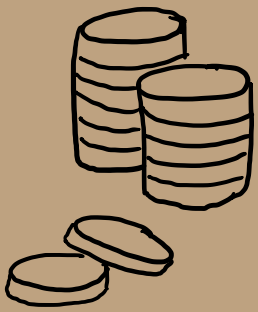
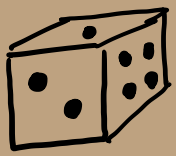
Probability





John Rachlin
CS 5002
Northeastern

* Antoine Gombaud
1607 - 1684

(A) Introduction

1. Probability theory is the foundation for modern statistics, but interest in probability was originally motivated by gambling, and thus we'll encounter problems about coins, dice, and playing cards!

2. Historians trace probability theory to the French gambler Chevalier de Mere who would bet:
   - Roll at least one 6 in four rolls of a single die. (Winning ≈ 52% of the time)
   - Roll double-6 in twenty-four rolls of a pair of dice. (Losing ≈ 51% of the time)

3. Chevalier de Méré* brought these problems to Blaise Pascal (binomial theorem, pascal's triangle) who also communicated with Pierre de Fermat, (number theory), leading to modern probability theory.

4. Today, probability and statistics are key to many areas of CS/AI: machine learning, robotics, expert systems, NLP, simulation.

(B) Definitions

1. A <u>trial</u> refers to an event whose outcome is unknown.
   - Pick a card
   - flip a coin.
   - test a drug on one patient
2. An <u>experiment</u> involves multiple trials.
   - multiple coin flips
   - clinical trials: an experiment to test a drug on many patients.
3. <u>Sample Space</u> defines all possible outcomes or alternatives

   Coin flips: $S = \{H, T\}$

   2 coin flips: $S = \{(H,H), (H,T), (T,H), (T,T)\}$

   Roll 1 die: $S = \{1, 2, 3, 4, 5, 6\}$

   Take an aspirin: $S = \{$headache goes away, headache doesn't go away$\}$
4. An <u>event</u>, E, specifies a particular outcome or set of outcomes.

   Flipped a coin and got heads: $E = \{H\}$

   Rolled a die and got odd: $E = \{1, 3, 5\}$

   Rolled two dice and got less than 6: $E = \{2, 3, 4, 5\}$ or $E = \{\text{sum} < 6\}$

   n people in room: $E = \{$number with same b-day $\geq 2\}$

5. The probability of an event, $P(E)$, is the liklihood that the event occurs in repeated trials.

Flipping a fair coin $P(E = \{H\}) = 0.5$

"50% probability"

Rolling a 6 in a fair die: $P(E = \{6\}) = \frac{1}{6}$

When we say "30% chance of rain tomorrow", we mean: If I simulate the complex dynamics of the atmosphere, with some randomness built in to reflect uncertainties, 30% of the time it rains the next day.**

6. Some basic facts:

$0 \leq P(E) \leq 1$

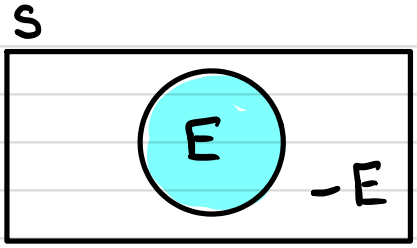
$P(S) = 1$ (all possible outcomes)

$P(E) + P(\neg E) = 1$ Event happens, or it doesn't

$P(\neg E) = 1 - P(E)$

** Due to turbulence, chaos, and sensitivity to initial conditions, local forecasts out beyond about 10-14 days are pretty useless.

Probability visualized w/ Venn diagrams



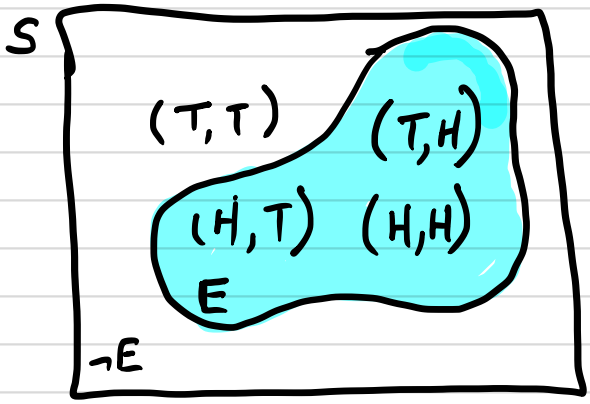
Events are a subset of outcomes drawn from the sample space.

Experiment: Flip coin twice

$S = \{(H,H), (H,T), (T,H), (T,T)\}$

$E = \{\text{At least one head}\} = \{(H,H), (H,T), (T,H)\}$

$\therefore P(E) = \frac{3}{4} = 0.75$



Note: reducing probabilities to counting problems usually makes sense only when each outcome is equally likely.

In that case:

$P(E) = |E| / |S|$

But an extreme (and silly) counter-example, if I take an aspirin for my headache and

$S = \{\text{headache cured}, \text{headache remains}\}$

$P(E = \{\text{headache cured}\}) \neq 0.50$!

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$
$$= (x+y)(x+y)(x+y)(x+y)$$

Flipping 4 Coins:

$S = \{TTTT, TTTH, TTHT, TTHH, \ldots, HHHH\}$

$|S| = 2^4 = 16$ possibilities.

a) $P(E = \{\text{exactly two heads}\})$

$$|E| = \binom{4}{2} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6, \text{ so } P(E) = \frac{6}{16} = \frac{3}{8}$$

b) $P(E = \{\geq 1 \text{ head}\})$

$$= 1 - P(E = \{\text{No Heads}\})$$
$$= 1 - P(\{TTTT\} = 1 - 1/16 = 15/16$$

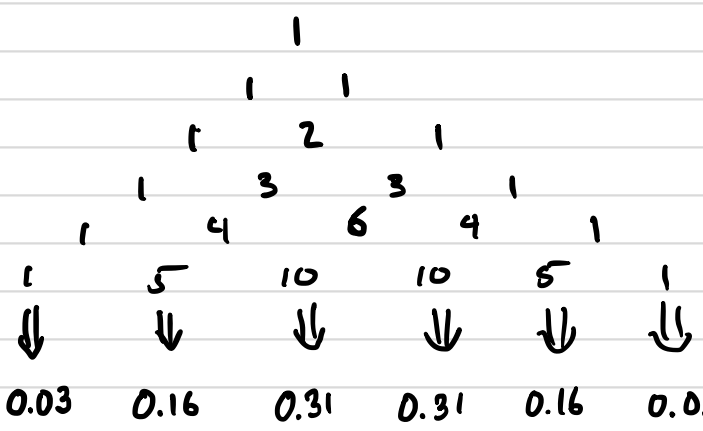
c) Flipping n coins: $|S| = 2^n$

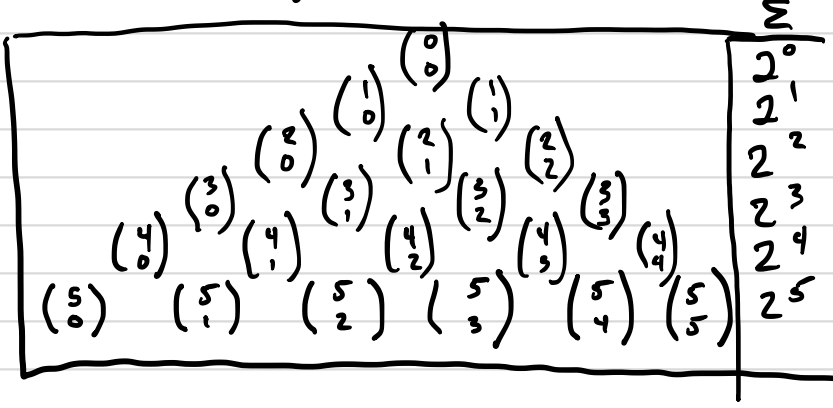
$P(E = \{k \text{ heads}\}) = \binom{n}{k}$ } # ways to get k heads

$\overline{2^n}$ } # of possible sequences.

OPTIONAL

This should remind us of Pascal's $\triangle$:





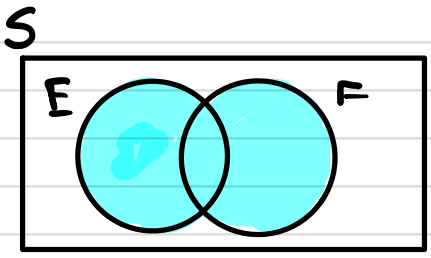
And in general: $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots \binom{n}{n} = 2^n$

Probability of Multiple events

Two Events E, F.

Either event can occur

$$P(E \cup F) = P(E \text{ or } F)$$
$$= P(E) + P(F) - P(E \cap F)$$
$$= P(E) + P(F)$$ if E, F are "mutually exclusive" (sets are disjoint)



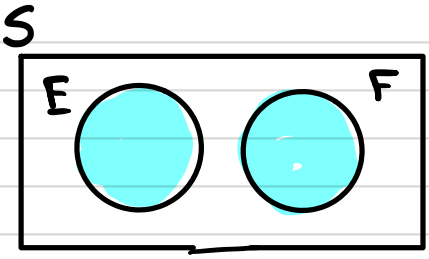
$P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Suppose roll 1 die

$E = \{\text{Roll odd}\}$ $|E| = 3$

$F = \{\text{Roll} \geq 4\}$ $|F| = 3$

$$P(E \cup F) = \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}$$



$P(E \cup F) = P(E) + P(F)$

Suppose double majors aren't allowed.

$P(\{\text{CS}\}) = 0.6$

$P(\{\text{art history}\}) = 0.05$

$P(\text{CS} \cup \text{Art}) = 0.65$

Probability of Multiple events – continued

Two events E, F $\quad P(E \cap F)$ = "Joint Probability"

All events occur. Now it depends on whether the two events are linked conditionally or independent.

Independence: Two events are independent if the outcome of one event has no impact on the outcome of the other.

Then, $P(E \cap F) = P(E) \cdot P(F)$

E.g $\quad E = \{\text{heads on flip \#1}\}$

$F = \{\text{heads on flip \#2}\}$

Then: $P(\text{two heads}) = P(E \cap F) = P(E) \cdot P(F)$

$= 0.50 \times 0.50 = 0.25$

(Coin flips are independent trials.)

P(Roll 4-sixes on 4 rolls of a die)

$= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{1296} \approx 0.00077$

Revisiting Chevalier de Mere's Bets:

Roll 4 Dice and get at least one

$S = \{1111, 1112, 1113, \ldots 6665, 6666\}$

$|S| = 6^4 = 1296$ outcomes.

$E = \{\text{at least one 6}\}$

$P(E) = 1 - P(\neg E) = 1 - P(\{\text{no 6}\})$

$= 1 - P(\{1111, 1112, \ldots, 5554, 5555\})$

$= 1 - \frac{|\{\text{no 6}\}|}{|S|} = 1 - \frac{5^4}{6^4}$

$= 1 - \frac{625}{1296} = 0.5177$

OR:

$= 1 - \underbrace{\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)}$

dont roll a 6 in all 4 attempts

$= 1 - (5/6)^4 = 0.5177$ (Winning Bet)

Roll 2 Dice 24 times and get at least one:

E_i = roll double six on ith roll

$P(E) = 1 - P(\neg E) = 1 - P(\neg E_1) \cdot P(\neg E_2) \cdot P(\neg E_3) \cdots P(\neg E_{24})$

$P(\geq 1 \text{ double six}) = 1 - P(0 \text{ double six}) = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914$ (Losing Bet)

Conditional Probability

We want to know the probability of one event given that another event has occurred:

$P(\text{lung cancer})$: overall liklihood of having lung cancer in the general population.

But we want to try to understand the cause or factors that contribute to Risk:

$P(\text{lung cancer} \mid \text{smoke})$ = Probability of getting lung cancer given that person also smokes.

Studies have shown:

$$P(\text{lung cancer} \mid \text{smoke}) > P(\text{lung cancer}) > P(\text{lung cancer} \mid \text{don't smoke})$$

For non-independent events:

$$\boxed{P(E \cap F) = P(E) \cdot P(F \mid E)}$$

For independent events we had

$$P(E \cap F) = P(E) \cdot P(F)$$

∴ Saying that Events E and F are independent means E has no impact on F, or:

$$\boxed{P(F \mid E) = P(F)}$$

Example: Draw two spades from a 52 card deck.

E = draw spade on 1st draw
F = draw spade on 2nd draw

<u>With replacement</u>: Draws are independent

$$P(E \cap F) = P(E) \cdot P(F) = \underset{13/52}{0.25} \times \underset{13/52}{0.25} = 0.0625$$

<u>Without replacement</u>. 2nd draw affected by 1st

$$P(E \cap F) = P(E) \cdot P(F|E)$$

$$= \frac{13}{52} \times \frac{12}{51} \approx 0.0588$$

Again, we can think of this as a counting problem:

2-card hands: $\binom{52}{2}$

2-spade combinations $\binom{13}{2}$

$$P(E \wedge F) = \binom{13}{2} \Big/ \binom{52}{2} = \frac{13 \cdot 12}{52 \cdot 51} \approx 0.0588$$

Revisiting:

Independent -vs- Non Independent Events.

