)^c = A^c \cap B^c$
2. $(A \cap B)^c = A^c \cup B^c$

These identities are useful for calculations and can be extended to unions/intersections of n events.

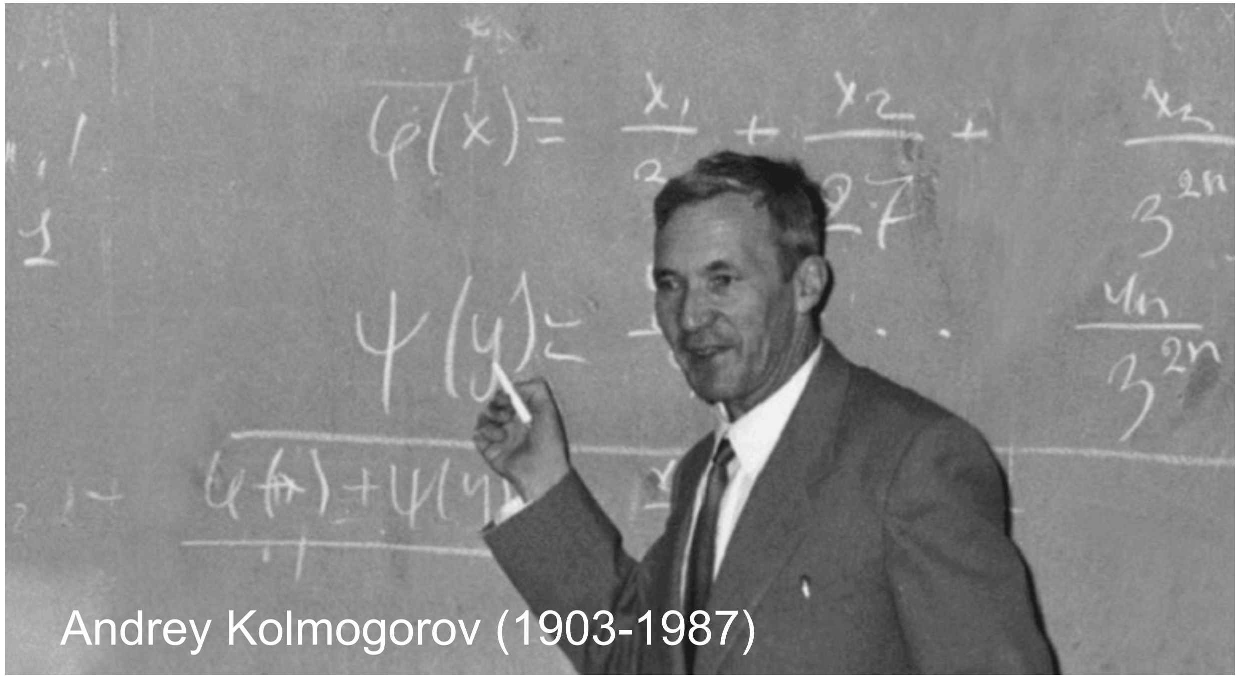




Probability axioms (AKA Kolmogorov axioms)

Def: A probability function P assigns to every event A a number $P(A)$, called the probability of A, satisfying the following axioms:

1. $P(A) \geq 0$, for every event A (nonnegativity)
2. $P(\Omega) = 1$ (normalization)
3. If A, B are two disjoint events, $P(A \cup B) = P(A) + P(B)$ (additivity)


Andrey Kolmogorov (1903-1987)

Example: How everything works together?

- Consider rolling a fair die. For this experiment we have

$\Omega = \{1,2,3,4,5,6\}$ and $P(\{\omega_i\}) = \frac{1}{6}$.

- What is the probability of getting an even number?
- We define the event $A = \{2, 4, 6\}$ and then

$$P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}$$

- Consider now tossing a fair coin twice. For this experiment we have

$\Omega = \{H, T\}^2$ and $P(\{x_1, x_2\}) = \frac{1}{4}$ for all $\{x_1, x_2\} \in \Omega$.

