HTF: … 2 …
B: Ch 2
RN: Ch 13

Probability 101

Thanks to R Parr, C Guesterin

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Probability: Who needs it?

- Learning without probabilities is possible
  - Version spaces
  - Explanation-based learning

  but rare…
- Learning almost always involves
  - Noise in data (training, testing)
  - Prediction about the future
- Learning systems
  that don't use probability in some way
  tend to be very, very brittle

Probabilities

- Natural way to represent uncertainty
- ∃ intuitive notions about probabilities
  - Many notions are wrong or inconsistent
  - Many people don't get what probabilities mean

⇒ Have FORMAL description,
that is consistent and useful

  - Overall framework is understood
  - Fine details of "meaning" still debated

- Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

- Rank the following by probability
  (1 = most probable; 8 = least probable)
  a. Linda is a teacher in elementary school.
  b. Linda works in a bookstore and takes yoga classes.
  c. Linda is an active feminist.
  d. Linda is psychiatric social worker.
  e. Linda is a member of the League of Women Voters.
  f. Linda is a bank teller.
  g. Linda is an insurance salesperson.
  h. Linda is a bank teller and is an active feminist.

Understanding Probabilities

- Probabilities have dual meanings
  - Relative frequencies (frequentist view)
  - Degree of belief (Bayesian view)
- Neither is entirely satisfying
  - No two events are truly the same (reference class problem)
  - Statements should be grounded in reality in some way

Probability as Relative Frequency?

- What is probability of *event* E ?
- Over long sequence of experiments, ratio of
  - (# of times E occurred)
    number of times *E* occurs in sequence, to
  - (# of trials)
    total number of experiments
- Estimate:
  P(E) ≈ (# of times E occurred) /(# of trials)

- As (# of trials) $\rightarrow \infty$,
  ratio approaches true probability
  - given std assumptions

Examples…

- P(… swimmer succeeds …)
  - Swimmer S …
    - tries **100** times to swim 50’ in 15 secs.
    - succeeds **20** occasions
  - Estimate: probability that
    S can swim 50’ in 15 seconds   is:
    - P(S can swim 50’ in 15 seconds) ≈ 20/100 = 0.2
- For probability to be meaningful, must clearly defined
  - experiments
  - sample space
  - events
- What is the probability of a *nuclear accident*?

Interpretations of probability – A can of worms!

- Frequentists
  - $P(\alpha)$ = the frequency of α in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event "it will rain tomorrow" ? … "nuclear war tomorrow" ?
- Subjective interpretation
  - $P(\alpha)$ = my degree of belief that α will happen
  - Where "degree of belief" means…