Independence means that the outcome of one event doesn't affect the outcome of another.

Independent

- Separate rolls of dice
- Separate coin flips
- Card draw with replacement
- Weather on Mars -vs- Weather on Venus
- Two unrelated people's birthday

$P(E \cap F) = P(E) \cdot P(F)$

E and F are distinct
E has no affect on F and vice versa

Not Independent

E(roll is even), F(roll > 3)
(If I know outcome is even, liklihood of outcome >3 increases $\frac{1}{2} \to \frac{2}{3}$)

Card draw without replacement

Weather in Boston today
Weather in Boston tomorrow

△ stock price of two stocks
(This might be hard to find).

$P(E \cap F) \neq P(E) \cdot P(F)$

$= P(E) \cdot P(F \mid E)$

↑ conditional probability

E 'informs' F
E influences F
Possible causal connection

Are Two events Independent?

Sometimes its unclear; so we calculate:

If $P(E) \cdot P(F) = P(E \cap F)$

Then independent, otherwise not.

Roll Two Dice

E = roll doubles
F = roll at least 1 six

$P(E) = 1/6$ (6 possible doubles / 36 outcomes)

$P(F) = 1 - P(\neg F) = 1 - P(\text{no sixes}) = 1 - \left(\frac{5}{6}\right)^2 = \frac{36}{36} - \frac{25}{36} = \frac{11}{36}$

$P(E \cap F) = P(\text{double 6}) = \frac{1}{36} \neq \frac{1}{6} \cdot \frac{11}{36} \quad \left(1 \neq \frac{11}{6}\right)$

$\therefore$ not independent. Knowing that we rolled doubles reduces chances of at least 1 six.

But in this example, E and F are independent:

E = roll doubles — $P(E) = 1/6$
F = roll 6 on die #1 — $P(F) = 1/6$

$P(E \cap F) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6}$ ✓ $(\therefore$ independent$)$

Knowing that we rolled doubles doesn't inform us as to the outcome of the 1st die.

Monopoly: Roll at least 1 double in three rolls to get out of jail for free.

What are your chances?

The rolls are <u>independent</u>.

$$P(\geq 1 \text{ double in 3 rolls})$$
$$= 1 - P(\text{no doubles in three rolls})$$
$$= 1 - P(\text{no double on one roll})^3$$
$$= 1 - (1 - P(\text{double on one roll}))^3$$
$$= 1 - (1 - 1/6)^3 = 1 - (5/6)^3 = 1 - \frac{125}{216} = \frac{91}{216} \approx 0.42$$

Full House:

3 cards of 1 Rank

2 cards of a different rank

example: K♡ K◇ K♤ 5◇ 5♧

P(dealt full house) = ?

S = 5 card hands, $|S| = \binom{52}{5} = 2598960$

$|E|$ = # full houses

Pick first rank : 13 $(2, 3, \ldots, 10, J, Q, K, A)$

Pick 2nd rank : 12

Pick suits of 1st Rank: $\binom{4}{3} = 4$

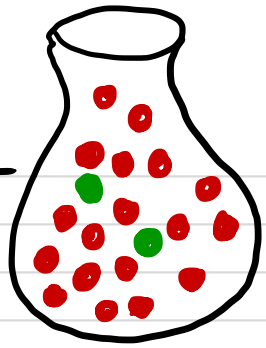
Pick suits of 2nd Rank $\binom{4}{2} = 6$

$3744 = 13 \cdot 12 \cdot 4 \cdot 6$

$\therefore P(E) = |E| / |S| \simeq 0.0014$

(0.14%)

Colored Balls in an Urn



20 balls
18 red
2 green

With replacement: After each draw, we put the ball back (we might choose it again)

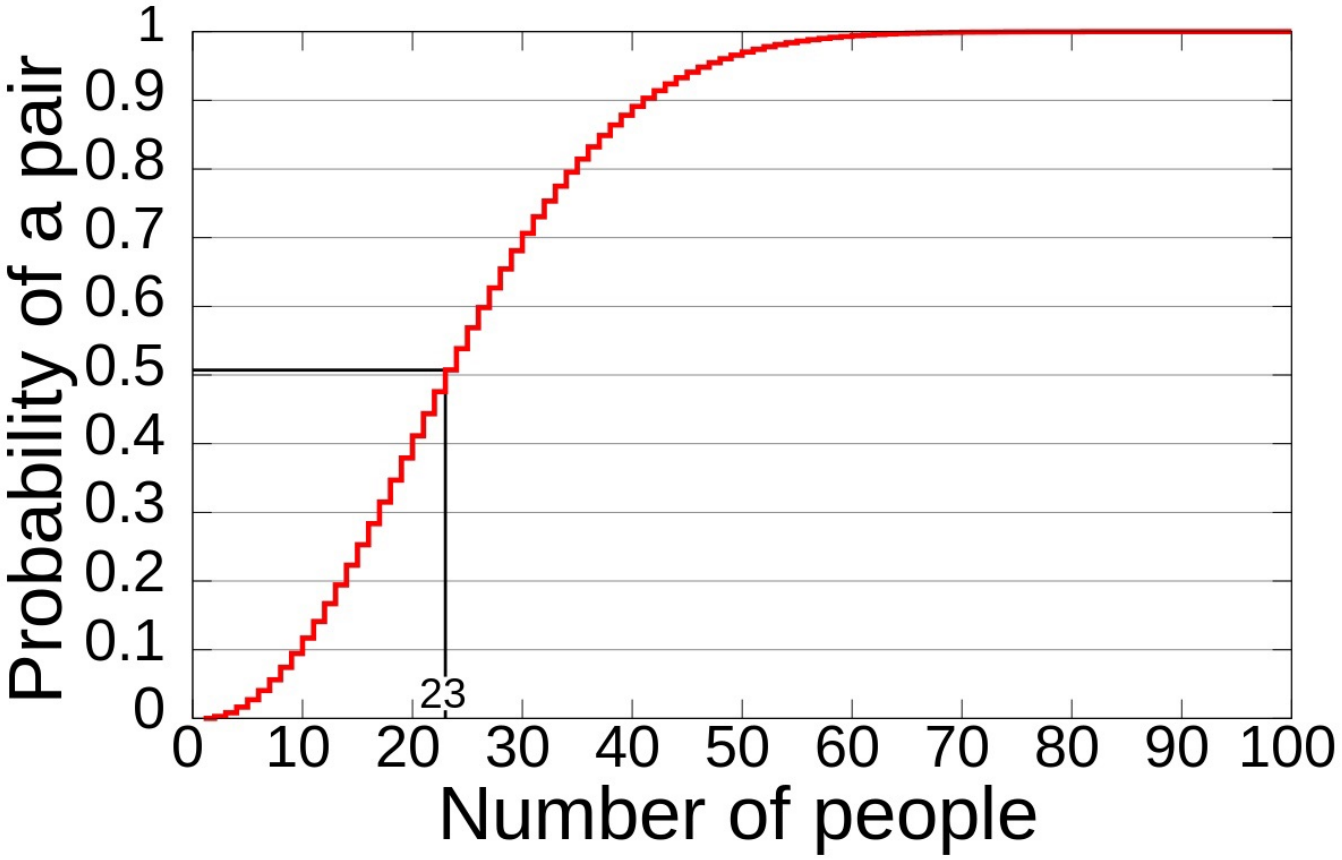
Without replacement: Once drawn, a ball is kept out of the urn, and can't be.

Pick 3 balls.

n_{Green}	With Replacement $S = 20 \times 20 \times 20$ $\quad |S| = 8000$	Without Replacement $|S| = \binom{20}{3} = 1140$
0	$\frac{18}{20}\frac{18}{20}\frac{18}{20} = 0.729$	$\binom{18}{3} / \binom{20}{3} = .716$ OR: $\frac{18}{20} \cdot \frac{17}{19} \cdot \frac{16}{18}$
1	R R G G-Location $\frac{18 \times 18 \times 2 \times 3}{8000} = 0.243$	$\binom{18}{2}\binom{2}{1} / \binom{20}{3} = 0.268$
2.	R G G R-location $\frac{18 \times 2 \times 2 \times 3}{8000} = 0.027$	$\binom{18}{1}\binom{2}{2} / \binom{20}{3} = \frac{18}{1140} = 0.016$
3.	$\frac{2}{20} \cdot \frac{2}{20} \cdot \frac{2}{20} = \frac{8}{8000} = 0.001$	$\emptyset$ (only two greens)

OPTIONAL

The birthday paradox: If $n > 23$ people in a room, chances are at least two people have the same birthday.



It seems counter-intuitive that it only requires 23 people. Most people are thinking: "How many people are needed to make it likly someone has the same B-Day as me"

Think of calendar days being filled gradually:

OPTIONAL

Let P_k = probability of <u>no</u> common birthdays with k random people in the room.

Then:

$$\left(\frac{365}{365}\right) = 1.0$$ (only one person, so there are 365/365 choices)

$$P_2 = \left(\frac{365}{365}\right) \cdot \left(\frac{364}{365}\right)$$

↑ 2nd person has 364/365 choices to ensure no common birthday

$$P_3 = \left(\frac{365}{365}\right) \cdot \left(\frac{364}{365}\right) \cdot \left(\frac{363}{365}\right)$$

↑ 3rd Person has 363 choices

$$P_k = \left(\frac{365}{365}\right) \cdot \left(\frac{364}{365}\right) \cdot \left(\frac{363}{365}\right) \cdots \left(\frac{365-k+1}{365}\right)$$

$$= \frac{{}_{365}P_k}{365^k}$$ ← k-permutation

e.g. $P_{23} = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right) \cdots \left(\frac{343}{365}\right) \approx 0.4927 < 50\%$

→

Probability Examples

1. Roll 3 6-sided dice.

$P(\Sigma = 18) = ?$ $\quad E = \{(6,6,6)\} = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{6^3} = \frac{1}{216}$

Note: $|E| = 1$, $|S| = 6^3$ and all outcomes are equally likely so $P(E) = |E| / |S|$

2. $P(\Sigma = 17) = ?$

$E = \{(5,6,6), (6,5,6), (6,6,5)\}$

$|E| = 3$

$\therefore P = 3/216 = 1/72$

3. $P(\text{Roll} = \{1,2,3\})$ (in any order)

$|E| = 3! = 6 \quad \therefore P = 6/216 = 1/36$

4. Dealt 2 cards in range 2..10 ?

$|S| = \binom{52}{2}$ (# ways to be dealt 2 cards)

$|E| = \binom{36}{2}$ 9 ranks (2..10) $\times$ 4 suits/rank $= 36$

$\therefore |E|/|S| = \binom{36}{2} / \binom{52}{2} \approx 0.475$

=

OPTIONAL

5. Deck with 7 cards: A B C 1 2 3 4

After shuffling, letters and numbers are in order but not necessarily contiguous.

Valid: A B 1 2 C 3 4 ⇒ L L N N L N N

1 2 3 A 4 B C ⇒ N N N L N L L

There are 7! orderings.

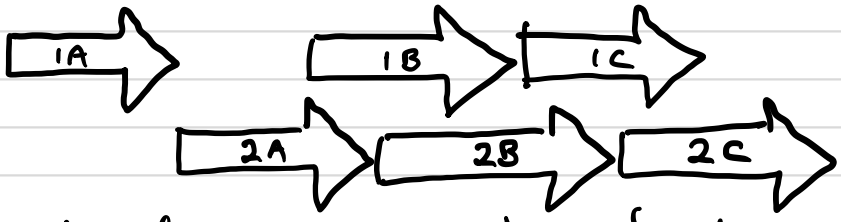
The number of valid possibilities = # of choices for where the letters (or numbers) go:

$$\binom{7}{3} \equiv \binom{7}{4}$$

$$\therefore P = \frac{\binom{7}{3}}{7!} = \frac{\frac{7!}{3!\,4!}}{7!} = \frac{1}{3!\,4!}$$

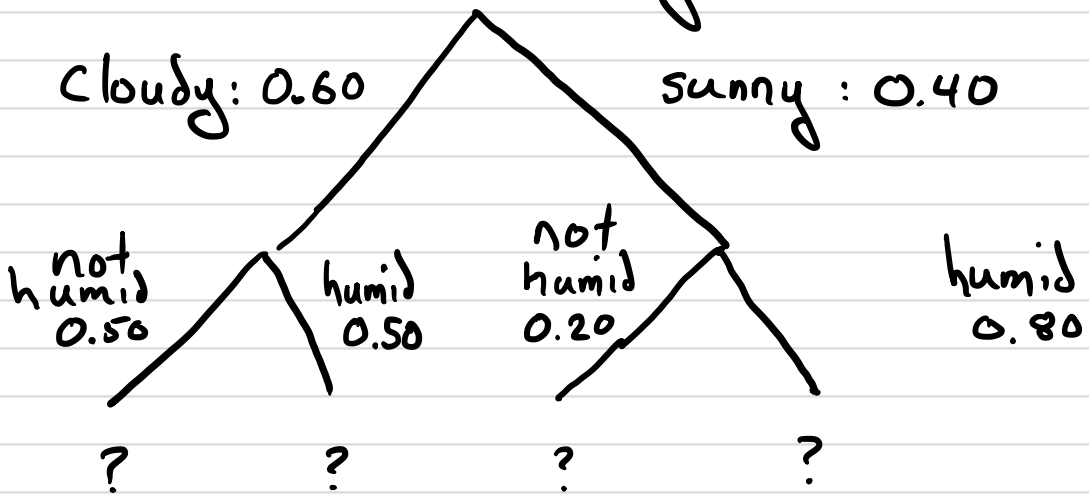
Question: How would you extend this to 3 classes of cards??

