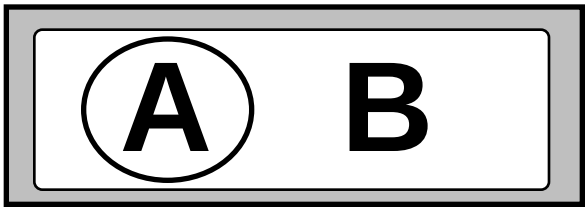
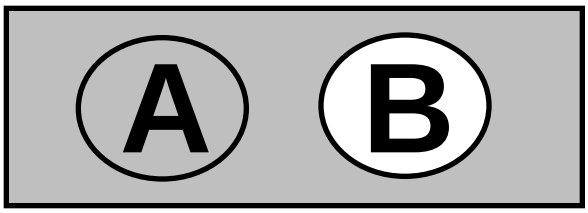
Chapter 2 Conditional Probability

- Suppose that we have *partial knowledge* of the result of a trial of the experiment
  — we know that event A occurred on the trial
  — but we do not know *which* outcome in A occurred; only that the outcome is *some* member of event A
- Did the event B occur?
- We cannot say for sure: All we know is that the actual outcome x is a member of A, and all we can say is that B occurred if and only if x is a member of B *as well*, that is, B occurred if x AB, and B did not occur if x AB^c
- Event A is known to have occurred; but we *don't know which outcome* in A occurred
- Event B has occurred if the actual outcome belongs to AB
- Event B has not occurred (i.e B^c has occurred) if the actual outcome belongs to AB^c
- Since we cannot say for sure whether B occurred or not, we would like a probabilistic description of the situation
- P(B) and P(B^c) should be *updated* in view of the partial knowledge that the event A occurred on this particular trial of the experiment
- Special cases:
- If A B, then x A x B; hence B occurs whenever A does. We should *update* the probability of B from P(B) to 1
- If AB = ø, x A x B; hence B cannot occur when A does. We should *update* the probability of B from P(B) to 0
- More generally, how should we *update* P(B) in light of the *partial knowledge* that event A occurred?
- We need nomenclature and notation to distinguish between the probability that we originally assigned to the event B and the updated probability of the event B
- P(B), the number originally assigned as the probability of B, is called the *unconditional* or *a priori* probability of B
- The updated probability of B is called the *conditional* probability of B *given* A (or the probability of B *given* A) and is denoted by P(B|A)
- Assume P(A) > 0
- P(B|A), the *conditional probability* of B *given* A, is $P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(BA)}{P(A)}$
- Similarly, $P(B^c|A) = \frac{P(B^cA)}{P(A)}$
- A is the conditioning event
- $P(B|A) = \frac{P(BA)}{P(A)}$
- $P(B^c|A) = \frac{P(B^cA)}{P(A)}$
- These definitions also work in the special cases considered above.
- **Example:** If A B BA = A, and B^{c}A = ø



- Hence, P(B|A) = 1; P(B^c|A) = 0

- **Example:** If $BA = \emptyset$, then $P(B|A) = 0$. Also $B^cA = A$ $P(B^c|A) = 1$



- **Example:** Two fair dice are rolled. What is the (conditional) probability that the sum of the two faces is 6 *given* that the two dice are showing different faces?
  Code phrase: *given* = we are asked to find conditional probability

  $\Omega = \{(i, j): 1 \le i \le 6, i \le j \le 6\}$. $B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$; $P(B) = \frac{5}{36}$

  $A = \{(i, j): (i, j) \in \Omega, i \ne j\}$; $P(A) = \frac{30}{36}$; $BA = \{(1,5), (2,4), (4,2), (5,1)\}$; $P(BA) = \frac{4}{36}$

  Hence, $P(B|A) = \frac{P(BA)}{P(A)} = \frac{(4/36)}{(30/36)} = \frac{2}{15}$. Note that $P(B|A) < P(B)$
- **Example** (continued): What is the conditional probability that the sum of the dice is 7 ?

  $C = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$; $P(C) = \frac{6}{36} = \frac{1}{6}$

  Here, C A so that $CA = C$. Thus, $P(C|A) = \frac{P(CA)}{P(A)} = \frac{(6/36)}{(30/36)} = \frac{1}{5}$.
  Note that $P(C|A) > P(C)$
- **Example** (continued): What is the conditional probability that the first die shows a 6?
  $D = \{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$; $P(D) = 6/36 = 1/6$.
  $DA = \{(6,1), (6,2), (6,3), (6,4), (6,5)\}$; $P(DA) = 5/36$.

  Thus, $P(D|A) = \frac{P(DA)}{P(A)} = \frac{(5/36)}{(30/36)} = \frac{1}{6}$. Note that $P(D|A) = P(D)$.
- Summary: $P(B|A) = \frac{2}{15} < P(B) = \frac{5}{36}$

  $P(C|A) = \frac{1}{5} > P(C) = \frac{1}{6}$

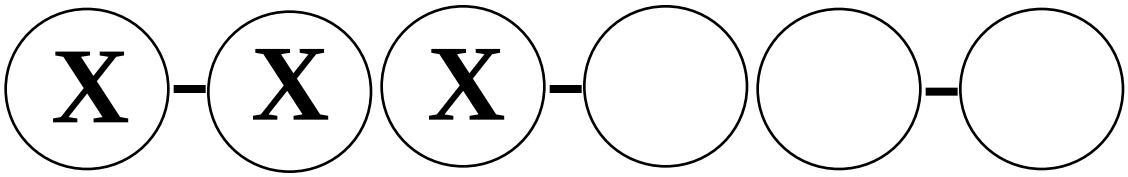
  $P(D|A) = \frac{1}{6} = P(D) = \frac{1}{6}$
- **Moral:** The conditional probability of an event can be smaller, larger, or the same, as the unconditional probability of the event

Relative frequencies and conditional probability

- Suppose that N independent trials of the experiment are performed and the event A occurs on N_A of these N trials
- Suppose that event B occurs on N_B of these N trials
- Suppose that the event BA occurs on N_{BA} of the N trials
- The *relative frequencies* of A, B, and BA are N_A/N, N_B/N, and N_{BA}/N respectively
- $P(A) \approx \frac{N_A}{N}$ $P(B) \approx \frac{N_B}{N}$ $P(BA) \approx \frac{N_{BA}}{N}$
- Consider *only those* N_A *trials on which A occurred*
- Event BA will have occurred on M of these N_A trials
- Event B^cA will have occurred on the remaining $N_A - M$ trials
- Claim: $M = N_{BA}$
  - BA occurred A occurred

- Consider *only those* N_A *trials on which A occurred*
- What is the relative frequency of B on these N_A trials (on which A is known to have occurred)?
- If B occurred, then the event BA must have occurred!
- Relative frequency of B on the trials on which A is known to have occurred
  $= \frac{N_{BA}}{N_A} = \frac{N_{BA}/N}{N_A/N} \approx \frac{P(BA)}{P(A)} = P(B|A)$
- The conditional probability of B given A is approx. the relative frequency of B on the trials on which A occurred
- B and B^c partition Ω, and hence $P(B) + P(B^c) = 1$
- *One* of the events B and B^c always occurs
- BA and B^cA partition event A, and hence $P(BA) + P(B^cA) = P(A)$
- If A occurred, *one* of BA and B^cA *must have* occurred
- $P(B|A) + P(B^c|A) = 1$
- But $\frac{P(BA)}{P(A)} + \frac{P(B^cA)}{P(A)} = 1$ follows from $P(BA) + P(B^cA) = P(A)$
- The statement "the odds in favor of B are m:n" means $P(B) = \frac{m}{m+n}$ and $P(B^c) = \frac{n}{m+n}$
- m and n are usually integers
- The ratio m:n is the ratio $P(B):P(B^c)$
- Knowing that event A has occurred should change the odds in favor of B
- The *a priori* odds in favor of B are $P(B):P(B^c)$
- Given that event A occurred, B occurs if BA does and B^c occurs if B^cA does
- The odds in favor of B seem to be $P(BA):P(B^cA)$ since one of these two events must occur
- It seems we should change the "Probability of B" to $\frac{P(BA)}{P(BA)+P(B^cA)} = \frac{P(BA)}{P(A)} = P(B|A)$
- For any event A with $P(A) > 0$, the conditional probabilities $P(\bullet|A)$ are a probability measure, that is, they satisfy the Axioms of Probability
- Axiom I: For all events B, $0 \le P(B|A) \le 1$
  Since BA A, $P(BA) \le P(A)$
- Axiom II: $P(\Omega|A) = 1$
  Since A Ω, $\Omega A = A$ $P(\Omega|A) = 1$
- Axiom III also holds as do all the various consequences of the axioms
- $P(\emptyset|A) = 0$
- $P(B^c|A) = 1 - P(B|A)$
- If B C, then $P(B|A) \le P(C|A)$
  If $BC = \emptyset$, then $P(B \cup C|A) = P(B|A) + P(C|A)$
- More generally, $P(B \cup C|A) = P(B|A) + P(C|A) - P(BC|A)$
- Condition on A throughout and use any of the rules learned in Chapter 1!
- Everything to the right of the vertical bar | is the *conditioning event* which is known to have occurred
- Everything to the left of the vertical bar | is the *conditioned event* whose conditional probability we are finding

- $P(C \cup D|A \cup B)$ is the conditional probability of event $C \cup D$ given event $A \cup B$
- Even if $BC = \emptyset$, $P(B \cup C|A) \neq P(B) + P(C|A)$
- Even if $AB = \emptyset$, $P(C|A \cup B) \neq P(C|A) + P(C|B)$
- **Example:** 3 poker chips are marked as follows: one chip has X on both sides, one chip has X on one side only, and one chip is blank on both sides. You are shown one randomly chosen side of a randomly chosen chip. What should you bet there is on the other side?