- What is the probability of getting the same result?
- Define the event $A = \{\{H, H\}, \{T, T\}\}$ and then $\mathrm{P}(A) = \frac{1}{2}$

More properties of $P(\cdot)$

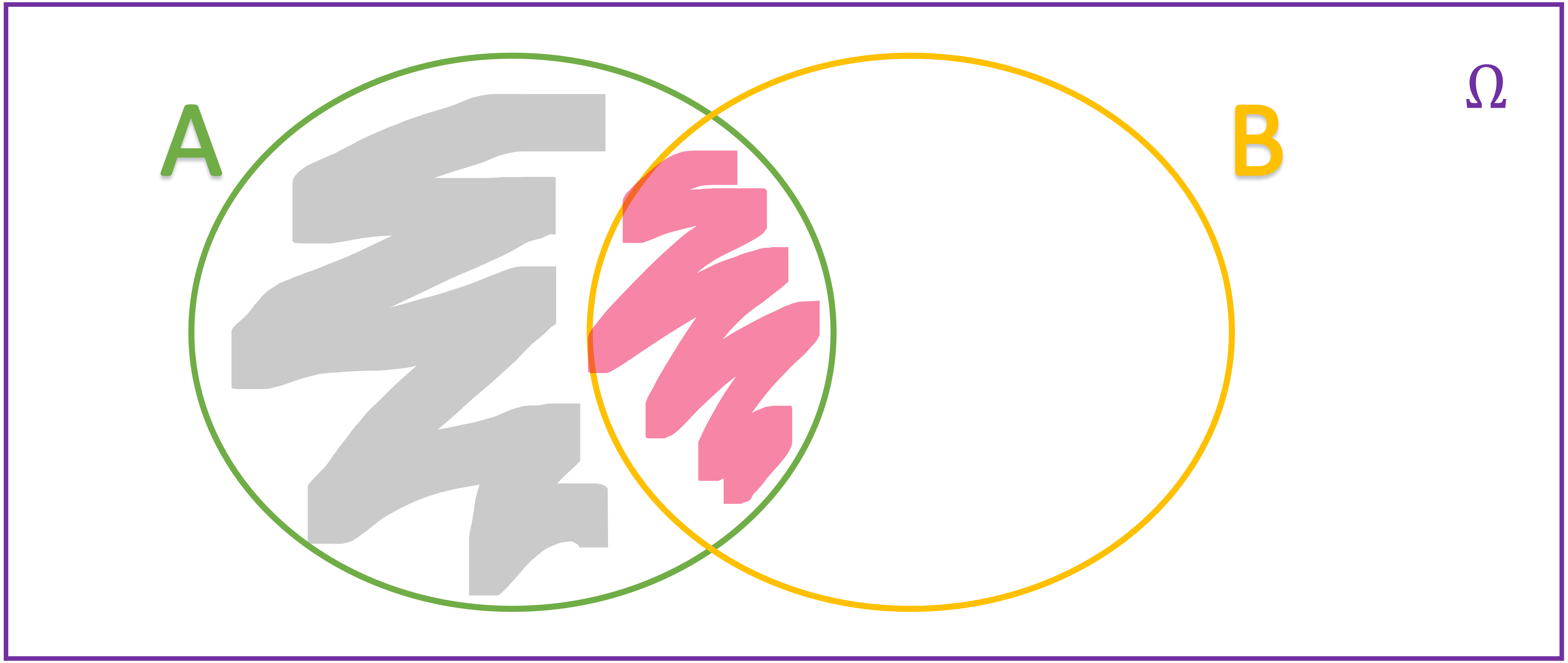
- Theorem: If the sample space Ω is finite, then the probability function is determined by the probability of elementary events:

$$P(A) = \sum_{i:\omega_i \in A} P(\{\omega_i\}) \;\; for \; A \subset \Omega$$

More properties:

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A) \leq 1$
- If $A \subseteq B$ then $\mathrm{P}(A) \leq P(B)$
- $P(A) = P(A \cap B) + P(A \cap B^c)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Why does $P(A) = P(A \cap B) + P(A \cap B^c)$? An intuitive explanation (but not a proof!)



