STATISTICS & PROBABILITY

LECTURE 1

STATISTICS IS THE LOGIC OF UNCERTAINTY.

A **SAMPLE SPACE** S: THE SET OF ALL POSSIBLE OUTCOMES OF SOME EXPERIMENT (INTERPRET THAT AS ANYTHING W/ TRIALS, I THINK). AN **EVENT** A IS A SUBSET OF THE SAMPLE SPACE.

WARNING: THIS SUBJECT IS EXTREMELY ANTI-INTUITIVE (WHICH MAKES IT FUN. AND MEANS WE NEED TO DO LOTS OF PRACTICE)

THE NAIVE DEFINITION OF PROBABILITY: THE PROBABILITY OF EVENT A $(P(A)) = \dfrac{\#\text{ favourable outcomes}}{\#\text{ possible outcomes}}$

$\Downarrow$

$\|A\| / \|S\|$:)

USING NAIVE DEFINITION AND FLIPPING COIN TWICE, THERE ARE FOUR OUTCOMES HH, HT, TH, TT. SO $P(A_{HH}) = 1/4$. (OBVIOUSLY THIS ONLY WORKS ON AN EVEN DISTRIBUTION; ASSUMES ALL OUTCOMES EQUALLY LIKELY & FINITE SAMPLE SPACE).

NAIVE DEFINITION IS JUSTIFIED IN PROBLEMS W/ TRUE SYMMETRY.

WHAT'S THE PROBABILITY OF LIFE ON NEPTUNE?

$P(L_N) = \dfrac{\text{life}}{\text{life} + \text{no life}} = 0.5!$ SO YEAH DON'T USE THIS DEFINITION WHEN NOT JUSTIFIED :)

WE'LL NEED TO GO BEYOND NAIVE DEFINITION SOON, BUT FIRST WE'RE GOING TO DO SOMETHING IMPORTANT! COUNTING!

COUNTING: MULTIPLICATION RULE — if experiment w/ n_1 outcomes, and for each outcome of 1st exp there are n_2 outcomes for 2nd exp, ..., there are n_R outcomes for R^{th} exp. THEN $n_1 n_2 \cdots n_R$ overall possible outcomes.

ICE CREAM

cone choice (1 of 2)

flavour choice (1 of 3)

note that this is commutative.

also note that these trees grow v. fast.

BINOMIAL COEFFICIENT $\binom{n}{k} = \dfrac{n!}{(n-k)!\,k!}$ ① ; 0 if $k > n$

subset k in set size n; order agnostic

① $[n \cdot (n-1) \cdot (n-2) \ldots (n-k+1)] / k!$
← any order

EXAMPLE: PROBABILITY OF FULL HOUSE, 5-CARD DRAW

$\binom{52}{5}$ is $\|S\|$

HOW DO WE GET NUMERATOR FOR NAIVE CALC?

$13 \binom{4}{3} \qquad 12 \binom{4}{2}$

different possible cards → 13

$\binom{4}{3}$ ↑ need 3 of first card

12 ↑ possible cards after first used

$\therefore \quad P(A_{FH}) = 13\binom{4}{3}12\binom{4}{2} \Big/ \binom{52}{5}$

$\binom{n}{k}$ is a form of <u>sampling</u>

SAMPLING TABLE: CHOOSING k OBJECTS OUT OF n

	order matters	order doesn't
replace	n^k	①
don't replace	$n(n-1) \vdots n(n-k+1)$	$\binom{n}{k}$

FULL HOUSE EXAMPLE

① SURPRISINGLY SUBTLE.

BUT $\binom{n+k-1}{k}$

if you have $\|S\| = n$; $\|A\| = k$, we need to divide possible choices w/ $n-1$ dividers. → positions then $(k+n-1)$; k from this is $\binom{n+k-1}{k}$ ✓

LECTURE 2: STORY PROOFS, AXIOMS OF PROBABILITY.

HOMEWORK ADVICE: 1) DON'T LOSE YOUR COMMON SENSE

2) DO CHECK ANSWERS, ESPECIALLY SIMPLE AND EXTREME CASES.

3) LABEL PEOPLE, OBJECTS, ETC. IF WE HAVE n PEOPLE, LABEL 1...N :>

(REMEMBER TO DO HW #1 & 2 AFTER THIS)

4) IF WE HAVE 10 PEOPLE & WE WANT TO SPLIT INTO TEAM OF 6 & TEAM OF 4, THERE ARE $\binom{10}{4} = \binom{10}{6}$ WAYS.

2 TEAMS OF 5 DIFFERENT :> => $\binom{10}{5}/2$

LAST TIME WE LOOKED @ SAMPLING TABLE. LET'S LOOK AT REPLACE, ORDER DOESN'T MATTER:

PICK k TIMES FROM SET OF n OBJECTS.

SHOW $\binom{n+k-1}{k}$ ways

CHECK FIRST! EXTREME CASE $k=0 \Rightarrow \binom{n-1}{0} = 1$ $(0! = 1)$

$k=1 \Rightarrow \binom{n}{1} = n$

ANOTHER CASE $n=2$ (simplest non-trivial)

THEN $\binom{k+1}{k} \Rightarrow \binom{k+1}{1} = k+1$

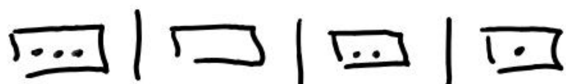
1 2

if we know items taken from 1 object we know items taken from other box

dots is $\{0, 1, \dots, k\} \Rightarrow k+1$

EQUIVALENT TO GENERAL QUESTION: HOW MANY WAYS ARE THERE TO PUT k INDISTINGUISHABLE PARTICLES INTO N DISTINGUISHABLE BOXES?

FOUR BOXES



$\hookrightarrow n=4;\ k=6$

note $n-1$ dividers :>

$$\underbrace{\bullet\ \bullet\ \bullet\ |\ |\ \bullet\ \bullet\ |\ \bullet}_{k+(n-1)}$$

$k+(n-1)$ items and we're done after choosing $k \Rightarrow \binom{k+(n-1)}{k} = \binom{k+n-1}{n-1}$

(BOSE PROPOSED MODEL ABOUT INDISTINGUISHABILITY IN PHYSICS, USED TO PREDICT BOSE-EINSTEIN CONDENSATE.)

COUNTING, CONTINUED.

STORY PROOF - PROOF BY INTERPRETATION

① $\binom{n}{k} = \binom{n}{n-k}$

② $n\binom{n-1}{k-1} = k\binom{n}{k}$

pick k people out of n; w/ one designated with a label. choose who's in the club, then choose the labelled person. or choose the labelled person and then pick the club

$n\binom{n-1}{k-1}$

③ $\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$ VANDERMONDE'S IDENTITY

NONTRIVIAL ALGEBRAICALLY. BUT STORY PROOF IS MORE STRAIGHTFORWARD 

PICK SOME NUMBER j FROM m & SOME $k-j$ FROM N

from $m \rightarrow \binom{m}{j}$

from $n \rightarrow \binom{n}{k-j}$

then add up for all j with the sum

GENERAL DEFINITION OF PROBABILITY: A PROBABILITY SPACE CONSISTS OF 2 INGREDIENTS S & P, WHERE S IS SAMPLE SPACE AND P IS A FUNCTION WHICH TAKES EVENT $A \subseteq S$ & MAPS TO $P(A) \in [0,1]$

P MUST SATISFY

① $P(\phi) = 0$; $P(S) = 1$ ← SUM OF POSSIBLE EVENTS HAS PROBABILITY 1

IMPOSSIBLE EVENTS HAVE PROBABILITY 0

② $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ if A_1, A_n are disjoint.

FROM THESE TWO RULES WE CAN DERIVE EVERY RESULT IN PROBABILITY!

PROBLEM SET #1

1. FOR EACH PART, DECIDE WHETHER BLANK SHOULD BE FILLED W/ =, <, or >, WITH A SHORT EXPLANATION

a) PROBABILITY THAT TOTAL AFTER ROLLING 4 FAIR DICE IS 21 __ TOTAL AFTER ROLLING 4 FAIR DICE IS 22

> — THERE ARE MORE WAYS OF GETTING 21 THAN 22 FROM 4 ROLLS

b) PROBABILITY OF RANDOM 2-LETTER WORD BEING A PALINDROME __ RANDOM 2-LETTER WORD BEING A PALINDROME

= (MIDDLE LETTER OF ODD # WORDS IRRELEVANT)

2. A RANDOM 5-CARD POKER HAND IS DEALT FROM A STANDARD DECK OF CARDS, FIND PROBABILITY OF EACH OF THE FOLLOWING IN TERMS OF BINOMIAL COEFFICIENTS

a) A NON-ROYAL FLUSH $\rightarrow \left[\binom{13}{5}-1\right] / \binom{52}{5}$ PER SUIT

misread this one, oops $\quad \therefore \left[4\binom{13}{5}-1\right] / \binom{52}{5}$

b) TWO PAIR & AN ACE $\rightarrow \left[12\binom{4}{2}+11\binom{4}{2}+4\right] / \binom{52}{5}$

3. a) HOW MANY PATHS ARE THERE FROM (0, 0) TO (110, 111) SUCH THAT EACH STEP CONSISTS OF GOING UP ONE UNIT TO THE RIGHT OR LEFT?

221 STEPS, EACH HAS BRANCHING CHOICE w/ 110 AS A MAX., SO $\binom{221}{110}$

b) HOW MANY PATHS FROM (0,0) TO (210, 211) w/ SAME STEP CONSTRAINTS PLUS GOES THROUGH (110, 111)?

EQUIV. TO $\binom{221}{110} \times \#(0,0) \rightarrow (100,100)$ SO $\binom{211}{110} \cdot \binom{200}{100}$

4. A "NOREPEATWORD" IS A SEQUENCE OF AT LEAST ONE (AND POSSIBLY ALL) OF THE USUAL 26 LETTERS, WITH REPETITIONS NOT ALLOWED. ORDER MATTERS i.e. "COURSE" != "SOURCE"

4 (cont). A NOREPEATWORD IS CHOSEN RANDOMLY, WITH ALL NOREPEATWORDS EQUALLY LIKELY. SHOW $P(w_{26}) \approx 1/e$

OUR NOREPEATWORD CAN BE ONE OF $26!$ CHOICES

GENERIC WORD W/ LENGTH k IS $\binom{26}{k}k!$, CHOOSING k LETTERS FROM 26 & ARRANGING RANDOMLY.