- Naive argument: If an X is showing, then you are seeing the X side of either the X-O or the X-X chip. Thus, the other side is equally likely to be an X or an O. Similarly, if an O is showing, the other side is equally likely to be an X or O
- Correct argument:: A = "this side shows an X"; B = "other side is an X"
  $P(B|A) = P(BA)/P(A) = P(\text{X-X chip was picked})/P(A) = (1/3)/(1/2) = 2/3$
  Similarly, $P(B^c|A^c) = 2/3$ also
- Another argument leading to the same result
- There are two chips that are "doubles" in the sense that both sides are the same. Thus, P(double chosen) = 2/3, and hence 2/3rds of the time, you will win by betting that the other side is the same as the one you are seeing
- Conditional probabilities are often more meaningful or useful than unconditional probabilities
- Radar system transmits a pulse and listens for an echo
- Receiver input = $\begin{cases} \text{echo + noise} \\ \text{noise} \end{cases}$. Is a target present?
- P(wrong decision) = P(E) = ?
- Different wrong decision that the receiver could make have different consequences and economic costs
- Announce target present when it is not, or announce target absent when it is actually out there
- P(false alarm) = P(E|target not present)
- P(missed detection) = P(E|target present)
- If T denotes the event that target is present, then $P(E|T)$ and $P(E|T^c)$ are separately more interesting than P(E)

Product Rule

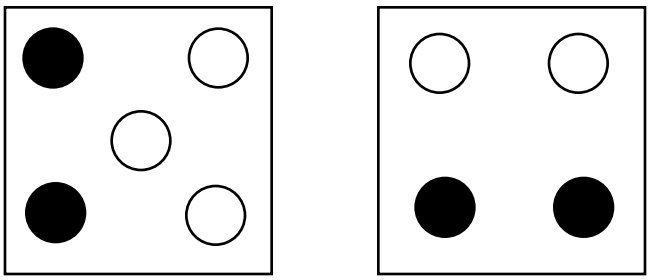
- Assume $P(A) > 0$
- $P(B|A)$, the *conditional probability* of B *given* A, is $P(B|A) = \dfrac{P(B \cap A)}{P(A)} = \dfrac{P(BA)}{P(A)}$
- Product rule for probability of an intersection: $P(BA) = P(AB) = P(B|A)P(A)$
- The product rule is useful in computing the probability of a complicated event
- Generalized product rule for probability of an intersection
  $P(A_1A_2A_3\ldots A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \times \ldots \times P(A_k|A_1A_2A_3\ldots A_{k-1})$
- Idea behind Rule: $P(A_1)P(A_2|A_1) = P(A_1A_2)$. Hence,
  $P(A_1)P(A_2|A_1)P(A_3|A_1A_2) = P(A_1A_2A_3)$ and so on

- **Example:** A sample of size k is drawn without replacement from a set of size n. What is the probability that a specified subset is obtained?
- There are $\binom{n}{k}$ subsets that could be obtained; hence P(specified subset)= $\left[\binom{n}{k}\right]^{-1}$
- A_i = "i-th element drawn belongs to the given subset"
- $P(A_1) = k/n$; $P(A_2|A_1) = (k-1)/(n-1)$; $P(A_3|A_1A_2) = (k-2)/(n-2)$;
  $P(A_k|A_1A_2\ldots A_{k-1}) = 1/(n-k+1)$ $P(A_1A_2\ldots A_k) = \frac{k(k-1)\ldots(1)}{n(n-1)\ldots(n-k+1)}$
- **Example:** Find the probability that at least two persons in a room with N people in it have the same birthday. ($N < 366$)
- Assumption: All days are equally likely to be birthdays
- It is easier to compute the probability that no two have the same birthday
- Let A_m = "m-th person has a different birthday than 1st, 2nd, 3rd, … , (m–1)th person"
- $A_2A_3\ldots A_N$ = "all N persons have different birthdays
- $P(A_2A_3\ldots A_N) = P(A_2)P(A_3|A_2)P(A_4|A_2A_3)\ldots = \frac{364}{365}\times\frac{363}{365}\times\frac{362}{365}\times\ldots\times\frac{365-(N-1)}{365}$
- P(all different birthdays) $= \frac{364}{365}\times\frac{363}{365}\times\frac{362}{365}\times\ldots\times\frac{365-(N-1)}{365}$ $< 1/2$ for $N \geq 23$
  $< .11$ for $N = 40$
  $< .03$ for $N = 50$

Theorem of Total Probability

- $P(B|A) = \frac{P(BA)}{P(A)}$
- $P(B|A^C) = \frac{P(BA^C)}{P(A^C)}$
- $P(BA) = P(B|A)P(A)$
- $P(BA^C) = P(B|A^C)P(A^C)$
- BA and BA^C is a partition of B
- $P(B) = P(BA) + P(BA^C) = P(B|A)P(A) + P(B|A^C)P(A^C)$
- The conditional probabilities on the right hand side are conditioned on two <u>different</u> events
- Theorem of total probability (Ross, (3.1), p.72) allows us to compute the probability of an event by imagining special cases (what if A occurred? what if A^C occurred instead?) and combining them as a weighted average
- Weights are $P(A)$ and $P(A^C)$
- **Example:** 10% of the ICs from chipmaker #1 are faulty, while only 5% of the ICs from chipmaker #2 are faulty. You buy 40% of your ICs from #1, and 60% from #2. What is the probability that a randomly chosen IC chosen is faulty?
- A = IC made by #1; A^C = IC made by #2 $P(A) = 0.4$; $P(A^C) = 0.6$
  B = chip is faulty; $P(B|A) = 0.1$ $P(B|A^C) = 0.05$
  $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C) = 0.4\times0.1 + 0.6\times0.05 = 0.07$
- Consider 100 ICs
  40 of 100 ICs are from #1 and 10% = 4 of these might be expected to be defective
- 60 of 100 ICs are from #2 and 5% = 3 of these might be expected to be defective
- Thus, 7 of 100 ICs might be expected to be defective

- **Example:** Box I has 3 white and 2 black balls. Box II has 2 white and 2 black balls. A ball is drawn at random from Box I and transferred to Box II (without looking; so you don't know what color ball was transferred). Then a ball is drawn at random from Box II.
  P(ball from Box II is black) = ?



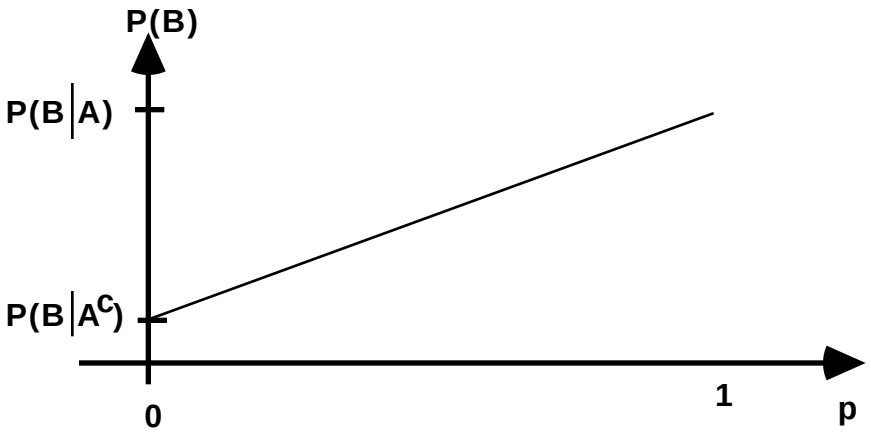
  A = black ball transferred to II

  B = ball from Box II is black $P(B|A) = 3/5$ $P(B|A^C) = 2/5$ $P(A) = 2/5$

  $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C) = 12/25$
- More on the Theorem of total probability