This has a real world application: A database must interleave operations of two transactions while maintaining the relative order of each transaction's operations.



There are, however, additional constraints that further restrict transactional concurrency.

Conditional Probability and Weather



$$P(\text{humid}) = P(\text{cloudy} \cap \text{humid}) + P(\text{sunny} \cap \text{humid})$$

$$= P(\text{humid} \mid \text{cloudy}) \cdot P(\text{cloudy}) + P(\text{humid} \mid \text{sunny}) \cdot P(\text{sunny})$$

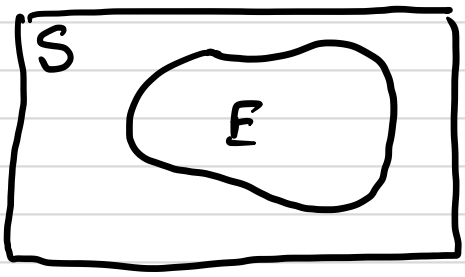
$$= (.50)(.60) + (.80)(.40)$$

$$= .30 + .32$$

$$= .62 \quad (62\% \text{ of the time})$$

Conditional Probability: Interpretation

$$P(E) = |E| / |S|$$



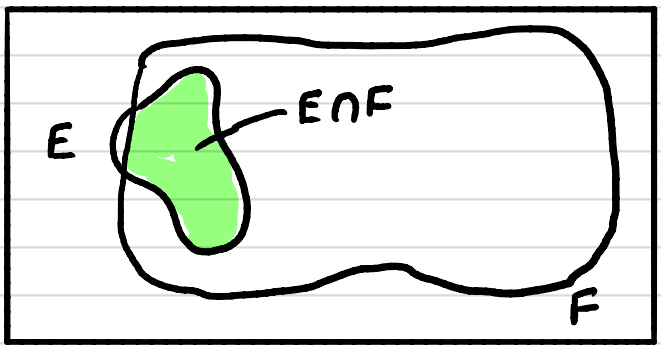
With dependent events E, F:

$$P(E \cap F) = P(E) \cdot P(F|E)$$

OR

$$P(F|E) = \frac{P(E \cap F)}{P(E)}$$

$$= |E \cap F| / |E|$$



Read: "what fraction of E is also in F?

$P(F|E) \sim 1.0$

$P(E|F) \ll 1$

It follows that $P(E|F) = P(E \cap F) / P(F) = \frac{|E \cap F|}{|F|}$

And therefore:

$$P(F|E) \cdot P(E) = P(E|F) \cdot P(F)$$

So $\boxed{P(F|E) = \frac{P(E|F) \cdot P(F)}{P(E)}}$ } Bayes' Theorem (LATER!)

Bayes' Theorem / or Bayes' Rule

- The reverend Thomas Bayes (1702-1761) was an English minister
- Famous for an essay published posthumously by the Royal Society of London concerning probability theory.
- Lead to the entire field of Bayesian Statistics, and two schools of thought:

"Frequentist": Probability as frequency of occurrence

⇓

"Bayesian": Probability as a statement about the strength of a belief.

- Applications involving Bayes Theorem are geared towards trying to determine the best (most likely) hypothesis to explain some observation or evidence.

Application	Evidence	Hypothesis
Medicine	symptoms diagnostic tests	disease or condition
Spam Detection	words, phrases, addresses	spam or not spam?
Speech Recognition	Sounds	Words
Self Driving Cars	Images, Video Signal	Intersection ahead
Criminal Justice, e.g CODIS: FBI's Combined DNA Index System	DNA ~ a particular pattern of repeats at multiple chromosomal loci	He's the killer!

The main idea of Bayes Theorem is that:

$$P(\text{hypothesis}|\text{evidence}) \propto P(\text{hypothesis}) \times P(\text{evidence}|\text{hypothesis})$$

↑ "posterior probability"

↑ "prior probability" "Priors"

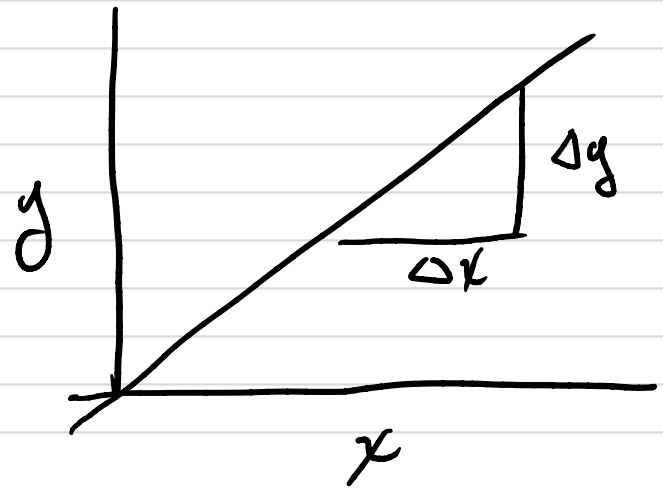
↑ "conditional probability" likelihood of observing evidence given that hypothesis is true

OR:

$$P(\text{hypothesis}|\text{evidence}) = k \cdot P(\text{hypothesis}) \cdot P(\text{evidence}|\text{hypothesis})$$

↑ constant of proportionality

Proportional Relationship



$$k = \frac{y}{x}$$

$$y = kx$$

$$y \propto x$$

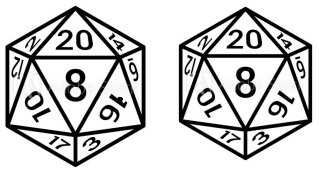
OPTIONAL

Example:

I have a bag containing

1 six-sided dice:

2 twenty-sided dice:



I pick one die at random from bag and roll a five (the evidence).

Which hypothesis is more likely:

a) I drew a six-sided die

b) I drew a twenty-sided die

$$P(\text{six-sided}) = 1/3$$
$$P(\text{twenty-sided}) = 2/3$$
} "Priors"

$$P(\text{six-sided} \mid \text{rolled 5}) \propto P(\text{rolled 5} \mid \text{six-sided}) \cdot P(\text{six-sided})$$
$$\propto \frac{1}{6} \cdot \frac{1}{3} = \frac{1}{18} = 0.0556$$

$$P(\text{twenty-sided} \mid \text{rolled 5}) \propto P(\text{rolled 5} \mid \text{twenty-sided}) \cdot P(\text{twenty-sided})$$
$$\propto \frac{1}{20} \cdot \frac{2}{3} = \frac{2}{60} = \frac{1}{30} = 0.033$$

Its more likely I drew a six-sided die.

Note: This is <u>not</u> the probability of drawing one die or another, it only tells us which is more likely.

Bayes' Theorem

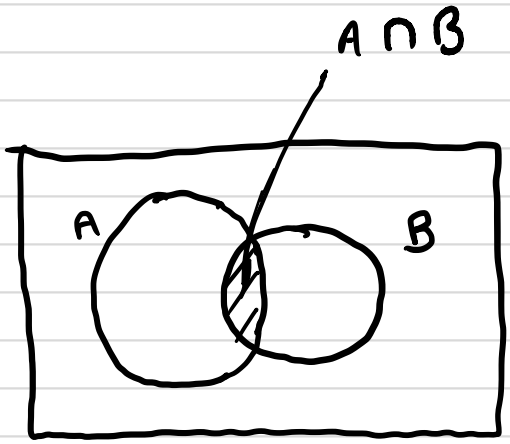
(a) **Simple Form:** $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

This follows from the fact that

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

$$\therefore P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$



Also,

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Here we see $\frac{1}{P(B)}$ is our constant of proportionality mentioned above.

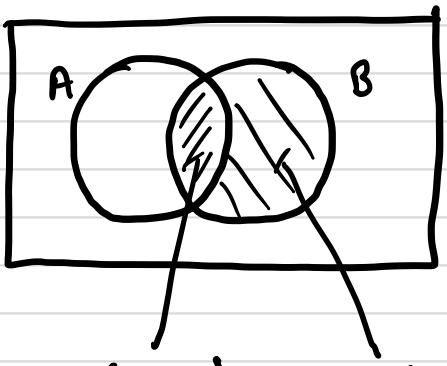
Bayes' Theorem

(b) Explicit Form

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

Since $P(B) = P(B \cap A) + P(B \cap \bar{A})$

$= P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})$



$P(B \cap A) + P(B \cap \bar{A}) = P(B)$

OPTIONAL

(c) General Form

$A_1, A_2, A_3 \ldots A_n$ are a complete and mutually exclusive set of events.

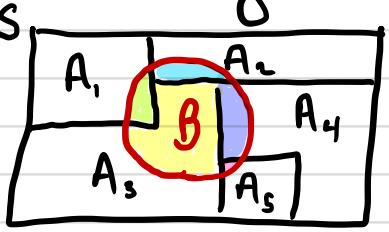
Complete: $A_1 \ldots A_n$ includes all possible outcomes.

mutually exclusive: Every outcome is part of exactly one A_i

$$P(A_k|B) = \frac{P(B|A_k) \cdot P(A_k)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

$P(B|A_k) \cdot P(A_k) = P(A_k \cap B)$

$\sum_i P(B|A_i) \cdot P(A_i) = P(B)$



$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i) \cdot P(A_i)$$

Fair and Loaded Dice

Given two dice

Fair : 1, 2, 3, 4, 5, 6

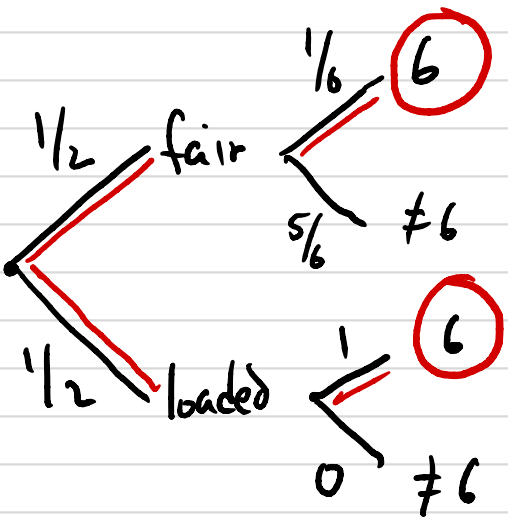
Loaded : 6, 6, 6, 6, 6, 6

Choose die at random. Then roll 6

$P(\text{fair} \mid 6) = ?$

$$P(\text{fair} \mid 6) = \frac{P(\text{fair} \cap 6)}{P(6)}$$

$$= \frac{P(6 \mid F) \cdot P(F)}{P(6)} = \frac{P(6 \mid F) \cdot P(F)}{P(6 \mid F) \cdot P(F) + P(6 \mid \bar{F}) P(\bar{F})}$$

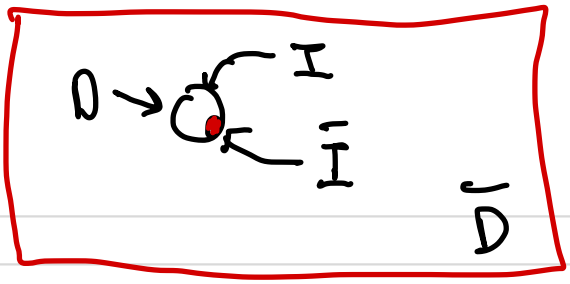


$$= \frac{(1/6)(1/2)}{(1/6) \cdot (1/2) + (1)(1/2)}$$

$$= \frac{\frac{1}{12}}{\frac{1}{12} + \frac{6}{12}} = \frac{1}{7}$$

$$\frac{1/6}{1/6 + 1} = \frac{1/6}{1/6 + 6/6} = \frac{1/6}{7/6} = \frac{1}{7}$$

Prosecutor's Fallacy



In a city of 1 million, detectives arrest a person who fits the description of an eye witness

✓ 1. Only 1:10,000 fit description. $P(D) \ll 1$

TRUE (Given) $P(D) = \frac{1}{10000}$

✓ 2. It is highly unlikly that an innocent person fits description.

TRUE:

$$P(D|I) = \frac{P(D \cap I)}{P(I)} \approx \frac{99}{999\,999} \approx \frac{1}{10\,000} \quad P(D|I) \ll 1$$

since in a city of 1 million we expect

$\frac{1}{10000} \times 1\,000\,000 = 100$ people to fit description

99 are innocent
1 is guilty

(More about expectation later!)

✗ 3. Therefore, highly unlikly that the defendant is innocent ∴ $P(\bar{I}|D) \overset{?}{\approx} 1$

FALSE!