USING NAIVE DEFINITION,

$$\underset{26!}{\#\text{ FAV EVENTS}} \Big/ \underset{\sum_{k=1}^{26}\binom{26}{k}k!}{\#\text{ TOTAL EVENTS}}$$

SO
$$P(w_{26}) = 26! \Big/ \left[\sum_{k=1}^{26} \frac{26!}{k!(26-k)!}k!\right]$$

$$= \cancel{26!} \Big/ \left[\cancel{26!}\left(1/25! + 1/24! + \ldots + 1\right)\right]$$

$$= 1 \Big/ \left(\underbrace{1 + \ldots 1/24! + 1/25!}_{\text{TAYLOR EXPANSION FOR } e^x \text{ @ } x=1!}\right)$$

$$= 1/e \checkmark$$

5. GIVE A STORY PROOF THAT $\sum_{k=0}^{n}\binom{n}{k} = 2^n$

WE PICK A k-SIZED SUBSET FROM POPULATION n. WORK THROUGH IN-OR-OUT CHOICES — BINARY SO 2^n GROUPINGS

BUT WE KNOW THERE ARE $\sum_{k=0}^{n}$ choices of k, and WE HAVE $\binom{n}{k}$ ALREADY; TWO SIDES ARE EQUAL

6. GIVE A STORY PROOF THAT

$$\frac{(2n)!}{2^n \cdot n!} = (2n-1)(2n-3)\ldots 3 \cdot 1$$

LEFT SIDE: PAIRINGS IN POP SIZE $2n$ = $(2n)!$, divided by order of pairings & order within pairings

RIGHT SIDE: FIRST PERSON CAN PICK FROM $2n-1$

SECOND " $2n-3$

$n-1^{TH}$ " 3

N^{TH} " 1

BOTH DESCRIBE PAIRINGS IN POP $2n$

7. SHOW THAT FOR ALL POSITIVE INTEGERS n & k w/ $n \geq k$,

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}$$

a) ALGEBRAICALLY: $\dfrac{n!}{k!(n-k)!} + \dfrac{n!}{(k-1)!(n-k+1)!}$

$$= \frac{n!}{k!(n-k)!} + \frac{kn!}{k!(n-k+1)!}$$

$$= \frac{(n-k+1)n! + kn!}{k!(n-k+1)!} = \frac{n!(n+1)}{k!(n-k+1)!}$$

7. a. (cont) $\dfrac{n!(n+1)}{k!(n-k+1)} \Rightarrow \dfrac{(n+1)!}{k!((n+1)-k)!} \Rightarrow \binom{n+1}{k}$

b. VIA STORY PROOF, IMAGINING $n+1$ PEOPLE, WITH ONE DESIGNATED "PRESIDENT"

RIGHT SIDE : K PEOPLE CHOSEN OUT OF $n+1$ PEOPLE

LEFT SIDE : THERE'S ONE PERSON DESIGNATED PRESIDENT. OUR GROUPING OPTIONS ARE TO INCLUDE THEM OR NOT.

w/ PRESIDENT : $\binom{n}{k-1}$

w/o " : $\binom{n}{k}$

THEY SUM TO EQUAL $\binom{n+1}{k}$

HOMEWORK #1

1. A CERTAIN FAMILY HAS 6 CHILDREN, 3 BOYS & 3 GIRLS. ASSUMING EVEN DISTRIBUTION, WHAT IS THE PROBABILITY THE ELDEST 3 CHILDREN ARE GIRLS?

$\|A\| / \|S\| \rightarrow \|A\| = (3!)^2; \ \|S\| = 6! \quad P(A) = (3!)^2 / 6!$

2. a) HOW MANY WAYS ARE THERE TO SPLIT 12 PEOPLE INTO THREE TEAMS, WHERE ONE TEAM HAS TWO PEOPLE & THE OTHERS HAVE FIVE EACH?

$$\binom{12}{2} \cdot \binom{10}{5}/2$$

TAKE TEAM OF TWO FIRST

THEN TAKE A TEAM OF FIVE. REMEMBER THAT THIS COULD BE ONE OF TWO TEAMS!

b) 12 INTO THREE TEAMS AGAIN, BUT EACH TEAM HAS 4 PEOPLE

$$\binom{12}{4}/3 \cdot \binom{8}{4}/2$$

3. A COLLEGE HAS 10 TIME SLOTS FOR COURSES, AND ASSIGNS COURSES TO TIME SLOTS RANDOMLY. IF A STUDENT RANDOMLY CHOOSES THREE CLASSES TO ENROLL IN, WHAT IS THE CHANCE THERE IS A CONFLICT IN THEIR SCHEDULE?

$$1 - P(A) = \frac{n!}{k!(n-k)!} \Rightarrow P(A) = \frac{10!}{3!\,\cancel{7!}} \Rightarrow \frac{10 \cdot 9 \cdot 8}{3!}$$

4. CITY WITH 6 DISTRICTS HAS 6 ROBBERIES IN A WEEK. TAKE ROBBERIES TO BE RANDOM & INDEPENDENT

WHAT IS THE PROBABILITY SOME DISTRICT HAD MORE THAN ONE ROBBERY?

4. (cont) 6^6 configurations.

$P(A) = 1 - P(A')$; $A' \Rightarrow$ one robbery per district.

$\Downarrow 6!/6^6$

$P(A) = 1 - 6!/6^6$

5. ELK DWELL IN A CERTAIN FOREST. THERE ARE N ELK, OF WHICH A SIMPLE RANDOM SAMPLE OF SIZE n ARE CAPTURED AND TAGGED. THE CAPTURED ELK ARE RETURNED, AND A NEW SAMPLE OF SIZE m IS DRAWN.

WHAT IS THE POSSIBILITY THAT EXACTLY k ELK IN GROUP m HAVE BEEN PREVIOUSLY TAGGED?

% TAGGED: $\binom{N}{n}$

||s||

% IN NEW SAMPLE: $\binom{N}{m}$

WHAT IS ||A||? $\binom{n}{k} \cdot \binom{N-n}{m-k}$ untagged

k tagged

$\therefore \binom{n}{k}\binom{N-n}{m-k} \Big/ \binom{N}{m}$

6. A JAR CONTAINS r RED BALLS & g GREEN BALLS, WHERE r & g ARE FIXED POSITIVE INTEGERS. A BALL IS DRAWN FROM THE JAR RANDOMLY. A SECOND IS ALSO DRAWN RANDOMLY.

6 (CONT) a) EXPLAIN WHY $P(B_2 G) = P(B_1 G)$

$$P(B_1 G) = g/(g+r)$$

$$P(B_2 G) = \left(1 - g/(g+r)\right)^{①} \cdot g/(g+r-1)) + \left(g/(g+r)\right)^{②} \cdot \left(g-1/(g+r-1)\right)$$

THESE HAVE TO SUM TO $P(B_1 G)$; WE ONLY GET TO BRANCH ONCE FIRST BALL IS PICKED

b). THIS IS JUST REDUCING I THINK?

①

$$\left(1 - \frac{g}{g+r}\right)\left(\frac{g}{g+r-1}\right)$$

$$\left(\frac{r}{g+r}\right)\left(\frac{g}{g+r-1}\right) \Rightarrow \frac{rg}{(g+r)(g+r-1)}$$

②

$$\left(\frac{g}{g+r}\right)\left(\frac{g-1}{g+r-1}\right) \Rightarrow \frac{g^2-g}{(g+r)(g+r-1)}$$

$$① + ② = \frac{g\cancel{(g+r-1)}}{(g+r)\cancel{(g+r-1)}} = g/g+r \quad ✓$$

c) SUPPOSE 16 BALLS TOTAL. ALSO SUPPOSE PROBABILITY OF TWO BALLS BEING THE SAME COLOUR IS THE SAME AS THE PROBABILITY THEY ARE DIFFERENT COLOURS. LIST ALL VALUES OF g & r. $P(A) = P(A) - 1 \Rightarrow P(A) = 0.5$

$$0.5 = 2gr/(g+r)(g+r-1) \Rightarrow \overset{16}{g+r} = (g-r)^2$$

$|g-r| = 4$; $g = 10|6$; $r = 10|6$ (OBV WHEN $g=10$ $r=6$)

7. a) SHOW USING A STORY PROOF THAT

$$\binom{k}{k} + \binom{k+1}{k} + \binom{k+2}{k} + \dots + \binom{n}{k} = \binom{n+1}{k+1}$$

WHERE n & k ARE POSITIVE INTEGERS WITH $n \geq k$

<u>HINT</u>: IMAGINE ARRANGING A GROUP OF PEOPLE BY AGE, THEN THINK ABOUT THE OLDEST PERSON IN A CERTAIN SUBGROUP.

RIGHT SIDE: WE PICK $k+1$ PEOPLE OUT OF $n+1$ OPTIONS. WE TAKE THE OLDEST PERSON IN THAT SUBGROUP. WHAT IS THEIR POSITION RELATIVE TO THE FULL GROUP $n+1$?

IF THEY ARE THE OLDEST IN FULL SET, REST OF SUBGROUP IS $\binom{n}{k}$. OR, IF 2^{nd}-OLDEST, $\binom{n-1}{k}$. OR, ETC, ..., ALL THE WAY DOWN TO SECOND <u>YOUNGEST</u>, WHERE THERE IS ONLY $\binom{k}{k}$ CHOICE.

$$\sum_{i=k}^{n} \binom{i}{k} = \binom{n+1}{k+1}$$

GOOD TAKEAWAY IDENTITY

7. (cont) b) SUPPOSE A LARGE PACK OF HARIBO GUMMY BEARS CAN HAVE $30 \leq N \leq 50$. THERE ARE FIVE COLOURS: CLEAR, RED, ORANGE, GREEN, AND YELLOW. HOW MANY DIFFERENT POSSIBILITIES ARE THERE FOR THE COMPOSITION OF A PACK? DO NOT LEAVE YOUR ANSWER AS A SUM!

NARROWING DOWN... REPLACEMENT, ORDER-AGNOSTIC, SO COUNT IS... $\binom{k+n-1}{n}$

$$\Downarrow$$

$$\binom{n+4}{n} \Rightarrow \binom{n+4}{4}$$

SO OUR SUM IS $\sum_{n=30}^{50} \binom{n+4}{4} \Rightarrow \sum_{n=34}^{54} \binom{n}{4}$

$$\sum_{n=34}^{54} \binom{n}{4} = \sum_{n=4}^{54} \binom{n}{4} - \sum_{n=4}^{33} \binom{n}{4} \; ; \quad \text{apply } 7(a) \ldots$$

$\sum \binom{n}{k} = \binom{n+1}{k+1} \Rightarrow \binom{55}{5} - \binom{34}{5}$ COMBINATIONS

LECTURE 3

THE BIRTHDAY PROBLEM

YOU HAVE A GROUP OF PEOPLE. HOW LIKELY IS IT THAT AT LEAST ONE PAIR SHARE A BIRTHDAY?

k people, find prob that 2 have same birthday?