  $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C) = P(B|A)\times p + P(B|A^C)\times(1-p)$

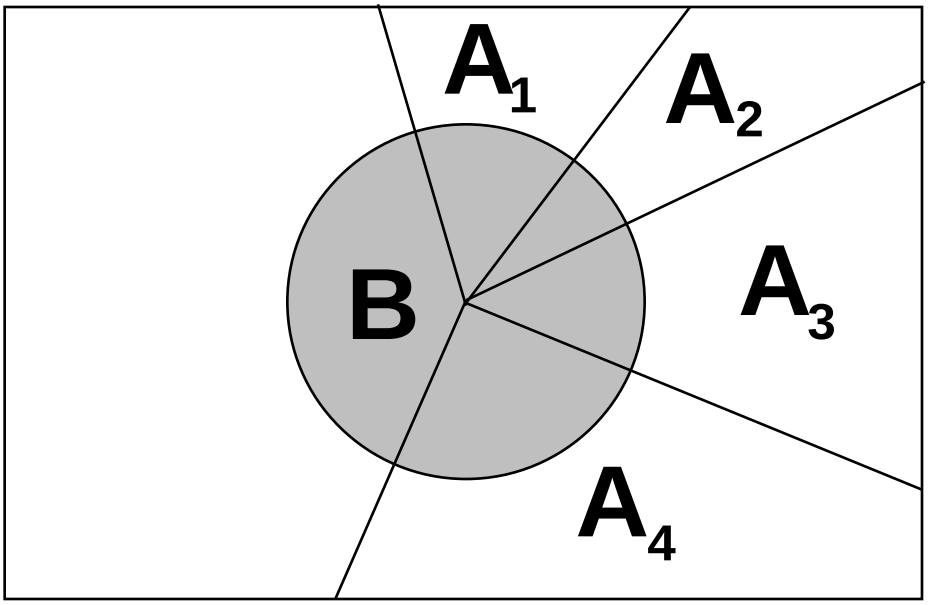
  $= P(B|A^C) + \left[P(B|A)-P(B|A^C)\right]\times p$ is a linear function of $p = P(A)$
- Value is $P(B|A^C)$ at $p = 0$; Value is $P(B|A)$ at $p = 1$



- $P(B) \le \max\{P(B|A), P(B|A^C)\}$
- $P(B) \ge \min\{P(B|A), P(B|A^C)\}$
- P(B) lies between $P(B|A)$ and $P(B|A^C)$
- P(B) is no larger than its largest conditional value and is no smaller than its smallest conditional value
- Generalizations of the theorem of total probability
- $P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$
- Hence, $P(B|C) = P(B|AC)P(A|C) + P(B|A^CC)P(A^C|C)$
  (Ross, (5.3), p. 101)
- Another generalization of the theorem of total probability
- If the *countable* sequence $A_1, A_2, A_3, \ldots, A_n, \ldots$ partitions Ω,

  $$P(B) = \sum_{k=1}^{\infty} P(B|A_k)P(A_k) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n) + \ldots$$

- This works even if the countable sequence is a partition of a superset of B



Bayes' Formula

- Bayes' formula is
  – also called Bayes' theorem,
  – also (mistakenly) called Bayes' rule
  is a means of "turning the conditioning around"
- Given $P(B|A)$, what is $P(A|B)$?
- Assume $P(A) > 0$, $P(B) > 0$
- We know that $P(B|A) = \frac{P(BA)}{P(A)}$
- Similarly, $P(A|B) = \frac{P(AB)}{P(B)}$
- Hence $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- This is the simplest version of Bayes' formula
- Generally, P(B) is expressed using the theorem of total probability
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$
- Note that the numerator is one of the terms in the denominator
- When is $P(B|A)P(A) \neq P(B|A)P(A)$?
- When you compute it twice in evaluating the Bayes' formula expression

  $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$
- **Example:** 10% of the ICs from chipmaker #1 are faulty, while only 5% of the ICs from chipmaker #2 are faulty. You buy 40% of your ICs from #1, and 60% from #2. If a randomly chosen IC chosen is faulty, find the probability that it was manufactured by chipmaker #1
- A = "IC made by #1"; A^C = "IC made by #2"
  $P(A) = 0.4$ $P(A^C) = 0.6$

B = chip is faulty: $P(B|A) = 0.1$ $P(B|A^c) = 0.05$

$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.4\times0.1 + 0.6\times0.05 = 0.07$

- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)} = \dfrac{0.4\times0.1}{0.07} = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$

  $= \dfrac{0.4\times0.1}{0.4\times0.1 + 0.6\times0.05} = \dfrac{4}{7}$
- Earlier, we saw that 4 of the 7 likely bad chips were from #1
- $P(A|B) = 4/7$ $P(A^c|B) = 3/7$
- Even though only 40% of the chips are made by maker #1 (i.e., $P(A) = 0.4$), a faulty chip is more likely to have been made by maker #2
- Whom do you choose to blame?

Decision Making under Uncertainty

- **Example:** Radar system transmits a pulse and listens for an echo to determine if a "target" is present
- Receiver input waveform consists of echo + noise if target is present
- Receiver input is noise alone if no target is present
- Receiver input = $\begin{cases}\text{echo + noise}\\ \text{noise}\end{cases}$
- System must decide if the receiver input waveform includes an echo (target is present) or not (there is no target present)
- How should a decision be made?
- Receiver measures the energy in its input waveform
- Receiver input waveform (usually) has more energy when target is present than when it is absent
- Compare measured energy to a threshold
- If energy > threshold, *declare* target is present
- If energy ≤ threshold, *declare* that no target is present
- Later, we shall consider how the threshold is set and how it affects performance
- Comparing the energy with the threshold results in quantization into HI and LO
- Decision rule:
  - decide target is present if energy is HI
  - decide target is absent if energy is LO
- If target is present, energy is usually *(but not always)* HI
- If target is absent, energy is usually *(but not always)* LO
- System is tested on known targets to see how often they are detected, and by counting the *false alarms* when targets are known to be absent
- Test results are described in terms of various probabilities
- **If the target is present,** P(HI) = 0.9; P(LO) = 0.1
  If the target is absent, P(HI) = 0.2; P(LO) = 0.8
- Validate the decision rule
- Suppose that the energy is HI. We can decide that the target is present or that the target is not present
- In actuality, the target is either present or not present; but ***we don't know which of these conditions holds***
- When we observe that energy is HI, either we have observed an event whose probability is 0.9 (assuming that the target is present) or we have observed an event whose probability is 0.2

(assuming that the target is absent). By *choosing* to decide that the target is present, we are choosing the condition that maximizes the probability (likelihood) of the observation HI

- Similarly for observation LO
- **If the target is present,** P(HI) = 0.9; P(LO) = 0.1
- **If the target is absent,** P(HI) = 0.2; P(LO) = 0.8
- The decision that no target is present when energy is LO chooses the condition which maximizes the probability (likelihood) of the observation LO
- This methodology for making decisions is called a maximum-likelihood decision rule. We decide that the "actual state of nature" is that which maximizes the probability (likelihood) of the observation
- *General version of* ***maximum-likelihood (ML) decision rule***
- One of two (or possibly many more) different conditions might exist (e.g. target is present, or target is absent), and we wish to determine which of these conditions actually exists. These conditions are referred to as *hypotheses* and we denote them by H_0, H_1, … etc. Corresponding to these hypotheses, we have probabilities that can be associated with observable events. More formally, we have different *probability measures* that are applicable depending on which hypothesis is correct.
- **Example (continued):** H_0: target is not present
  H_1: target is present
  are the two hypotheses in our radar example. If hypothesis H_0 is true, the probability of the observation HI is 0.2 while the probability of observation LO is 0.8. On the other hand, if hypothesis H_1 happens to be true, the probability of the observation HI is 0.9 while the probability of observation LO is 0.1.
- Remember that a probability measure is an assignment of numbers (called probabilities) to events, that is, it is a *function* that specifies the probabilities of events. Depending on which hypothesis happens to be true, we have a *different* function. Thus, let $P_i(A)$ denote the probability of an event A under the assumption that hypothesis H_i is the true hypothesis.
- **Example (continued):** $P_0(HI) = 0.2$ and $P_0(LO) = 0.8$,
  while $P_1(HI) = 0.9$ and $P_1(LO) = 0.1$
- When we have only two hypotheses H_0 and H_1, they are often referred to as the *null* hypothesis and the *alternative* hypothesis respectively. Generally, the alternative hypothesis is the one that we are "trying to prove" or establish. Such modeling is appropriate in a variety of situations.
- **Example:** If we *suspect* that a coin turns up heads 2/3rds of the time (i.e. that it is not a fair coin,) the null hypothesis is that the coin is fair while the alternative hypothesis is that the coin has P(heads) = 2/3.
- More formally, we wish to decide whether or not the alternative hypothesis should be accepted. If the alternative hypothesis is accepted, then we have agreed that the evidence supports the alternative hypothesis, and we are rejecting the null hypothesis. However, if the alternative hypothesis is not accepted (that is, it is rejected), then we are not necessarily accepting the null hypothesis, even though that is the seemingly blindingly obvious interpretation. All we are saying is that the evidence is insufficient for us to accept the alternative. It is by no means an assertion that the null hypothesis is valid. In our example above, if we do not accept the alternative hypothesis, we are not insisting that the coin is fair, but only that the evidence is insufficient to accept that the coin is biased.
- As a legal version of this, note that in a criminal trial, the court is considering two hypotheses:
  H_0: defendant is innocent of the crime *null hypothesis*
  H_1: defendant is guilty of the crime *alternative hypothesis*
  in accordance with the principle taught to all schoolchildren that a person is innocent of a crime until found guilty by a jury. However, what the jury actually decides is whether the defendant is *guilty* (accept the alternative hypothesis) or *not guilty* (reject the alternative). A not guilty