NO!

$$P(I \mid D) = \frac{P(D \mid I) \cdot P(I)}{P(D)}$$

$$= \frac{P(D \mid I) \cdot P(I)}{P(D \mid I) \cdot P(I) + P(D \mid \bar{I}) \cdot P(\bar{I})}$$

$$= \frac{\left(\frac{99}{999999}\right)\left(\frac{999999}{1000000}\right)}{\frac{99}{1000000} + 1 \cdot \frac{1}{1000000}}$$

$$= \frac{\frac{99}{1000000}}{\frac{100}{1000000}} = \frac{99}{100} = 99\%!$$

OPTIONAL

Example:

A medical screening test

Even "accurate" tests can be very misleading, particularly when dealing with (inherently) rare diseases. Bayes' Theorem shows us why:

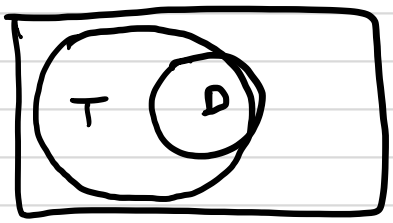
What makes a good diagnostic test?

Sensitivity: Identifies patients that have a disease.

e.g. $P(\text{positive} \mid \text{disease}) \sim 1.0$

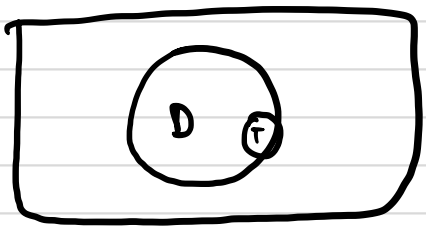
Specificity: Identifies patients that dont have a disease.

e.g $P(\text{negative} \mid \sim\text{disease}) \sim 1.0$



high sensitivity
low specificity

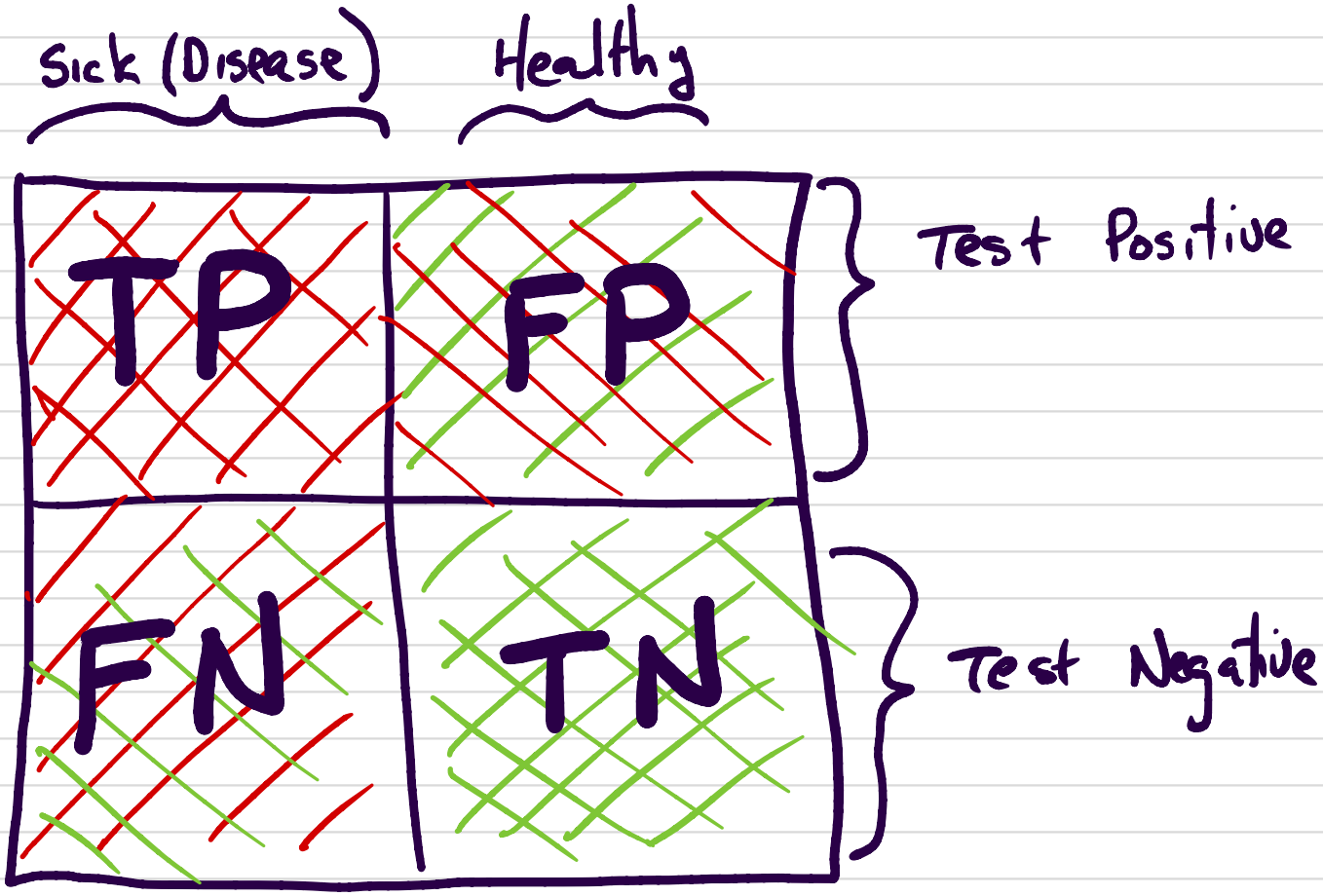
You catch most cases but your false positive rate is high:
Test often says patient has the disease when actually they dont.



low sensitivity
high specificity

You miss a lot of cases, but you can more reliability trust a positive result.
Test says patient doesnt have the disease when really they do!

Medical Testing



Sensitivity: What fraction of sick do you detect?

$$\frac{TP}{TP+FN} \equiv \frac{\text{Positive} \cap \text{Sick}}{\text{Sick}}$$

Specificity: What fraction of healthy do you screen out?

$$\frac{TN}{TN+FP} \equiv \frac{\text{Neg} \cap \text{Healthy}}{\text{Healthy}}.$$

In medical testing we distinguish between:

True Positives (TP): Sick people who test positive for a disease.

False Positives (FP): Healthy people who test positive

True Negatives (TN): Healthy people who test negative

False Negatives (FN): Sick people who test negative.

Then: sensitivity (aka: Recall) $= \frac{TP}{TP + FN}$ } Fraction of sick who test positive

Goal: Minimize False Negatives
DETECTION

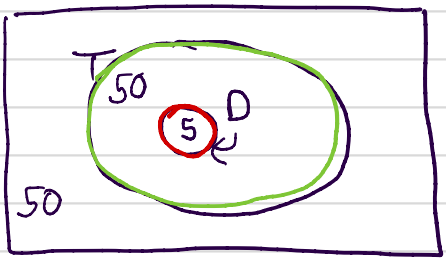
specificity $= \frac{TN}{TN + FP}$ } Fraction of healthy who test negative

Goal: Minimize False Positives
SCREENING

precision $= \frac{TP}{TP + FP}$ } Fraction of Positives who are sick.

Suppose a random sample of n=100 people

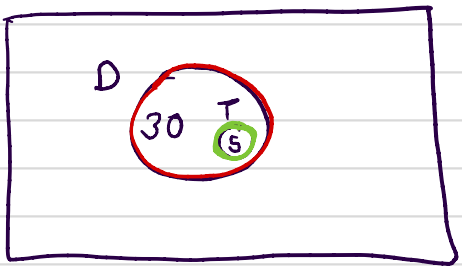
Scenarios



Test casts a wide net. It catches all cases, but many false positives

TP = 5 FP = 45
TN = 50 FN = 0

Sensitivity = $TP/(TP+FN) = 5/(5+0) = 100\%$
Specificity = $TN/(TN+FP) = 50/(50+45) = 53\%$



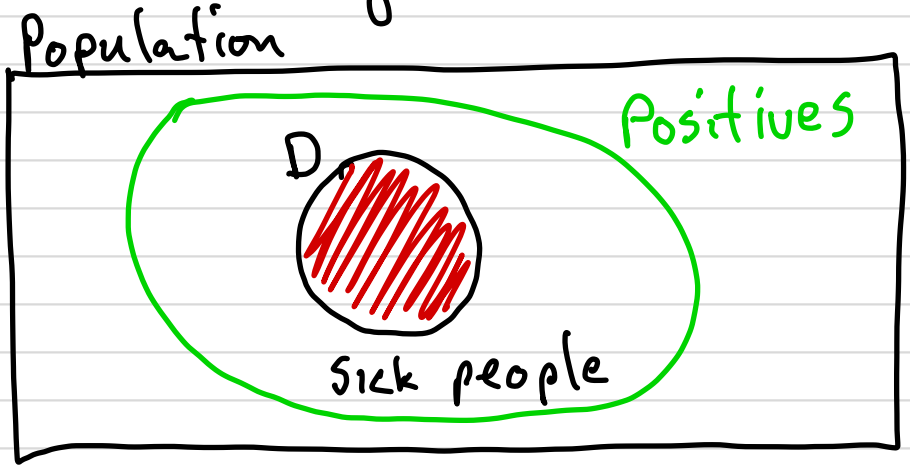
Test is narrowly focussed, identifying only a fraction of diseased patients. But no false positives.

TP = 5 FP = 0
TN = 70 FN = 25

$P(T|D) =$ Sensitivity $= \frac{5}{5+25} = 17\%$

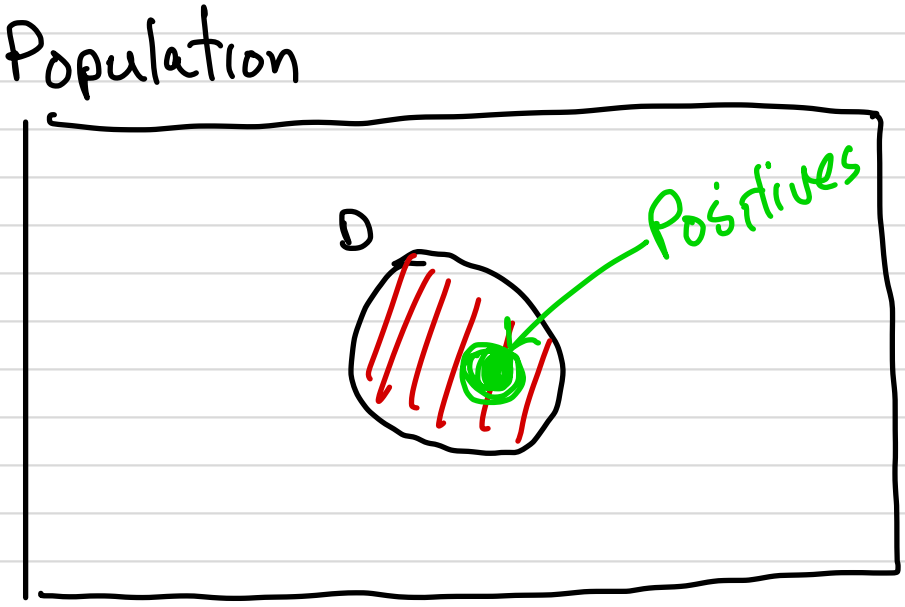
$P(\bar{T}|\bar{D}) =$ Specificity $= \frac{70}{70+0} = 100\%$

Medical Testing Summary



Sensitive ↑ Detect All Sick People

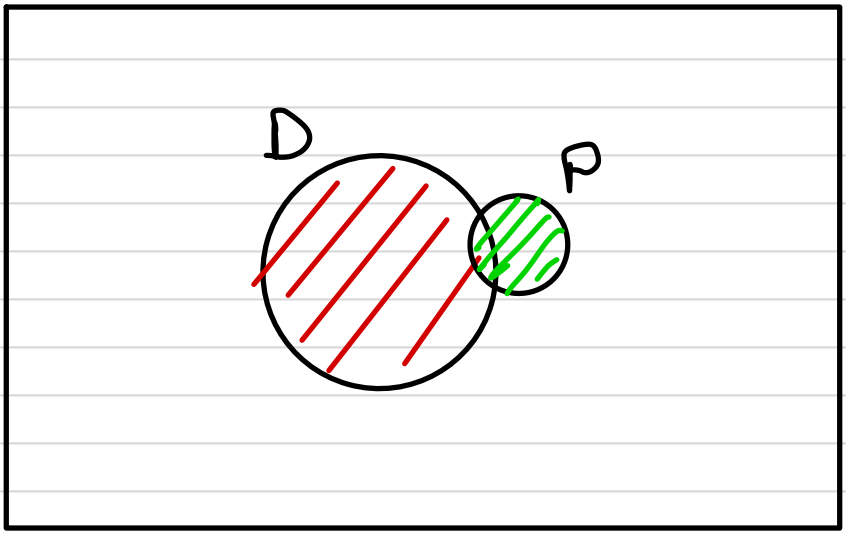
but not specific ↑ Lots of false positives