Example:

- In order to pass from first year to second year in the department of statistics, a student must pass both calculus and basic probability. It is known that 30% fail in calculus, 20% fail in basic probability and 10% fail in both of them. What is the probabaility to pass to the second year?
- Define the events of interest: $A = pass\ calculus$,

$B = pass\ basic\ probability$.

- We know that $\mathrm{P}(A^c) = 0.3$, $\mathrm{P}(\mathrm{B}^c) = 0.2$ and $\mathrm{P}(A^c \cap B^c) = 0.1$.
- We need to compute $\mathrm{P}(A \cap B)$.
- In this case:

$$P(A \cap B) = 1 - P\big((A \cap B)^c\big) = 1 - P(A^c \cup B^c)$$
$$= 1 - [P(A^c) + P(B^c) - P(A^c \cap B^c)] = 1 - 0.4 = 0.6$$

Example: Letters

- A letter is taken from the word "mathematics" and a letter from the word "statistics". What is the probability that they are the same letter?
- **Solution:** Assuming that all combinations have the same probability, then it is 14/110 (the fraction of such combinations.

	M	A	T	H	E	M	A	T	I	C	S
S											+
T			+					+			
A		+					+				
T			+					+			
I									+		
S											+
T			+					+			
I									+		
C										+	
S											+

Discrete uniform probability law

- To generalize the intuitive approach from the previous example, we have the following probability law.
- **Theorem:** If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{n}$$

Where $|A|$ is the cardinality* of A .

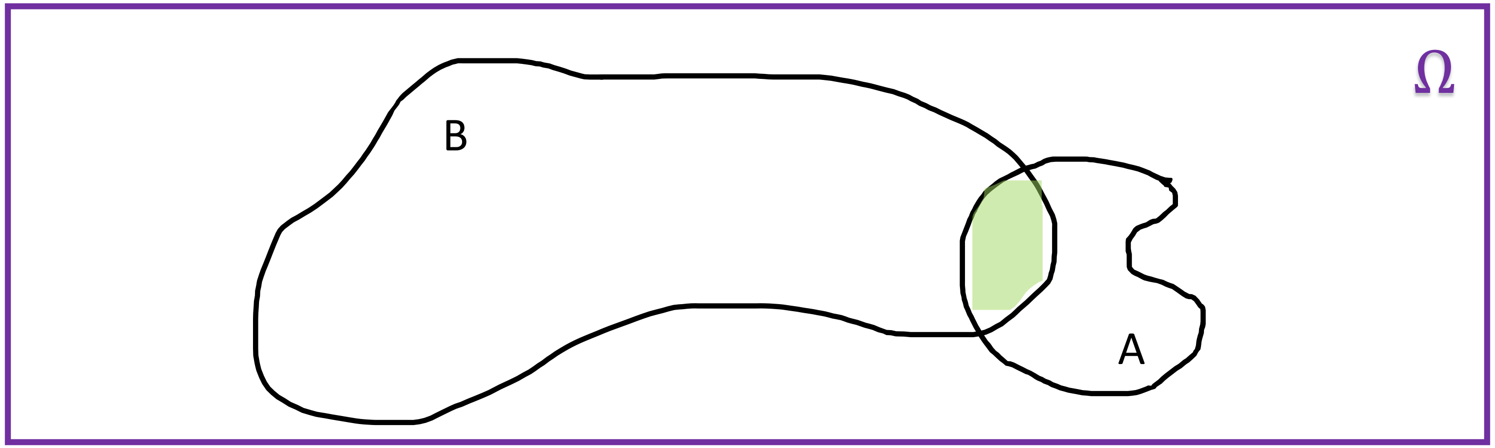
* The cardinality of a finite set is the number of elements in the set.