    If I say $P(\alpha)=0.8$, then I am willing to bet!!!
- For this class…

  we (mostly) don't care what camp you are in

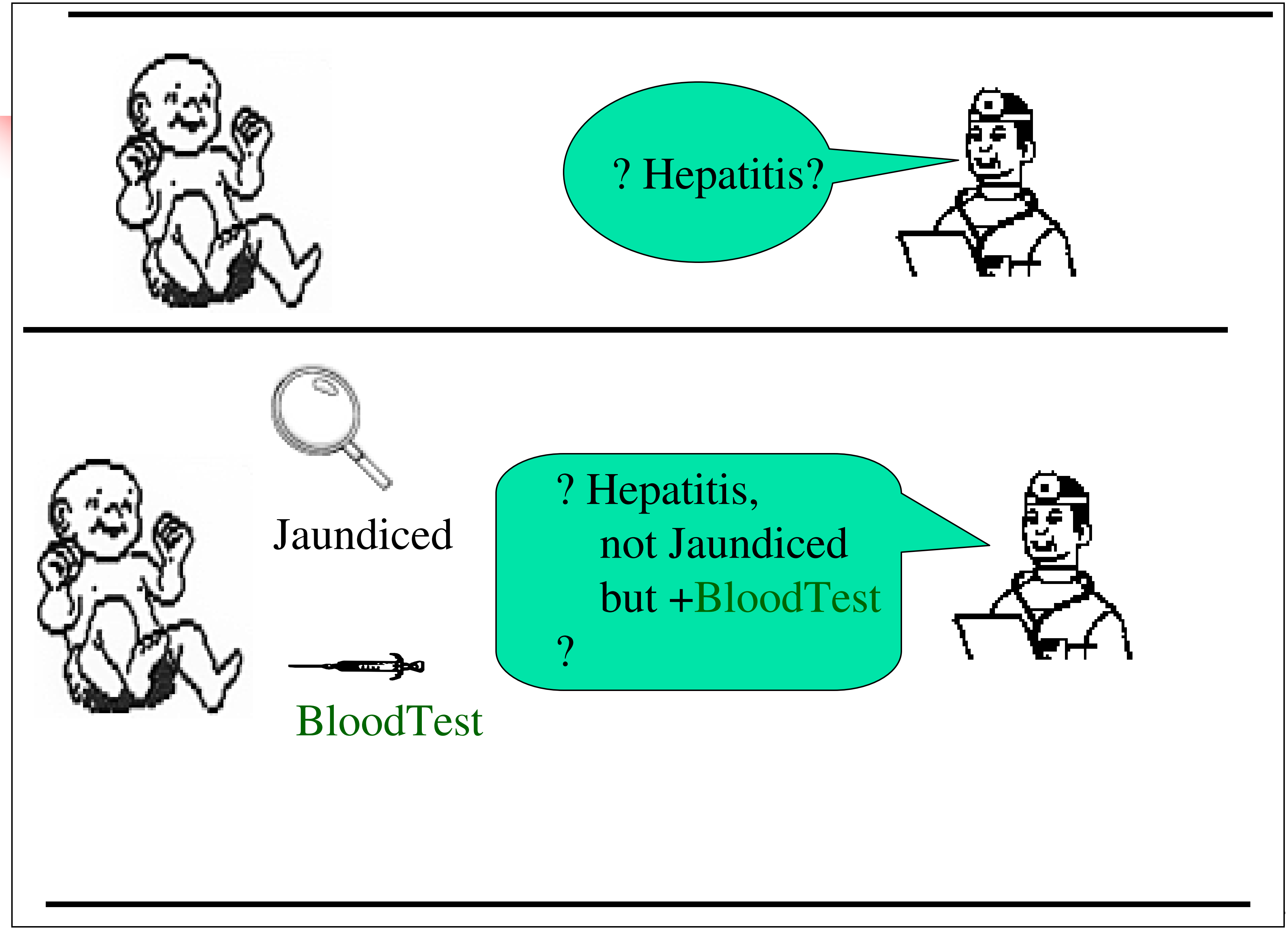
Why Probabilities are Good … despite difficulties

- Subjectivists: *probabilities are degrees of belief*
- Is any *degree of belief* ≡ *probability*?
  - AI has used many notions of belief:
    - Certainty Factors
    - Fuzzy Logic
- NO!!
  - Dutch book
  - If you follow doesn't follow probability theory, you will lose… see below.

Terms from Probability Theory

- **Random Variable**:
  Weather $\in$ { Sunny, Rain, Cloudy, Snow }
- **Domain**: Possible values a random variable can take.
  (... finite set, $\Re$, functions…)
- Probability distribution:
  mapping from domain to values $\in$ [0, 1]
- P(Weather) = $\langle$ 0.7, 0.2, 0.08, 0.02 $\rangle$

  means
  $$\left\{ \begin{array}{ll} P(\text{ Weather = Sunny }) & = 0.7 \\ P(\text{ Weather = Rain }) & = 0.2 \\ P(\text{ Weather = Cloudy }) & = 0.08 \\ P(\text{ Weather = Snow }) & = 0.02 \end{array} \right\}$$
- Event:
  Each assignment (eg, Weather = Rain) is “event”


? Hepatitis?
Jaundiced
BloodTest
? Hepatitis,
not Jaundiced
but +BloodTest
?

Typical Task

- Given observations $\{O_1=v_1, \ldots O_k=v_k\}$

  (J=No, B=Yes [symptoms, history, test results, …])

  what is best DIAGNOSIS Dx_i for patient?

  (Hep=Yes, Hep=No)

- Compute Probabilities of Dx_i

  given observations $\{O_1=v_1, \ldots O_k=v_k\}$

$$P(\text{Dx} = u \mid O_1 = v_1, \ldots, O_k = v_k)$$

General Events

- **Atomic Event**: "Complete specification"
  Conjunction of assignments to EVERY variable [PossibleWorld]
- **Joint Probability Distribution**:
  Probability of every possible atomic event

n binary variables: 2^n entries
($2^n - 1$ independent values, as sum = 1)
A huge table!