verdict rejects the alternative but it does not say the defendant is *innocent of the crime*. All a not guilty verdict means is that the charge has not been proved beyond a reasonable doubt, that is, the evidence tendered by the prosecution is insufficient to support the charge.

- In other situations, both hypotheses are equally interesting, e.g. in a digital communications situation, H_0 and H_1 respectively may be the hypotheses that 0 and 1 has been transmitted, and the nomenclature of null hypothesis and alternative hypothesis is not very appropriate. In this case, rejecting H_1 is equivalent to accepting H_0.
- Which interpretation is to be used is very dependent on the actual situation. Careful modeling and much thought should be used in deciding on the interpretation to be placed on the results of statistical decision-making processes.
- In what follows, we shall simply say "decide in favor of H_0" or "choose H_0" even though in some instances, a more appropriate interpretation would be that we are rejecting H_1 without necessarily embracing H_0.
- Returning to the fundamental notions, one of two (or possibly many more) different conditions (states of nature) might exist, and we wish to determine which of these conditions actually exists. These conditions are referred to as *hypotheses* and we denote them by H_0, H_1, … etc. We use the phrase "Hypothesis H_i is true" to mean that the state of nature corresponds to hypothesis H_i. Corresponding to these hypotheses, we have probabilities that can be associated with observable events. More formally, we have different *probability measures* $P_0(\bullet)$, $P_1(\bullet)$, … etc that should be used depending on which hypothesis is true.
- We perform an experiment whose sample space Ω is *partitioned* into n different events A_1, A_2, … , A_n. Thus, one and only one of these events can be observed on any trial. (If you don't like events, think of them as n possible outcomes of the experiment; but **do not** assume that the n outcomes are equally likely — we may not be working with a classical sample space here)
- We observe that event A_i has occurred. Which hypothesis is true?
- The likelihood of the observed event, that is, the probability of the observed event is $P_0(A_i)$, or $P_1(A_i)$, or … depending on which hypothesis is true.
- The maximum-likelihood decision rule states that if we observe event A_i, we should compare the likelihoods $P_0(A_i)$, $P_1(A_i)$, …, and if $P_k(A_i)$ is the largest of these numbers, then we should decide that hypothesis H_k is true.
- Naturally, if we observe A_j, we should compare the likelihoods $P_0(A_j)$, $P_1(A_j)$, or …, and make the decision based on whichever is the largest of these numbers.
- If the largest likelihood is not unique, say $P_k(A_i) = P_l(A_i)$ are the two largest numbers among $P_0(A_i)$, $P_1(A_i)$, …, then we can choose either hypothesis H_k or hypothesis H_l as being the true hypothesis. The maximum-likelihood decision rule is not very helpful in deciding between the two cases. In this course, however, we shall always choose the lower numbered hypothesis, so that, for example, we shall choose H_0 over H_1 so that we do not accept the alternative hypothesis when there is insufficient reason to do so.

- Maximum-likelihood decision rule: *Decide* that the hypothesis that maximizes the likelihood of the observation is the true hypothesis.

- It is convenient to specify the maximum-likelihood decision rule in terms of the likelihood matrix L which has m rows corresponding to the m hypotheses H_0, H_1, … H_{m-1}, and n columns corresponding to the n events A_1, A_2, … , A_n. Note that m, n ≥ 2. The entries in the matrix L are the likelihoods of the events under the various hypothesis. Thus, the i-jth entry in L is $P_i(A_j)$. The entries on the i-th row ($0 \le i \le m-1$) are the likelihoods (probabilities) of the n events A_1, A_2, … , A_n. Since these events are a partition of Ω, their probabilities sum to 1. Thus, the likelihood matrix L has the property that all the row sums are equal to 1. Of course, since the entries are probabilities, they are all nonnegative also.
- The entries in the i-th column are precisely the numbers $P_0(A_i)$, $P_1(A_i)$, …, $P_{m-1}(A_i)$ that need to be compared to determine the decision when A_i is observed.

- **Example:** For our radar problem and the biased coin problem discussed above, the likelihood matrices are as shown below:

	HI	LO
H_0	0.2	0.8
H_1	0.9	0.1

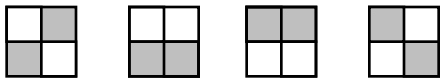
	Tails	Heads
H_0	1/2	1/2
H_1	1/3	2/3

- The maximum-likelihood decision rule is then easily obtained by looking for the maximum number in each column. The maximum-likelihood decision rule can be most easily described by shading the appropriate entries in the matrix as shown below.

	HI	LO
H_0	0.2	**0.8**
H_1	**0.9**	0.1

	Tails	Heads
H_0	**1/2**	1/2
H_1	1/3	**2/3**

- We have already enunciated the maximum-likelihood decision rule for the radar problem as decide that H_1 is true (a target is present) if the returned signal energy is HI and to decide that H_0 is true (no target is present) if the energy is LO.
- For the biased coin example, the maximum-likelihood decision rule is to decide that we have a biased coin if a toss results in a Head, and to decide that the coin is fair if it turns up Tails. Note that the interpretation might well be that we are only rejecting the insinuation that the coin has probability 2/3 of turning up Heads without necessarily agreeing that the coin is fair. For example, if we had assumed that the null hypothesis was that $P_0(\text{Heads}) = 0.4$ and $P_0(\text{Tails}) = 0.6$, we would still have the same decision rule (accept H_1 if Heads, reject H_1 if Tails)! Thus, in interpreting the results, we should remember that rejection of the alternative does not necessarily mean acceptance of the null hypothesis. Even if you firmly believe that if the shoe doesn't fit, you must acquit, you must keep in mind that the verdict is "not guilty" rather than "innocent."
- The likelihood matrix L allows us to describe arbitrary decision rules (that is, not necessarily in accordance with the principle of maximum likelihood) as well. Since we must specify a decision for each of the observations $A_1, A_2, \ldots, A_n$, all we need to do is to shade one (and only one) entry in each column.
- Clearly, if $m > n$, then some of the hypotheses will never be chosen as the true hypothesis. On the other hand, if $n > m$, then some of the hypotheses will enjoy the privilege of being chosen as the true hypothesis for more than one observation.
- **Exercise:** How many different decision rules are there (as a function of m and n)?
- For the case of two hypotheses and two observations, there are 4 different decision rules according to the formula that you found in the exercise (Is the formula $m \times n$? or $m+n$? or m^n? or n^m?). These are illustrated below. In the radar example, these correspond to the four broad divisions of the populace into normal people, hawks, doves, and the lunatic fringe.



- **Error probabilities**

  One of the m hypotheses describes the state of nature. However, we are observing the outcome of an experiment and making a decision as to which hypothesis is true based on this observation. It is our fondest hope that our decision is correct, but we might well be wrong and end up *deciding* that H_j is the true hypothesis when in fact H_j is the true hypothesis. Such errors should be avoided as much as possible.
- There are different *kinds* of errors that might be made.
- For the simplest case of 2 hypotheses H_0 and H_1 (respectively the *null* hypothesis and the *alternative* hypothesis), we might accept the alternative hypothesis when it is not true.

Statisticians call this a Type I error. On the other hand, we might reject the alternative hypothesis even though it happens to be true. This is called a Type II error. More descriptive terminology comes from the radar literature: a Type I error is called a *falss alarm* because we raise an alarm by announcing that a target is present, but this is a false alarm because in fact there is no target present. Similarly, a Type II error is called a *false dismissal* or *missed detection* because we do not detect the target that happens to be present out there.