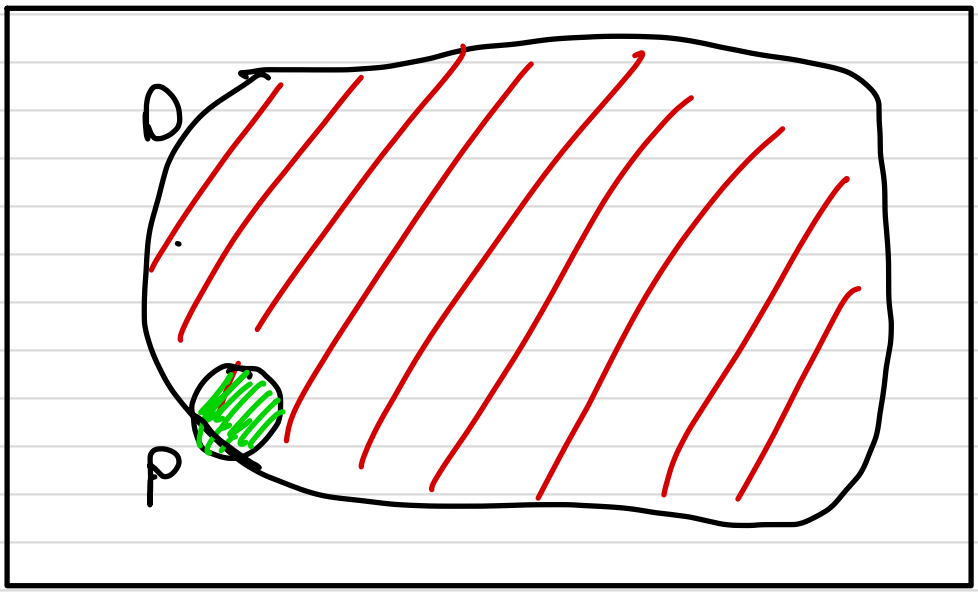
specific ↑ all healthy people test negative

but not very sensitive ↑ many sick people go undetected.

Also Precise: All positives are sick

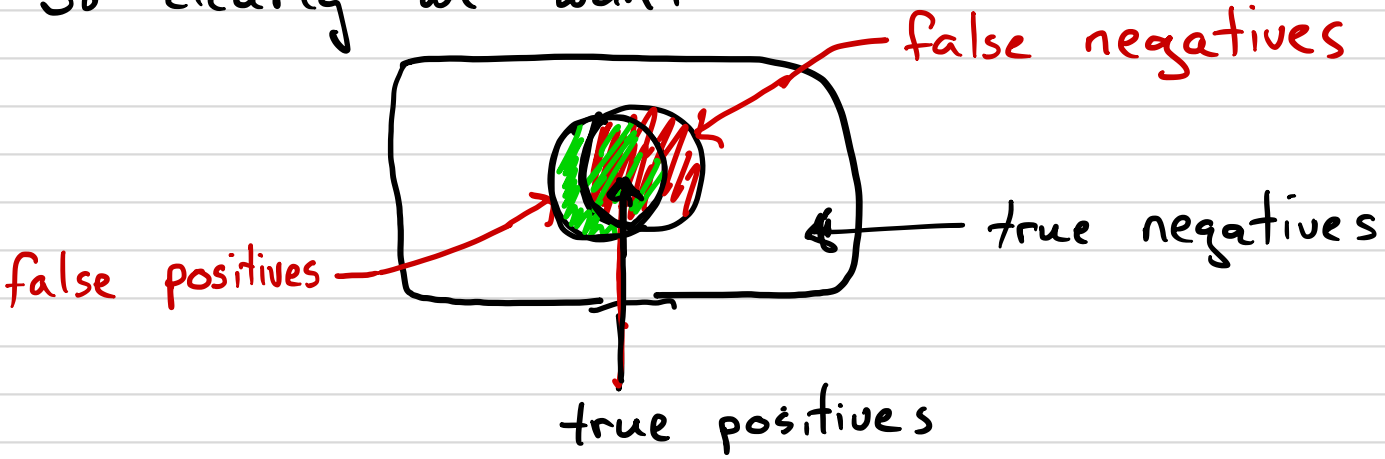


specific: Most Negatives are healthy
but not precise: Most Positives are not sick



precise: Most Positives are sick
but not specific: Most negatives are not heathy.

So clearly we want:



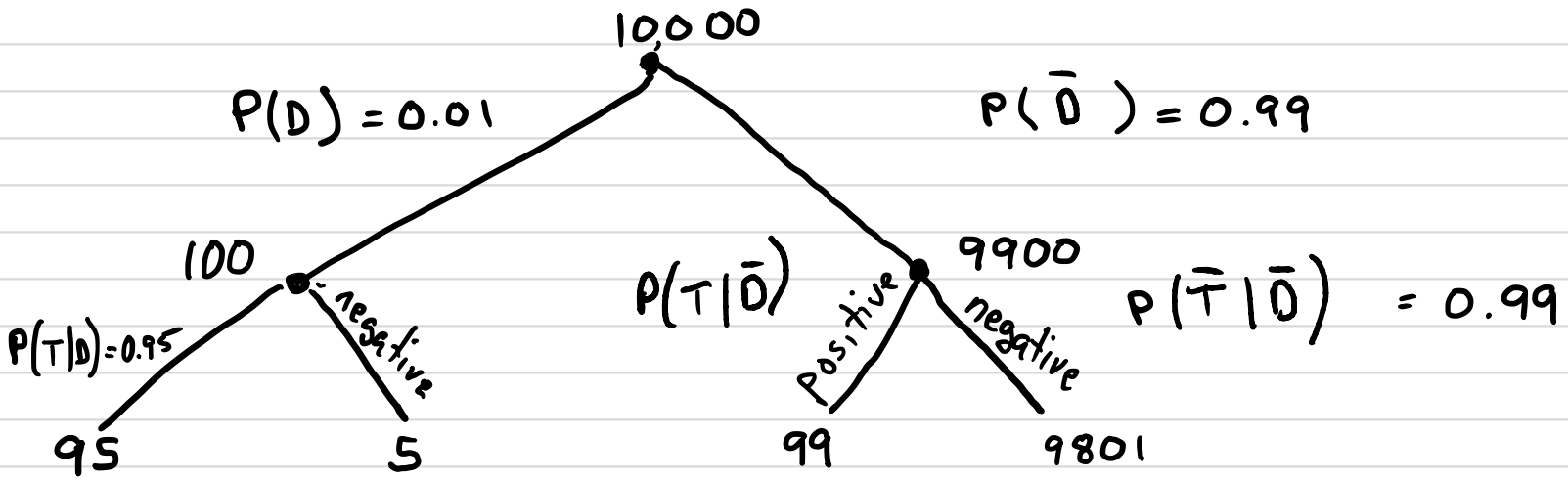
But now consider a relatively rare disease coupled with a reasonably accurate diagnostic test:

$P(D) = 0.01$
$P(T|D) = 0.95$ (sensitivity) $= \frac{P(D \cap T)}{P(D)}$

$P(\bar{T}|\bar{D}) = 0.99$ (specificity) $= \frac{P(\bar{D} \cap \bar{T})}{P(\bar{D})}$

What is $P(D|T)$? i.e, What is the probability that we actually have the disease (the hypothesis) give that we tested positive (evidence).

Imagine 10,000 take the test. What happens?



true positives | false negatives | false positives | true negatives

Corona Virus Test: 30% sensitivity, 90% specific

30 | 70 | 990 | 8910

$$P(D|T) = \frac{\#\text{ true positives}}{\#\text{ positives}}$$

$$= \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}}$$

$$= \frac{95}{95 + 99}$$

$$= 0.4897$$

About half the people tested are false positives!

Formally plugging into Bayes' Theorem: (Explicit Form)

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D)\,P(D)}{P(T)}$$

$$= \frac{P(T|D)\,P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

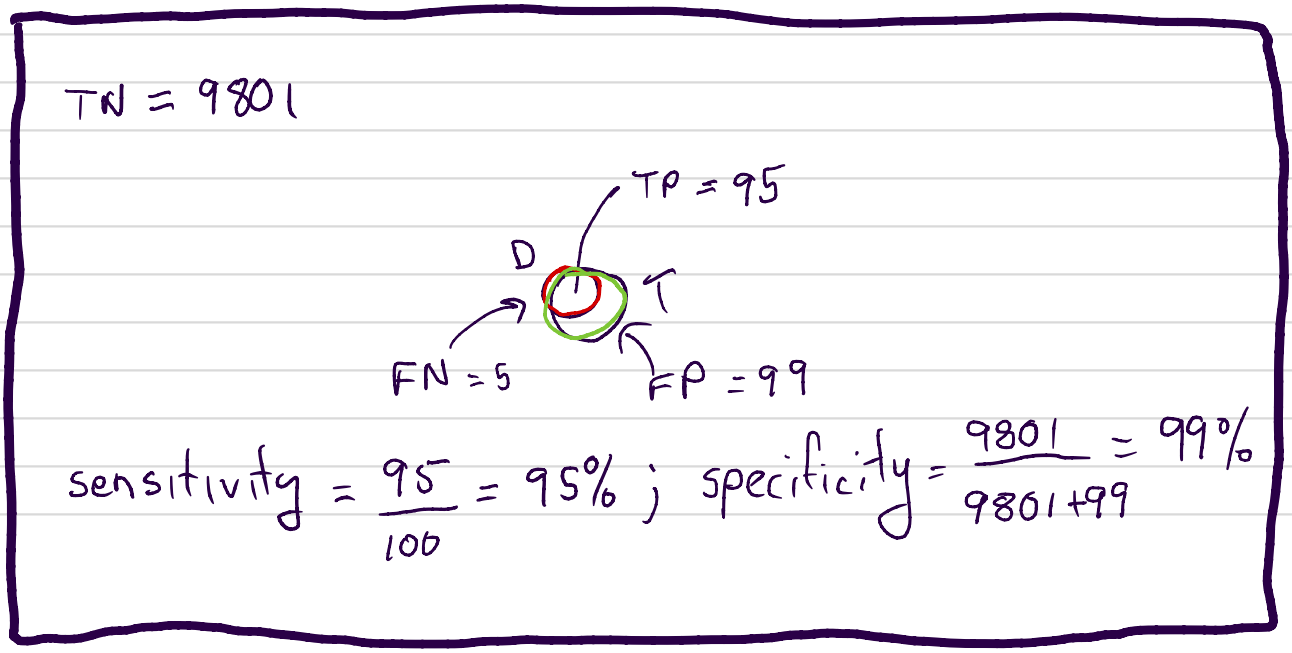
$P(T|D) = 0.95$ = sensitivity

$P(\bar{D}) = 1 - P(D) = 0.99$

$P(T|\bar{D}) = 1 - P(\bar{T}|\bar{D}) = 0.01$ (false positive rate)

$$\therefore\ P(D|T) = \frac{(0.95)(0.01)}{[(0.95)(0.01) + (0.01)(0.99)]} = \frac{0.0095}{0.0095 + 0.0099} = 0.4897$$

The diagram might look like this:



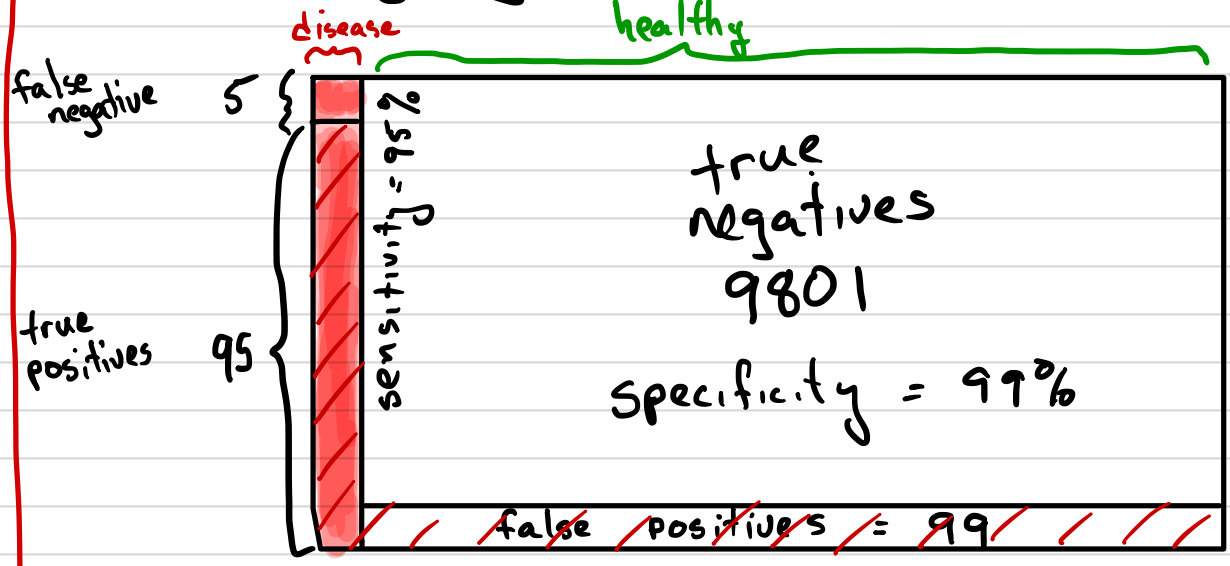
OPTIONAL

If $P(D) = 0.005$ (Even rarer!)

$$P(D|T) = \frac{(0.95)(0.005)}{[(0.95)(0.005) + (0.01)(0.995)} = 0.3231$$

only $1/3$ of the positives are true positives.