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

H	Hepatitis
J	Jaundice
B	(positive) Blood test

Inference by Enumeration

- Using only joint probability distribution:

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

- For any proposition φ, add the atomic events where it is true:

  $$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$$

- $P(+j)$

  $= 0.01455 + 0.038 + 0.00045 + 0.722$

  $= 0.775$

Cost of Marginalization

- Called "marginal"

$$P(X_n) = \sum_{x_1,\ldots,x_{n-1}} P(x_1, \ldots, x_{n-1}, X_n)$$

- To compute marginal distribution P(X_n):
  If all binary, 2^{n-1} additions
  - one term for each value of $x_1, \ldots, x_{n-1}$

H	Hepatitis
J	Jaundice
B	(positive) Blood test

Inference by Enumeration

- Using only joint probability distribution:
- For any proposition φ, add the atomic events where it is true:

$$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega)$$

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

- P(-j v +b)

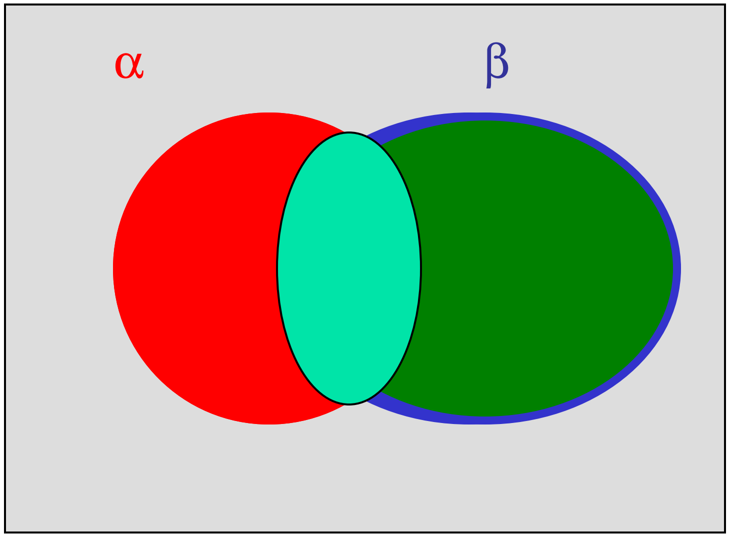
= .03395 + .0095 + .0003 + .1805 + .00045 + .722 = 0.9467

Conditional Probabilities

- After learning that β is true,
  how do we feel about α?
- If roll EVEN, what is chance of rolling 2?
- If have hepatitis, what is chance of jaundice?

β α

$$P(\alpha \mid \beta)$$



Conditional Probability

- Conditional Probability:
  $P(\alpha \mid \beta)$ = Probability of event α,
  given that event β has happened
- P(Jaundice | Hepatitis) = 0.8

- In gen'l:

$$P(\alpha \mid \beta) = \frac{P(\alpha \,\&\, \beta)}{P(\beta)}$$

$$P(\alpha \,\&\, \beta) = P(\alpha \mid \beta)\, P(\beta)$$

Conditional Probability

$$P(\alpha \mid \beta) = \frac{P(\alpha \,\&\, \beta)}{P(\beta)}$$

$$P(\alpha \,\&\, \beta) = P(\alpha \mid \beta)\; P(\beta)$$

- **Unconditional (prior) Probability**:
  - Probability of event before evidence is presented
  - P(Jaundice) = 0.04

    prob that someone (from this population) is jaundiced is 4 in 100
- **Evidence**: Percepts that affects degree of belief in event

- **Conditional (posterior) Probability**:
  - Probability of event after evidence is presented
  - N.b., posterior prob can be COMPLETELY different than prior prob!