- More generally, an error of type $i \to j$ is said to have occurred if hypothesis H_i happens to be true but we announce that our experiment has led us to conclude that hypothesis H_j is true. Thus, a Type I error (false alarm) is of type $0 \to 1$, while a Type II error (false dismissal) is of type $1 \to 0$.
- For *any* decision rule (not necessarily a maximum-likelihood decision rule), the various error probabilities can be read off from the likelihood matrix L.

	HI	LO
H_0	0.2	0.8
H_1	0.9	0.1

	Tails	Heads
H_0	1/2	1/2
H_1	1/3	2/3

- For example, suppose that H_0 happens to be the true hypothesis. Our decision rule says that we declare H_1 to be true whenever we observe HI (or Heads). The probability of this happening (remember that H_0 happens to be the true hypothesis) is 0.2 (1/2 in the other case.) Thus, the Type I error probability or false alarm probability is 0.2 (1/2 in the other case.) Similarly, suppose that H_1 happens to be the true hypothesis. Our decision rule says that we declare H_0 to be true whenever we observe LO (or Tails). The probability of this happening (remember that H_1 happens to be the true hypothesis) is 0.1 (1/3 in the other case.) Thus, the Type II error probability or missed detection probability is 0.1 (1/3 in the other case.)
- More generally, suppose that the decision rule declares H_j to be true whenever the observation is *any* of the events A_k, A_l, …. If the true hypothesis happens to be H_i, then an error of type $i \to j$ occurs whenever any of A_k, A_l, … occur. Hence, the decision rule mistakenly announces that hypothesis H_j is true when hypothesis H_i happens in fact to be true with probability $P(i \to j) = P_i(A_k) + P_i(A_l) + \ldots$ Note that all these numbers are found in the i-th row of the likelihood matrix, but which numbers are to be included in the sum is determined by looking in the columns to see which A's lead to an announcement that H_j is true.
- **Example:** For the likelihood matrix shown, $P(0 \to 1) = 0.15 + 0.02 = 0.17$, $P(0 \to 2) = 0.08$; $P(1 \to 0) = 0.15 = P(1 \to 2)$; and $P(2 \to 0) = 0.05$, $P(2 \to 1) = 0.1 + 0.2 = 0.3$.
- **Exercise:** Is this a maximum-likelihood decision rule?

	A	B	C	D
H_0	0.75	0.15	0.08	0.02
H_1	0.15	0.6	0.15	0.1
H_2	0.05	0.1	0.65	0.2

- We can also read off the probability of *making a correct decision* whenever H_i happens to be true from the likelihood matrix. $P(i \to i)$ is just the sum of the shaded entries in the i-th row of the likelihood matrix L. The decision rule shown in the example above makes correct decisions with probability 0.75 if H_0 happens to be true, probability $0.6 + 0.1 = 0.7$ if H_1 happens to be true, and probability 0.65 if H_0 happens to be true.
- Finally, the probability of making a wrong decision is just the probability that the decision rule announces anything other than H_i (that is, not i) and is given by $P(i \to \neg\, i) = 1 - P(i \to i)$.

- **Exercise:** If there are *no* shaded entries in the i-th row, what are $P(i \rightarrow i)$ and $P(i \rightarrow \neg i)$? What are $P(i \rightarrow i)$ and $P(i \rightarrow \neg i)$ if *all* the entries in the i-th row are shaded?
- Different wrong decisions that the receiver could make have different consequences or economic costs, so such lumping together of all wrong decisions is not always appropriate.
- **Bayesian decision-making**
- The maximum-likelihood decision rule attempts to make rational decisions as to which of m hypotheses represents the true state of nature. One of these states exists in nature, and all that we seek is to decide which one it is. But what if we have some prior knowledge about the these states?
- **Example:** If we know that we have been given a randomly chosen coin from a box containing 999 fair coins and 1 coin with P(Heads) = 2/3, should we take this information into account in making our decision as to whether the coin is fair or biased? And if so, how? Clearly, the coin is very likely to be a fair coin, and thus it seems difficult to justify saying it is a biased coin with P(Heads) = 2/3 just because we tossed it once and it came down Heads. Note that the maximum-likelihood decision rule insists that there is no such prior information available. All that we know is that we have been handed a coin that *is* either a fair coin or a coin that turns up Heads with probability 2/3, and we are asked to decide which coin it is.
- Bayesian decision rules take such prior information into account. They are a subject of considerable conrtoversy because the prior information may be quite subjective or not easily quantifiable. In our coin example above, we could say that the probability that H_0 is true is 0.999. But what of the radar target example? Does it make sense to talk of the probability of a target being present? What number can (or should) we choose as the probability of a target being present if this event has never been observed in the past? For example, there are many military radars that have never ever observed a reflection off an actual hostile target!
- Another complaint against Bayesian decision rules is that the "answer can be made to come out to be whatever you want by suitable choice of the prior information." It is, of course, possible for unscrupulous decision makers to use Bayesian techniques to justify their preconceived notions and prejudices, but that does not mean that the methods should be discarded by one and all. Bayesian methods are very useful and very valuable tools though, like all tools, they can be dangerous in the wrong hands. Remember that no one has mounted campaigns to outlaw chainsaws (not even in Texas!)
- We shall see a little later that a maximum-likelihood decision rule is actually a Bayesian decision rule for the special case when all the hypotheses are equally likely.
- What is not controversial, of course, is the Bayesian decision rule when everyone is agreed on the prior information available. In particular, we assume that the various hypotheses have probabilities associated with them, possibly arising from purely subjective considerations, or possibly from relative frequency studies, or in some other fashion, and that we are all agreed about these numbers. The m hypotheses occur with probabilities $P(H_0)$, $P(H_1)$, … , $P(H_{m-1})$. These are often called the *a prioir* probabilities of the hypotheses. This is the only case we consider in the rest of this chapter.
- Suppose that we have to make a series of decisions as to which hypothesis is true rather than a single decision as envisioned in the maximum-likelihood case. For example, we are studying the transmission of a large number of bits in a digital communication system. As each bit is transmitted, the receiver must make a decision as to whether the transmitted bit was a 0 or a 1. Letting H_0 and H_1 respectively denote the hypotheses that the transmitted bit was a 0 or a 1, the receiver has to make a large number of decisions, and for some of these decisions, H_0 happens to be the true hypothesis, while for the rest of the decisions, H_1 happens to be the true hypothesis. In fact, if N bits are being transmitted, then, roughly $NP(H_0)$ of these are 0's and $NP(H_1)$ of these are 1's. Let α and β denote the false alarm and false dismissal probabilities for the decision rule being used. Then, in the course of making the roughly $NP(H_0)$ decisions

during which H_0 is true, the receiver makes $\alpha NP(H_0)$ errors — deciding that a 1 has been received when in fact a 0 has been transmitted — and similarly in the course of the roughly $NP(H_1)$ decisions during which H_1 is true, the receiver makes $\beta NP(H_1)$ errors — deciding that a 0 has been received when in fact a 1 has been transmitted. Thus, roughly $\alpha NP(H_0) + \beta NP(H_1)$ of the N bits are in error, and the *relative frequency* of the errors (the bit error rate or BER) is $\alpha P(H_0) + \beta P(H_1)$.

- We can cast this argument in purely probabilistic terms by noting that α can be viewed as the *conditional* probability of bit error given that hypothesis H_0 is true, that is, $\alpha = P(E|H_0)$, while β is the *conditional* probability of bit error given that hypothesis H_1 is true, that is, $\beta = P(E|H_1)$. Here, E is the event that the transmitted bit is received in error. The theorem of total probability tells us that P(E), the unconditional (or average) probability of bit error is given by $P(E|H_0)P(H_0) + P(E|H_1)P(H_1)$. As we noted above, the relative frequency of the number of bit errors can be expected to be roughly this number.
- More generally, if we are agreed that we can treat the occurrence of the various states of nature (hypotheses) as occurring at random with probabilities $P(H_0)$, $P(H_1)$, … , $P(H_{m-1})$, then the likelihood $P_i(A_j)$, which we had defined to be the probability of the event A_j when hypothesis H_i is the true state of nature, can be thought as $P(A_j|H_i)$, the *conditional* probability of the event A_j given that the hypothesis H_i is the true state of nature on this particular trial of the experiment. The likelihood matrix L, whose i-j^{th} entry in L is $P_i(A_j)$, can be viewed as a matrix of conditional probabilities with entries $P(A_j|H_i)$.
- For any decision rule, the probability of making a correct decision when H_i is the true state of nature is the sum of the shaded entries in the i-th row of the likelihood matrix L. This can be viewed as $P(C|H_i)$, the conditional probability of making a correct decision given that H_i is the true state of nature. From the theorem of total probability, the unconditional (or average) probability of making a correct decision is

