7.2.6 Bernoulli trails & binomial distribution

Experiment with 2 outcomes

- success prob p
- failure prob $1-p$

Repeat experiment n times

Theorem 2 The probability of having exactly k successes in n independent Bernoulli trials with success prob p is $\binom{n}{k} p^k (1-p)^{n-k}$

P: Let $(x_1, x_2, \ldots, x_n)$ be the ordered set of outcomes. So $x_i \in \{\underset{S}{\text{success}}, \underset{F}{\text{failure}}\}$

We can choose k experiments among the n in $\binom{n}{k}$ ways. The probability that a fixed choice of k experiments are all successes while the remaining $n-k$ are all failures is $p^k(1-p)^{n-k}$. Hence the desired probability is

$$\binom{n}{k} p^k (1-p)^{n-k}$$ $\square$

Define $b(k,n,p)$ as probability of exactly k successes in n independent B-trials so $b(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k}$

Let $q = 1-p$ so $b(k,n,p) = \binom{n}{k} p^k q^{n-k}$

NB: $\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n$ by binomial formula

$= (p+(1-p))^n = 1$

So $b(k,n,p)$ is a probability distribution called the binomial distribution

7.2.7 Random variables

Def 6 A random variable associated with a sample space S is a function $f: S \to \mathbb{R}$

So each event $s \in S$ is assigned a value $f(s)$

Example 10+11 Flip fair coin 3 times $X(t) = \#\text{heads}$ $t \in S$

	$P(X=t)$
$X(HHH) = 3$	$1/8$
$X(HHT) = X(HTH) = X(THH) = 2$	$3/8$
$X(HTT) = X(THT) = X(TTH) = 1$	$3/8$
$X(TTT) = 0$	$1/8$

Def 7 The distribution of a random variable X on a sample space S is

$$\{(r, p(X=r)) \mid r \in X(S)\}$$ where

$X(S)$ is the set of values taken by X on S

Example 13 Birthday problem

Find min # persons in a room such that probability of having 2 with the same birthday is at least $1/2$.

Assumptions

- birthdays independent
- all days equally likely as a birthday

We find p_n = prob. all distinct

when all distinct there are 366 possibilities for person 1

365 — — — 2

$366-(i-1)$ — — — i

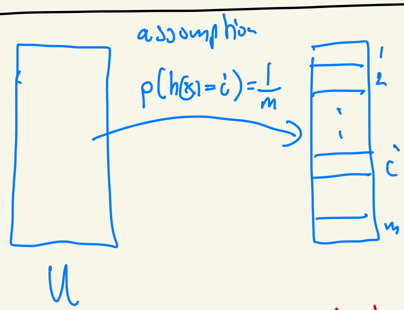
366^n outcomes in total so

$$p_n = \frac{366}{366} \cdot \frac{365}{366} \cdot \ldots \cdot \frac{367-n}{366}$$

when $n = 22$ p_n is still larger than $1/2$

$\boxed{n = 23}$ $p_n \sim 0.494$ so $1-p_n \sim 0.506$

Example 14 Probability of collisions in hashing



probability that n diffent keys from U map to n distict cell in hash table is

$$P_n = \frac{m}{m} \cdot \frac{m-1}{m} \cdot \ldots \cdot \frac{m+1-n}{m}$$

probability of a collision is $1-P_n$

smallest n such that $1-P_n \geq \frac{1}{2}$ is $n \sim 1.177\sqrt{m}$

If $m = 10^6$ it means that n should be larger than 1177

7.2.9 Monte Carlo Algorithms

- probabilistic algorithm that always outputs an answer true / unknown (or false) running time bounded (fixed)

- The answer may be wrong
  - true-biased : always correct when true is returned
  - false-biased : always correct when false is returned

Suppose $p(\text{test returns 'true'} \mid \text{answer} = \text{true}) = p$
Then $p(\text{test returns 'unknown'} \mid \text{answer} = \text{true}) = 1-p$

$\Big\}$ assume true-biased

Assume that the test only returns 'true' if this is the correct answer (can be deduced from the execution)

⇒ If correct answer is 'false' then the test will correctly return false

Amplifying probability of a correct answer:

- run the algorithm n times
- Assume the correct answer is 'true'

Then the n runs of the algorithm will result in at least one true with probability $1-(1-p)^n \to 1$ as n increases

Here we need that the n runs of the test are independent.

Example (not in book) Majority element

$S = \{x_1, x_2, \dots, x_n\}$ $x_i \in \mathbb{Z}$ $n = 2k$

Question is there a value $x \in \mathbb{Z}$ s.t. $|\{x_i \mid x = x_i\}| > k$?