Conditional probability

- **Def:** The probability of an event A given an event B with $\mathrm{P(B)} > 0$ is

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

- The function $\mathrm{P}(\cdot\ |B)$ is a probability function.
- Interpretation: If we know that B happened, what is the probability that A happened? In a sense, B is our new universe.

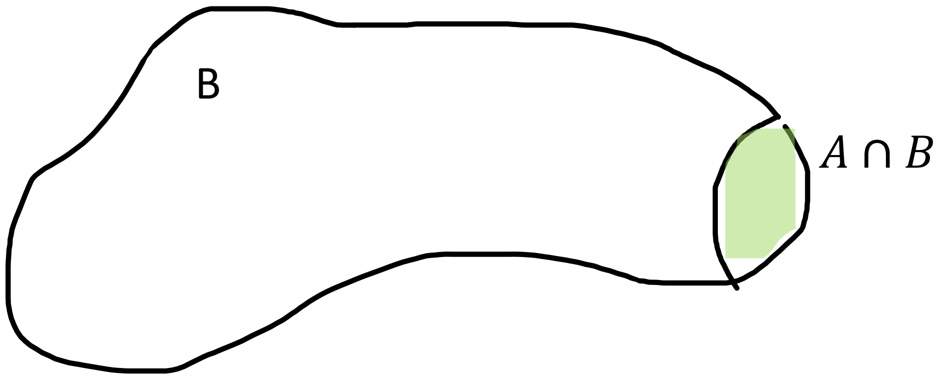


Conditional probability

- **Def:** The probability of an event A given an event B with $\mathrm{P(B)} > 0$ is

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

- The function $\mathrm{P}(\cdot\ |B)$ is a probability function.
- Interpretation: If we know that B happened, what is the probability that A happened? In a sense, B is our new universe.



Example: Coins

- We toss a fair coin 3 successive times. We wish to find the conditional probability $\mathrm{P}(\mathrm{A}|B)$ when A and B are the events $A = \{more\ heads\ than\ tails\ come\ up\}, B = \{1st\ toss\ is\ a\ head\}$
- The sample space consists of 8 (equally likely) sequences
$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$
- The probability of the first toss being H is $\mathrm{P}(B) = \frac{4}{8} = \frac{1}{2}$
- The event $\mathrm{A} \cap B$ consists of the first three elements in the sample space, so $\mathrm{P}(A \cap B) = 3/8$.
- To sum up,
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3}{4}\left(= \frac{|A \cap B|}{|B|}\right)$$

Multiplication law

- The multiplication law is a corollary of the conditional probability function:

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

- In general, for $A_1, \ldots, A_n$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots$$

- Proof: We apply the definition of conditional probability to the right-hand side

$$\cancel{P(A_1)} \frac{\cancel{P(A_1 \cap A_2)}}{\cancel{P(A_1)}} \frac{P(A_1 \cap A_2 \cap A_3)}{\cancel{P(A_1 \cap A_2)}} \cdots \frac{P(A_1 \cap A_2 \cap \cdots \cap A_n)}{P(A_1 \cap A_2 \cap \cdots \cap A_{n-1})}$$

