You can play around with $P(A)$ as a function of k for fixed $n = 365$ & find that $P(A) > 0.5$ for $k \geq 23$, which many people find surprisingly low.

(2 Aug 17)

Generalizing the binomial coefficients

(p. 33)

$\binom{n}{y}$

What if there are more than 2 possible outcomes in a generalization of the Toy-Sacks case study (T, N)?

T: fixed by us; N: not fixed by us

we want n distinct elements to be divided into k different groups ($k \geq 2$) so that n_j elements fall into group j, $\sum_{j=1}^{k} n_j = n$,

Q in how many different ways can this be done? Follow the argument in DS pp. 42-43, which generalizes the line of reasoning leading to the binomial coefficients $\binom{n}{k}$ when $k=2$:

$$\frac{n!}{k!(n-k)!}$$

Definition: A multinomial coefficient is of the form

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

$n \geq 1$
$k \geq 2$
$1 \leq n_j \leq k$
$\sum_{j=1}^{k} n_j = n$

this answers the how many different ways question Q above

Example: (2016 presidential election)
(see DS pp. 333-334) (with replacement)

Imagine randomly sampling n eligible prospective voters from all such people in the U.S. ("the population" ⊛ (TAn, 2016));

possible outcome ($k=5$)
① Clinton (Democrat)
② Trump (Republican)
③ Johnson (Libertarian)
④ Stein (Green)
⑤ Undecided

let X_i = # people in sample who say they will vote for candidate i, $i = 1, \ldots, k=5$ as in this table.

Suppose (unknown to us) that the proportion of voters who favor candidate i in the population ⊛ above is p_i,

where $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i = 1$.

Because the people are chosen with independent identically distributed (IID) sampling (ie, at random with replacement), each person's outcome will be independent of all the other outcomes. Thus

P(1st person favors candidate i_1, (2) 2nd person favors i_2, $\cdots$, (2) person n favors i_n) = $p_{i_1} p_{i_2} \cdots p_{i_n}$

listed in a pre-specified order

Therefore P(the sample has x_1 people favoring candidate 1, x_2 people choosing candidate 2, $\ldots$, x_k people favoring candidate k) = $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$,

with $0 \le x_i \le n$ and $\sum_{i=1}^{k} x_i = n$. Thus (42)

$P($ exactly x_1 people favor Clinton (2), ... (4)

x_k people "favor" Undecided $) = \boxed{?}\, p_1^{x_1} \cdots p_k^{x_k}$,

where $\boxed{?}$ is the total # of different ways the order of the n people in the sample can be listed. But this $\boxed{?}$ is precisely

$$\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!},$$

the multinomial coefficient defined on p. (39) above. Thus

(7) ... (2)

$$P(\Sigma_1 = x_1, \ldots, \Sigma_n = x_n) = \frac{n!}{x_1!, \ldots, x_k!}\, p_1^{x_1} \cdots p_k^{x_k}.$$

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

How to work with OR when you have more than 2 events ($\cup_{i=1}^{n}$)

We already worked out that

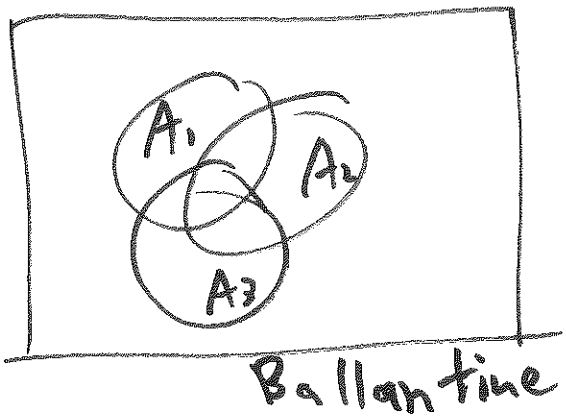
$$P(A_1 \text{ or } A_2) = P(A_1 \cup A_2)$$

(and)

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

We also know from Kolmogorov's 3rd Axiom that if events $A_1, \ldots, A_n$ 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 can probably see how this generalizes (or guess) for any events $A_1, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$
$$+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots +$$
$$(-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

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♠

(practically speaking)

Shuffle deck 2 until all 52! orderings are equally likely.

Now turn the first card of each deck over & do they match? Continue through all 52 cards;

$P(\text{at least one match}) = ?$

(let $n=52$)

Let A_i = (a match occurs on card i);

we want $P\left(\bigcup_{i=1}^{n} A_i\right)$, which can be computed with the complicated formula on the previous page.