What's going wrong?



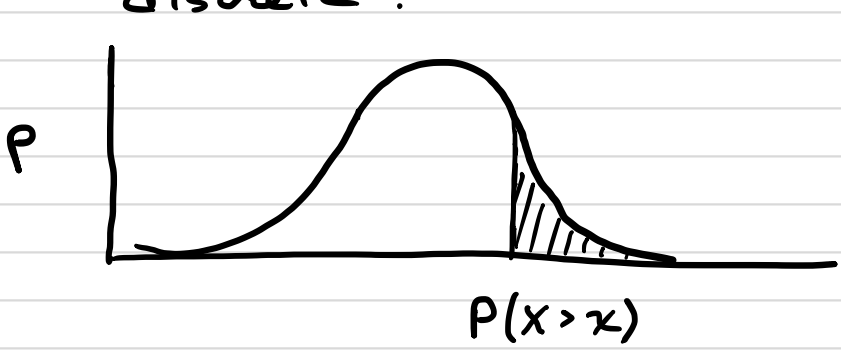
$P(D) = 0.01$
$P(T|D) = 0.95$
$P(\bar{T}|\bar{D}) = 0.99$
$P(D|T) = 0.49$

Even with high sensitivity and specificity, <u>false</u> positives might be a high fraction of the positives.

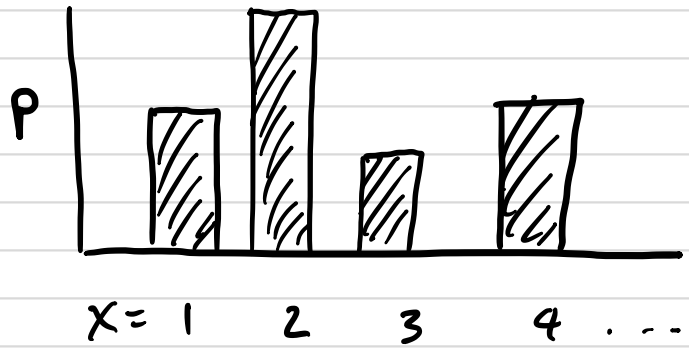
Random Variables.

A random variable, X, is a quantity having various possible outcomes, (its domain), each associated with a particular probability

Random variables may be continuous or discrete:



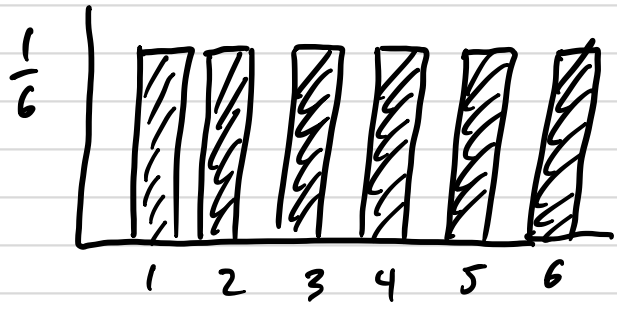
continuous



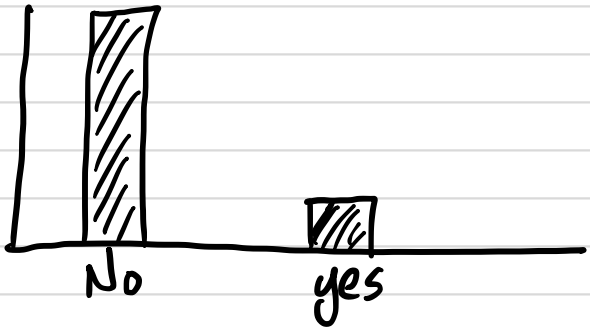
discrete

Examples:

D = roll a die



C = Crash your car



Expectation

$E[X]$ = the expected outcome of a random variable whose outcomes have various probabilities.

$$= \sum_x P(X = x) \cdot x$$, where x are the possible outcomes

D = outcome of a 6-sided die.

$$E[D] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \dots + \frac{1}{6} \cdot 6$$

$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 21/6 = 7/2 = 3.5$$

= average value (all outcomes equally probable)

Lottery (Mega-Millions)

Price of ticket : \$2

Jackpot : \$140 million

Odds : 1 : 300 million

$$E[L] = \$140{,}000{,}000 \cdot \left(\frac{1}{300{,}000{,}000}\right) + \$0 \cdot \left(\frac{299{,}999{,}999}{300{,}000{,}000}\right)$$

$$= \$0.46$$

So at \$2/ticket we are wasting money (Probably!)

Expectation and Insurance

When buying car insurance, should you pay extra for a lower deductable?

Choice	Rate	Deductable
A	$100/mo	$500
B	$80/mo	$1000

My total annual expense depends on liklihood of an accident, let's assume 5%.

A: $E[C] = 0.95(12 \times 100) + 0.05(12 \times 100 + 500)$

$= \$1225$

B: $E[C] = 0.95(12 \times 80) + 0.05(12 \times 80 + 1000)$

$= \$1010$

Usually you are better off accepting the highest deductable you can afford in the unlikly event of an accident.

Linearity of Expectation

Given Random variables X, Y, Z

If $Z = X + Y$

Then $E[Z] = E[X] + E[Y]$

Example:

Roll Two Dice D_1 and D_2

S = sum

$E[S] = E[D_1] + E[D_2]$

$= 3.5 + 3.5 = 7$

The hard way:

S =	1	2	3	4	5	6	7	8	9	10	11	12
N =	$\emptyset$	1	2	3	4	5	6	5	4	3	2	1
	—	1,1	1,2	1,3	1,4	1,5	1,6	2,6	3,6	4,6	5,6	6,6
			2,1	2,2	2,3	2,4	2,5	3,5	4,5	5,5	6,5	
				3,1	3,2	3,3	3,4	4,4	5,4	6,4		
					4,1	2,4	4,3	5,3	6,3			
						1,5	5,2	6,2				
							6,1					

$$E[S] = \frac{1}{36}\cdot 2 + \frac{2}{36}\cdot 3 + \frac{3}{36}\cdot 4 + \frac{4}{36}\cdot 5 + \frac{5}{36}\cdot 6 + \frac{6}{36}\cdot 7 +$$

$$+ \frac{5}{36}\cdot 8 + \frac{4}{36}\cdot 9 + \frac{3}{36}\cdot 10 + \frac{2}{36}\cdot 11 + \frac{1}{36}\cdot 12$$

$= 7$ (Applying Linearity is usually easier.)

Drawing Aces with Expectation/Linearity

A = # Aces in a 5 card hand.

$= 0, 1, 2, 3,$ or 4

$E[A] = ?$

Let A_x = # of Aces in x^{th} card dealt.

So, $A = A_1 + A_2 + A_3 + A_4 + A_5$

$0/1 \quad 0/1 \quad 0/1 \quad 0/1 \quad 0/1$

"Indicator variables": Outcomes are 0 or 1

$$E[A] = E[A_1] + E[A_2] + E[A_3] + E[A_4] + E[A_5]$$

$$= \sum_{x=1}^{5} E[A_x] = 5 \cdot E[A_1]$$

Although the probability of being dealt an A depends on whether we already received an Ace, any given card still has the same probability of being an Ace.

$$= 5\left(1 \cdot \left({}^{4}/_{52}\right) + 0\left({}^{48}/_{52}\right)\right) = \frac{20}{52} \approx 0.385$$

OPTIONAL

Drawing Aces (The hard way)

Again we could have done it the long way, which we show just to emphasize the benefits of leveraging linearization.

Like the urn example with red and green balls, think of the aces as the green balls. We are drawing 5 cards without replacement, so:

$$p(A=0): \binom{48}{5}\binom{4}{0}\Big/\binom{52}{5}$$

$$p(A=1): \binom{48}{4}\binom{4}{1}\Big/\binom{52}{5}$$

$$p(A=2): \binom{48}{3}\binom{4}{2}\Big/\binom{52}{5}$$

$$\vdots$$

etc.

$$E[A] = \sum_{k=0}^{4} p(A=k)\cdot k = \frac{1}{\binom{52}{5}} \sum_{k=0}^{4} \binom{48}{5-k}\binom{4}{k} k$$

$$= \frac{1}{2598960}\left[(1712304)(1)(0) + (194580)(4)(1) + (17296)(6)(2) + (1128)(4)(3) + (48)(1)(4)\right]$$

$$= \frac{1}{2598960}(999600) \simeq 0.385 \quad \text{(as before)}$$

OPTIONAL

Proof of Linearization

Let $X_1, X_2, \ldots X_n$ be a finite collection of discrete random variables, and

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

S = sample space

Prove:

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = E\left[\sum_{i=1}^{n} X_i\right]$$

Proof:

$$E[X] = \sum_{x \in S} X(x)\, P(x) = \sum_{x \in S} \left(X_1(x) + X_2(x) + \cdots + X_n(x)\right) P(x)$$

$$= \sum_{i=1}^{n} \sum_{x \in S} X_i(x)\, P(x)$$

$$= \sum_{i=1}^{n} E[X_i]$$

Handing back Exams.

$n = 45$ students

I give back the exams randomly.

C = # of exams that go to the correct student.

$C = \{0, 1, 2, \ldots, n\}$

$E[C] = ?$ (How many exams would I expect to be returned to the correct student?)

S_1 = student 1 received their actual exam

= 0 (no) or 1 (yes)

$$E[S_1] = 0 \cdot \left(\frac{n-1}{n}\right) + 1\left(\frac{1}{n}\right) = \frac{1}{n}$$

$$E[C] = E[S_1] + E[S_2] + \ldots + E[S_n]$$

But $E[S_1] = E[S_2] = E[S_3] = \ldots$

So $E[C] = n\,E[S_1] = n \cdot \frac{1}{n} = \underline{1}$

On average I expect to return 1 exam to the correct student independent of how many students I have.

Jacob Bernoulli : 1654 - 1705
Uncle of Daniel Bernoulli (1700-1782), "Bernoulli Principle" in physics

OPT

Bernoilli Trials

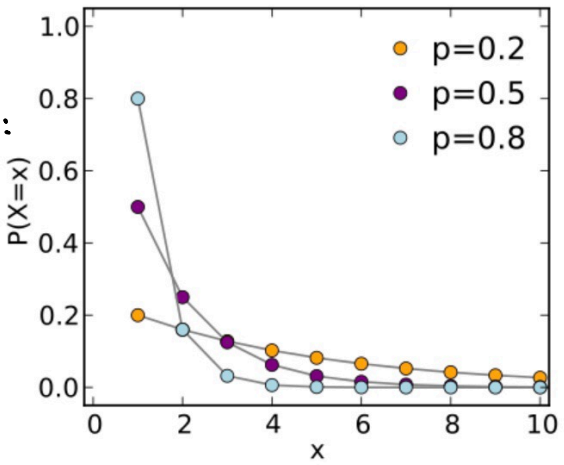
We have a discrete random variable with two possible outcomes : {success, failure}

$P(\text{success}) = p$ { yes , no }

$P(\text{failure}) = (1-p)$ etc.

Probability of success on k^{th} attempt after $k-1$ failures.

If follows a geometric distribution:



Trying to roll a 6 :

$p = 1/6 \quad (1-p) = 5/6$

$k=1$: 6 $\quad (1-p)^0 p^1 = p = 1/6$

$k=2$: x6 $\quad (1-p)^1 p^1 = 5/6 \cdot 1/6 = 5/36$

$k=3$: xx6 $\quad (1-p)^2 p$

$\vdots$

k : $\underbrace{xx\ldots x}_{k-1}6 \quad = (1-p)^{k-1} p$

X = # trials until first success
= 1, 2, 3, ...

$$E[X] = \sum_{k=1}^{\infty} k\,(1-p)^{k-1} p = \frac{1}{p}$$

OPT

Dice and Baseball Cards

If p is fixed, $E[X] = \frac{1}{p}$

e.g., It takes an average of 6 rolls to roll a 6 because

$1/(1/6) = 6$

Collecting n unique baseball cards, we assume all cards are equally probable.