H	Hepatitis
J	Jaundice
B	(positive) Blood test

Inference by Enumeration

- Using only joint probability distribution:

J	B	H	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

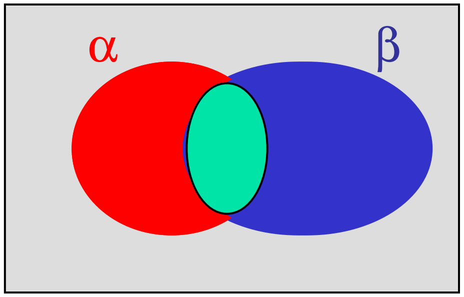
- Can compute *conditional probabilities*:

$$P(-b \mid +j) = \frac{P(-b \wedge +j)}{P(+j)} = \frac{0.01455 + 0.038}{0.01455 + 0.038 + 0.00045 + 0.722}$$

$$\approx 0.0678$$

Useful Rule #1: The chain rule

- $P(\alpha, \beta) = P(\alpha)\, P(\beta | \alpha)$



- More generally:

$$P(\alpha_1, \ldots, \alpha_k) = P(\alpha_1)\, P(\alpha_2 | \alpha_1) \cdots P(\alpha_k | \alpha_1, \ldots, \alpha_{k-1})$$

  - … any order …
    $P(\alpha_1, \ldots, \alpha_k) = P(\alpha_3)\, P(\alpha_7 | \alpha_3)\, P(\alpha_{14} | \alpha_3, \alpha_7) \cdots$

Useful Rule #2: Bayes rule

- $$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}$$

- More generally, external event γ:

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$$

Bayes' Rule and its Use

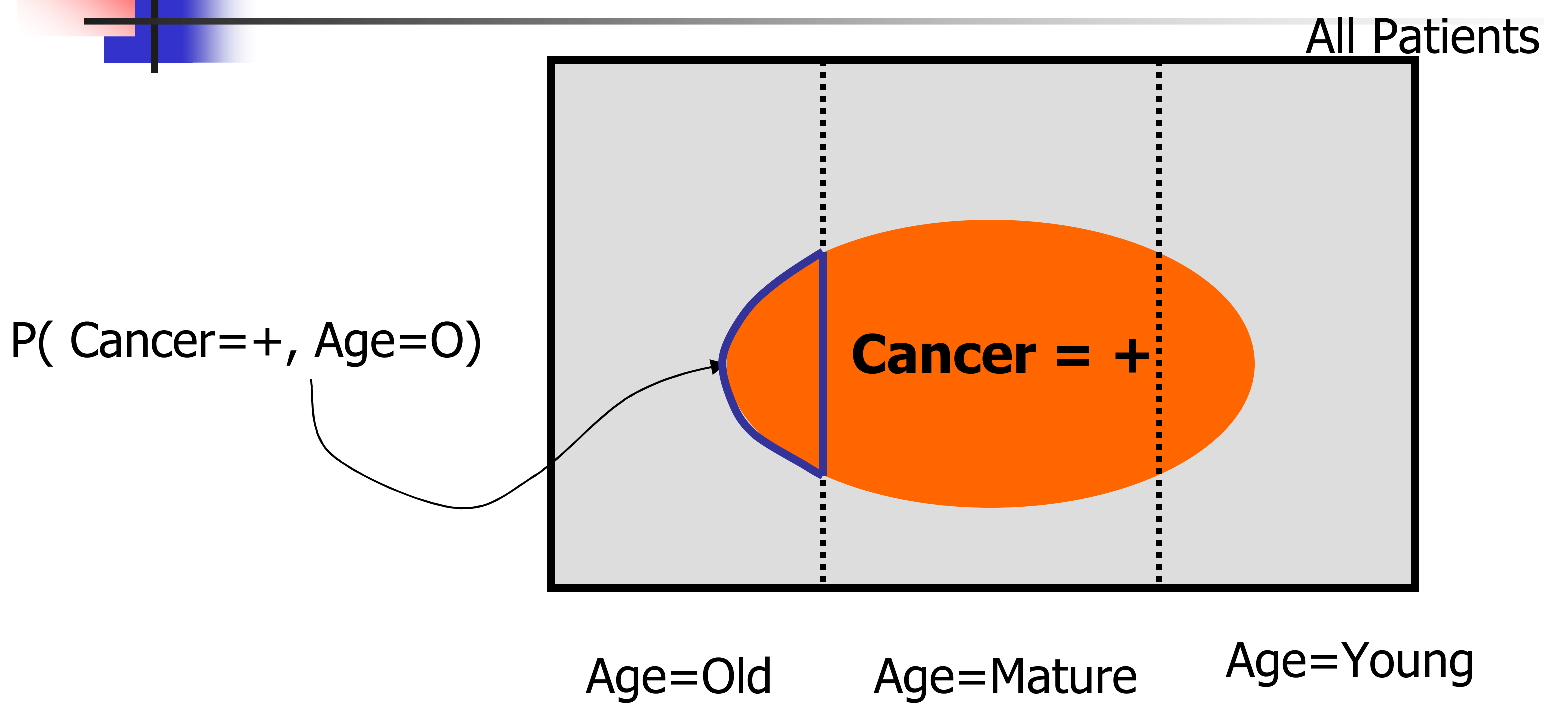
- **Diagnosis** typically involves computing P(Hypothesis | Symptoms)
  What is P(Meningitis | StiffNeck) ?
  ≡ prob that patient A has meningitis, given that A has stiff neck?
- Typically have . . .
  - Prior prob of meningitis P(+m) = 1/50,000
  - Prior prob of having a stiff neck P(+s) = 1/20
  - Prob that meningitis causes a stiff neck P(+s | +m) = 1/2
- Bayes' Rule:

