# Lecture 2

Cezar Ionescu

DEPARTMENT FOR CONTINUING EDUCATION

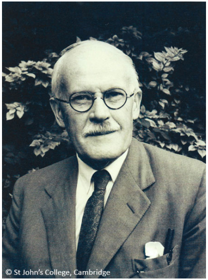
UNIVERSITY OF OXFORD

18/05

# Probability theory

Richard von Mises (1883-1953):
*The term probability is meaningful for us only with regard to a clearly defined collective (or population).*

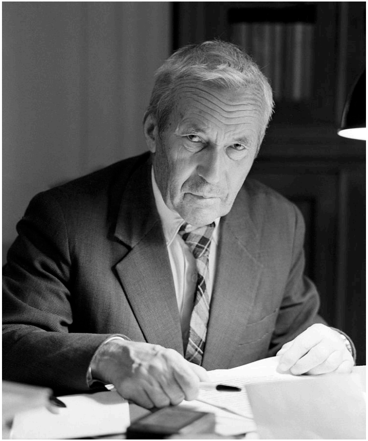
Sir Harold Jeffreys (1891-1989):
*[T]here is a valid primitive idea expressing the degree of confidence that we may reasonably have in a proposition.*



# Mathematical model

A. N. Kolmogorov (1903-1987):

- $\Omega$, the set of all possible outcomes
- `Event`, the set of all events
  `Event ⊆ ℙ(Ω)`
- `p : Event → [0, 1]`, the probability measure, satisfying
  1. `p(∅) = 0, p(Ω) = 1`
  2. for all events $X_1, \ldots, X_n, \ldots$ pairwise disjoint,
     $\Sigma\, p(X_i) = p(\bigcup X_i)$



# Examples

1. Drawing a card from a standard deck.
   - `Ω = {2♣, 2♦, 2♥, 2♠, ...}`
   - `Event = ℙ(Ω)`
2. Choosing a point of the unit disc.

- `Ω = {(x, y) | x, y ∈ ℝ, x² + y² ≤ 1}`
- We would like to have `Event = ℙ(Ω)` … but we can't
- We can take `Event = {X : Ω → {0, 1} | X computable}`

# The classical model of probability

`Ω` finite, `Event = ℙ(Ω)`, and `p` defined by

```
p(X) = card(X) / card(Ω)
```

Interpretations:

- Frequentist: every elementary outcome `ω ∈ Ω` turns up with approximately the same frequency as any other.
- Bayesian: we have no reason to prefer one elementary outcome over any other.

# Example

Rolling a die:

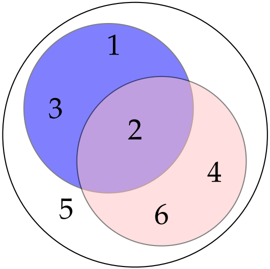
- `Ω = {1, 2, 3, 4, 5, 6}, card(Ω) = 6`
- rolling an even number: `even = {2, 4, 6}, card(even) = 3`
- probability of rolling an even number:
  `p(even) = card(even)/card(Ω) = 3/6 = 0.5`

# Operations with events

- union: `A ∪ B`
- intersection: `A ∩ B`
- complement: `¬ A = Ω - A`

---

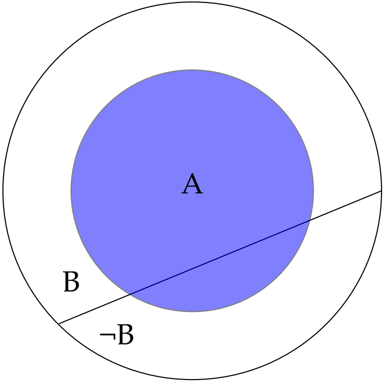
- result is even **and** ≤ `3`
- result is even **or** ≤ `3`
- result is **not** ≤ `3`



# Exercises

1. Derive the formula for $p(A \cap B)$.
2. Derive the formula for $p(A \cup B)$
3. Assume that $B_1, \ldots, B_n$ are pairwise disjoint. Show that

$$p(A) = p(A \cap B_1) + \ldots + p(A \cap B_n)$$



# Conditional probability

**Definition**: Let `(Ω, Event, p)` as above, `X ∈ Event` s.t `p(X) ≠ 0`. Let `Y ∈ Event`. The **conditional probability** of `Y` given `X` is defined by

```
p(Y | X) = p(Y ∩ X) / p(X)
```

# Notational problems

- is $|$ a set-theoretical operation?
- is $|$ commutative?
- Kolmogorov's original notation: $\mathtt{p}_X(\mathtt{Y})$

# The law of total probability

$B_1, \ldots, B_n$ pairwise disjoint, $p(B_i) \neq 0$ for all $i$. Show that

$$p(A) = p(A \mid B_1) * p(B_1) + \ldots + p(A \mid B_n) * p(B_n)$$

# Bayesian learning

- Set of hypotheses, `H`
- Data, `d ∈ D`
- Find the "best" hypothesis that fits the data
  - e.g., `Find-S`, "best" = the most specific hypothesis consistent with the data.
- Bayesian learning: "best" = *the most probable* hypothesis, given the data.