$$P(C) = \sum_{i=0}^{m-1} P(C|H_i)P(H_i)$$

- The likelihood matrix L has i-j^{th} entry given by $P(A_j|H_i)$. If we multiply the entries in the i-th row of the likelihood matrix by $P(H_i)$, then $P(A_j|H_i)P(H_i) = P(A_j \cap H_i)$ is the joint probability of the events A_j and H_i. We formalize this by defining the joint probability matrix J as the matrix with i-j^{th} entry given by $P(A_j \cap H_i)$. Note that J is found by multiplying each entry in the likelihood matrix L by the probability of the hypothesis. Thus, all entries on the i-th row are multiplied by $P(H_i)$. In matrix terms, let K denote an m×m diagonal matrix with entries $[P(H_0), P(H_1), \ldots, P(H_{m-1})]$ on the diagonal. Then, the joint probability matrix J is related to L via the equation J = KL.
- The joint probability matrix is essentially a Venn diagram. The n events $A_1, A_2, \ldots, A_n$ are a partition of Ω; exactly one event occurs on a given trial. Similarly, the m hypotheses $H_0, H_1, \ldots H_{m-1}$ als are a partition of Ω since exactly one of the hypotheses is true on any given trial.
- **Exercise:** Show that the row sums of J are the *a priori* probabilities of the hypotheses, and that the column sums are the probabilities of the events $A_1, A_2, \ldots, A_n$. It follows that the sum of all the entries in J equals 1.

- Both L and J are m×n matrices. A decision rule can be specified by shading one entry in each column of L (or J if you prefer). The advantage of J is that P(C) can be read off from J as the sum of the shaded entries in J. Note that there is one shaded entry in each column.
- Which decision rule has the maximum probability of making correct decisions? Obviously, the probability of making a correct decision is maximized by picking the largest entry in each column of the joint probability matrix J.
- In contrast, the maximum-likelihood decision rule picks the largest entry in each column of L, the likelihood matrix.
- The decision rule obtained when the largest entry in each column of J is chosen is called the *minimum-probability-of-error* (MEP) decision rule or Bayes' rule.
- Bayes' rule is often stated in a slightly different form. To derive this other version, note that when we are looking at the m entries $P(A_j \cap H_0)$, $P(A_j \cap H_1)$, … , $P(A_j \cap H_{m-1})$ in the j-th column of the joint probability matrix to determine which is the largest, this process is unaffected if all the entries are multiplied by the same (positive) number. Thus, if we *imagine* that all the entries in the j-th column have been multiplied by $1/P(A_j)$, then instead of comparing $P(A_j \cap H_0)$, $P(A_j \cap H_1)$, … , $P(A_j \cap H_{m-1})$, we are comparing $P(A_j \cap H_0)/P(A_j) = P(H_0|A_j)$, $P(A_j \cap H_1)/P(A_j) = P(H_1|A_j)$, … , $P(A_j \cap H_{m-1})/P(A_j) = P(H_{m-1}|A_j)$ to see which is the largest. If $P(H_k|A_j)$ is the largest of these m numbers, then, whenever the event A_j occurs, Bayes' decision rule announces that H_k is the true hypothesis.
- The probabilities $P(H_0|A_j)$, $P(H_1|A_j)$, … ,$P(H_{m-1}|A_j)$ are called the *a posteriori* probabilities of the hypotheses. Here, *a posteriori* means "after the fact," that is, after an experiment has been performed and the event A_j has been observed. If $P(H_k|A_j)$ is the largest of the *a posteriori* probability when A_j occurs, the decision is that H_k is the true hypothesis. For this reason, Bayes' rule is also called the maximum *a posteriori* probability (MAP or MAPP) decision rule.
- **Exercise:** Show that the sum of th *a posteriori* probabilities is 1.
- We contrast the maximum-likelihood decision rule and the Bayes' or MAP decision rule, writing both in terms of conditional probabilities

- **Maximum-likelihood decision rule:** For each event A_i, find the likelihood of A_i under each hypothesis, that is, find

$$P(A_i|H_0), \quad P(A_i|H_1), \quad \dots, \quad P(A_i|H_{m-1}).$$

If $P(A_i|H_k)$ is the largest of these m numbers, then declare that H_k is the true hypothesis whenever A_i is observed.

- It should be kept in mind that in most applications, the likelihoods are included as part of the data, and thus very little computation is needed to deternine the maximum-likelihood decision rule.

- **Maximum *a posteriori* probability or Bayes' decision rule:** For each event A_i, find the *a posteriori* probabilities of each hypothesis, that is, find

$$P(H_0|A_i), \quad P(H_1|A_i), \dots, \quad P(H_{m-1}|A_i)$$

If $P(H_k|A_i)$ is the largest of these m numbers, then declare that H_k is the true hypothesis whenever A_i is observed.

- Of course, the *a posteriori* probabilities $P(H_j|A_i)$ are related to the likelihoods $P(A_i|H_j)$ via Bayes' formula: $P(H_j|A_i) = P(A_i|H_j)P(H_j)/P(A_i)$
- There are several important points to be made here. First, in order to determine Bayes' decision rule, it is ***not*** necessary to actually find the *a posteriori* probabilities. It suffices to work with the joint probability matrix J = KL defined earlier and look for the largest entry in

each column. Second, in contrast to the maximum-likelihood decision rule for which the entries in the likelihood matrix are specified as part of the problem statement, we do need to do *some* computations to find J, but we don't need to go all the way and find the $P(H_j|A_i)$'s explicitly. Thirdly, the probabilities of error and correct decision can be read off from the shading in the J matrix, and complicated expressions involving the theorem of total probability should be avoided. Finally, the maximum-likelihood rule is finding maxima in the columns of the likelihood matrix, i.e., the largest of $P(A_i|H_0)$, $P(A_i|H_1)$, … , $P(A_i|H_{m-1})$. In contrast, the Bayes' decision rule is finding maxima in the columns of J, i.e., the largest of $P(A_i|H_0)P(H_0)$, $P(A_i|H_1)P(H_1)$, … , $P(A_i|H_{m-1})P(H_{m-1})$. If the *a priori* probabilities $P(H_0)$, $P(H_1)$, … , $P(H_{m-1})$ are all equal (to 1/m naturally!), then the Bayes' decision rule is the same as the maximum-likelihood decision rule.

- **Conclusion:** The maximum-likelihood decision rule is a special case of the Bayes' decision rule for the case when all the hypotheses H_0, H_1, … , H_{m-1} are equally likely.
- Bayes' decision rule treats the hypotheses H_0, H_1, … , H_{m-1} probabilistically which is controversial. The "answer" depends on the choice of the *a priori* probabilities.
- **Example:** We consider the radar system with likelihood matrix as shown below on the left. The maximum-likelihood decision rule is as shown by the shading.

	HI	LO
H_0	0.2	0.8
H_1	0.9	0.1

Likelihood matrix L

	HI	LO
H_0	0.18	0.72
H_1	0.09	0.01

Joint probability matrix J = KL

Now suppose that $P(H_0) = 0.9$ and $P(H_1) = 0.1$. The joint probability matrix J is as shown on the right, and the Bayes' rule is to always reject H_1 and declare that there is no target present. Is this really a minimum-probability-of error decision rule? Yes. The probability of a correct decision is 0.9 and the probability of error is 0.1. In contrast, the maximum-likelihood decision rule has a probability of correct decision = 0.09 + 0.72 = 0.81 and error probability 0.19.

- **Exercise:** Find the error probability for the hawk and the lunatic fringe rules and show that they are both larger than 0.1.
- If you insist on closure by finding the *a posteriori* probabilities first and then looking for the maximum values, you can do this by dividing the entries in each column of J by the column sum. In terms of matrices, let M be the n×n diagonal matrix whose diagonal entries are $[1/P(A_1), 1/P(A_2), \ldots, 1/P(A_n)]$. Then, the matrix JM = KLM has i-jth entry $P(H_i|A_j)$, the *a posteriori* probability of the hypothesis H_i given that A_j has been observed. However, let me speak to you as a Dutch uncle and say to you "Don't do this unless you absolutely have to!"
- **Exercise:** Show that the column sums of KLM equal 1.

