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

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

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

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

(29 Jul 16)

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 <u>single</u> argument A.

To include the extremely useful idea of <u>conditional</u> probability in his setup, Kolmogorov has to <u>define</u> it using P_K.

<u>Definition</u> Given any two events A, B in $\mathbb{C}$, the conditional probability of A given B is

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

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 <u>2</u> 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

in practice, all probabilities are conditional on background Assumptions, Information and Judgments (AIJ) | Example | (Tay-Sachs)

we actually computed not

P(at least 1 T-S baby) but

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

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 $P_K(B|AIJ)$.

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

then

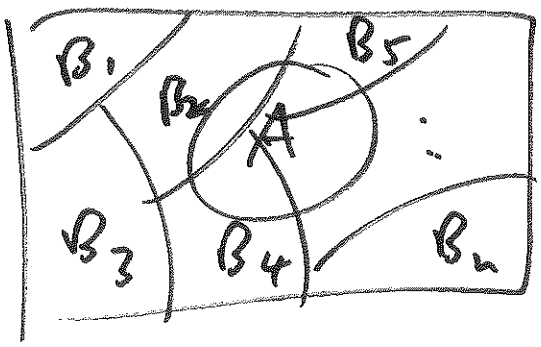
$\boxed{\text{chain rule for } \cap = \text{and}}$

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

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 <u>exhaustive</u> $\left(\bigcup_{i=1}^{n} B_i = S\right)$, then you have found a <u>partition</u> $(B_1, \ldots, B_k)$.



(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 $\{B_1, \ldots, B_k\}$ 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:

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

Extension of the LTP (4)

Assuming all conditional probabilities are defined in what follows, if C is in $\mathbb{C}$ then

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

Definition

Events A, B are <u>independent</u> if

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

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

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

and $P(B|A) = P(B)$.

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 <u>(mutually) independent</u> if, for every subset $A_{i_1}, \dots A_{i_j}$ of $(A_1, \dots, A_k)$ $(j = 2, \dots, k)$,

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

Interpretation (Bayesian) of independence

A, B independent $\overset{\text{iff}}{\longleftrightarrow}$ information about A doesn't change the chances associated with B, and vice versa.

Definition Another useful extension of independence

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

product

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

Example Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with $P(H) = \theta$, any one toss, 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 <u>unknown</u> to you and starts producing coin tosses $Y_1, Y_2, \ldots$ Suppose the first 10 tosses come out

HTHHHTTHHH (7 H, 3 T).

$\boxed{Q:}$ Is there information in these first 10 tosses that helps you to predict Y_{11}?

$\boxed{A:}$ Yes, definitely: it looks like θ is around $\frac{2}{10}$, so you would predict $I_{11} = H$. Thus I_{11} depends on $I_1 \ldots I_{10}$ probabilistically.

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

$\boxed{A:}$ <u>No</u>; you know that $\theta = 0.81$, so there's no information in any of the I_v

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.

Bayes's Theorem for events

Suppose that the events $B_1, \ldots, B_k$ partition the sample space (a finite partition) 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 $i = 1, \ldots, k$

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

and, by the LTP, this is

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

How this theorem is used in Bayesian statistics: The B_i represent unknown states of the world: they're all possible — $P(B_i) > 0$ — and only one of them is true, but you don't know which one. A represents <u>data</u>: information that will help you identify the most probable B_i.

Before (a priori) the dataset A arrives, you have background information about the plausibility of the B_i that you can represent with prior probabilities $P(B_i)$. After (a posteriori) the dataset A arrives, you can use Bayes's Theorem to update your prior probabilities to posterior probabilities $P(B_i \mid A)$.

The probabilities $P(A \mid B_i)$ represent how likely the dataset A would be if B_i were the actual unknown state; this is often called likelihood information.