Test : pick a random x_i
check if x_i occurs more than k times
if yes return 'true'
else return false/unknown

Observation at most one majority element
($\geq k+1$ copies)

so $p(\text{test returns true} \mid \exists \text{ majority}) > \frac{1}{2}$
and test always returns false if no majority element

A: found $\Leftarrow$ false, count $\Leftarrow 0$
while not found and count $\leq n$
count $\Leftarrow$ count $+1$
pick random $i \in [n]$
if x_i majority
found $\Leftarrow$ true
end
return found

- If S has no majority the algorithm will return the correct answer 'false'
- if S has a majority the probability that $\mathcal{A}$ returns true is $1 -$ prob that all n tests fail. Each test fails with prob $< \frac{1}{2}$ (when there is a majority) so probability that all n test fail is less than $\left(\frac{1}{2}\right)^n$

with $n = 10$ we have prob(wrong answer) $< \frac{1}{2}^{10} < \frac{1}{1000}$

$n = 20$ $\quad < \frac{1}{10^6}$

Example 15 Quality control

Testing chips

- assume that if a batch of n chips has not been tested, there is a $\frac{1}{10}$ chance of bad chip in batch and if it has been tested and passed they are all good

Q: how many of the n chips in an unchecked batch should we check to be very sure they are all good?

MC test: pick random chip and test it
repeat $\leq k$ times until all passed or bad chip found

probability that there is a bad chip in batch but we did not find it is $\left(\frac{9}{10}\right)^k$ (independent of n !!)

$$\left(\frac{9}{10}\right)^{132} < \frac{1}{10^6} \qquad \left(\frac{9}{10}\right)^{264} < \frac{1}{10^{12}}$$

$$\boxed{\text{So by running just a small number of tests we can become very sure that the batch is good!}}$$

Example 16 Primality testing

Miller Rabin test $MR(n,b)$ uses $0<b<n$
to test whether n is a prime.
probability that test says 'prime' for composite n is $< \frac{1}{4}$

MC Algorithm to test if n is composite

Repeat r times
    pick random $b \in]0,n[$
    run $MR(n,b)$
        if 'composite' output this and stop

end
output unknown

probability of answer 'unknown' for composite n
is at most $\left(\frac{1}{4}\right)^r$

The Probabilistic Method (Erdös - Spencer)

$\boxed{\text{Basic idea: If } p(\text{some element in } S \text{ has property } P) < 1 \text{ then } \exists x \in S \text{ without property } P}$

Very strong tool to prove existence of configurations

Theorem $\forall k \geq 2 \quad R(k,k) \geq 2^{k/2}$

($R(k,k)$ is min n s.t $\forall$ 2-col of K_n $\exists$ either red K_k or blue K_k)

P: $R(2,2) = 2$, $R(3,3) = 6 \geq 2^{3/2}$ ✓

assume $k \geq 4$

Consider a random 2-col r/b of the edges of K_n

$(p(\text{col } r) = p(\text{col } b) = \frac{1}{2})$

Consider the $\binom{n}{k}$ k-subsets of the vertices of K_n

and denote them $S_1, S_2, \ldots, S_{\binom{n}{k}}$

E_i: all edges in S_i have same colour

$$p(\text{monochromatic } K_k) = p(\cup E_i) \leq \sum p(E_i)$$ Union bound Boole

$$p(E_i) = 2 \cdot \left(\frac{1}{2}\right)^{\binom{k}{2}}$$ red or blue

so $p(\text{monochromatic } K_k) \leq \binom{n}{k} 2 \cdot \left(\frac{1}{2}\right)^{\binom{k}{2}}$

$$p(\text{monochromatic } K_k) \leq \binom{n}{k} 2 \cdot \left(\frac{1}{2}\right)^{\binom{k}{2}}$$

$$\leq \frac{n^k}{2^{k-1}} \cdot 2 \cdot \left(\frac{1}{2}\right)^{\binom{k}{2}} \qquad \left(\binom{n}{k} \leq \frac{n^k}{2^{k-1}} \right)$$