	HI	LO
H_0	0.18	0.72
H_1	0.09	0.01

Joint probability matrix J = KL

	HI	LO
H_0	2/3	72/73
H_1	1/3	1/73

a posteriori probability matrix KLM

- It can be argued that $P(H_1)$ is a purely subjective quantity; it reflects the bias of the problem solver
- It is not possible to assign probabilities to the condition of target being present. Either a target is present or not; it is a fact of nature that has already occurred
- Bayesians counter that maximum-likelihood decision rules are also subjective since the maximum-likelihood rule implicitly assumes that all the hypotheses H_0, H_1, … , H_{m-1} are equally likely

- Also, Bayesian decision rules include maximum-likelihood decision rules as a special case.
- **Decisions involving costs**
- Suppose that whenever hypothesis H_j happens to exist, and your decision rule mistakenly declares that hypothesis H_i is the true state of nature, you have to pay $\$C_{ij}$ as a penalty.
- There may be costs involved even if you correctly declare that hypothesis H_i is the true state of nature. $\$C_{ii}$ could represent the fixed costs of making a decision. Of course, we could allow for the possibility that $C_{ii} < 0$, in which case it would be a bonus or cash prize for getting the right answer!)
- The costs that you have to pay depend on which event $A_1, A_2, \ldots, A_n$ you observe, and what you declare to be the true state of nature, and what the actual state of nature is. Each trial may require you to pay a different cost. The cost you pay is called the *risk*. However, over a long sequence of trials, we have an *average* cost, and we seek to make this average cost (called thhe *average risk*) as small as possible.
- A modification of Bayesian decision strategy allows us to minimize the average cost instead of the average error probability
- Suppose that the event A_j is observed
- We have to *declare* one of the hypotheses $H_0, H_1, \ldots, H_{m-1}$ as the state of nature. The *a posteriori* probabilities are $P(H_0|A_j), P(H_1|A_j), \ldots, P(H_{m-1}|A_j)$.
- Let H_i denote the hypothesis that we choose when we observe A_j.
- Consider N trials on each of which A_j happens to be observed. On these N trials, hypothesis H_0 happens to be true roughly $NP(H_0|A_j)$ times, H_1 happens to be true roughly $NP(H_1|A_j)$ times, $\ldots, H_{m-1}$ happens to be true roughly $NP(H_{m-1}|A_j)$.
- Our decision rule declares that H_i is the true hypothesis whenever we observe A_j, and thus we pay $\$C_{i0}$ on $NP(H_0|A_j)$ trials, $\$C_{i1}$ on $NP(H_1|A_j)$ trials, $\ldots$, $\$C_{ii}$ on $NP(H_i|A_j)$ trials, $\ldots$, $\$C_{i(m-1)}$ on $NP(H_{m-1}|A_j)$ trials.
- Total risk over N trials on each which we observe A_j is $\sum_{k=0}^{m-1} \$C_{ik} \times N \times P(H_k|A_j)$ and the average risk is $\sum_{k=0}^{m-1} \$C_{ik} \times P(H_k|A_j)$
- We can choose whichever value of i we like. It is *our* decision rule, after all!
- Calculate $\sum_{k=0}^{m-1} \$C_{ik} \times P(H_k|A_j)$ for each i, $0 \le i \le m-1$, and declare in favor of whichever H_i gives the least average risk. We can formulate this in matrix terms. If C is an m×m matrix with i-j^{th} entry C_{ij}, the cost of declaring that H_i happens to exist when hypothesis H_j is the true state of nature, then the average risk of declaring that H_i happens to exist upon observing A_j is the i-j^{th} entry of CJM = CKLM where JM = KLM is the matrix of *a posteriori* probabilities. Once again, we look in each column and pick the *smallest* entry as the decision. This is also called a Bayes' (minimum average cost or minimum risk) decision rule
- The minimum average risk is called the *Bayes' risk*, and is the sum of the entries in the shaded cells of CKL (not CKLM)
- **Example:** In our radar problem, assume that the costs (in thousands of dollars) are as follows: $C_{00} = 3$, $C_{10} = C_{11} = 30$, and $C_{01} = 3{,}000$. Note that C_{01} is the cost of failing to detect a target, and consequently being bombed back to the Stone Age. On the other hand, it costs the

same to send an interceptor up to investigate when it is a false alarm as when it is a real attack, and C_{00} is essentially the base cost of running the whole facility. We thus get the risk matrix shown below, and note that the minimum cost solution is to always declare that there is a target present!

	HI	LO
H_0	2/3	72/73
H_1	1/3	1/73

a posteriori probability matrix KLM

	HI	LO
H_0	1002	44.05
H_1	30	30

Risk matrix CKLM

- This example is skewed because of the high cost C_{01}. As might be expected, the use of costs is even more controversial than the use of probabilities. For example, while C_{00}, C_{10}, and C_{11} may be readily quantifiable from budgetary considerations, C_{01} might be pure speculation. No American city has been bombed. If C_{01} were halved, we would have the same rule as the maximum-likelihood decision rule.
- **Exercise:** Verify this and calculate
- The simple (minimum-error-probability) version of Bayes' rule is obtained if $C_{ij} = 1$ for $i \neq j$ and $C_{ii} = 0$ (uniform cost)
- $\sum_{k=0}^{m-1} \$C_{ik} \times P(H_k|A_j)$ reduces to $1 - P(H_i|A_j)$, and we declare in favor of the maximum $P(H_i|A_j)$
- **Exercise:** Verify this claim.
- **Likelihood ratio**
- The likelihood ratio is a very useful concept. All our decision rules can be formulated in terms of testing whether the likelihood ratio exceeds an appropriately specified *threshold*. We shall only consider the case of two hypotheses H_0 and H_1.
- **Definition:** If event A_i is observed, the likelihood ratio is defined to be $\Lambda(A_i) = P_1(A_i)/P_0(A_i)$, the ratio of the likelihoods of the observation under the two hypotheses.
- **Maximum-likelihood decision rule:** Here, if we observe A_i, we compare $P_0(A_i)$ and $P_1(A_i)$, the two entries in the i-th column of the likelihood matrix L and declare that H_1 is true if $P_1(A_i) > P_0(A_i)$ or declare that H_0 is true if $P_1(A_i) < P_0(A_i)$. This can be stated in terms of $\Lambda(A_i)$ as follows: If $\Lambda(A_i) > 1$, declare that H_1 is true while if $\Lambda(A_i) < 1$, declare that H_0 is true Here, the *threshold* is 1, and our declaration depends on whether the likelihood ratio exceeds the threshold or not.
- **MAP or Bayes' decision rule:** Here, if we observe A_i, we compare $P_0(A_i)P(H_0)$ and $P_1(A_i)P(H_1)$, the two entries in the i-th column of the joint probability matrix J, and declare that H_1 is true or H_0 is true according as $P_1(A_i)P(H_1) > P_0(A_i)P(H_0)$ or $P_1(A_i)P(H_1) < P_0(A_i)P(H_0)$. This can be stated in terms of $\Lambda(A_i)$ as follows: If $\Lambda(A_i) > P(H_0)/P(H_1)$, declare that H_1 is true while if $\Lambda(A_i) < P(H_0)/P(H_1)$, declare that H_0 is true. Here, the threshold is $P(H_0)/P(H_1)$ and our declaration depends on whether the likelihood ratio exceeds the threshold or not.
- **Minimum cost Bayes' decision rule:** Here we compare the two entries in the i-th column of the risk matrix CKLM. Once again, the decision can be formulated in terms of the likelihood ratio $\Lambda(A_i)$ being compared to the threshold $\frac{(C_{10} - C_{00})P(H_0)}{(C_{01} - C_{11})P(H_1)}$ with the decision favoring H_1 if $\Lambda(A_i)$ exceeds the threshold. This reduces to the previous case if the costs are uniform.
- **Exercise:** Work the radar target problem with this formulation.

Maximum-likelihood estimation

- The principle of maximum likelihood can also be applied to problems of estimating the numerical value of a parameter p (say). Suppose that we observe an event A. The probability of this event may depend on the unknown value of the parameter p. Thus, we ask "Which value of the parameter p maximizes the probability of A?" The answer is the maximum-likelihood estimate of the parameter p — it is the value of p that maximizes the likelihood of the observed event A.
- **Example:** K = "student knows the answer to multiple-choice question"
  C = "student answers question correctly"
- Five choices of answer for each question
- $P(C|K) = 1$ and $P(C|K^c) = 0.2$
- If $P(K) = p$, what is $P(K|C)$?

$$P(K|C) = \frac{P(C|K)P(K)}{P(C)} = \frac{P(C|K)P(K)}{P(C|K)P(K) + P(C|K^c)P(K^c)}$$

$$= \frac{1\times p}{1\times p + 0.2\times(1-p)} = \frac{p}{0.2 + 0.8p}$$ Note that $P(C) = 0.2 + 0.8p$