ASSUME 365 BIRTHDAYS POSSIBLE

ASSUME BIRTHDAYS ARE EVENLY DISTRIBUTED*

ASSUME INDEPENDENCE OF BIRTHS

*NOT REALLY TRUE BUT TRUE ENOUGH

if $k > 365$; $P(A) = 1$ by inspection (pidgeonhole principle)

$\sqcup \sqcup \ldots \sqcup$

1/1 1/2 12/31

NB: THIS PROBLEM IS IMPORTANT IN COMPUTER SCIENCE; MEMORY IS IN PHYSICAL CHUNKS & ONE NEEDS TO AVOID COLLISIONS

complement

if $k \leq 365$ $\quad P(\text{no match}) = \dfrac{365 \cdot 364 \ldots (365-k+1)}{365^k}$

$\Downarrow$

$P(\text{match}) = 50.7\% \text{ @ } k=23$

$97\% \text{ @ } k=50$

$>99.99\% \text{ @ } k=100$

WHY IS THIS? BECAUSE WE ARE NOT LOCKED INTO SPECIFIC DAYS. $\binom{k}{2} = \dfrac{k(k-1)}{2}$, $\binom{23}{2} = \dfrac{23 \cdot 22}{2} = 253$ pairs in 23 ppl

BACK TO NON-NAIVE DEFINITION OF PROBABILITY

AXIOMS

① $P(\emptyset) = 0$; $P(S) = 1$ ← SCALING

② $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$ IF $A_1, A_2 \ldots A_n$ DISJOINT

IMAGINE PROBABILITY AS AREA IN SAMPLE SPACE FOR NOW. ② BECOMES OBVIOUS

SOME SIMPLE CONSEQUENCES OF THESE AXIOMS

① $P(A^c) = 1 - P(A)$ FROM $P(S) = P(A + A^c)$

② IF $(A \subseteq B)$; $P(A) \leq P(B)$ FROM $P(B) = P(A) + P(B \cap A^c)$

③ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

"disjointification"

FROM $P(A \cup B) = P(A \cup (B \cap A^c))$

$= P(A) + P(B \cap A^c)$

$P(A) + P(B) - P(A \cap B) \overset{?}{=}$

$\Downarrow$

equiv $P(A \cap B) + P(B \cap A^c) = P(B)$ ✓

$A \cap B$ & $A^c \cap B$ disjoint; $\cup$ is B

EXTEND TO > 2 EVENTS

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap C) - P(A \cap B) - P(B \cap C)$
$+ P(A \cap B \cap C)$

PROOF SIMILAR TO ③; JUST TEDIOUS (INDUCTIVE)

GENERAL FORM OF INCLUSION/EXCLUSION OF N EVENTS

$$P(A_1 \cup A_2 \cup \ldots A_N) = \sum_{j=1}^{N} P(A_j) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap k)$$

$$- \ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots A_N)$$

EXAMPLE OF IN/EX APPLICATION

deMONTMORT'S PROBLEM (1713), MATCHING PROBLEM

DECK OF CARDS LABELLED 1->N

GAME: SHUFFLE CARDS, COUNT "1 → N"

WIN IF CARD MATCHES COUNT; e.g. 7^{TH} CARD IS "7"

WHAT IS PROBABILITY OF WINNING?

LET A_j BE THE EVENT THAT J^{TH} CARD MATCHES

$$P(A_1 \cup A_2 \cup \ldots A_N) = n \cdot \frac{1}{n} - \frac{n(n-1)}{2!}\left(\frac{(n-2)!}{n!}\right) + \frac{n(n-1)(n-2)}{3!} \cdot \left(\frac{(n-3)!}{n!}\right) + \ldots$$

$P(A_j) = 1/n$ since all positions equally likely

$$P(A_1 \cap A_2) = \frac{(n-2)!}{n!}$$

$$\vdots$$

$$P(A_1 \cap \ldots A_k) = \frac{(n-k)!}{n!}$$

SMELLS LIKE A TAYLOR SERIES TO ME...

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots + (-1)^{n+1}\frac{1}{n!}$$ YEP :)

$$\approx \underline{\underline{1 - 1/e}}$$

LECTURE 4: INDEPENDENCE & CONDITIONAL PROBABILITY

DEMONTMORT / MATCHING CONTINUED.

$1 - 1/e \rightarrow$ really interesting thing here is that we're independent of n @ reasonable n

($n=3 \sim 0.667$; $n=4 = 0.625$, $n=7 \sim 1 - 1/e$)

NOTE on MATCHING - we use first k & use symmetry

$$P(A_1 \cap A_2 \cap \ldots A_k) = \frac{(n-k)!}{n!}$$, there are $\binom{n}{k}$ terms $= \frac{n!}{k!(n-k)!}$

all the same by symmetry

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

$$P(\text{no-match}) = P\left(\bigcap_{j=1}^{n} A_j^c\right) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{n} \frac{1}{n!}$$

$$= 1/e$$

COMMON TAYLOR SERIES: GEOMETRIC & e^x
(IN PROBABILITY)

INDEPENDENCE: EVENTS A & B ARE INDEPENENT IF

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

note: THIS IN <u>NOT</u> THE SAME AS DISJOINT! DISJOINT PROBABILITIES SAY A <u>PRECLUDES</u> B

A, B, C ARE INDEPENDENT if $P(A \cap B) = P(A)P(B) \cap P(A \cap C) = P(A)P(C)$

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

(needs parallel ind. & all 3 at once) $P(A \cap B \cap C) = P(A)P(B)P(C)$

SIMILARLY FOR EVENTS $A_1, \dots, A_N$

any 2 have to be independent, then any 3, etc

"independent means multiply" ← (a paraphrase of definition)

EXAMPLE: NEWTON-PEPYS PROBLEM

HAVE 6-SIDED FAIR DICE. WHICH OF THESE IS MOST LIKELY:

TRUTH → (A) AT LEAST ONE DICE GETS A SIX w/ SIX ROLLS

(B) " TWO w/ 12 "

PEPYS'S INTUITION → (C) " THREE w/ 18 "

$$P(A) = 1 - \left(5/6\right)^6 \quad \text{[complement; independent]}$$

$$\approx 2/3$$

$$P(B) = 1 - \left(5/6\right)^{12} - 12\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{11}$$

$$\approx 0.619$$

$$P(C) = 1 - \underbrace{\sum_{k=0}^{2} \binom{18}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{18-k}}_{\text{BINOMIAL PROBABILITY}}$$

$$\square\ \square \cdots \square\ \square$$
1 2 17 18

$$\approx 0.597$$

NEWTON'S CALC WAS RIGHT BUT HIS ARGUMENT WAS WRONG & CONFUSING :) YOU CAN TELL B/C HIS ARGUMENT IS INVARIANT IF YOU USE UNFAIR DICE

CONDITIONAL PROBABILITY

HOW SHOULD UPDATE YOUR BELIEFS WHEN YOU RECIEVE NEW EVIDENCE?

"CONDITIONING IS THE SOUL OF STATISTICS"

DEFINITION: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, if $P(B) > 0$

A given B

INTUITION #1: PEBBLEWORLD

S

A

EVENT IS A SUBSET "B"

ASSUME FINITE POSSIBLE OUTCOMES, EACH REPRESENTED BY A PEBBLE. TOTAL MASS OF PEBBLES = 1

$P(A|B)$: B occurred, so get rid of pebbles in B^c

our new universe is B

then renormalise so total mass of pebbles is one again

INTUITION #2: FREQUENTISTWORLD repeat many times

1001 01101

00100111

⋮

ONE INTERPRETATION OF PROBABILITY IS LONG-RUN FREQUENCY. SO KEEP LIST OF RESULTS

CIRCLE REPS w/ B

AMONG THOSE, WHICH FRACTION OF TIME DID A OCCUR?

THEOREM 1

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A) \quad ①$$

THEOREM 2 ($n!$ THEOREMS)

$$P(A_1, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1, \dots, A_{n-1}) \quad ②$$

THEOREM 3

$$P(A|B) = P(B|A)P(A) / P(B) \quad \text{BAYES'S RULE}$$

① SAYS PROBABILITY OF A & B OCCURING (INTERSECTION) IS $P(B) \cdot P(A|B)$ & $P(A) \cdot P(B|A)$; JUST COMMUTATIVE MULTIPLICATION GAMES. IF A & B INDEPENDENT, CONDITIONING ON B DOES NOTHING

② DO ① LOTS :) THIS IS THE CHAIN RULE OF CONDITIONAL PROBABILITY.

LECTURE 5: CONDITIONAL PROBABILITY, CONTINUED

THINKING: PROBABILITY IS HOW TO THINK ABOUT UNCERTAINTY

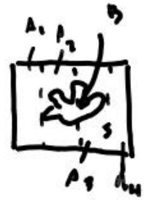
HOW DO WE THINK CONDITIONALLY? "THINKING CONDITIONALLY IS A CONDITION FOR THINKING."

HOW TO SOLVE A PROBLEM

1) TRY SIMPLE & EXTREME CASES

2) DECOMPOSE THE PROBLEM INTO SMALLER ONES

2) cont.



BREAK B INTO PIECES

LET $A_1, \dots, A_n$ PARTITION S

THEN $P(B) = P(B \cap A_1) + \dots + P(B \cap A_n)$

$= P(B|A_1)P(A_1) + \dots + P(B_n|A_n)P(A_n)$ ①

① THE LAW OF TOTAL PROBABILITY

UTILITY DEPENDENT ON CHOICE OF PARTITION

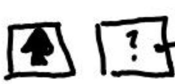
EXAMPLE: SUPPOSE WE TAKE TWO RANDOM CARDS FROM STD DECK.

FIND PROB. BOTH CARDS ARE ACES ② GIVEN ONE IS AN ACE. FIND PROB BOTH CARDS ③ ARE ACES GIVEN THAT WE HAVE THE ACE OF SPADES.

REDUNDANT CONDITIONAL

② $\dfrac{P(\text{both aces} | \cancel{\text{ace}})}{P(\text{ace})} = \dfrac{\binom{4}{2} / \binom{52}{2}}{1 - \binom{48}{2} / \binom{52}{2}} \approx 1/33$

③ $P(\text{both aces} \mid \text{have ace} \spadesuit) = 3/51 = 1/17$



BY SYMMETRY, IMMEDIATELY OBVIOUS THIS IS EQUALLY LIKELY TO BE ANY OF 51 CARDS

WHY IS ② DIFFERENT TO ③?. WHAT INFORMATION DOES THE SUIT OF CARDS REALLY GIVE YOU?.