$$P(M \mid SN) = \frac{P(SN \mid M)\ P(M)}{P(SN)}$$

- Eg:
  P(+m | +s) = P(+s | +m) P(+m) / P(+s) = 0.5 × 0.00002 / 0.05 = 0.0002

- Only 1 in 5000 stiff necks have meningitis...
  even though SN is major symptom of M...

Factoids



$$P(\ +c\ ) = \sum_a P(\ +c, A = a\ )$$

Important concept: (a) Independence

- Coin tosses:
  - H_1: the first toss is a head;
    T_2: the second toss is a tail
  - $P(T_2 \mid H_1) = P(T_2)$

- α and β ***independent*** iff $P(\beta|\alpha) = P(\beta)$
  - In distribution P, α indep of β

- **Proposition:** α and β *independent*
  if and only if
  $P(\alpha, \beta) = P(\alpha) P(\beta)$

Independence

- Events α and β are independent *iff*
  - $P(\alpha, \beta) = P(\alpha) P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\alpha \vee \beta) = 1 - (1 - P(\alpha))(1 - P(\beta))$

- Variables independent

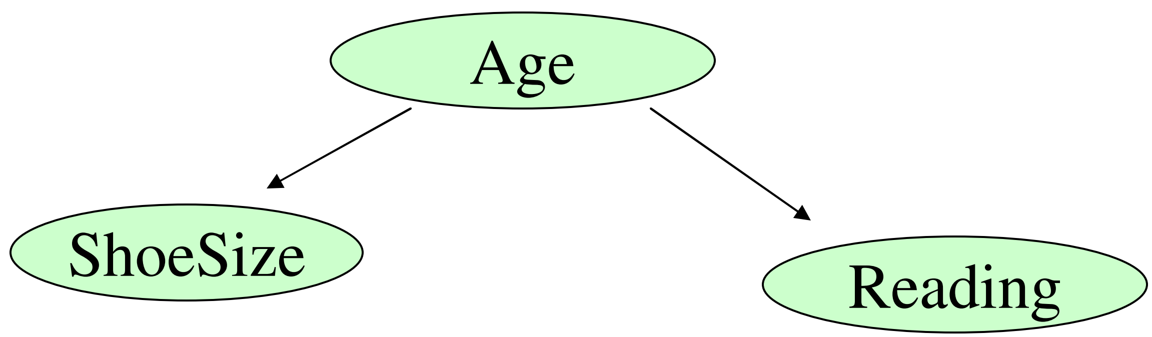
  $\Leftrightarrow$ independent for all values

$\forall a, b \quad P(A = a, B = b) = P(A = a)\, P(B = b)$

Conditional Independence

- **ReadingAbility** and **ShoeSize** are dependent,
  $P(\text{ReadAbility} \mid \text{ShoeSize}) \neq P(\text{ReadAbility})$

- but become independent, given Age
  $P(\text{ReadAbility} \mid \text{ShoeSize, Age}) = P(\text{ReadAbility} \mid \text{Age})$



Conditional Independence

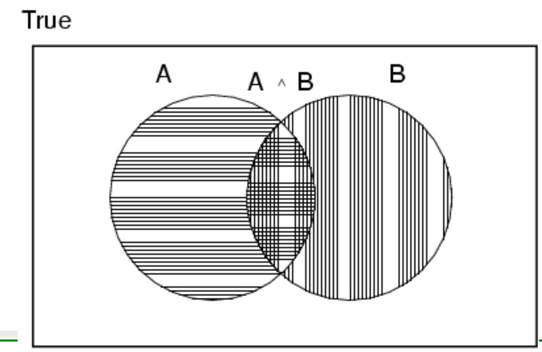
- Events A and B are conditionally independent given E iff