- Values:

p	P(K|C)
0	0
1/4	5/8
1/2	5/6
3/4	15/16
1	1

- $p = 1/2$ $P(K|C) = 5/6$. Does this make sense?
- Student knows 50% of the answers, guesses on the rest
- Correct response *(usually)* to 60% = 50% (knows this stuff) + 10% (1/5th guesses)
  Thus, 5/6ths of the correct answers are based on *knowledge*, 1/6th on guessing
- What's this "usually" stuff?
- *Guessing* on N problems can result in correct answers to 0, or 1, or 2, or … , or all N of the problems
- Of these N+1 possibilities, it is *most probable* that 0.2N of the answers will be guessed correctly. We will prove this result later in the course.
- max P{"m correct"} = P{"0.2N correct"}
- On *average*, 0.2N answers will be guessed correctly
- Average number of answers that will be guessed correctly

$$= \sum_{m=0}^{N} m\times P\{\text{m guessed correctly}\} = 0.2N$$

- P{"m correct"} = *likelihood* of m correctly guessed answers
- The event "0.2N correct" has the largest probability — it has the *maximum likelihood*, but we cannot be *sure* that it occurred
- How many correct answers resulted from lucky guesses?
- Given a student's answer sheet, we cannot say for *sure*
- But we can *estimate* how many correct answers were the result of guesswork, and be reasonably confident that the estimate is accurate
- In an average or probabilistic sense, we can be confident that our methodology gives accurate estimates

- But we cannot say for sure how accurate our estimate is for a *particular* answer sheet
- How many correct answers did the student know?
- How many correct answers resulted from lucky guesses?
- Maximum-likelihood estimate of p is the one that maximizes the likelihood of the observed number of correct answer.
- *Assume* that the most likely event (the one with maximum likelihood) occurred
- Thus, *assume* that 0.2N correct answers were the result of lucky guesses
- But what is N?
- The *assumption* that 0.2N correct answers were the result of lucky guesses is equivalent to the *assumption* that 0.8N incorrect answers (these are *all* the incorrect answers!) were the result of unlucky guesses!
- Count how many answers were incorrect and "equate" that to 0.8N
- Maximum-likelihood *estimate* of N is denoted by $\hat{N}$
  $\hat{N}$ = (# incorrect answers)/0.8 = 1.25×(# incorrect answers)
- $\hat{N}$ = 1.25×(# incorrect answers) is our *estimate* of N; we hope it is close to N — maybe even exactly the same as N
- We *estimate* that $0.2\hat{N}$ = 0.25×(# incorrect answers) were answered correctly due to lucky guesses
- 0.25×(# incorrect answers) is the maximum-likelihood (ML) *estimate* of how many of the correct answers were the result of lucky guesses
- The number of correct answers is reduced by 25% of the number of wrong answers as a "guessing penalty"
- A probabilistic version of all this begins with the fact that P(C) = 0.2 + 0.8p
- $P(C|K) = 1$ and $P(C|K^c) = 0.2$ are natural assumptions
- We want to estimate p, i.e. P(K), but we can only count how many problems were answered (in)correctly
- ML estimate of P(C) is $\hat{P}(C) = \dfrac{\text{\# answered correctly}}{\text{\# problems attempted}}$
- This is a *relative frequency estimate* of P(C) based on what actually happened on a particular answer sheet
- Probability theory gives us that P(C) = 0.2 + 0.8p
- *Assuming* P(C) equals its ML *estimate* gives $\hat{P}(C) = 0.2 + 0.8\hat{p}$ where $\hat{p}$ is the ML estimate of p
- The ML estimate of p is given by $\hat{p} = \dfrac{\hat{P}(C) - 0.2}{0.8} = 1.25\hat{P}(C) - 0.25$
- The ML *estimate* of p is $\hat{p} = 1.25\hat{P}(C) - 0.25 = \hat{P}(C) - 0.25\times\left[1 - \hat{P}(C)\right]$
- $\hat{P}(C) = \dfrac{\text{\# answered correctly}}{\text{\# problems attempted}}$; • $1 - \hat{P}(C) = \dfrac{\text{\# answered incorrectly}}{\text{\# problems attempted}}$
- $\hat{p} = \hat{P}(C) - 0.25\times\left[1 - \hat{P}(C)\right] = \dfrac{\#\text{correct} - 0.25\times(\#\text{incorrect})}{\text{\# problems attempted}}$
- As before, we see that a "guessing penalty" equal to 25% of the number of incorrectly answered problems is assessed
- The guessing penalty is always applied in the manner indicated

- If the maximum-likelihood event actually occurs, the guessing penalty correctly compensates for this
- Otherwise, the student is lucky (or unlucky!)
- A student who knows 80 answers and guesses the remaining 20 usually gets 84 = 80 + (20% of 20) of the problems right
- Guessing penalty reduces the score by 4 = 25% of the 16 incorrect answers and thus compensates perfectly
- If the student is lucky and gets 88 answers correct, the guessing penalty reduces the score by only 3 = 25% of 12
- Getting 4 extra correct answers by sheer dumb luck has boosted the post-penalty score from 80 to 85
- The unlucky student who answers all 20 problems incorrectly suffers the further indignity of getting a post-penalty score of only 75
- Moral: it is better to be born lucky than unlucky
- Is there any hope for unlucky test-takers?
- If on each problem to which the answer is unknown, *some* of the choices are obviously wrong and can be eliminated, then choosing between fewer alternatives boosts the score
- Foe example, suppose that $P(C|K^c) = 1/3$ instead of 1/5

  $P(C) = P(C|K)P(K) + P(C|K^c)P(K^c) = 1\times p + (1/3)\times(1-p) = (1/3) + (2/3)\times p$

  $= \frac{2p + 1}{3} > 0.2 + 0.8p$
- $\hat{p} = \frac{\hat{P}(C) - 0.2}{0.8} = \frac{5p+1}{6} > p$
- Thus, student is rewarded for "partial knowledge"
- Guessing penalties are applied using the number of wrong answers only. There is no penalty if questions are not answered at all
- MORAL: If you must guess, guess intelligently
- The principle of maximum-likelihood estimation says that the "best" estimate of the value of a parameter is the number that maximizes the likelihood (the probability) of the observation(s)
- This is the classical approach or the frequentist approach to parameter estimation
- **Example:** A plant makes N widgets (N very large) each day. M of these are defective. The quality control engineer tests n << N of the widgets and finds that k of the n are defective. What should she estimate the value of M to be?
- Roughly speaking, $M \approx (k/n)N$
- Obviously, $k \le M \le N-(n-k)$
- A random subset of size n is drawn (without replacement) from the set of size N that contains M defective widgets
- $P(k \text{ of } n \text{ defective}) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$
- ML estimate of M = whatever value of M maximizes this probability
- $P(k \text{ of } n \text{ defective}) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ is a function f(M) of known and fixed parameters N, n, k, and the unknown quantity M. Maximize this w.r.t. M

- It is a bad idea to differentiate with respect to M in order to find the maximum. Instead just systematically evaluate f(r) for r = k, k+1,…,N–(n–k)
- We shall see that successive values of f(•) increase for a while, peak at some $\hat{M}$ (which is the ML estimate of M!) and then decrease

  $f(k) < f(k+1) < \cdots\cdots < f(\hat{M}) > f(\hat{M}+1) > \cdots > f(N-(n-k))$
- If $\frac{f(r)}{f(r+1)} < 1$, $f(r) < f(r+1)$; $r < \hat{M}$
- If $\frac{f(r)}{f(r+1)} = 1$, $f(r) = f(r+1)$
- If $\frac{f(r)}{f(r+1)} > 1$, $f(r) > f(r+1)$; $r > \hat{M}$
- Since $\binom{L}{m} = \frac{L!}{m!(L-m)!}$, we get after some substitutions and cancellations that

  $\frac{f(r)}{f(r+1)} = \frac{r-k+1}{r+1} \times \frac{N-r}{N-r-n+k}$
- Is (r–k+1)(N–r) larger, smaller, (or possibly the same as) (r+1)(N–r–n+k)?

  $\frac{f(r)}{f(r+1)} < 1$ if k(N+1)–(r+1)n > 0, that is, if r < (k/n)(N+1) – 1
- $\frac{f(r)}{f(r+1)} > 1$ if k(N+1)–(r+1)n < 0, that is, if r > (k/n)(N+1) – 1
- $\frac{f(r)}{f(r+1)} = 1$ if k(N+1)–(r+1)n = 0, that is, if r = (k/n)(N+1) – 1
- k, n, and N are fixed so that (k/n)(N+1)–1 is some fixed number
- If (k/n)(N+1)–1 = 10.23 (say), then f(10)/f(11) < 1 while f(11)/f(12) > 1 so that f(•) peaks at 11
- If (k/n)(N+1)–1 = 11, then f(10) < f(11) = f(12) > f(13)
- Conclusion: f(r) is maximum at r = é(k/n)(N+1)–1ù, the least integer ≥ (k/n)(N+1)–1
- Conclusion: the ML estimate of M is $\hat{M} = é(k/n)(N+1)–1ù \approx \frac{k}{n}(N+1) - 1$

  $\approx \frac{k}{n}N - \left(1 - \frac{k}{n}\right) \approx \frac{k}{n}N$