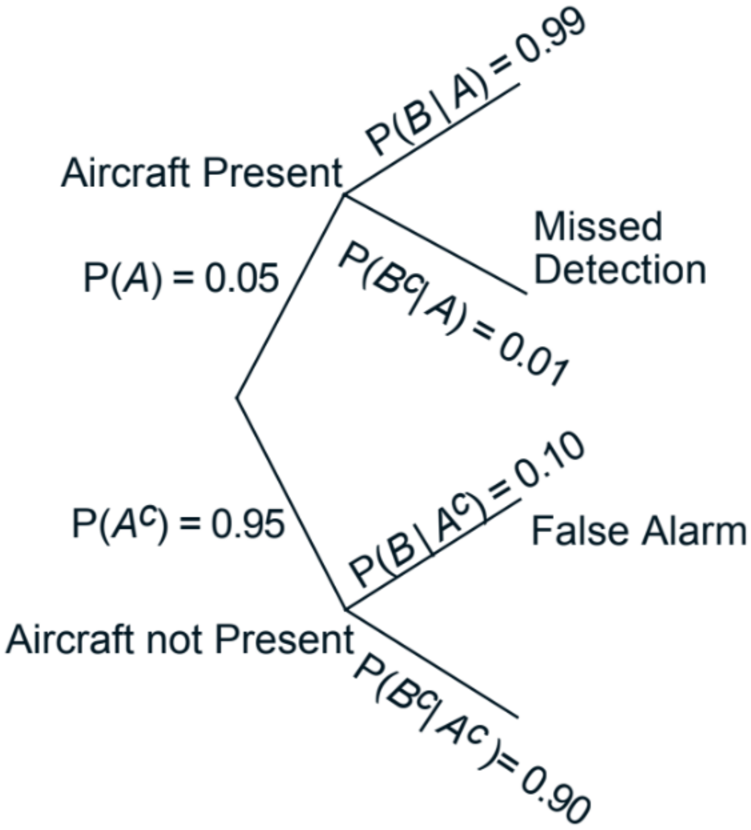
Example: Radar detection

- If an aircraft is present in a certain area, a radar detects it and generates an alarm signal with probability 0.99. If an aircraft is not present, the radar generates a (false) alarm w.p. 0.1. We assume that an aircraft is present w.p. 0.05 What is the probability of aircraft presence and no detection? (false negative)
- Define the events

$$A = \{an\ aircraft\ is\ present\},$$
$$B = \{the\ radar\ generates\ an\ alarm\}$$

- $P(A \cap B^c) = P(B^c|A)P(A) = 0.01 \cdot 0.05 = 0.0005$

Tree-based sequential description



- The event $A \cap B$ uccurs if and only if A and B have occurred.
- The occurrence of $A \cap B$ is viewed as the occurrence of A followed by the occurrence of B and it is visualized as a path on the tree with two branches.
- This can be generalized to the intersection of n events viewed as a tree with n branches.

Law of Total Probability

- The total probability law is given by the formula
$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
- **Proof:** Directly from properties of probability function + the multiplication law.

- Generalized version: Let $B_1, B_2, \dots, B_n$ be a partition of the sample space. Then,
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Example: Alice as a Ydata student

Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind.

If she is up-to-date in a given week, the probability that she will be up-to-date in the next week is 0.8. If she is behind in a given week, the probability that she will be up-to-date in the next week is 0.4.

Alice is up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Solution: Let U_i denote the event that Alice is up-to-date on week i. We know that $P(U_{i+1}|U_i) = 0.8$ and $P(U_{i+1}|U_i^c) = 0.4$. We want to compute $P(U_3)$.

Alice as a Ydata student (solution)

Solution: Let U_i denote the event that Alice is up-to-date on week i. We know that $P(U_{i+1}|U_i) = 0.8$ and $P(U_{i+1}|U_i^c) = 0.4$. We want to compute $P(U_3)$.

$$P(U_3) = P(U_3|U_2)P(U_2) + P(U_3|U_2^c)P(U_2^c)$$
$$= 0.8 \cdot P(U_2) + 0.4 \cdot P(U_2^c)$$

$$P(U_2) = P(U_2|U_1)P(U_2) + P(U_2|U_1^c)P(U_1^c) = 0.8$$
$$\Rightarrow P(U_2^c) = 1 - 0.8 = 0.2$$

And we can complete the calculation: $P(U_3) = 0.72$.