Let X_i = # baseball cards we collect going from i to $i+1$ unique cards

X = total cards collected to go from 0 to n unique cards

By linearity: $E[X] = \sum_{i=0}^{n-1} E[X_i]$

X_i	i_{start}	i_{finish}	p	$E[X_i] = 1/p$
X_0:	0	1	1	1
X_1:	1	2	$(n-1)/n$	$n/(n-1)$
X_2:	2	3	$(n-2)/n$	$n/(n-2)$
X_{n-1}	$n-1$	n	$\frac{1}{n}$	n

harmonic series!

$$E[X] = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=0}^{n-1} \frac{1}{n-i} = n \sum_{i=1}^{n} \frac{1}{i} \approx n \ln n$$

OPT

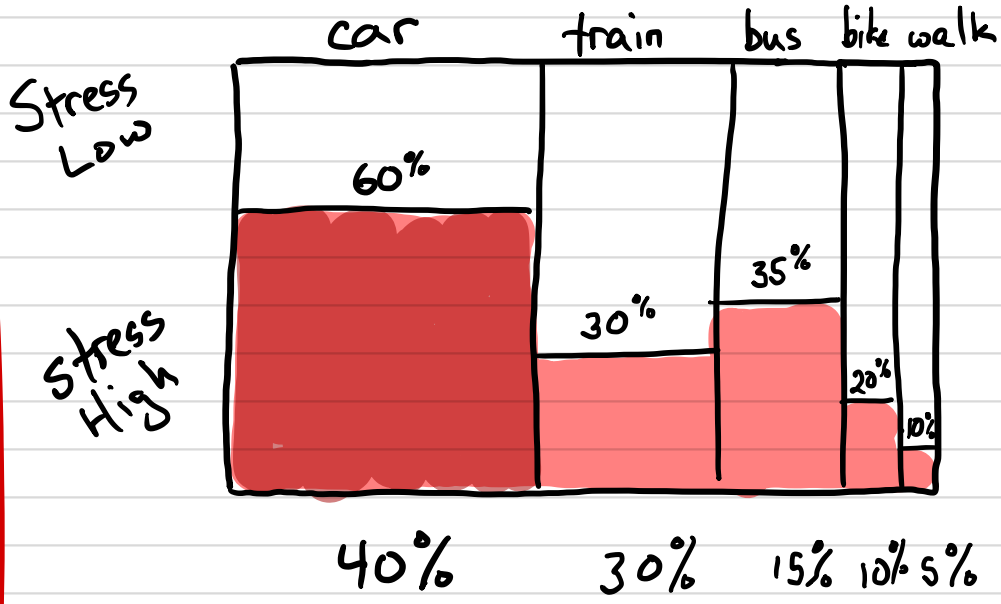
Example

You are studying rates of stress as a function of mode of transport during a commute.

Stress Level : High or Low

Commute Mode : Walk, Bike, Car, Bus, Train.

The picture might look like this:

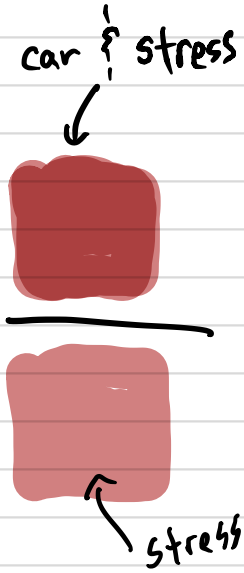


e.g

$P(\text{High} \mid \text{Car}) = 0.60$

$P(\text{Car}) = 0.40$

$$P(\text{Car} \mid \text{High}) = \frac{P(\text{High} \mid \text{Car}) \cdot P(\text{Car})}{\sum_{\text{commute}} P(\text{High} \mid \text{Commute}) \cdot P(\text{Commute})}$$



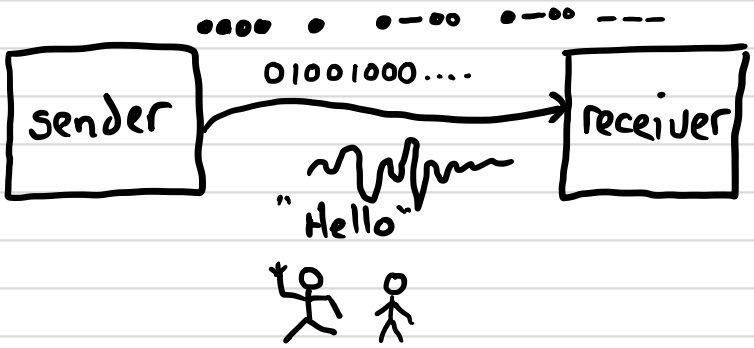
$$= \frac{(0.60)(0.40)}{(0.60)(0.40) + (0.30)(0.30) + (0.35)(0.15) + (0.20)(0.10) + (0.10)(0.05)}$$

$= .24 / .4045 = .59$

OPTIONAL

Information and Signals

Information is a message communicated via some signal: words, sounds, light, bits...



In general, messages with more "information content" are longer: more words, more sounds, more bits, and it takes more time to communicate that message.

shorter = less information

"Its a cold day in Boston today, November 30, 2018"

"Its a cold day in Boston"

"Its cold today!"

"Brrr..."

Computer scientists are very interested in communication
- email, instant messaging, photo sharing, etc.
- network communications between browsers & servers
- electrical signals between RAM, CPU, Disk, I/O

requires communication for a functioning computer.

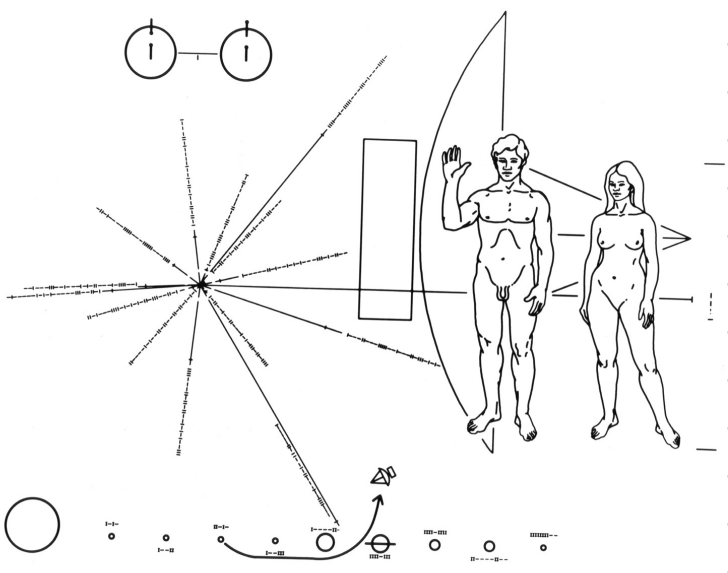
Message Encoding

In a computer, everything ultimately boils down to 1's and 0's, or bits. Binary is used for numbers, letters, sound, pictures, video, everything!

The conversion from message to bits requires a specific encoding. How many bits will we need? It depends on how long our message is, and how we do the encoding.

Dec	Bin	Hex	Char	Dec	Bin	Hex	Char	Dec	Bin	Hex	Char	Dec	Bin	Hex	Char
0	0000 0000	00	[NUL]	32	0010 0000	20	space	64	0100 0000	40	@	96	0110 0000	60	`
1	0000 0001	01	[SOH]	33	0010 0001	21	!	65	0100 0001	41	A	97	0110 0001	61	a
2	0000 0010	02	[STX]	34	0010 0010	22	"	66	0100 0010	42	B	98	0110 0010	62	b
3	0000 0011	03	[ETX]	35	0010 0011	23	#	67	0100 0011	43	C	99	0110 0011	63	c
4	0000 0100	04	[EOT]	36	0010 0100	24	$	68	0100 0100	44	D	100	0110 0100	64	d
5	0000 0101	05	[ENQ]	37	0010 0101	25	%	69	0100 0101	45	E	101	0110 0101	65	e
6	0000 0110	06	[ACK]	38	0010 0110	26	&	70	0100 0110	46	F	102	0110 0110	66	f
7	0000 0111	07	[BEL]	39	0010 0111	27	'	71	0100 0111	47	G	103	0110 0111	67	g
8	0000 1000	08	[BS]	40	0010 1000	28	(	72	0100 1000	48	H	104	0110 1000	68	h
9	0000 1001	09	[TAB]	41	0010 1001	29	)	73	0100 1001	49	I	105	0110 1001	69	i
10	0000 1010	0A	[LF]	42	0010 1010	2A	*	74	0100 1010	4A	J	106	0110 1010	6A	j
11	0000 1011	0B	[VT]	43	0010 1011	2B	+	75	0100 1011	4B	K	107	0110 1011	6B	k
12	0000 1100	0C	[FF]	44	0010 1100	2C	,	76	0100 1100	4C	L	108	0110 1100	6C	l
13	0000 1101	0D	[CR]	45	0010 1101	2D	-	77	0100 1101	4D	M	109	0110 1101	6D	m
14	0000 1110	0E	[SO]	46	0010 1110	2E	.	78	0100 1110	4E	N	110	0110 1110	6E	n
15	0000 1111	0F	[SI]	47	0010 1111	2F	/	79	0100 1111	4F	O	111	0110 1111	6F	o
16	0001 0000	10	[DLE]	48	0011 0000	30	0	80	0101 0000	50	P	112	0111 0000	70	p
17	0001 0001	11	[DC1]	49	0011 0001	31	1	81	0101 0001	51	Q	113	0111 0001	71	q
18	0001 0010	12	[DC2]	50	0011 0010	32	2	82	0101 0010	52	R	114	0111 0010	72	r
19	0001 0011	13	[DC3]	51	0011 0011	33	3	83	0101 0011	53	S	115	0111 0011	73	s
20	0001 0100	14	[DC4]	52	0011 0100	34	4	84	0101 0100	54	T	116	0111 0100	74	t
21	0001 0101	15	[NAK]	53	0011 0101	35	5	85	0101 0101	55	U	117	0111 0101	75	u
22	0001 0110	16	[SYN]	54	0011 0110	36	6	86	0101 0110	56	V	118	0111 0110	76	v
23	0001 0111	17	[ETB]	55	0011 0111	37	7	87	0101 0111	57	W	119	0111 0111	77	w
24	0001 1000	18	[CAN]	56	0011 1000	38	8	88	0101 1000	58	X	120	0111 1000	78	x
25	0001 1001	19	[EM]	57	0011 1001	39	9	89	0101 1001	59	Y	121	0111 1001	79	y
26	0001 1010	1A	[SUB]	58	0011 1010	3A	:	90	0101 1010	5A	Z	122	0111 1010	7A	z
27	0001 1011	1B	[ESC]	59	0011 1011	3B	;	91	0101 1011	5B	[	123	0111 1011	7B	{
28	0001 1100	1C	[FS]	60	0011 1100	3C	<	92	0101 1100	5C	\	124	0111 1100	7C	|
29	0001 1101	1D	[GS]	61	0011 1101	3D	=	93	0101 1101	5D	]	125	0111 1101	7D	}
30	0001 1110	1E	[RS]	62	0011 1110	3E	>	94	0101 1110	5E	^	126	0111 1110	7E	~
31	0001 1111	1F	[US]	63	0011 1111	3F	?	95	0101 1111	5F	_	127	0111 1111	7F	[DEL]

ASCII Character Encoding



Plaque on the Pioneer Spacecraft

Fixed - vs - Variable Length Encoding.

Fixed-Length: Every code is the same length. e.g, ASCII dedicates 8 bits for each character.

- No ambiguity: we know where each code begins and ends
- e.g HI = 01001000 01001001
  H I

Variable - Length

- More commonly used characters are assigned fewer bits ~ giving us better compression.
- Example

  BANANA

  Code 1: 00 = A 01 = B 10 = N

  010010001000 = (12 bits)

  Code 2: 0 = A 1 = N B = 01

  0101010 = (7 bits) but we've introduced ambiguity

  B B B A ?
  A N B B A ?
  A N B A N A ?

Code 3: 0 = A 10 = N 110 = B

B A N A N A
110 0 10 0 10 0 = (10 bits)

We've compromised:

a) We've saved some bits by assigning common characters to shorter codes

b) Its unambiguous as long as we decode starting at the left and work our way right.

Why did this work? Our code satisfies the prefix property: No code is the beginning (prefix) of any other code, so we always know when to stop one code and begin the next.

Morse Code doesn't have the prefix property So we need time-wasting spaces to disambiguate

E	•	T	—
A	•—	N	—•
R	•—•	K	—•—
L	•—•—	C	—•—•

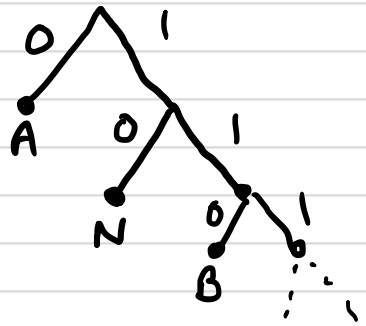
•—•— ⇒ L ?
⇒ A A ?
⇒ E K ?
⇒ E T E T ?

Pauses

S.O.S: ••• ––– ••• (this might come in handy some day - don't forget the pauses!)

