

Chapter 3 Independence

- Suppose that it is known that event A has occurred. We update the probability of the event B from $P(B)$ to $P(B|A) = \frac{P(BA)}{P(A)}$
- $P(B|A)$ can be larger, smaller, or the same as $P(B)$
- If $P(B|A) = P(B)$, then the occurrence of A does not change our estimate of the probability of B
- $P(B|A) = P(B)$ $P(B^c|A) = P(B^c)$
- The occurrence of A does not change the odds favoring B — $P(B):P(B^c) = P(B|A):P(B^c|A)$
- Equivalently, the occurrence of B does not change the odds favoring A
- In such cases, A and B are said to be *independent events*
- The formal definition of independence uses A and B in more symmetric fashion
- If $P(B|A) = P(AB)/P(A) = P(B)$, then $P(AB) = P(A)P(B)$

• **Definition:** Events A and B are said to be (stochastically) independent if

$$P(AB) = P(A)P(B)$$

- Do not confuse the notions of independent events and mutually exclusive events
- $P(AB) = P(A)P(B) = 0$ for independent events, so they *cannot be* mutually exclusive
- $P(AB) = 0 \neq P(A)P(B)$ for mutually exclusive events, so they *cannot be* independent
- Magic mantra to memorize
- Independent events cannot be mutually exclusive
Mutually exclusive events cannot be independent
- Note: There are trivial uninteresting exceptions when either $P(A)$ or $P(B)$ equals 0
- If A and B are independent events, then so are A and B^c , A^c and B , and A^c and B^c
- $P(AB^c) = P(A) - P(AB) = P(A) - P(A)P(B) = P(A)[(1 - P(B))] = P(A)P(B^c)$
- The other two results can be proved similarly (Try them!)
- Independence of events is of great help in calculations;
 $P(AB)$ is just $P(A)P(B)$ for independent events A and B
- For example, $P(A \cup B) = P(A) + P(B) - P(AB) = P(A) + P(B) - P(A)P(B)$
- $P(A \cup B^c) = P(A) + P(B^c) - P(AB^c) = P(A) + P(B^c) - P(A)P(B^c)$
- Even though independence is a great help in calculations, it cannot and should not be used indiscriminately (e.g., whenever you are stuck in solving a problem and can't figure out a way to proceed)
- **DO NOT ASSUME** that events are independent unless the problem explicitly says so
- On homework and exams, if you are asked to determine whether or not A and B are independent events, just find $P(AB)$ and compare it to the product $P(A)P(B)$
- But this is not the way the concept of independence is used in probability theory
- **Physical independence versus stochastic independence**
- It might be reasonable to assume on the basis of study and preliminary analysis of the physical phenomenon under consideration that two events are *physically independent*
- Physical independence is an *assumption* that we justify based purely on physical considerations
- We conclude from physical principles that occurrence (or non-occurrence) of one event seems to have no influence on the occurrence of another event
- **Example:** two successive tosses of a fair coin. Does the occurrence of a head on the first toss influence the result of the second toss?
Is the second more likely to be head? or is it more likely to result in a tail to “balance” the results?