IN ③ WE KNOW WE HAVE A SPECIFIC CARD. IN ② WE KNOW WE HAVE AT LEAST ONE OF A GROUP OF CARDS. THIS IS A RESTRICTION ON HANDS w/ ONLY ONE ACE!

EXAMPLE: DISEASE

PATIENT IS TESTED FOR A DISEASE AFFLICTING 1% OF SAMPLE POPULATION. PATIENT TEST POSITIVE. SUPPOSE TEST ADVERTISED AS BEING "95% ACCURATE." SUPPOSE THAT MEANS...

D: patient has disease

T: patient tests positive

$P(T|D) = 0.95 = P(T^c|D^c)$

$P(D|T)$ is what patient actually cares about.

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T^c|D^c)P(D^c)}$$

$1/(1+x)$ HMMM...

DENOMINATOR OF BAYES' RULE OFTEN TRICKY; WE'LL USE LAW OF TOTAL PROB. HERE OFTEN

$\approx \boxed{0.16}$

THERE IS A COMPETITION BETWEEN RARITY OF DISEASE AND RARITY OF TEST INACCURACY

WE'RE PSCHOLOGICALLY PRIMED TO AVOID FALSE POSITIVES!

BAYES' RULE IS COHERENT. UPDATE IN ANY ORDER, BREAK UPDATES INTO CHUNKS, ETC.

COMMON MISTAKES

1. CONFUSING $P(A|B)$, $P(B|A)$ "prosecutor's fallacy"

   EX. SALLY CLARK CASE. TWO SIDS CASES, CONVICTED OF MURDER. EXPERT CLAIMED "1/8500 CHANCE OF RANDOM DEATH; $1/8500^2$ OF TWO". EVEN ASSUMING INDEPENDENCE WE WANT $P(\text{innocence} \mid \text{evidence})$, not $P(\text{evidence} \mid \text{innocence})$

2. CONFUSING $P(A)$ "PRIOR" WITH $P(A|B)$ "POSTERIOR"

   $P(A|A) = 1$. BUT $P(A)$ GIVEN <u>A != 1</u>!!

3. CONFUSING INDEPENDENCE WITH CONDITIONAL INDEPENDENCE.

CONDITIONAL INDEPENDENCE

EVENTS A, B ARE CONDITIONALLY INDEPENDENT GIVEN C IF $P(A \cap B|C) = P(A|C)P(B|C)$. DOES <u>NOT</u> IMPLY UNCOND. IND.

EXAMPLE: CHESS OPPONENT OF RANDOM STRENGTH

- GAMES INDEPENDENT INDIVIDUALLY
- BUT THEY GIVE YOU EVIDENCE OF PLAYER STRENGTH

game outcomes conditionally independent given strength of opponent but not unconditionally independent

DOES INDEPENDENCE IMPLY CONDITIONAL INDEPENDENCE?

NO. WHY? SUPPOSE FIRE ALARM GOES OFF. CAUSED BY F (fire) OR C (popcorn). $P(F \cap C) = P(F)P(C)$

BUT $P(F \mid A \cap C^c) = 1$

← NOT CONDITIONALLY INDEPENDENT; NO C GIVEN A MEANS B!

LECTURE 6 CONDITIONAL PROBABILITY SOME MORE

MONTY HALL! THREE DOORS

1 DOOR HAS CAR
2 DOOR HAVE GOATS

CONTESTANT CHOOSES DOOR w/o KNOWLEDGE OF CONTENTS

MONTY KNOWS WHICH DOOR HAS A CAR

CONTESTANT DOESN'T WANT GOATS

ASSUME DOOR 1 CHOSEN

MONTY OPENS EITHER DOOR $\underbrace{\text{2 or 3, REVEALING GOAT.}}_{}$ always open door w/ goat ②

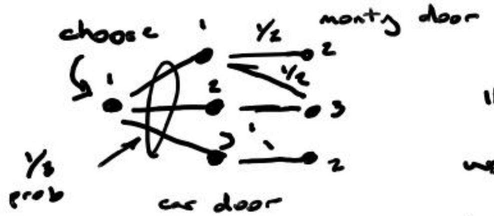
CONTESTANT HAS OPTION TO SWITCH.

① IF MONTY HAS A CHOICE OF WHICH DOOR TO OPEN, HE DOES W/ EQUAL PROBABILITY.

SHOULD CONTESTANT SWITCH? YES! 2/3 vs. 1/3

MANY WAYS TO LOOK AT IT. IF MONTY OPENS DOOR 2, WE KNOW DOOR 2 HAS A GOAT, <u>AND</u> MONTY OPENED DOOR 2.

APPROACH #1: TREE DIAGRAM



IF MONTY OPENS DOOR 2, WE DELETE EVERY BRANCH INCONSISTENT WITH THAT.

SO TOP BRANCH IS $1/6$, BOTTOM $1/3$, RENORMALISING GIVES $2/3$ SUCCESS IF SWITCHING; $1/3$ IF NOT

INTUITIVELY: $1/3^{rd}$ of the time your guess is good. $2/3^{rds}$ is wrong.

APPROACH #2: Law of total probability

we wish we knew where the car is

S: succeed (assuming switch)

D_j: door D_j has car $(j \in \{1,2,3\})$

$$P(S) = P(S|D_1)\underbrace{P(D_1)}_{1/3} + P(S|D_2)\underbrace{P(D_2)}_{1/3} + P(S|D_3)\underbrace{P(D_3)}_{1/3}$$

$$= 0 + 1\cdot\tfrac{1}{3} + 1\cdot\tfrac{1}{3} = 2/3$$

$P(S)$ is unconditional probability that strategy is successful. by symmetry $P(S|\text{monty opens } 2) = 2/3$. if monty doesn't like to open doors equally, conditional changes

APPROACH #3: EXTREME CASE

1 MILLION DOORS: MONTY OPENS 999,998. YOU'RE CONFIDENT YOUR FIRST GUESS IS WRONG :)

SIMPSON'S PARADOX

IS IT POSSIBLE TO HAVE TWO DOCTORS WHERE 1ST IS MORE SUCCESSFUL AT EVERY TYPE OF SURGERY YET 2nd DOCTOR HAS HIGHER SUCCESS RATE? (OBVIOUSLY YES LOL)

DOCTOR HIBBERT VS DOCTOR NICK

	♥	BAND AID REMOVAL	
✓	70	10	≈ 80%
×	20	0	

	♥	BAND AID REMOVAL	
✓	2	81	≈ 83%
×	8	9	

WE WANT TO CONDITION ON TYPE OF SURGERY

ANOTHER EX.

TWO BASEBALL PLAYERS. 1 PLAYER HAS HIGHER AVG IN 1ST HALF AND 2ND HALF, BUT SECOND PLAYER HAS HIGHER OVERALL AVERAGE. HOW? .323 vs .300, .100 vs .50 ← BUT INJURED SECOND HALF

$\underbrace{}_{.250 \text{ vs } .290}$

MAPPING ONTO CONDITIONAL PROB

A: successful surgery

B: treated by Dr. Nick (B^c is Hibbert)

C: heart surgery

$$P(A \mid B, C) < P(A \mid B^c, C)$$
$$P(A \mid B, C^c) < P(A \mid B^c, C^c)$$

but $P(A \mid B) > P(A \mid B^c)$ C is a "confounder" and we need to control for it

LET'S TRY TO PROVE THAT $P(A \mid B)$ MUST BE LESS THAN $P(A \mid B^c)$ AND SEE WHAT GOES WRONG.

LOTP but conditional

$$P(A \mid B) = \underbrace{P(A \mid B, C)}_{< P(A \mid B^c, C)} \overset{①}{P(C \mid B)} + \underbrace{P(A \mid B, C^c)}_{< P(A \mid B^c, C^c)} \overset{①}{P(C^c \mid B)}$$

① "WEIGHTS" – we have no way of knowing/reducing in advance

ANOTHER EXAMPLE: LAWSUIT AGAINST UC BERKLEY – grad adm. men > women

BUT BERKLEY ADMITTED MORE WOMEN THAN AVG ACROSS EVERY DEPARTMENT

ANOTHER EXAMPLE:

 FOUR JARS WITH JELLY BEANS

 TWO FLAVOURS OF "

SUPPOSE THE PAIRS OF JELLY BEAN JARS ON LEFT ARE EACH "BETTER" THAN COUNTERPART ON RIGHT. WHAT HAPPENS IF YOU MERGE TWO BETTER JARS & TWO WORSE JARS?

EXTREME CASE:

$$\left.\begin{array}{l} J_{L1} = 2 \text{ vs } 1 \\ J_{L2} = 1 \text{ vs } 1E^{6} \end{array}\right\} J_L = 3 \text{ vs } 1{,}000{,}001$$

$$\left.\begin{array}{l} J_{R1} = 1 \text{ vs } 2 \\ J_{R2} = 0 \text{ vs } 1 \end{array}\right\} J_R = 1 \text{ vs } 3$$

IS IT POSSIBLE TO DO THIS UNDER THE CONSTRAINT THAT EACH PAIR OF JARS HAS SAME # OF BEANS? <u>ABSOLUTELY NOT</u> — SIMPSON'S PARADOX IS JUST "DIFFERENT DENOMINATORS EXIST" :)

PROBLEM SET #2

1-1. FOR A GROUP OF 7 PEOPLE, FIND THE PROBABILITY THAT ALL FOUR SEASONS OCCUR AT LEAST ONCE AMONG THEIR BIRTHDAYS, ASSUMING ALL SEASONS EQUALLY LIKELY.

at least 1

___ ___ ___ ___

sp su at wt

between 1 & 5 between 1 & 4 between 1 & 3

LET A_n BE PROBABILITY OF NO BIRTHDAY IN N^{TH} SEASON

THEN USE COMPLEMENT.
$$\sum_{n=1}^{4} P(A_n) - \sum_{n=1}^{3}\sum_{m>n} P(A_n \cap A_m) + \sum_{n=1}^{3}\sum_{m>n}\sum_{l>m} P(A_n \cap A_m \cap A_l)$$

$$P(A_n) = (3/4)^7 \qquad \therefore \qquad 4P(A_1) - 6P(A_1 \cap A_2) + 4P(A_1 \cap A_2 \cap A_3)$$
$$= 4(3/4)^7 - 6\left(\frac{1}{2^7}\right) + 4\left(\frac{1}{4}\right)^7$$
$$\approx 49\%$$

$1 - 49\% \approx 51\%$ ✓

1-2. ALICE CHOOSES BETWEEN 30 NON-OVELAPPING CLASSES, EACH MEETING ONCE A WEEK, WITH 6 POSSIBLE CLASSES ON EACH WEEKDAY. IF SHE PICKS 7 CLASSES AT RANDOM, WHAT ARE THE CHANCES SHE HAS CLASS EVERY DAY?