Follow the logic detailed on DF pp. 49-50 to obtain

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

calculus result:

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

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

DF ch. 2

Conditional probability

Note that Kolmogorov's probability axioms defined the function $P_k(A)$, where A is a set in the

collection $\mathcal{C}$ 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 $\mathcal{C}$, the conditional probability of A given B is

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

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

were responding to the reality that

in practice, all probabilities are conditional on background (A)ssumptions, (I)nformation and (J)udgments (AIJ)

Example (Toy-Sachs)

we actually computed not

P(at least 1 T-S baby) but

P(at least 1 T-S baby | family of 5, mother and father both carriers (and))

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

Consequences of the conditional probability definition (theorems)

① A, B events in $\mathcal{C}$:

if $P(B) > 0$ then

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

and if $P(A) > 0$

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

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

then **chain rule for $\cap$ = and**

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

$\ldots P(A_n | A_1 \cap \ldots \cap A_{n-1})$

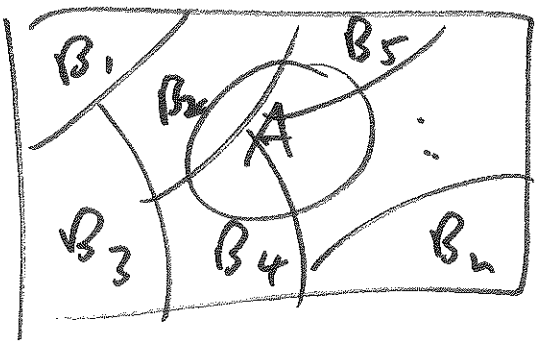
Recall previous Definition

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

such that the B_j are disjoint and (5)

exhaustive ($\bigcup_{i=1}^{n} B_i = S$), then you

have found a partition $(B_1, \ldots, B_k)$ 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

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

This is the **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 $\{B_1, \ldots, B_k\}$ with known $P(B_j)$ of S and ② A depends 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)

Extension of the LTP (4) Assuming all conditional probabilities are defined in what follows, if C is in the then

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

Definition Events A, B are independent iff ← (freq)

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

which (as long as $P(A)>0$, $P(B)>0$) is equivalent to ← (Bayesian)

$$\left.\begin{aligned} P(A|B) &= P(A) \\ \text{and } P(B|A) &= P(B) \end{aligned}\right\}.$$

Consequences of the definition of independence

① If A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c.

② Extension of the definition to more than 2 events:

Definition:

Given events $A_1, \dots, A_k$, they are (mutually) independent if, for every subset $A_{i_1}, \dots A_{i_j}$ of $(A_1, \dots, A_k)$ $(j = 2, \dots, k)$,

$$P(A_{i_1} \cap \dots \cap A_{i_j}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_j})$$

(Bayesian) Interpretation of independence

A, B independent iff information about A doesn't change the chances associated with B, and vice versa.

Definition Another useful extension of independence

Events $\{A_1, \ldots, A_k\}$ are conditionally independent given event B if for every subset $\{A_{i_1}, \ldots, A_{i_j}\}$ of $\{A_1, \ldots, A_k\}$ $(j = 2, \ldots, k)$

$$P(A_{i_1} \cap \cdots \cap A_{i_j} \mid B) = \prod_{\ell=1}^{j} P(A_{i_\ell} \mid B)$$

product

Statistical Example Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with $P(H) = \theta$, and θ can be set to any value in $[0,1]$ with a dial on the machine's control panel. Someone sets the dial to a θ value that's unknown to you and starts producing coin tosses $X_1, X_2, \ldots$

Suppose the first 10 tosses come out

1 0 1 1 1 0 0 1 1 1 ← "bits" (binary digits)

H T H H H T T H H H (7 H, 3 T).

Q: Is there information in these first 10 tosses that helps you to predict X_{11}?

A: Yes, definitely: it looks like θ is around $\frac{2}{10}$, so you would predict $X_{11} = H$.

Thus X_{11} depends on $X_1 \ldots X_{10}$ probabilistically.

Now, suppose instead that you watched the guy with the machine set the dial to $\theta = 0.81$, so that θ is now known to you. The next 10 tosses come out HHHTHTHHHH (8H, 2T). Q: Is there information in these 10 tosses that helps you to predict the next toss?

A: No; you know that $\theta = 0.81$, so there's no information in any of the X_i

that helps you to predict any of the other Y_j.

given θ, the Y_i are indep.

Thus the Y_i are unconditionally dependent but conditionally independent given θ.

(4 Aug 17)

Bayes's Theorem ~~for events~~ (a finite partition)

Suppose that the events $B_1, \ldots, B_k$ partition the sample space in such a way that $P(B_j) > 0$ for all $j = 1, \ldots, k$. If A is an event with $P(A) > 0$, then for each $i = 1, \ldots, k$

$$P(B_i \mid A) = \frac{P(B_i)\,P(A \mid B_i)}{P(A)}$$