- It is commonly assumed that the two tosses are *physically independent* and one toss does not influence the other
- This assumption need not be valid if (say) the coin is damaged when it is tossed the first time
- **Example:** Antennas pointed in different directions pick up noise voltages
- **Example:** Arrivals of packets, jobs, disk accesses, phone calls, cars, etc
- *Assumptions* of physical independence are usually made in most such cases
- Suppose that we know, or assume, or believe, that two events A and B are *physically independent*
- In this case, we *set* $P(AB) = P(A)P(B)$ that is, we insist that A and B be *stochastically independent* as well
- If we believe that physical independence is a reasonable assumption, we insist that this independence be *reflected* in the probability measure
- If we have measured $P(A)$ and $P(B)$ (e.g. via relative frequencies), we *assume* that $P(AB) = P(A)P(B)$
- This assumption can be tested via relative frequency measurements
- If $\text{rel. freq}(AB) / \text{rel. freq}(A) \times \text{rel. freq}(B)$, our assumption of physical independence may need to be re-examined
- However, if, on calculating $P(AB)$ and comparing it to $P(A)P(B)$, we discover that A and B are stochastically independent events, we *should not* automatically assume that A and B are physically independent as well
- **Example:** A and B are the events that input #1 and input #2 respectively to a 2-input XOR gate are logical 1's. We assume that A and B are independent and that $P(A) = P(B) = 1/2$. Let C be the event that the output is a logical 1.
- What is $P(C)$?
- The output is 1 if and only if exactly one of the two inputs is 1. Thus, $C = A \bar{B} \cup \bar{A} B$
 $= A\bar{B} \cup \bar{A}B$ and $P(C) = P(A\bar{B}) + P(\bar{A}B) = P(A)P(\bar{B}) + P(\bar{A})P(B) = 1/2$
- Are A and C independent events?
- $AC = A(A \bar{B} \cup \bar{A}B) = A(A\bar{B}) \cup A(\bar{A}B) = A\bar{B}$ and hence $P(AC) = P(A\bar{B}) = P(A)P(\bar{B}) = 1/4 = P(A)P(C)$
- Hence, A and C are independent events!
- Does the output of an XOR gate not depend at all on one of the inputs?
- All we have found is that A and C are *stochastically independent* events
- We *cannot* conclude from this that A and C are *physically independent* events
- In fact, A and C are very much *physically dependent*
- Now suppose that $P(A) = P(B) = 0.5000001$. Repeating the calculations gives $P(C) = .4999991$ instead of 0.5. Also, $P(AC) \neq P(A)P(C)$ so A and C are no longer (stochastically) independent
- It would be very difficult to determine practically whether $P(A) = 0.5$ or 0.5000001
- The events A and C are not stochastically independent if $P(A) = 0.5000001$
- This illustrates the idea that stochastic independence is a property of the probability measure — minor changes in $P(\bullet)$ destroyed the equality
- In contrast, physical independence is a property of the events (that is, of the physical phenomenon that we are modeling)
- Physically independent events are always assumed to be stochastically independent
- **Summary:** Where physical independence is a reasonable assumption, we insist that this independence be *reflected* in the probability measure
- We *set* $P(AB) = P(A)P(B)$ whenever we believe that A and B are independent events

- If the probability measure is such that *calculation* shows $P(AB) = P(A)P(B)$, we *do not* jump to the conclusion that the events are also physically independent
- Now that the distinction is clear, we drop the adjective *stochastically* in our discussion of independence
- Consider three events A, B, and C. When should we say that they are independent?
- Obviously, we want $P(ABC) = P(A)P(B)P(C)$ to hold. It also seems reasonable that A and B also should be independent of each other

• **Definition:** A, B, and C are said to be independent events (mutually independent events) if all four of the following conditions hold:

$$P(AB) = P(A)P(B)$$

$$P(AC) = P(A)P(C)$$

$$P(BC) = P(B)P(C)$$

$$P(ABC) = P(A)P(B)P(C)$$

- The first three of these equations *do not* imply the fourth, and the fourth equation *does not* imply the first three either
- $\{A_1, A_2, \dots, A_n\}$ is a collection of independent events if

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n)$$
 and every subcollection that has two or more of the A_i 's in it is also a collection of independent events

$$P(A_i A_j \dots A_m) = P(A_i)P(A_j) \dots P(A_m)$$

- Alternatively, the events are independent if all 2^n of the following conditions hold:

$$P(A_1^* A_2^* \dots A_n^*) = P(A_1^*)P(A_2^*) \dots P(A_n^*)$$

where A_i^* denotes either A_i or A_i^c (same on both sides!)

- **Example:** Prop. 4.4, p. 35 states that $P(A_1 A_2 \dots A_n)$

$$= P(A_1) - \sum_{i=1}^{n-1} P(A_i A_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} P(A_i A_j A_k) - \dots + (-1)^{n+1} P(A_1 A_2 \dots A_n)$$

- For independent events, this gives $P(A_1 A_2 \dots A_n)$

$$= P(A_1) - \sum_{i=1}^{n-1} P(A_i)P(A_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} P(A_i)P(A_j)P(A_k) - \dots + (-1)^{n+1} P(A_1)P(A_2) \dots P(A_n)$$

- This is a **stupid** way of calculating the probability of the union. It is much better to proceed as follows:

- $P(A_1 A_2 \dots A_n) = 1 - P((A_1 A_2 \dots A_n)^c) = 1 - P(A_1^c A_2^c \dots A_n^c)$ by DeMorgan's law

$$= 1 - P(A_1^c)P(A_2^c) \dots P(A_n^c)$$
 by independence $= 1 - \prod_{i=1}^n (1 - P(A_i))$

- **Useful result:** Any Boolean function of $\{A_1, A_2, \dots, A_n\}$ is independent of any other Boolean function as long as they do not include any events in common