$P(A)$ does not depend on the B_i, and can therefore be regarded as a <u>normalizing constant</u>, put into Bayes's Theorem to make all the $P(B_i|A)$ add up to 1. Thus

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

is interpreted as

$$\begin{pmatrix}\text{posterior}\\ \text{information}\end{pmatrix} = \frac{\begin{pmatrix}\text{prior}\\ \text{information}\end{pmatrix}\cdot\begin{pmatrix}\text{likelihood}\\ \text{information}\end{pmatrix}}{\begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix}}.$$

Random variables and their distributions

Example: Tay-Sachs Disease

T = T-S baby
N = not

$\mathcal{S}$

NNNNN	0
TNNNN	
NTNNN	
NNTNN	1
NNNTN	
NNNNT	
TTNNN	
TNTNN	
TNNTN	
TNNNT	
NTTNN	2
NTNTN	
NTNNT	
NNTTN	
NNTNT	
NNNTT	
⋮	⋮
TTTTT	5

← # of T-S babies = X

Given a sample space $\mathcal{S}$ for an experiment $\mathcal{E}$, a (real-valued) random variable (rv) is a function from the non-weird collection $\mathcal{C}$ of subsets of $\mathcal{S}$ to the real number line $\mathbb{R}$.

Definition

In the T-S case study, the elements s of $\mathcal{S}$ look like NNNTN and the rv X counts how many Ts they contain.

For instance, $I(TNNTN) = 2$ and $I(NNNTT) = (0)(50)/2$ (i.e., I ignores the <u>order</u> of the children).

We can use the following notation to simplify things.

Notation $P(\underbrace{I = y}_{\text{T/F proposition}}) \triangleq P(\underbrace{\{s : I(s) = y\}}_{\text{set}})$

For example, $P(I = 1) = P(\{s \in S : I(s) = 1\})$
$= P(\{TNNNN, NTNNN, NNTNN, NNNTN, NNNNT\})$.

In general the values a random variable takes on could be just about anything, but in this course all of our rvs will be <u>real-valued</u>

In the T-S case study the rv I can only take on the values $0, 1, \ldots, 5$.

y	$P(Y=y)$
0	0.237
1	0.396
2	0.264
3	0.088
4	0.015
5	0.001

You can see that a rv Y is completely specified by two things: the values it can take on, and the probabilities for those values.

(see p. 23)

Definition The (probability) distribution of a random variable Y is the collection of all probabilities of the form $P(Y \in A)$ for all sets A of real numbers in the non-weird collection $\mathcal{C}_{\mathbb{R}}$ of subsets of the real number line $\mathbb{R}$.

The rv Y in the T-S case study has a finite set of possible values —

this is true of some, but not all, rvs.

Definition A random variable X has a <u>discrete distribution</u>, or equivalently X is a <u>discrete</u> rv, if the set of possible (distinct) values of X is finite or at most countably infinite; rvs for which the set of possible values is uncountable are called <u>continuous random variables</u>

Examples ① The rv $X = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{otherwise} \end{cases}$ (with $X = \#$ T-s tossing) is discrete, taking on only the values $\{0, 1\}$ – such rvs are called <u>dichotomous</u> {yes, no} or <u>binary</u> {1, 0}.

② Imagine a scale for weighing things that has a dial you can set to specify how many significant figures (sigfigs) of precision you want. Buy a "1 pound" package of butter at your favorite market and weigh it!

possible weights (ounces)
16
16.0
15.99
15.993
15.9928
⋮

If there's no <u>conceptual</u> limit to the number of sigfigs you could get, a rv $Y = \left(\begin{array}{c}\text{the actual (true)} \\ \text{weight of the package}\end{array}\right)$ should be modeled as continuous, having values (e.g.) on $(0, \infty)$, the positive part of $\mathbb{R}$.

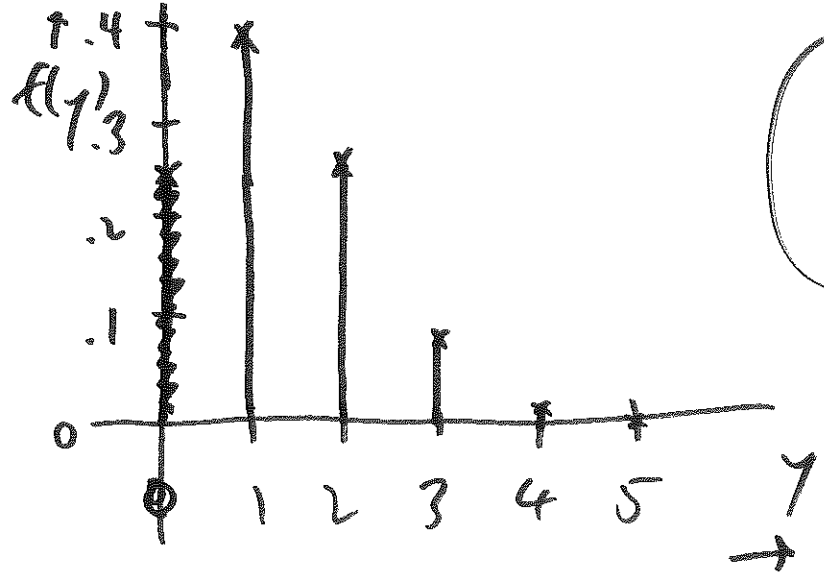
$\boxed{\text{Reality check:}}$ Infinite precision is impossible in practice;