Huffman Coding

Codes that obey Prefix property can be modeled as Trees:



0 = A

10 = N

110 = B

$X = x_i \in \{A, B, C, D, E, F, G\}$

A message generated from X: $x_1, x_2, x_3 \ldots x_n$

$X = x$ Character	$X = x$ Frequency	Fixed-width ASCII Encoding (8 bits)	Total Bits
A	16	8 = 01000001	128
B	4	8 = 01000010	32
C	4	8	32
D	2	8	16
E	1	8	8
F	1	8	8
G	4	8 = 01000111	32
Total	32	x 8 =	256 bits

*

Or we could use 3 bits/character and only require 96 bits but

a) This is a non-standard character encoding

b) Even still, we can improve on this

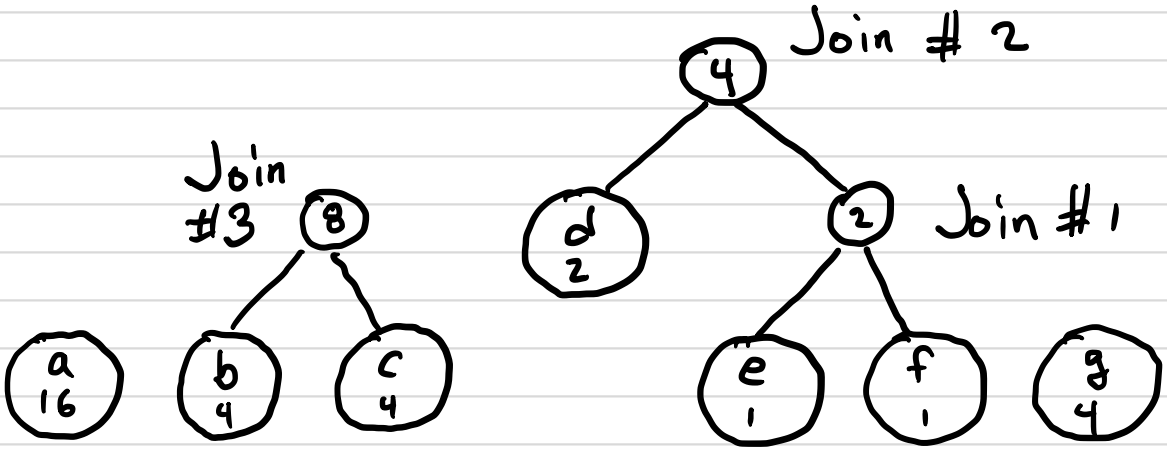
Derivation of Huffman Codes

Note that the codes assigned to each character depends on the character frequency in a particular message. (Different messages ⇒ Different encodings.)

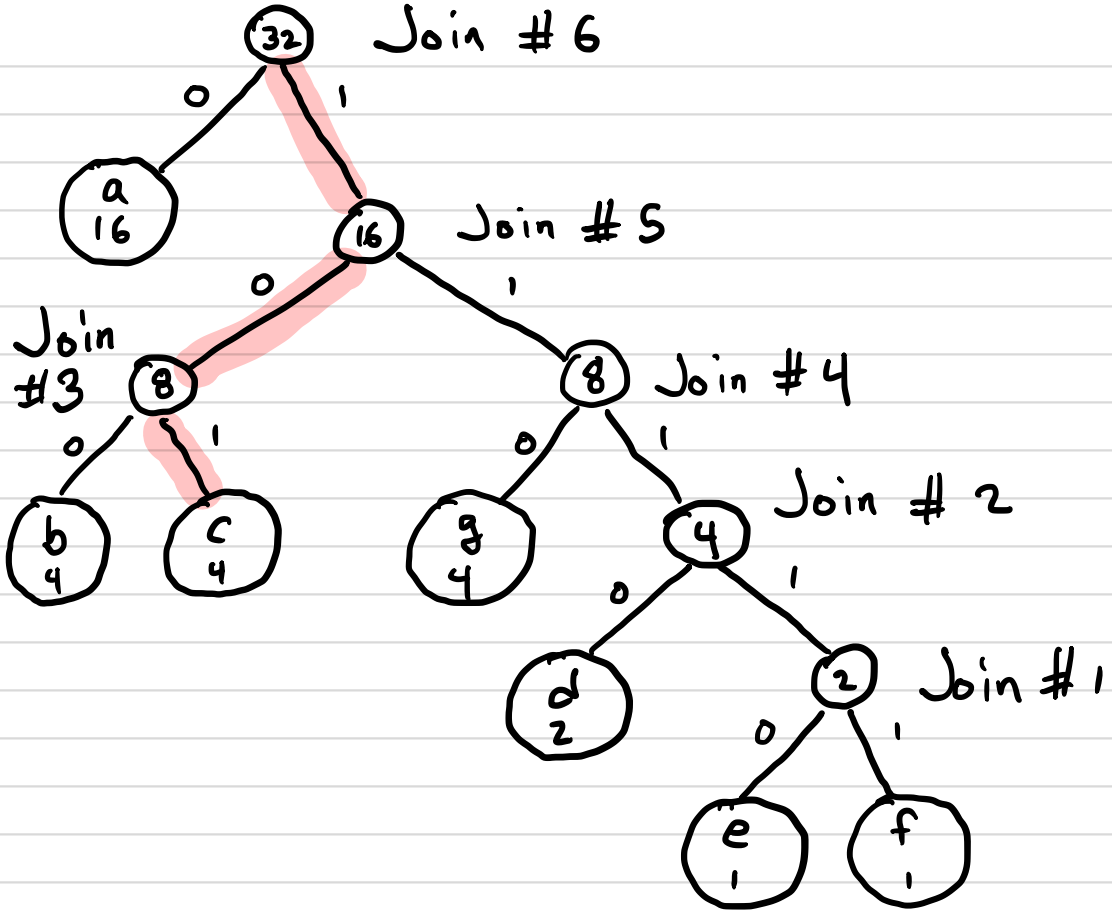
Huffman Code Algorithm builds the tree:

1. Each character is a one-node "tree"

2. Repeatedly join two trees with smallest counts, and Repeat.



3. Completing the tree:



		Fixed		Variable (Huffman) Encoding		
Char	Freq	Ascii Bits/char	Total Bits	Huffman Code	Bits/char	Total Bits
A	16	8	128	0	1	16
B	4	8	32	100	3	12
C	4	8	32	101	3	12
D	2	8	16	1110	4	8
E	1	8	8	11110	5	5
F	1	8	8	11111	5	5
G	4	8	32	110	3	12
	32 char		256 bits			70 bits

Compression Ratio = $\frac{256}{70} = 3.7$ (original message is 3.7x as large as compressed)

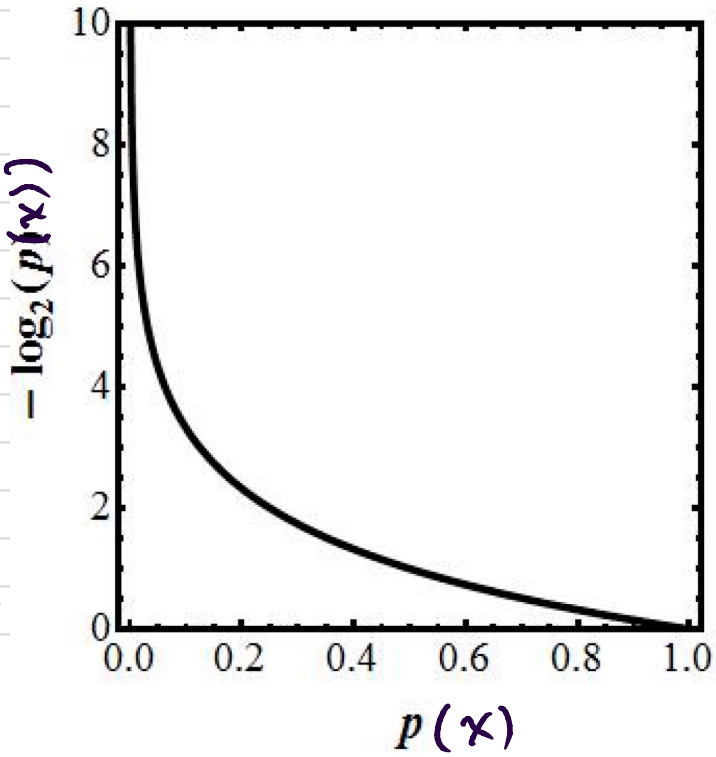
Char	x_i Freq.	Code	$P(x_i)$	# Bits $-\log_2 P$	Total # of Bits
A	16	0	0.500 00	1	16
B	4	1 00	0.12500	3	12
C	4	1 0 1	0.12500	3	12
D	2	1 1 1 0	0.06250	4	8
E	1	1 1 1 1 0	0.03125	5	5
F	1	1 1 1 1 1	0.03125	5	5
G	4	1 1 0	0.12500	3	12
	32 chars				70 bits

In general, when we generate a Huffman tree, the resulting # bits required to encode a character is inversely related to the frequency/liklihood of that character.

$$\#\text{ bits required}(x) \geq \log_2\left(\frac{1}{p(x)}\right)$$

$$= -\log_2(p(x))$$

$$= I(X=x)$$



Self-Information of event $X=x$

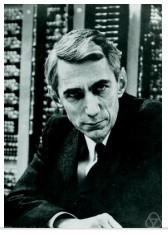
Shannon information

"Suprisal" — a measure of how suprised you are to encounter x in your message.

OPTIONAL

Information Theory

Claude Shannon
(1916-2001)
(Founder of Information Theory)



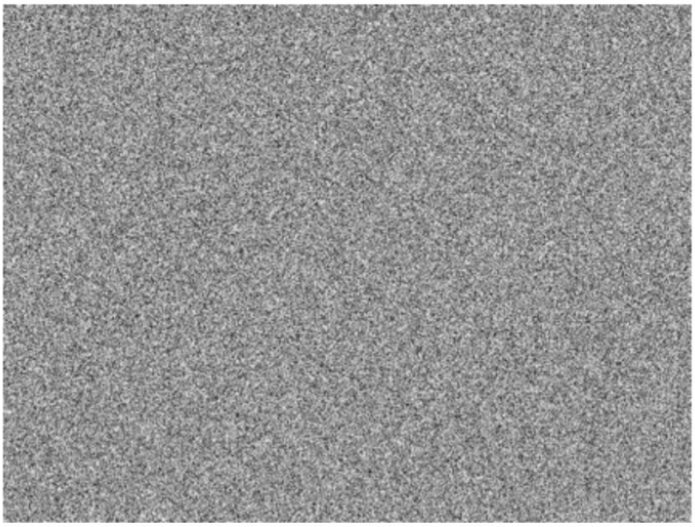
Consider a signal consisting of a stream of symbols, each with output with some probability. So the message is like a discrete random variable X, with symbols $x_1, x_2, \ldots x_n$.

$$I(X = x_i) = -\log_2 P(X = x_i)$$

Entropy is the expected information, and also measures how unpredictable each symbol of the message is as well as how many bits we need to use per symbol on average to encode the stream of symbols.

$$\text{Entropy} = H(X) = E[I(X)] = \sum_{x_i \in X} -P(x_i) \log_2 P(X_i)$$

147,467 bytes



322,147 bytes

More Predictable
Less suprising
Low Entropy
Fewer bits required
Less Information!

More unpredictable
More suprising
High Entropy
More bits required
More information!
(But admittedly less interesting.)

Entropy Examples

(a) Flipping Coins $E = \{H, T\}$

H T T T H T H H T H H H . . . .

$P(H) = 0.5$

$P(T) = 0.5$

$$H(E) = \underbrace{-\frac{1}{2}\log_2\left(\frac{1}{2}\right)}_{\text{Heads}} \underbrace{-\frac{1}{2}\log_2\left(\frac{1}{2}\right)}_{\text{Tails}}$$

$$= -\frac{1}{2}(-1) \; -\frac{1}{2}(-1) = 1 \text{ bit} \quad (H=0, T=1)$$

(b) Rolling Dice

$D = \{1, 2, 3, 4, 5, 6\}$

$P(D = x) = 1/6$

$$H(D) = \sum_{x \in D}\left(-\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right) = 2.58 \text{ bits}$$

↑

2 bits isn't enough: can only represent 4 outcomes

3 bits is too many!

Entropy Examples - continued

(c) 256 equally probable characters.

$$H(X) = \sum_{x \in X} -\frac{1}{256} \log_2 256 = \log_2 256 = 8 \text{ bits.}$$

We need all 8 bits.

(d) Consider our original example with symbols a - g:

Symbol	P
a	.5
b	.125
c	.125
d	.0625
e	.03125
f	.03125
g	.125

$$H(X) = \underbrace{-\frac{1}{2}\log\frac{1}{2}}_{a} \underbrace{- 3\cdot\frac{1}{8}\log\frac{1}{8}}_{b,c,g} \underbrace{- \frac{1}{16}\log\frac{1}{16}}_{d} \underbrace{- 2\cdot\frac{1}{32}\log\frac{1}{32}}_{e,f}$$

$$= -\frac{1}{2}(-1) - \frac{3}{8}(-3) - \frac{1}{16}(-4) - \frac{2}{32}(-5)$$

$$= \frac{1}{2} + \frac{9}{8} + \frac{4}{16} + \frac{10}{32} =$$

$$= \frac{16 + 36 + 8 + 10}{32} = \frac{70}{32} = 2.19 \text{ bits/symbol}$$

(same as our Huffman code.)