1-2 (cont): WE KNOW THAT IF ALICE HAS 1 DAY w/ 3 ① CLASSES OR 2 ② DAYS w/ 2 EACH THE CONDITION ISN'T MET.

① $\binom{5}{1}\binom{6}{3}6^4$

② $\binom{5}{2}\binom{6}{2}^2 6^3$

$$\frac{① + ②}{\binom{30}{7}}$$

2-1 IS IT POSSIBLE FOR AN EVENT TO BE INDEPENDENT OF ITSELF? IF SO, WHEN?

def: $P(A,B) = P(A)P(B)$

if $P(A)=0$ or $1 \Rightarrow P(A,B) = 0|1 = P(A)P(B)$?

2.2 IS IT ALWAYS TRUE THAT IF A & B ARE INDEPENDENT EVENTS THAT A^c & B^c ARE INDEPENDANT?

$P(A,B) = P(A)P(B)$

$P(A^c,B^c) = (1-P(A))(1-P(B))$

$= 1 - P(A) - P(B) + P(A)P(B)$

$= 1 - P(A) - P(B) + P(A,B)$

$= 1 - P(A \cup B)$

$= P(A^c, B^c)$

2-3) GIVE AN EXAMPLE OF 3 EVENTS A, B, C WHICH ARE PAIRWISE INDEPENDENT BUT NOT MUTUALLY INDEPENDENT.

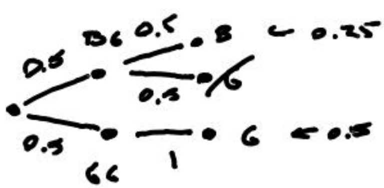
LOGICAL AND OPERATOR: C is known if & only if A & B known

2-4) GIVE AN EXAMPLE OF 3 EVENTS A, B, C WHICH ARE NOT INDEPENDENT YET SATISFY MUTUAL INDEPENDENCE

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

if $P(C) = 0$, mutual independence always satisfied even if A & B dependent

3-1) A BAG CONTAINS 1 MARBLE WHICH IS EITHER GREEN OR BLUE W/ EQUAL PROBABILITIES. A GREEN MARBLE IS PLACED IN THE BAG. A RANDOM MARBLE IS REMOVED AND REVEALED TO BE GREEN. WHAT IS THE CHANCE THE REMAINING MARBLE IS GREEN.



$\boxed{2/3}$

3-2) A SPAM FILTER IS DESIGNED BY LOOKING AT COMMONLY OCCURING PHRASES IN SPAM. SUPPOSE 80% OF EMAIL IS SPAM. SUPPOSE THE PHRASE "FREE MONEY" APPEARS IN 10% OF SPAM EMAILS & 1% OF NON-SPAM EMAILS. A NEW EMAIL WHICH CONTAINS THAT PHRASE WAS ARRIVED; WHAT IS THE PROBABILITY THAT IT IS SPAM?

$$P(S|FM) = \frac{\overset{0.1}{P(FM|S)}\,\overset{0.8}{P(S)}}{\underset{①}{P(FM)}} \qquad ①\ P(FM|S)P(S) + P(FM|S^c)P(S^c)$$

$$= 0.08 / (0.08 + 0.002) = 98\%$$

3-3) LET G BE THE EVENT THAT A CERTAIN INDIVIDUAL IS GUILTY OF A CERTAIN ROBBERY. IN GATHERING EVIDENCE, IT IS LEARNED THAT AN EVENT E_1 OCCURRED. A LITTLE LATER IT IS ALSO LEARNED THAT E_2 OCCURED.

a) IS IT POSSIBLE THAT INDIVIDUALLY THESE PIECES OF EVIDENCE INCREASE THE CHANCE OF GUILT BUT TOGETHER REDUCE IT? i.e.

$$P(G|E_1) > P(G) \quad \& \quad P(G|E_2) > P(G)$$

but $$P(G|E_1,E_2) < P(G)$$

3-3) a cont.) YES – THEY COULD INTERACT IN A WAY THAT MAKES G LESS LIKELY OR EVEN IMPOSSIBLE

b) SHOW PROBABILITY OF GUILT GIVEN EVIDENCE IS THE SAME WHETHER WE UPDATE ALL A ONCE OR IN MULTIPLE STEPS.

① ALL AT ONCE: $P(G \mid E_1, E_2)$

$$= \frac{P(G \cap E_1 \cap E_2) / P(E_1)}{P(E_1 \cap E_2) / P(E_1)} = \frac{P(G \cap E_2 \mid E_1)}{P(E_2 \mid E_1)}$$

"all at once" decomposes into steps

3-4) A CRIME IS COMMITTED BY ONE OF TWO SUSPECTS, A & B. THERE IS EQUAL EVIDENCE AGAINST BOTH. FURTHER INVESTIGATION REVEALS BLOOD TYPE MATCHING A AT THE CRIME SCENE. B's BLOOD TYPE IS UNKNOWN. A's BLOOD TYPE OCCURS IN 10% OF POPULATION.

a) WHAT IS PROBABILITY OF A BEING GUILTY?

$$P(A \mid \text{BLOOD}) = \frac{P(\text{BLOOD} \mid A) P(A)}{P(\text{BLOOD} \mid A) P(A) + P(\text{BLOOD}^c \mid A^c) P(A^c)}$$

$$= \frac{1/2}{1/2 + 1/20} = \frac{10}{11}$$

3-4) CONT b.) GIVEN THIS INFORMATION, WHAT IS THE PROBABILITY THAT SUSPECT B MATCHES THE SAMPLE BLOOD.

$1/11$ CHANCE OF GUILT

$+ \; 10/11 \cdot 1/10$ CHANCE OF INNOCENCE + RANDOM MATCH

$= 2/11$

3-5) YOU ARE GOING TO PLAY TWO GAMES OF CHESS AGAINST AN OPPONENT OF UNKNOWN STRENGTH. EQUALLY LIKELY TO BE BEGINNER (90% V), INT (50% V) OR MASTER (30% V)

a) WHAT IS YOUR PROBABILITY OF WINNING THE FIRST GAME?

$(0.9 + 0.5 + 0.3)/3 \Rightarrow 17/30$

b) YOU WON GAME ONE! GIVEN THIS INFORMATION HOW LIKELY ARE YOU TO WIN GAME TWO?

from (a)

$P(V|B) = 0.9$

$P(V|I) = 0.5$

$P(V|M) = 0.3$

$P(V_2|V_1) = \overset{①}{P(V_2) P(V_1)} / P(V_1)$

$① = \frac{1}{3}\big(P(V_1, V_2|B) + P(V_1 V_2|I) + P(V_1 V_2|M)\big)$

$= \frac{1}{3}(0.9^2 + 0.5^2 + 0.3)^2$

$① / ⓐ = \frac{23/60}{17/30} = 23/34$

3-5) cont c) EXPLAIN THE DISTINCTION BETWEEN ASSUMING THE OUTCOMES OF THE GAMES ARE INDEPENDENT & ASSUMING THEY ARE CONDITIONALLY INDEPENDENT ON OPPONENT'S SKILL LEVEL. WHICH OF THESE ASSUMPTIONS IS MORE REASONABLE. WHY?

TRUE INDEPENDENCE: NO CORRELATION BETWEEN GAMES; CHESS COMPLETELY RANDOM

CONDITIONAL INDEPENDENCE: OPPONENT'S SKILL LEVEL CONSTRAINS RANDOMNESS TO SOME EXTENT.

LATTER IS MORE REASONABLE SINCE CHESS IS KNOWN GAME OF SKILL

LECTURE #7 GAMBLER'S RUIN & RANDOM VARS

THIS COURSE IS ABOUT

1) CONDITIONING: THE SOUL OF STATISTICS

2) RANDOM VARIABLES & THEIR DISTRIBUTIONS

FOR NOW WE'LL STICK WITH #1 BUT #2 IS COMING SOON

GAMBLER'S RUIN

TWO GAMBLERS BETTING \$1/ROUND IN SEQUENCE

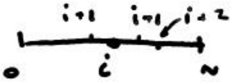
GAME ENDS IF ONE RUNS OUT OF MONEY.

FIND PROBABILITY PLAYER A WINS GAME

$p = P(\text{A wins round})$, $q = 1-p$

A starts with $\$i$; B with $\$(n-i)$

RANDOM WALK: same structure



$p = P(\text{right})$
$q = P(\text{left})$

w/ absorbing states at 0 & N

SEEMS DIFFICULT BUT STRUCTURE IS RECURSIVE, WHICH GIVES US A NICE HINT - CONDITION ON FIRST STEP

FIRST STEP ANALYSIS - Let $P_i = P(\text{A wins game} \mid \text{A starts w/ } \$i)$

$$P_i = pP_{i+1} + qP_{i+1} \quad (1)$$

(BRANCHING STATES W/ LAW OF TOTAL PROBABILITY)

b.c. $\begin{cases} P_0 = 0 \\ P_N = 1 \end{cases}$

(1) A DIFFERENCE EQUATION (DISCRETE VERSION OF DIFFERENTIAL)

SOLVE (1): TRY ANSATZ ANALOGUE

$P_i = x^i$ ← GENERALLY A GOOD GUESS FOR DIFFERENCE EQS.

$$x^i = px^{i+1} + qx^{i-1} \Rightarrow px^2 - x + q = 0$$

$$x = \frac{1 \pm \sqrt{1-4pq}}{2p}$$ plug back in

$$x = \frac{1 \pm \sqrt{4 - pq}}{2p}$$

$$1 - 4pq = 1 - 4p(1-p)$$
$$= 1 - 4p + 4p^2$$
$$= (2p-1)^2 \quad ②$$

$$x = \frac{1 \pm (2p-1)}{2p} \Rightarrow 1, \ q/p$$

GENERAL SOLUTION IS LINEAR COMBO OF BOTH ROOTS

$P_i = A1^i + B(q/p)^i$; $p \neq q$ (no repeated root)

B.C.: $P_0 = 0$; $B = -A$

$P_N = 1$; $1 = A(1 - q/p)$

if we repeat for B we get 1-A, which rules out eternal oscillation

$P_i = \left(1 - \left(\frac{q}{p}\right)^i\right) \Big/ \left(1 - \left(\frac{q}{p}\right)^n\right)$; if $p \neq q$

$x = q/p \quad \lim_{x \to 1} (1 - x^i)/(1 - x^n) \Rightarrow \lim_{x \to 1} = (i x^{i-1})/(N x^{N-1})$

$= i/N$ if $p = q$ fair case

EXAMPLES.

let $i = N - i$, $p = 0.49$ ← slightly unfair

$N = 20 \Rightarrow P(A) = 0.4$

$N = 100 \Rightarrow P(A) = 0.12$

$N = 200 \Rightarrow P(A) = 0.02$ THE HOUSE WINS :)

RANDOM VARIABLES

OUR NOTATION IS GETTING A BIT UNWEILDY. WE NEED NEW TERMINOLOGY TO HELP US.