  P(A | E, B) = P(A | E)

- Given E, knowing B does not change the probability of A
- Equivalent formulations:

  P(A,B | E) = P(A | E) P(B | E)

  P(B | E,A) = P(B | E)



Probability Theory

- Axioms:

$0 \le P(A) \le 1$

$P(\text{ True }) = 1, \quad P(\text{ False }) = 0$

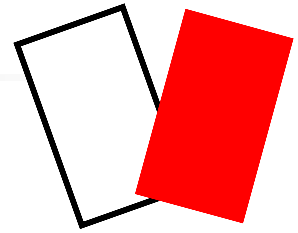
$P(A \vee B) = P(A) + P(B) - P(A \,\&\, B)$

$P(A) + P(\neg A) = 1$

- Not arbitrary:
  - If Agent1 use probabilities that violate axioms, then

    ∃ betting strategy s.t.
    Agent1 guaranteed to lose $
  - “Dutch book”

The Three-Card Problem



- Three cards
  - RR = red on both sides
  - WW = white on both sides
  - RW = red on one side, white on the other
- Draw single card randomly and toss it into the air.
- What is the probability …

a. … of drawing red-red? P(D_RR)

b. … that the drawn cards lands white side up? P(W_up)

c. … that the red-red card was not drawn,
assuming that the drawn card lands red side up.
P(not-D_RR | R_up)

Fair Bets

B believes

- P(D_RR) = 1/3
- P(W_up) = 1/2
- P(not-D_RR | R_up) = ~~1/3~~ 1/2

- A bet is *fair* to an individual B if,
  - according to B's probability assessment,
  - the bet will break even in the long run.
- B thinks these 3 bets are fair :

Bet **(a)** : Win $4.20 if D_RR;
lose $2.10 otherwise. [B believes P(D_RR)=1/3]

Bet **(b)**: Win $2.00 if W_up;
lose $2.00 otherwise. [B believes P(W_up)=1/2]

Bet **(c)**: Win $4.00 if R_up and not D_RR;
lose $4.00 if R_up and D_RR;
win $0 if not-R_up.

[B believes P(not-D_RR | R_up)=1/2]

Possible Outcomes

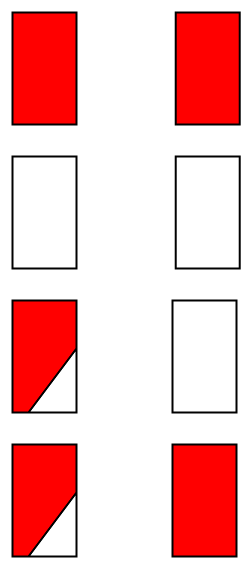
(a): Win \$4.20 if D_RR;
lose \$2.10 otherwise.

(b): Win \$2.00 if W_up;
lose \$2.00 otherwise

(c): Win \$4.00 if R_up and not D_RR;
lose \$4.00 if R_up and D_RR;
win \$0 if not-R_up

1. R_up & D_RR:
   Draw RR,
   which lands red side up.
2. W_up & not-D_RR:
   Draw non-RR card,
   which lands *white* side up.
3. R_up & not-D_RR:
   Draw non-RR card,
   which lands *red* side up.

Select **Observe**



(a)	(b)	(c)	Total
+4.20	-2.00	-4.00	-1.80
-2.10	+2.00	±0.00	-0.10
-2.10	-2.00	+4.00	-0.10

Possible Outcomes

(a): Win $4.20 if D_RR;
lose $2.10 otherwise.

(b): Win $2.00 if W_up;
lose $2 ...

... and not D_RR;
... and D_RR;

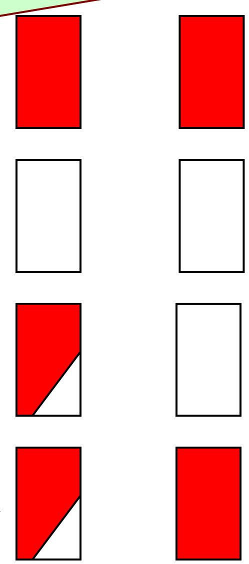
B is always guaranteed to lose money…
- whichever card is drawn, &
- however it lands !