# Maximum a posteriori hypothesis

The aim of Bayesian learning is to determine the **maximum a posteriori (MAP) hypothesis**:

```
hmap = argmaxh p(h | d)
```

What does `p(h | d)` mean?

# Bayes' theorem

```
X, Y ∈ Event, p(X) * p(Y) ≠ 0.

p(X | Y) = p (X ∩ Y) / p (Y)
    therefore p(X ∩ Y) = p(X | Y) * p(Y)
p(Y | X) = p (Y ∩ X) / p (X)
    therefore p(Y ∩ X) = p(Y | X) * p(X)
```

`X ∩ Y = Y ∩ X ⇒ p(X ∩ Y) = p(Y ∩ X)`, therefore:

```
p(X | Y) = p(Y | X) * p(X) / p(Y)
```

# Example

Mitchell, pages 157-158:

- `H = {healthy, ill}`
- test results: $\ominus$, $\oplus$
- `0.008` of the population have the disease.
- patient ill $\Rightarrow$ test positive in `98%` of the cases
- patient healthy $\Rightarrow$ test positive in `3%` of the cases
- A patient's test has come back positive.

What is the MAP hypothesis?

# Translation to probability-speak

- $\Omega$ = the set of people in the population
- `Event = ℙ(Ω)`
- classical model for `p`: `p(X) = card(X)/card(Ω)`
- the events of interest: `Healthy`, `Ill`, `Pos`, `Neg`
- `p(healthy) = 0.992`, `p(ill) = 0.008`, `p(⊕ | ill) ) = 0.98`, `p(⊕ | healthy) = 0.03`

Compute `p(ill | ⊕)`.

# Applying Bayes' theorem

```
    p(ill | ⊕)
=   p(⊕ | ill) * p(ill) / p(⊕)
=   0.98 * 0.008 / p(⊕)
=   0.00784 / p(⊕)

    p(healthy | ⊕)
=   p(⊕ | healthy) * p(healthy) / p(⊕)
=   0.03 * 0.992 / p(⊕)
=   0.02976 / p(⊕)
```

Therefore, the MAP hypothesis is `healthy`.

# A general argument

```
hmap = argmaxh p(h | d)
     = argmaxh (p(d | h) * p(h) / p(d))
     = argmaxh (p(d | h) * p(h)) -- p(d) is constant w.r.t. h
```

# Likelihood

If all hypotheses are equally likely:

$$
\begin{aligned}
h_{map} &= \text{argmax}_h \; p(h \mid d) \\
&= \text{argmax}_h \; p(d \mid h)
\end{aligned}
$$

$p(d \mid h)$ is called the **likelihood** of $d$ given $h$.

A hypothesis that maximises the likelihood is a *maximum likelihood hypothesis (ML)*. Therefore

$$h_{ml} = \text{argmax}_h \; p(d \mid h)$$

In general, $h_{map} \neq h_{ml}$!

# Exercise

What is `h m l` in the example above?

# Homework

Apply Bayes' theorem to answer the following question (Elmer Mode 1966, page 53):

> *A class in advanced mathematics contains 10 juniors, 30 seniors, and 10 graduate students. Three of the juniors, 10 of the seniors, and 5 of the graduate students received an A in the course. If a student is chosen at random from this class and is found to have earned an A, what is the probability that he is a graduate student?*

# Concept learning revisited

Concept:

`c : X → {0, 1}`

`h ∈ H ⇒ h : X → {0, 1})`

We assume `c ∈ H`

The data:

$d = \{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$

# “Brute-force” Bayesian concept learning

```
given H, p(h)
for every h ∈ H:
  compute p(h | d) = p(d | h) * p(h) / p(d)
  return h_map = argmax_h p(h | d)
```

The problem is computing, for every `h`, `p(d | h)`. We also want to compute `p(d)`.

# Computing the likelihood

Assumption: the data is completely correct, and hypotheses are not "noisy" functions.

```
p(d | h) = if for all (x_i, b_i) ∈ d  h(x_i) = b_i
              then 1
              else 0
```

# Computing `p(d | h)`

We can now compute `p(d)` by applying the law of total probability:

```
p(d) = Σₕ p(d | h) * p(h)
```

In a certain sense, the data and the hypothesis must have the same nature.

# Consistent learners

Assume all hypotheses are equally likely. Then the brute-force Bayesian concept learning algorithm will return an ML hypothesis:

$$h_{map} = h_{ml} = \text{argmax}_h \; p(d \mid h)$$

Since $p(d \mid h) = 1$ if $h$ is consistent with the data, and $p(d \mid h) = 0$ otherwise, the algorithm will return any one of the hypotheses in $H$ consistent with $d$.

# Bayesian concept learning and Find-S

Find-S returns a consistent hypothesis ⇒ it returns a MAP (and an ML) hypothesis!

Even if p is not constant w.r.t. hypotheses, Find-S can still return a MAP hypothesis, if there is a unique most specific hypothesis consistent with d in H and

$h_i \subseteq h_j \Rightarrow p(h_i) \geq p(h_j)$