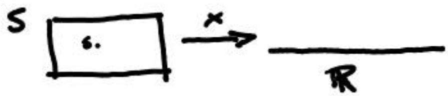
WHAT IS A RANDOM VARIABLE? WHAT IS A VARIABLE?

VARIABLES: $x+2=9$; $x=7$ ← 7 is a constant, not a variable

x doesn't actually change in value; it's not really 'variable': need functions for that

RANDOM VARIABLE -> a function!

A FUNCTION FROM THE SAMPLE SPACE S TO $\mathbb{R}$



THINK OF X AS A NUMERICAL "SUMMARY" OF AN ASPECT OF EXPERIMENT

SIMPLE EXAMPLE

BERNOULLI DISTRIBUTION: A R.V. X IS SAID TO HAVE BERN. DIST. IF X HAS ONLY TWO POSSIBLE VALUES, 0 AND 1, AND $P(\underbrace{X=1}_{\text{event } \{s:X(s)=1\}}) = p$; $P(X=0) = 1-p$

BINOMIAL (n,p) DIST: THE DISTRIBUTION OF THE NUMBER OF SUCCESSES (X) IN n INDEPENDENT BERN. TRIALS. GIVEN BY $P(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}$ ← PROBABILITY MASS FUNCTION

1110000 ← $n=7$, $k=3$

$X \sim Bin(n,p)$; $Y \sim Bin(m,p)$ THEN $X+Y \sim Bin(m+n, p)$

PROOF: consider n trials, then m more trials.

X = # successes in n, y = # successes in m

LECTURE 8 RANDOM VARIABLES CONTINUED

BINOMIAL DISTRIBUTION

$X \sim \text{Bin}(n,p)$ ← family depending on n, $p \in \{0,1\}$

THREE WAYS TO THINK ABOUT THIS

1) STORY: n independent Bern. trial; each results in success "X" or failure. p is prob success.

2) INDICATOR RANDOM VARIABLES: $X = X_1 + X_2 + \dots + X_n$;

BREAKS COMPLICATED DISTRIBUTION INTO SIMPLE CHUNKS

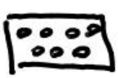
$$x_j = \begin{cases} 1 & \text{if } x_j \text{ success} \\ 0 & \text{otherwise} \end{cases}$$

$x_1, \dots, x_n \rightarrow$ i.i.d. $\text{Bern}(p)$

← independent & identically distributed

3) PMF $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

RANDOM VARIABLES RV

S 

RV is a function that assigns number to each pebble in pebbleworld.

$x = 7$ is an <u>event</u>

CDF: $X \le x$ is an event

$F(x) = P(X \le x)$ then F is the cumulative distribution function of X. describes distribution of X

$a_1, a_2, \ldots,$ and $a_1, a_2, \ldots$

integer values for our purposes

PMF - only for discrete random variables

$P(X = a_j)$ for all j

① $p_j \geq 0$, $\sum_j p_j = 1$ ②

PMFs are easier to work with; CDFs more general

BACK TO BINOMIAL

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k \in \{0, \ldots, n\}$ ① obv satisfied

$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p+q)^n = 1^n = 1$, by binomial theorem

② satisfied

SHOW $X \sim \text{Bin}(n,p)$; $Y \sim \text{Bin}(m,p) \rightarrow X+Y \sim \text{Bin}(n+m, p)$

make sure they're in the same domain :>

X # successes in n trials

Y " m "

note $p = p$ throughout!

we've done story proof already but...

$X = X_1 + \ldots + X_N$; $Y = Y_1 + \ldots + Y_M$

$X + Y = \sum_{j=1}^{N} X_j + \sum_{j=1}^{M} Y_j \rightarrow$ sum of $n+m$ iid Bern. Ps

$\Downarrow$

$\text{Bin}(m+n, p)$

WHAT ABOUT FROM PMF PERSPECTIVE?

k because if $x > k$, $x + y > k$ always

$$P(X+Y=k) = \sum_{j=0}^{k} P(X+Y=k \mid X=j)\,P(X=j)$$

by independence

$$= \sum_{j=0}^{k} P(Y=k-j \mid \cancel{X=j}) \binom{n}{j} p^{j} (1-p)^{n-j} \quad (1)$$

$$= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j} (1-p)^{m-k+j} \cdot (1)$$

simplify...

$$= p^{k} q^{m+n-k} \sum_{j=0}^{k} \binom{m}{k-j} \binom{n}{j}$$

vandermonde identity

$$= p^{k} q^{m+n-k} \binom{m+n}{k}$$

$= 1$ by binomial theorem :)

BE CAREFUL ABOUT ASSUMING BINOMIAL DISTRIBUTION

- TRIALS MUST BE INDEPENDENT
- P must be consistent

Ex: 5-card hand. find distribution of aces in hand (discrete)

Let X = # aces

$P(X=k)$. THIS IS ZERO UNLESS $k \in \{0,1,2,3,4\}$

NOT BINOMIAL B/C TRIALS ARE NOT INDEPENDENT!

$$P(X=k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}} \quad \text{for } k \in \{0, \dots, 4\}$$

SAME AS THE ELK PROBLEM FROM HOMEWORK #1

CARDS INSTEAD OF ELK; ACES "TAGGED"

HAVE JAR OF MARBLES, b BLACK, w WHITE,

PICK SIMPLE RANDOM SAMPLE OF SIZE n.

FIND DISTRIBUTION X OF # white MARBLES IN SAMPLE.

$$P(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{b+w}{n}}, \quad 0 \le k \le w, \; 0 \le n-k \le b$$

THIS DISTRIBUTION IS CALLED THE HYPERGEOMETRIC DISTRIBUTION. UNDERSTANDING THE STORY HELPS SPOT PROBLEMS WHERE IT'S USEFUL. SAMPLING <u>WITHOUT REPLACEMENT</u>

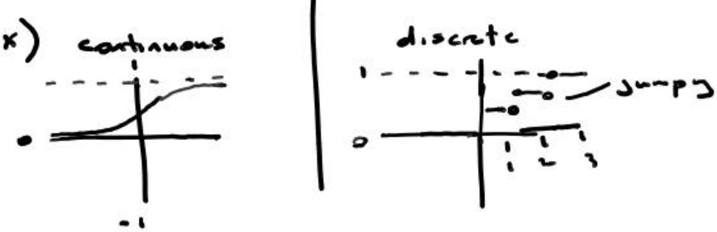
HYPERGEOMETRIC converges on BINOMIAL FOR very large n!

WE SHOULD CHECK HYPERGEOMETRIC IS A VALID PMF.

vandermonde $\binom{w+b}{n}$

$$\sum_{k=0}^{w} \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = 1 \checkmark$$

CDF $P(X \le x)$



PROBLEM SET #3

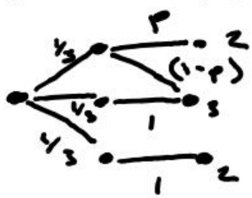
1. CONSIDER MONTY HALL EXCEPT MONTY LIKES OPENING DOOR 2 MORE THAN DOOR 3; WHEN GIVEN A CHOICE HE OPENS DR 2 w/ PROB p WHERE $\frac{1}{2} \le p \le 1$

a) FIND THE UNCONDITIONAL PROBABILITY THAT "ALWAY SWITCH" SUCCEEDS

$$P(w) = \cancel{P(w|C_1)}^{0} P(C_1) + P(w|C_2)P(C_2) + P(w|C_3)P(C_3)$$

$1/3 + 1/3 = 2/3$; unchanged ✓

b) FIND "ALWAYS SWITCH" SUCCESS RATE IF MONTY OPENS DOOR 2



DELETE BRACH 2

SUCCESS IS $\dfrac{1/3}{1/3 + p/3} \Rightarrow 1/(1+p)$ ✓

c) AS b) BUT WITH DOOR 3

$\Rightarrow 1/3 \,/ \left(1/3 + {1-p}/3\right) \Rightarrow 1/(2-p)$ ✓

2. CONSIDER THE FOLLOWING CONVERSATION

L: DAD, I THINK HE'S AN IVORY DEALER! HIS BOOTS ARE IVORY, HIS HAT IS IVORY & I'M PRETTY SURE THAT CHECK IS ...

H: LISA, A GUY WHO'S GOT LOTS OF IVORY IS LESS LIKELY TO HURT STAMPY THAN ONE WHOSE IVORY SUPPLIES ARE LOW.

2 cont a) DEFINE CLEAR NOTATION FOR THE VARIOUS EVENTS OF INTEREST HERE

$P(S)$ = Probability of person harming "stampy"

$P(I)$ = Probability of person being an ivory dealer

$P(Q)$ = Probability of person having extensive ivory supplies

b) EXPESS L & H'S ARGUMENTS AS CONDITIONAL W/ ABOVE NOTATION:

$$L: \quad P(S|I) > P(S|I^c)$$

$$H: \quad P(S|Q) < P(S|Q^c)$$

c) ASSUME H'S NOTION THAT HAVING HIGH IVORY SUPPLIES MAKES A PERSON LESS LIKELY TO HARM STAMPY IS TRUE. EXPLAIN WHAT'S WRONG W/ REASONING.

IF PERSON IS PART OF HIGH-HARM GROUP BEING LOWER HARM WITHIN THAT GROUP DOESN'T IMPLY LOW HARM!

