Later in the course we'll refer to this as the <u>multinomial</u> (probability) <u>distribution</u>

(16 Apr 19)

How to work with $\boxed{\text{OR}}$ when you have more than 2 events (union) $\updownarrow$ $\cup$

We already worked out that

$$P(A_1 \text{ or } A_2) = P(A_1 \cup A_2)$$

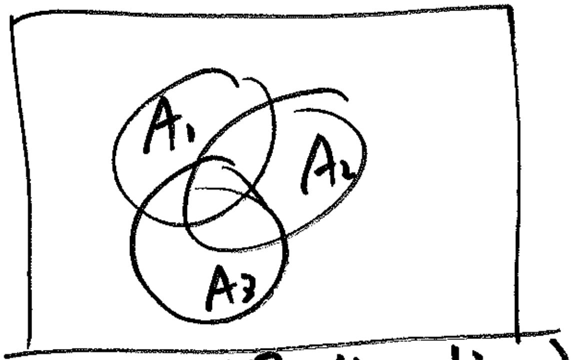
$$= P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

and

---

we also know from Kolmogorov's 3rd Axiom that if events $A_1, \dots, A_n$ are disjoint then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$.

How do these 2 things generalize?



(Ballantine)

By (tedious) enumeration you can show that with 3 events,

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$

$$- \left[P(A_1 \cap A_2) + P(A_1 \cap A_3) + P(A_2 \cap A_3) \right]$$

$$+ P(A_1 \cap A_2 \cap A_3).$$

You probably can now see (or guess) how this generalizes: for any events $A_1, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

$$+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots +$$

$$(-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

Example/ Get 2 decks of ordinary playing cards; order deck 1 from (1 to 52) using any sequence you like, e.g.

1 = 2♣
⋮
13 = A♣
14 = 2♦
⋮
26 = A♦
27 = 2♥
⋮
39 = A♥
40 = 2♠
⋮
52 = A♠

Shuffle deck 2 until (practically speaking) all 52! orderings are equally likely.

Now turn the first card of each deck over; do they match? Continue through all 52 cards; P(at least one match) = ?

let n=52

Let A_i = (a match occurs on card i); we want $P(\bigcup_{i=1}^{n} A_i)$, which can be computed with the complicated formula on the previous page.

Follow the logic detailed on DS pp. 45-50 to obtain

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n+1}\frac{1}{n!}$$

wolfram alpha

limit (sum (-1)^(i+1) / i!, i = 1 to n) as n -> infinity

calculus result :

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} = 1 - \frac{1}{e} \doteq 0.63$$

This sum approaches its limit quickly; already with $n = 7$ you have the first 4 significant figures : 0.6321

DS ch. 2

Conditional probability

Note that Kolmogorov's probability axioms defined the function $P_K(A)$, where A is a set in the collection $\mathbb{C}$ of subsets of the sample space S in which nothing weird can occur; in other words, $P_K(A)$ is a function of a single argument A. To include the extremely useful idea of conditional probability in his setup, Kolmogorov has to define it using P_K. **Definition** Given any two events A, B in $\mathbb{C}$, the conditional probability of A given B is

$$\boxed{P(A \mid B) = \begin{cases} \dfrac{P(A \cap B)}{P(B)} & \text{if } P(B) > 0 \\ \text{undefined} & \text{if } P(B) = 0 \end{cases}}.$$

There are other foundational theories
of probability – one by the Italian mathematician and actuary (deF) Bruno de Finetti (1906–1985), and another by the American physicists Richard T. Cox (1898–1991) and Edwin T. Jaynes (1922–1998) (CJ) – in which the probability function $P_{deF}(A \mid B)$ or $P_{CJ}(A \mid B)$ has $\underline{\underline{2}}$ inputs, not 1, so that <u>conditional</u> probability is the primitive concept, not <u>unconditional</u> probability as with Kolmogorov's $P_K(A)$. deF and CJ

were responding to the reality that

$$\boxed{\text{in practice, all probabilities are conditional on background Assumptions, Information and Judgments (AIJ)}}$$

$\boxed{\text{Example}}$ (Tay-Sachs)

we actually computed not

$$P(\text{at least 1 T-S baby}) \quad \text{but}$$

$$P\left(\begin{array}{c} \text{at least 1} \\ \text{T-S baby} \end{array} \middle| \begin{array}{c} \text{family of 5, and mother and} \\ \text{father both carriers} \end{array} \right)$$

This impulse, to be explicit about your AIJ, is <u>Bayesian</u>; Kolmogorov worked in the <u>frequentist</u> paradigm; in this course, focusing on $P_K(B)$, we need to remember that it should really be $P_K(B \mid AIJ)$.

Consequences of the conditional probability definition (theorems)

① A, B events in $\mathcal{C}$:

if $P(B) > 0$ then

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

and if $P(A) > 0$

then $P(A \cap B) = P(A)\, P(B \mid A)$.

② Direct generalization: if $A_1, \dots, A_n$ are events with $P(A_1 \cap \dots \cap A_{n-1}) > 0$;

then

chain rule for $\cap$ = and

$$P(A_1 \cap \dots \cap A_n) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \dots \cap A_{n-1})$$

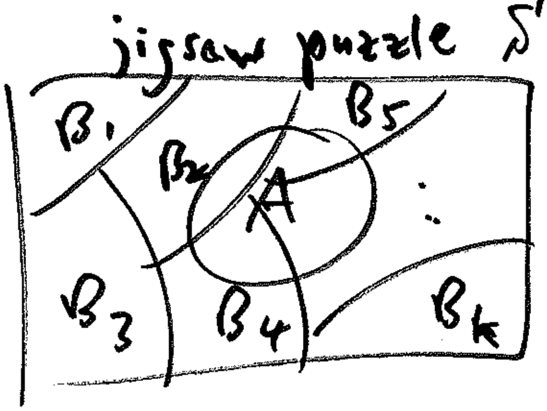
Recall previous Definition

S sample space; if you can find events $B_1, \dots, B_k$ in $\mathcal{C}$

such that the B_j are disjoint and 51

<u>exhaustive</u> $\left(\bigcup_{i=1}^{n} B_i = S \right)$, then you

have found a <u>partition</u> $(B_1, \ldots, B_k)$ of S.



③ If $(B_1, \ldots, B_k)$ is a partition of S

with $P(B_j) > 0$ for all $j = 1, \ldots, k$,

then for any event A in C

$$P(A) = \sum_{j=1}^{k} P(B_j)\, P(A \mid B_j) \text{ —}$$

this is the $\boxed{\text{Law of Total Probability}}$ LTP

When is the LTP useful?

You're trying to compute $P(A)$ and you find it hard to compute directly. If you can find some aspect B of the world satisfying 2 properties —

① B defines a partition $\{B_1, \dots, B_k\}$ with known $P(B_j)$ of S and ② A <u>depends</u> on B in such a way that the conditional probabilities $P(A \mid B_j)$ are easier to compute than $P(A)$ itself — then you can work out $P(A)$ indirectly:

$$P(A) = \sum_{j=1}^{k} \overbrace{P(B_j)\,P(A \mid B_j)}^{P(A \cap B_j)} .$$

(Bayesian mixture modeling)

(18 Apr 19)