Independence

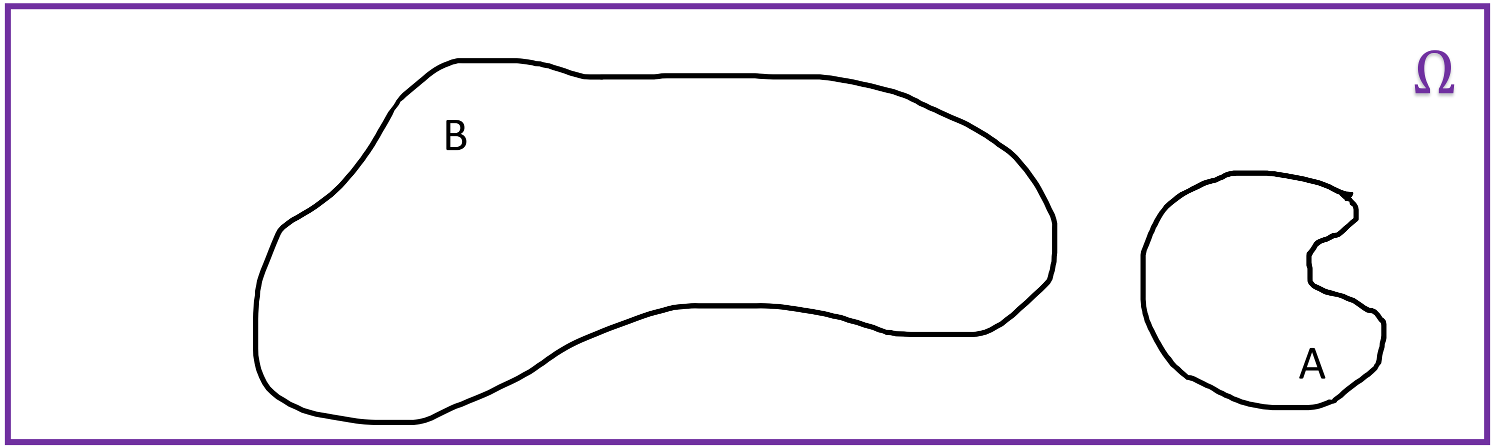
What if the occurrence of event B provides no information regarding the occurrence of event A? That is, $P(A|B) = P(A)$?

- In this case, we say that A and B are **independent** events.
- **Def:** The events A and B are independent if one (and then all) of the following conditions hold:

1. $P(A|B) = P(A)$ or $\mathrm{P}(B|A) = P(B)$ for $P(B) > 0$ or $P(A) > 0$, respectively (note the symmetry)
2. $P(A \cap B) = P(A)P(B)$
3. $P(A \cap B^c) = P(A)P(B^c)$
4. $P(A^c \cap B^c) = P(A^c)P(B^c)$

Independent vs disjoint events

These events are not independent!



In fact, two disjoint events (with positive probabilities) are never independent.

Be careful with your intuition!

Suppose that we toss 2 fair dice. Let A_1 denote the event that the sum of the dice is 6 and B the event that the first die equals 4. Then,

$$P(A_1 \cap B) = P(\{4,2\}) = \frac{1}{36}$$

whereas

$$P(A_1)P(B) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}$$
$$\neq P(A_1 \cap B)$$

=> A_1 and B are not independent.

Now let A_2 denote the event that the sum of the dice equals 7. In this case,

$$P(A_2 \cap B) = P(\{4,3\}) = \frac{1}{36}$$

whereas

$$P(A_2)P(B) = \frac{6}{36} \cdot \frac{1}{6} = \frac{1}{36}$$
$$= P(A_2 \cap B)$$

=> A_2 and B are independent.

Be careful with your intuition!

Suppose that we toss 2 fair dice. Let A_1 denote the event that the sum of the dice is 6 and B the event that the first die equals 4. Then,

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{1/36}{1/6}$$
$$= \frac{1}{6} > \frac{5}{36} = P(A_1)$$

The occurrence of B increases the probability that A_1 will occur.