3) A GAMBLER REPEATEDLY PLAYS A GAME WHERE IN EACH ROUND HE WINS A DOLLAR WITH PROBABILITY 1/3 & LOSES WITH PROBABILITY 2/3. HE WILL QUIT WHEN UP BY \$2. SUPPOSE HE STARTS W/ \$1,000,000. SHOW THE PROB HE REACHES +2 < 1/4

TAKE $p = 1/3$; $q = 2/3$; $i = 1{,}000{,}000$; $n = 1{,}000{,}002$

FROM EARLIER $P_i = \left(1-\left(\frac{q}{p}\right)^i\right) \Big/ \left(1-\left(\frac{q}{p}\right)^n\right)$; if $p \neq q$

MULTIPLY TOP & BOTTOM BY -1

$$\left[(q/p)^i - 1\right] \Big/ \left[(q/p)^{i+2} - 1\right] = (2^i - 1)/(2^{i+2} - 1)$$

$$1/4 > (2^i - 1)/(2^{i+2} - 1)$$

$$2^{i+2} - 1 > 4(2^i - 1)$$

$$2^{i+2} - 1 > 2^{i+2} - 4$$

$$-1 > -4 \checkmark$$

$$\therefore P(v) < 1/4$$

4-1) a) IN THE WORLD SERIES TWO TEAMS PLAY A SERIES OF GAMES AGAINST EACH OTHER & THE FIRST TO FOUR GAMES WINS. LET p BE THE PROBABILITY A WINS ANY INDIVIDUAL GAME & ASSUME GAMES ARE INDEPENDENT. WHAT IS PROBABILITY A WINS SERIES?

4-1) a) $X \sim \text{Bin}(7, p)$; $P(X \geq 4) = \sum_{k=4}^{7} P(X=k)$

$$= \binom{7}{4} p^4 q^3 + \binom{7}{5} p^5 q^2 + \binom{7}{6} p^6 q + p^7$$

b) DOES THIS RESULT CHANGE IF TEAMS DON'T PLAY DEAD-RUBBER GAMES? GIVE AN INTUITIVE EXPLANATION WHY OR WHY NOT.

NO, THIS WON'T CHANGE IF THERE ARE FEWER THAN 7 GAMES — WE CAN REMOVE POST-SERIES "GHOST" GAMES THAT WOULDN'T CHANGE THE OUTCOME WITHOUT ACTUALLY IMPACTING CALC.

4-2) A SEQUENCE OF n INDEPENDENT EXPERIMENTS IS PERFORMED. EACH IS A SUCCESS W/ PROBABILITY p AND FAILS W/ q-p. SHOW THAT CONDITIONAL ON # SUCCESSES, ALL POSSIBILITYS FOR THE LIST OF OUTCOMES FOR THE EXPERIMENT ARE EQUALLY LIKELY.

$$P(X_1 = \#_1, \dots, X_n = \#_n \mid X = k) \overset{\text{independent}}{=} P(X_1 = \#_1, \dots, X_n = \#_n) / P(X = k)$$

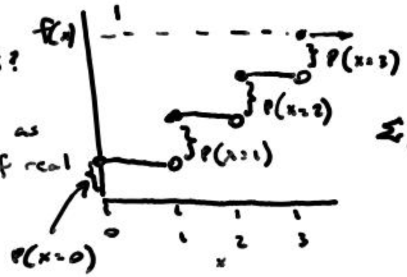
$$= \frac{p^k q^{n-k}}{\binom{n}{k} p^k q^{n-k}}$$

$\hookrightarrow$ INDEPENDENT OF $\#_i$; SYMMETRIC

LECTURE 9: RANDOM VARIABLES, CONT.

REMEMBER CDFS?

$F(x) = P(X \le x)$, as a function of real x



Σ jump sizes is PMF

find $P(1 < X \le 3)$ using CDF

$$P(X \le 3) = P(X \le 1) + P(1 < X \le 3) \Rightarrow P(1 < X \le 3) = P(X \le 3) - P(X \le 1)$$

generally $P(a < x < b) = P(X \le b) - P(X \le a)$

PROPERTIES of CDF

1. INCREASING
2. RIGHT-CONTINUOUS
3. $F(x) \to 0$ as $x \to -\infty$

   $F(x) \to 1$ as $x \to \infty$

if & only if these properties are satisfied we have a CDF

INDEPENDENCE of RANDOM VARIABLES

X, Y ARE IND. IF $P(X \le x, Y \le y) = P(X \le x)\,P(Y \le y)$ for all x, y

IN DISCRETE CASE:

$$P(X = x, Y = y) = P(X = x)\,P(Y = y)$$

AVERAGES (MEANS, EXPECTED VALUES)

$1,2,3,4,5,6 \to \sum \#s/6 = 3.5$ (remember $\frac{1}{n}\sum_{j=1}^{n} j = \frac{n+1}{2}$)

for arithmetic series

$1,1,1,1,1,3,3,5 \to 2$

TWO WAYS: ADD & DIVIDE

or GROUP & WEIGHT $\to 1\cdot\frac{5}{8},\ 3\cdot\frac{2}{8},\ 5\cdot\frac{1}{8} \to 2$

AVERAGE of DISCRETE R.V. X

$$E(X) = \sum_x \underbrace{x}_{\text{value}} \underbrace{P(X=x)}_{\text{PMF}},$$ summed over x w/ $P(X=x)$ positive

LET $X \sim \text{BERNOULLI}(p)$

$E(X) = 1\cdot P(X=1) + 0\cdot P(X=0) \Rightarrow p$

example: $X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$ (indicator r.v.)

$E(X) = P(A)$ ← THIS LOOKS SIMPLE BUT IS FUNDAMENTAL; BRIDGES BETWEEN EXPECTED VALUES & PROBABILITIES

LET $X \sim \text{Bin}(n,p)$

FROM EARLIER STORY PROOF

$$E(X) = \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1} p^{k} q^{n-k}$$

$$= np\sum_{k=1}^{n}\binom{n-1}{k-1} p^{k-1} q^{n-k} \qquad \text{let } j = k-1$$

$$= np\sum_{j=0}^{n-1}\binom{n-1}{j} p^{j} q^{n-j-1}$$

SIMPLIFY $\rightarrow$ $np \cdot 1$

a better way... use <u>linearity</u> of expectation

$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ even if X & Y are dependent

$\mathbb{E}(cX) = c\mathbb{E}(X)$ for constant c

each term of binomial dist has probability p; there are n of them, so $\mathbb{E}(X) = np$ $\leftarrow$ much easier!

$X_1, \dots, X_j$ $\quad X_j \sim \text{Bern}(p)$

HYPERGEOMETRIC EXAMPLE

5-card hand w/ $X = \#$ aces

let X_j be indicator of j^{th} card being an ace $1 \le j \le 5$

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X_1 + \dots + X_5) \\
&= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_5) \text{ by linearity} \\
&= 5\,\mathbb{E}(X_1) \text{ by symmetry} \\
&= 5\,P(1^{st} \text{ card is an ace}) = 5/13
\end{aligned}$$

, even though X_j's are not independent!

THIS GIVES EXPECTED VALUE OF ANY HYPERGEOMETRIC DISTRIBUTION

NOW LET'S LOOK AT A NEW DISTRIBUTION...

GEOM(p): independent Bern(p) trials -> # failures before first success

Let $X \sim \text{Geom}(p)$, $q = 1-p$

PMF: $P(X=k) = q^k p$; $k \in \{0,1,2\ldots\}$ FFFFFS $P(X=5) = q^5 p$

valid since $\sum_{k=0}^{\infty} pq^k = p/(1-q) = 1$

take derivative $\sum kq^{k-1} = \frac{1}{(1-q)^2}$ => $\frac{q}{p^2}$

$E(X) = \sum_{k=0}^{\infty} kpq^k = p\sum_{k=1}^{\infty} kq^k$ => q/p

STORY PROOF

LET $c = E(X)$

memoryless recursion

$c = 0$ w/ prob. p + $(1+c)q$

$= q + cq$ => q/p

LECTURE 10: MORE EXPECTED VALUES

START w/ PROOF OF LINEARITY.

- TWO METHODS OF AVERAGING - COUNT & DIVIDE
  GROUP & WEIGHT

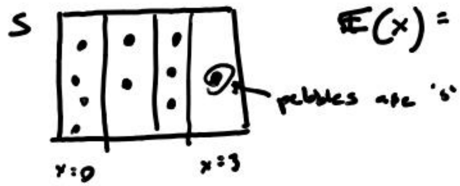
LET $T = X+Y$; show $E(T) = E(X) + E(Y)$

$\sum_t t P(T=t) \overset{?}{=} \sum_x x P(X=x) + \sum_y y P(Y=y)$

TRY TO SPLIT LEFT SIDE

$P(T=t) = \sum_x P(T=t \mid X=x) P(X=x)$ STRAIGHTFORWARD IF INDEPENDENT BUT INTRACTABLE IF NOT

TRY PEBBLEWORLD



$$\mathbb{E}(X) = \underbrace{\sum_x x P(X=x)}_{\text{GROUPED AVERAGE}} = \underbrace{\sum_s X(s) P(\{s\})}_{\text{UNGROUPED AVERAGE}}$$

proof of linearity from this:

$$\mathbb{E}(T) = \sum_s (X+Y)(s) P(\{s\})$$
$$= \sum_s (X(s) + Y(s)) P(\{s\})$$
$$= \sum_s X(s) P(\{s\}) + \sum_s Y(s) P(\{s\}) \quad \text{QED!}$$

KEY IS TO LOOK AT INDIVIDUAL EVENTS RATHER THAN TRYING TO GROUP THEM.

SIMILARLY $\mathbb{E}(cX) = c\mathbb{E}(X)$

DEPENDENCE $\rightarrow$ TAKE $X = Y$. THEN $\mathbb{E}(X+Y) = \mathbb{E}(2X) = 2\mathbb{E}(X)$
$= \mathbb{E}(X) + \mathbb{E}(Y)$

NEW DISTRIBUTION: NEGATIVE BINOMIAL parameters: r, p

not negative; not a binomial; actually a generalisation of geom.

Story: independent Bern(p) trials – number of failures before r^{th} success

PICTURE: 10001001011 $(r=5)$ $(n=6)$

note sequence ends on r^{th} success, first terms can be permuted at will w/ $r-1$ successes & n failures

PMF: $P(X=n) = \binom{n+r-1}{r-1} p^r (1-p)^n$

for $n = 0, 1, 2 \ldots$

FIND $\mathbb{E}(X)$ for NEG. BIN.

PROBABLY DON'T WANT TO TAKE THE SUM STRAIGHT UP.