every measurement you ever make is in actuality discrete, but it's useful to regard rvs that are <u>conceptually</u> continuous (i.e., no limit in principle to the precision of measurement) as continuous.

$\boxed{\text{Definition}}$ Given a <u>discrete</u> rv Y, the <u>probability</u> (mass) <u>function</u> (pmf or pf) of Y is the function f that keeps track of the probabilities associated with Y: $f(y) = P(Y = y)$.

The set $\{y : f(y) > 0\}$ is called the <u>support</u> of (the distribution of) Y.

(DS is almost unique in using pf, nearly everybody talks about the pmf.)



pf of I in the T-S case study

Definition

A rv I that only takes on the values {0, 1} — i.e., a binary rv — is said to have a Bernoulli distribution with parameter p — written Bernoulli(p) —

James Bernoulli Swiss (1655–1705)

if $f(y) = P(I = y) = \begin{cases} p & \text{for } y = 1 \\ 1-p & y = 0 \end{cases}$

Notation (I follows a Bernoulli(p) distribution) $\leftrightarrow$ I ~ Bernoulli(p)

is distributed as

Example In the powerball lottery (see homework 1 problem 2) 5 white numbered balls are drawn at random <u>with</u> replacement from a bin with balls numbered $\{1, 2, \ldots, 69\}$.

Let W_i = # on i^{th} drawn white ball.

Clearly $p(W_1 = w_1) = \begin{cases} \frac{1}{69} & w_1 = 1, 2, \ldots, 69 \\ 0 & \text{otherwise} \end{cases}$.

Less clearly (but true) $W_2, \ldots, W_5$ follow the same distribution if nothing is known about the previous draws.

Definition For any two integers $a \le b$, a rv Y that's equally likely to be any of the values $\{a, a+1, \ldots, b\}$ has the <u>uniform distribution</u> Uniform$\{a, b\}$. Evidently

its pf is $f(y) = P(Y=y) = \begin{cases} \frac{1}{b-a+1} & y = a, \dots, b \\ 0 & \text{else} \end{cases}$

$Y \sim \text{Uniform}\{a, b\} \leftrightarrow Y$ chosen at random from $\{a, a+1, \dots, b\}$.

Definition n random trials are performed (1 Aug 16), with each trial recorded as a success S or failure F. If each trial is independent of all the others and the chance p of success is constant across the trials, then Y = # of successes has the Binomial distribution

(not sample space)

$$\text{spf } f_Y(y) = P(Y=y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & \text{for } y = 0, 1, \dots, n \\ 0 & \text{else} \end{cases}$$

(with parameters n and p)

In shorthand $Y \sim \text{Binomial}(n, p)$.

Let $B_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \quad \ldots \quad \text{failure} \end{cases}$

for $i = 1, \ldots, n$; then under these assumptions

$B_i \overset{IID}{\sim} \text{Bernoulli}(p)$ and all the B_i are

independent. $\boxed{\text{Notation}}$ $X_i \overset{IID}{\sim} f_{X_i}(x_i)$

means that all of the rvs $X_1, X_2, \ldots$

are independent and identically distributed

draws from the distribution with pf

$f_{X_i}(x_i)$. Thus with the success/failure

trials, $B_i \overset{IID}{\sim} \text{Bernoulli}(p)$ and
$(i = 1, \ldots, n)$

$$Y = \sum_{i=1}^{n} B_i \sim \text{Binomial}(n, p).$$

This is our first example of the distribution of the <u>sum</u> of a bunch of IID rvs, a topic we'll examine in detail later.