Now let A_2 denote the event that the sum of the dice equals 7. In this case,

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{1/36}{1/6}$$
$$= \frac{1}{6} = P(A_2)$$

The occurrence of B does not change the probability that A_2 will occur.

Bayes Theorem

Theorem: Let $A_1, A_2, \dots, A_n$ be disjoint events that form a partition of the sample space and assume that $\mathrm{P}(A_i) > 0$ for all i. Then, for any event B such that $P(B) > 0$,

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

Example: Rare disease

A test for a certain rare disease is assumed to be correct 95% of the time: if a person has the disease, the test is positive with probability 0.95, and if the person does not have the disease, the test is negative with probability 0.95.

A random person drawn from the population has probability 0.001 of having the disease.

Given that a person tested positive, what is the probability of having the disease?

Let A be the event of having the disease and B the event of a positive test result.

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)} = \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} = 0.0187$$

Naïve Bayes classifier: Credit scoring example

$x = (x_1, x_2, \dots, x_k)$ – data about credit application of the client

$y \in \{bad, good\}$ – future behavior of the client

$P(y|x) = ?$

We can model the above probability as follows:

$$P(y|x) = \frac{1}{P(x)}\ P(y)P(x|y)$$

- We can estimate $P(y)$ with the available data.
- $P(x|y)$ is better than $P(y|x)$ in the sense that we can model the former using a simple model.
- We can treat $P(x)$ as a scaling factor.

Naïve Bayes classifier: Credit scoring example

To describe $P(x|y)$ we first rewrite it using the multiplication law

$$P(x|y) = P(x_1 \cap \cdots \cap x_k|y)$$
$$= P(x_1|y)P(x_2|x_1 \cap y)P(x_3|x_1 \cap x_2 \cap y) \cdots P(x_k|x_1 \cap \cdots \cap x_{k-1}y)$$

Now we make a **naïve assumption**.

We assume that all x_j's are independent conditional on y. That is,

$$P(x_1 \cap \cdots \cap x_k|y) = P(x_1|y)P(x_2|y)P(x_3|y) \cdots P(x_k|y)$$

This assumption allows us to estimate $P(x|y)$ easily.

This assumption is usually wrong, but the model may still be useful and does not require a large amount of data.

Combinatorics (some useful formulas)

- Inclusion-exclusion prnciple:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

- Number of **permutations** of n objects: $n!$
- Number of **ordered** samples of size r, **with** replacement, from n objects: n^r
- Number of **ordered** samples of size r, **without** replacement, from n objects:

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!} = {}_nP_r.$$

- Number of **unordered** samples of size r, **without** replacement, from a set of n objects (= number of subsets of size r from a set of n elements) (**combinations**):

$$\binom{n}{r} = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\dots(n-r+1)}{r!}.$$

Bernoulli trials

- **Bernoulli trials** are independet repeated trials of an experiment with exactly two possible outcomes.

Remember that example in which we toss a coin 3 times and calculated some probabilities after describing our sample space? This is a Bernoulli trial with $n = 3$ and $p = 0.5$.

Let’s try to use combinatorics for this problem with general n and general p.

$A_k = \{getting\ heads\ k\ times\}$, then

$$P(A_k) = \binom{n}{k} p^k (1-p)^{n-k}$$

References

- Tsitsiklis, John (Spring 2018). "Sample Spaces". Massachusetts Institute of Technology. Retrieved July 9, 2018.

- Bertsekas, Dimitri P. and Tsitsiklis, John N.. "Introduction to Probability." (2008) .

- Ross, Sheldon M. A First Course in Probability / Sheldon Ross. Eighth edition, global edition. Harlow: Pearson Education Limited, 2010.

- Lecture notes by A.J. Hildebrand https://faculty.math.illinois.edu/~hildebr/408/408combinatorial.pdf