1. R_up ...
Draw ...
which lands red side up.

2. W_up & not-D_RR:
Draw card not RR,
which lands *white* side up.

3. R_up & not-D_RR:
Draw card not RR,
which lands *red* side up.

...serve



(a)	(b)	(c)	Total
+4.20	-2.00	-4.00	-1.80
-2.10	+2.00	±0.00	-0.10
-2.10	-2.00	+4.00	-0.10

The Dutch Book Theorem

- Spse B accepts any bet it thinks is fair. Then…
- a Dutch book can be made against B

  iff

  B's assessment of probability violates Bayesian axiomatization.

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Expected Value

- Discrete
  - $E(X) = \sum_x x\, P(x)$
  - ≈ "average", "mean", arithmetic mean
  - P(X=1) = 1/6, P(X=2)=1/6, …, P(X=6) = 1/6
    $E[X] = (1 \times 1/6) + (2 \times 1/6) + \dots + (6 \times 1/6)$
    $= 21/6 = 3.5$
- Continuous
  - $E(X) = \int_x x\, P(x)\, dx$

Properties of Expectation

$$E(\, f(X)\,) = \sum_x f(x)\, P(x)$$

$$E(\, aX\,) = a\, E(X\,)$$

$$E(aX+b) = a\, E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

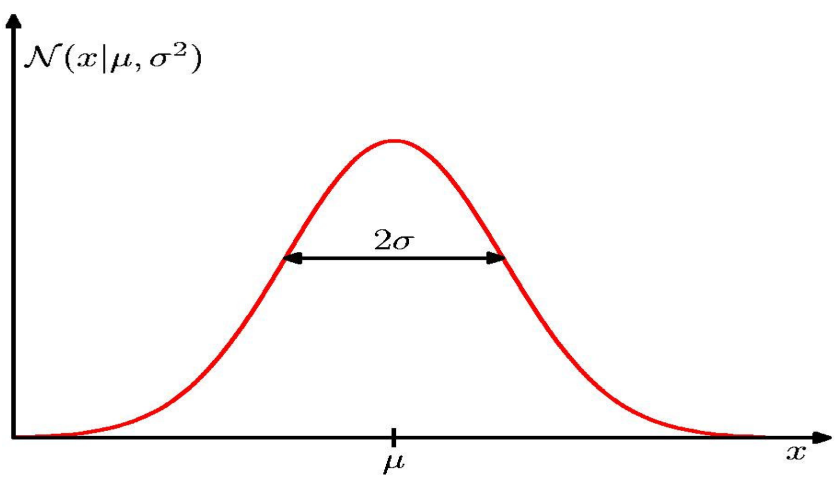
$$E(\, X\, Y\,) = \;???$$

If $X \perp Y$, then $E(X)\, E(Y)$

Variance

- ≈ "How much to *trust* the mean"
  … hard to define in words…

$$\mathrm{Var}(X) = E[\,X - E(X))^2\,]$$
$$= E(X^2) - E(X)^2$$



Properties of Variance

$$\mathrm{Var}(\ f(X)\ ) = E[\ X - E(X))^2\ ]$$

$$\mathrm{Var}(\ aX\ ) = a^2\ \mathrm{Var}(X\ )$$

$$\mathrm{Var}(aX+b) = a^2\ \mathrm{Var}(X)$$

$$\mathrm{Var}(X + Y) =$$

$$\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\ E[\ (X\text{-}E(X))\ (Y\text{-}E(Y)\ ]$$

If $X \perp Y$, then ... $= \mathrm{Var}(X) + \mathrm{Var}(Y)$

CoVariance

$$Var(X + Y) = Var(X) + Var(Y) + 2\,E[X-E(X))\,(Y-E(Y)]$$

- CoVariance captures the "leftover"

$$Cov(X,Y) = E[X-E(X))\,(Y-E(Y)]$$

$$Var(X + Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$$

- If $X \perp Y$, then $Cov(X, Y) = 0$

Standard Deviation

$$SD(X) = \sqrt{Var(X)}$$

- Sometimes more natural than variance:
  - SD(a X) = a SD(X)
- Sometimes, not:
  - X⊥Y, then SD(X + Y) =

$$SD(X+Y) = \sqrt{SD(X)^2 + SD(Y)^2}$$