WE'VE SOLVED $r=1$ ALREADY

USE RECURSION... $\mathbb{E}(X) = \mathbb{E}(X_1 + \dots + X_r)$

X_j is # failures between $(j-1)^{st}$ and j^{th} success

$X_j \sim \text{Geom}(p)$

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_r) = rq/p$$

BE CAREFUL ABOUT CONVENTIONS HERE – SOME END ON SUCCESS & SOME ON FAILURE BEFORE SUCCESS

$X \sim FS(p)$ counting until first success inclusive

let $Y = X-1$, then $Y \sim \text{geom}(p)$. $\mathbb{E}(X) = \mathbb{E}(Y)+1$

$$= \frac{q}{p} + 1 = \frac{1}{p}$$

MORE EXAMPLES WITH EXPECTATION

RANDOM PERMUTATION OF INTEGERS $1, 2, \dots, n$ where $n \geq 2$

FIND EXPECTED NUMBER OF LOCAL MAXIMA (any number larger than its neighbors)

$\underline{3}\ 2\ 1\ 4\ \underline{7}\ 5\ \underline{6}$ ①

Let I_j be indicator r.v. of position j being local max; $1 \leq j \leq n$

$\sum I_j = \#\text{ l.m}$

$\mathbb{E}(I_1 + \dots + I_n) = \mathbb{E}(I_1) + \dots + \mathbb{E}(I_n)$

for any triplet, $1/3$ chance largest # is in middle (int cas)

① OBV. TWO TYPES OF L.M.; INTERMEDIATE & ENDS

for a double obv $1/2$

$$\mathbb{E}(I_n) = \underbrace{\frac{n-2}{3}}_{\text{intermediate}} + \underbrace{1}_{\text{endpoints}} = \frac{n+1}{3}$$

ANOTHER EXAMPLE: St. Petersberg paradox

you're offered a game – coin is flipped over and over until it lands heads. if heads on first flip, you get \$2, 2^{nd} \$4, ... $\$2^x$, where x is # flips until heads (including H). HOW MUCH SHOULD YOU PAY TO PLAY GAME?

$$Y = 2^x$$

$$\mathbb{E}(Y) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2}^k = \sum_{1}^{\infty} 1 = \infty$$

← we should want to pay \$∞ to play??!

bound return at $x = 40$. THEN $\sum_{k=1}^{40} 2^k \frac{1}{2^k} = 40$!

$\mathbb{E}(2^x)$ IS INFINITY $\neq 2^{\mathbb{E}(x)}$. WE CAN'T BLINDLY MOVE $\mathbb{E}$ AROUND :) (OTHER PARADOX SOLUTION IS ABOUT UTILITY OF DOLLARS AS YOU SCALE UP)

PROBLEM SET 4

1-1) FIND AN EXAMPLE OF TWO RANDOM VARIABLES X & Y SUCH THAT THEY HAVE THE SAME DISTRIBUTION BUT X NEVER EQUALS Y.

- BIN DISTRIBUTION OF COINFLIPS WHERE X = # HEADS; Y = # TAILS, FLIPS = ODD

1-2) LET X BE A RANDOM DAY OF THE WEEK, CODED SO THAT MONDAY IS 1, etc. LET Y BE THE DAY AFTER X. DO X & Y ① HAVE THE SAME DISTRIBUTION? WHAT IS ② $P(X<Y)$?

① YES - EQUAL LIKLIHOOD OF $X_1 \dots X_7$ & $Y_1 \dots Y_7$

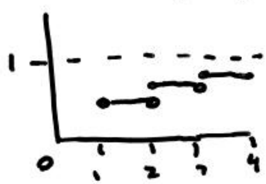
② $P(X<Y) = P(X \neq 7) = 6/7$

1-3) A COIN IS TOSSED REPEATEDLY UNTIL IT LANDS HEADS FOR THE FIRST TIME. X IS # FLIPS, p IS PROBABILITY OF HEADS, FIND THE CDF of X, & FOR $p = 1/2$ SKETCH ITS GRAPH.

$Y = X - 1 \sim \text{GEOMETRIC}(p)$

PMF: $P(X=k) = pq^{k-1}$ for $k \in \{1, 2, 3, \dots\}$

$\hookrightarrow$ CDF $= 1 - \frac{1}{2^x}$ for $x \geq 1$; 0 elsewhere.



$\leftarrow$ SKETCH

1-4) ARE THERE DISCRETE RANDOM VARIABLES X & Y SUCH THAT $E(X) > 100\, E(Y)$ BUT $Y > X$ w/ PROB of at least 0.99.

THIS IS SOMETHING LIKE SIMPSON'S PARADOX —

$X \rightarrow \begin{cases} 0 & @\ 0.99 \\ 1E^{1000} & @\ 0.01 \end{cases} \qquad Y = 0.01 \qquad E(X) >> E(Y)$

1-5) LET X BE A DISCRETE r.v. w/ POSSIBLE VALUES 1, 2, 3, ... LET $F(x) = P(X \le x)$ be the CDF of X. SHOW THAT

$$E(X) = \sum_{n=0}^{\infty} (1 - F(n)) \qquad 1 - F(n) = P(X > n)$$

$$P(X > n) = \sum_{n+1}^{\infty} P(X = k)$$

$$= \sum_{n=0}^{\infty} \sum_{n+1}^{\infty} \underbrace{P(X = k)}_{\text{EACH OF THESE APPEARS } k \text{ TIMES OVER SUM}}$$

$$= \sum_{k=1}^{\infty} k \cdot P(X = k) = E(X) \checkmark$$

1-6) JOB CANDIDATES $C_1, C_2, \ldots$ ARE INTERVIEWED ONE BY ONE, AND THE INTERVIEWER COMPARES THEM & KEEPS AN UPDATED LIST OF RANKINGS (IF n CANDIDATES HAVE BEEN INTERVIEWED SO FAR, THIS IS A LIST OF THOSE CANDIDATES FROM BEST TO WORST). ASSUME NO LIMIT ON NUMBER OF CANDIDATES AVAILABLE, THAT FOR ANY n THE CANDIDATES ARE EQUALLY LIKELY TO ARRIVE IN ANY ORDER, AND THAT THERE ARE NO TIES IN THE RANKINGS.

LET X BE THE INDEX OF THE FIRST CANDIDATE BETTER THAN CANDIDATE C_1. WHAT IS $E(X)$?

1-6 cont) FIND $P(X > n)$; THIS IS PROBABILITY OF FIRST CANDIDATE BEING BEST. THIS MUST BE $1/n$.

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = 1 + \sum_{n=1}^{\infty} \frac{1}{n} = 1 + \infty \,!!!$$

2-1) A GROUP OF 50 PEOPLE ARE COMPARING THEIR BIRTHDAYS. FIND THE EXPECTED NUMBER OF PAIRS & THE EXPECTED NUMBER OF DAYS OF THE YEAR IN WHICH AT LEAST TWO OF THESE PEOPLE WERE BORN.

indicator r.v. per pair: $\binom{50}{2}\frac{1}{365}$ pairs $\Rightarrow$ 3.4 pairs

" per day $\Rightarrow P(K \geq 2) = 1 - P(k=0) - P(k=1)$

$$= 1 - \left(\frac{364}{365}\right)^{50} - \frac{50}{365}\left(\frac{364}{365}\right)^{49} \times 365$$

$= 2.9$ days

2-2) A TOTAL OF 20 BAGS OF GUMMY BEARS ARE DISTRIBUTED TO 20 STUDENTS. EACH BAG IS OBTAINED BY A RANDOM STUDENT, OUTCOMES ARE INDEPENDENT. FIND THE AVERAGE NUMBER OF GUMMY BEARS THE FIRST THREE STUDENTS GET IN TOTAL, (a) AND THE AVERAGE NUMBER OF STUDENTS W/ AT LEAST 1 BAG (b)

2-2 cont) a) indicator r.v. S_j = # bags per student

binomial dist $(20, 1/20)$

$$\mathbb{E}(S) = np = 1$$

$$\mathbb{E}(S_1 + \ldots + S_3) = 3$$

b) $P_{bag} = 1 - (19/20)^{20}$

$$\sum P_{bag} = 20\left(1 - (19/20)^{20}\right) = \mathbb{E}(X)$$

2-3) THERE ARE 100 SHOELACES IN A BOX. AT EACH STAGE, PICK TWO RANDOM ENDS & TIE TOGETHER. EITHER THIS RESULTS IN A LONGER SHOELACE OR A LOOP. WHAT ARE THE EXPECTED # OF STEPS UNTIL EVERYTHING IS A LOOP? (a) WHAT IS THE EXPECTED NUMBER OF LOOPS? (b)

(a) OBV NEED 100 STEPS TO TIE 200 ENDS

CREATE i.r.v for each step $j \in \{1, 100\}$

$$\mathbb{E}(X_j) = \frac{n}{\binom{2n}{2}} = \frac{1}{2n-1} \text{; sum w/ linearity } \sum_{n=1}^{100} \frac{1}{2n-1} \quad (b)$$

2-4) A HASH TABLE IS A COMMONLY-USED DATA STRUCTURE IN COMPUTER SCIENCE. FOR EXAMPLE, SUPPOSE WE WANT TO STORE SOME PHONE NUMBERS. ASSUME THAT NO TWO PEOPLE HAVE THE SAME NAME. FOR EACH NAME x, A HASH FUNCTION h IS USED, WITH $h(x)$

2-4 cont) AS THE LOCATION STORING EACH PERSON'S PHONE NUMBER. AFTER THE TABLE HAS BEEN COMPUTED, TO LOOK IT UP ONE JUST RECOMPUTES $h(x)$.

h IS DETERMINISTIC, SINCE WE DON'T WANT DIFFERENT RESULTS EVERY TIME WE COMPUTE $h(x)$. FOR THIS PROBLEM, LET'S ASSUME RANDOMNESS. LET THERE BE k PEOPLE, WITH EACH PERSON'S NUMBER STORED IN A RANDOM LOCATION $i \in \{0, \dots, n\}$. COLLISIONS ARE POSSIBLE:

FIND THE EXPECTED # OF LOCATIONS w/ NO NUMBERS STORED (a), 1 NUMBER STORED (b), AND > 1 # STORED (c). SHOULD THESE ADD TO n? (d)

(d) YES, OBVIOUSLY

(a) LET $I_j = 1$ if j^{th} loc. empty
$\quad\quad 0$ otherwise

$$P(I_j = 1) = (1 - 1/n)^k \Rightarrow \mathbb{E}(\text{empties}) = \underline{n(1 - 1/n)^k}$$

(b) LET $I_j = 1$ if j^{th} loc has 1 stored item
$\quad\quad 0$ otherwise

$$P(I_j = 1) = \underset{\substack{\uparrow \\ \text{1 person} \\ \text{from } k \to \binom{k}{1}}}{k} \cdot \underset{\substack{\nwarrow \\ \text{person} \\ \text{lands in } j}}{\frac{1}{n}} \underset{\substack{\nwarrow \\ \text{nobody else does}}}{\left(1 - \frac{1}{n}\right)^{k-1}} \Rightarrow \mathbb{E}(\text{singletons}) = \underline{k\left(1 - \frac{1}{n}\right)^{k-1}}$$

(c) = $\underline{n - (a) - (b)}$