Suppon $n < 2^{k/2}$ then

$$< \frac{\left(2^{k/2}\right)^k}{2^{k-1}} \cdot 2 \cdot \left(\frac{1}{2}\right)^{\frac{k(k-1)}{2}}$$

$$= 2^{k^2/2 + 1 - k + 1 - \frac{k^2}{2} + \frac{k}{2}}$$

$$= 2^{2 - \frac{k}{2}}$$

$$\leq 1 \quad \text{as} \quad k \geq 4$$

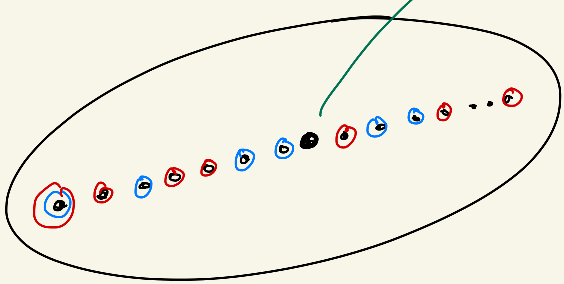
We have shown that when $n < 2^{k/2}$ the probability that a random 2-col leads to monochromatic K_k is < 1

Hence $\exists$ 2-col of K_n s.t no monochr. K_k !

$S = \{ \text{all 2-col of } E(K_n) \}$

$|S| = 2^{\binom{n}{2}}$

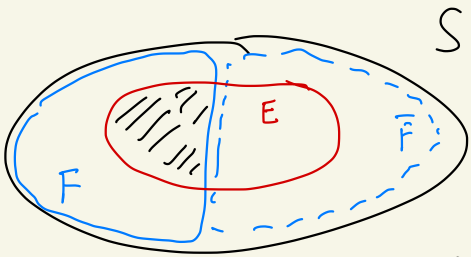
no red no blue



7.3 Bayes theorem

Let E and F be events s.t. $p(E), p(F) \neq 0$

then $$p(F|E) = \frac{p(E|F)\cdot p(F)}{p(E|F)\cdot p(F) + p(E|\overline{F})\cdot p(\overline{F})}$$



$F \cap E = ///$

We know by def of conditional prob $p(F|E) = \dfrac{p(F\cap E)}{p(E)}$

and $p(E|F) = \dfrac{p(E\cap F)}{p(F)}$ so $p(F|E)\cdot p(E) = p(E|F)\cdot p(F)$

$$\Rightarrow \quad p(F|E) = \frac{p(E|F)\cdot p(F)}{p(E)}$$

$$p(E) = p(E\cap S) = p(E\cap F) \cup p(E\cap \overline{F})$$
$$= p(E|F)\cdot p(F) + p(E|\overline{F})\cdot p(\overline{F})$$

$$\Rightarrow \quad p(F|E) = \frac{p(E|F)\cdot p(F)}{p(E|F)\cdot p(F) + p(E|\overline{F})\cdot p(\overline{F})} \qquad \square$$

Example 1

2 boxes

box 1	box 2
2G	4G
7R	3R

Experiment : 1. pick box with $p = \frac{1}{2}$ each
2. pick random ball from the chosen box

outcome red ball

Question : what is probability that we took from box 1?

E : outcome = red F : chosen box 1

we seek $p(F|E)$ and by Bayes theorem we know this is

$$p(F|E) = \frac{p(E|F)\cdot p(F)}{p(E|F)\cdot p(F) + p(E|\bar{F})\cdot p(\bar{F})}$$

$$= \frac{\frac{7}{9}\cdot\frac{1}{2}}{\frac{7}{9}\cdot\frac{1}{2} + \frac{3}{7}\cdot\frac{1}{2}} = \frac{\frac{7}{9}}{\frac{7}{9}+\frac{3}{7}} = \frac{49}{49+27} = \frac{49}{76}$$

Ex 2

1 in 10^5 have disease D

Test correct is $\frac{99}{100}$ if person has D

$\frac{995}{1000}$ if person does not have D

(a) Find prob (you are sick) if test = positive

F: person has D   E: positive test

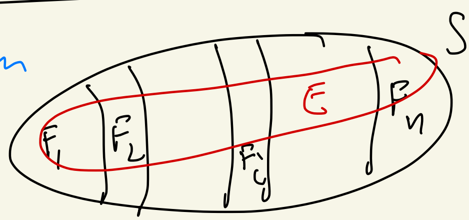
We seek $p(F|E)$

$$p(F|E) = \frac{p(E|F)\cdot p(F)}{p(E|F)\cdot p(F) + p(E|\bar{F})\cdot p(\bar{F})}$$

$$= \frac{\frac{99}{100}\cdot\frac{1}{10^5}}{\frac{99}{10^7} + \frac{5}{10^3}\left(1-\frac{1}{10^5}\right)} \sim 0{,}002$$

Conclusion prob of having disease is very small even if you test positive

Generalized Bayes thm



$$S = F_1 \cup F_2 \cup \ldots \cup F_n$$

$$p(F_j|E) = \frac{p(E|F_j)}{\sum_{i=1}^{n} p(E|F_i)\cdot p(F_i)}$$