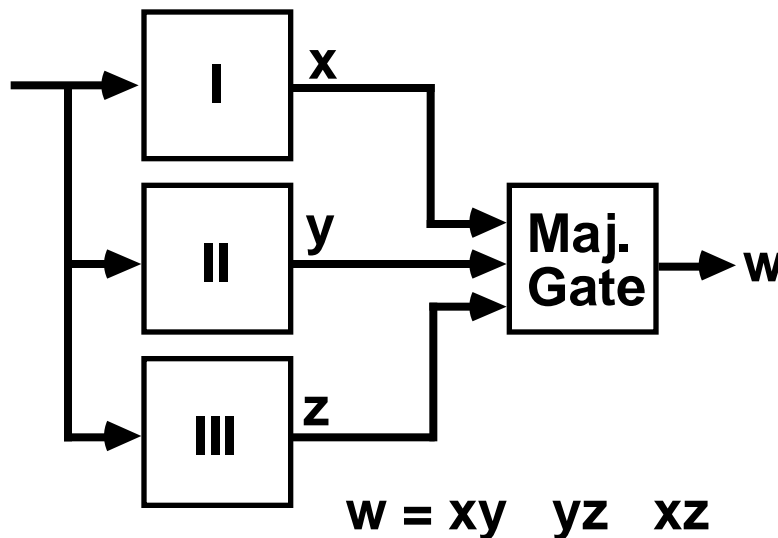
- **Example:** If A, B, C are independent events, then $A \cup B$ is independent of C

- **Example:** If $\{A, B, C, D, E, F, G, H\}$ are independent events, then $A \cup C$, $B \cup H$, D, and $E \cap F \cap G$ are independent events

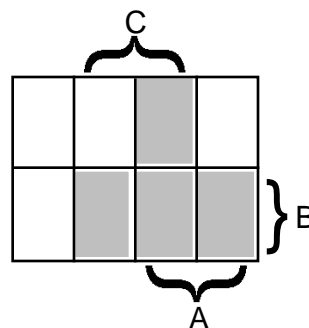
- $P((A \cap C)(B \cap H)D(E \cap FG)) = P(A \cap C)P(B \cap H)P(D)P(E \cap FG)$
- $P(E \cap FG) = P(E) + P(FG) - P(EFG) = P(E) + P(F)P(G) - P(E)P(F)P(G)$
- When events are shared among Boolean functions, independence cannot be guaranteed. However, problem analysis is still possible in conjunction with Karnaugh maps
- **Example:** If A, B, and C are independent, $A \cap C$ and $B \cap C$ are *not* independent events.
- $P((A \cap C)(B \cap C)) = P(C \cap AB) = P(C \cap ABC^c) = P(C) + P(ABC^c) = P(C) + P(A)P(B)P(C^c)$

Triple modular redundancy (TMR)

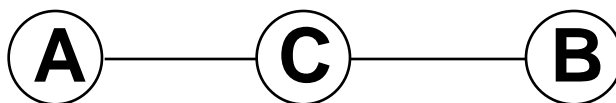
- Three identical copies of a logic circuit are supplied with identical inputs
- If *no more than one* circuit produces an incorrect output, then the *majority* of the three outputs is still correct



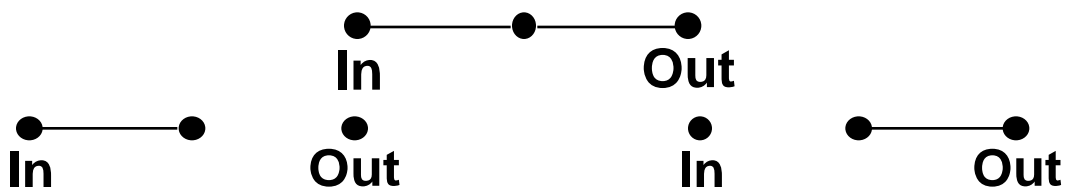
- **Example:** A, B, C denote the events that respectively x, y, and z are incorrect, i.e., *are not* the desired Boolean function of the input variables
- $P(A) = P(B) = P(C) = p$. We assume that A, B, C are independent events
- What is the probability that $w = xy + yz + xz$, the output of the majority gate, is incorrect?
- Incorrect = not the desired Boolean function of the input
- If outputs are *supposed* to be 1, then $w = 0$ if and only if at least two of x, y, z are 0 (i.e. are incorrect)
- If outputs are *supposed* to be 0, then $w = 1$ if and only if at least two of x, y, z are 1 (i.e. are incorrect)
- Thus, the majority gate output is incorrect (i.e. not the desired Boolean function) if and only if at least two of A, B, C occur, that is, if the event $AB \cup BC \cup AC$ occurs



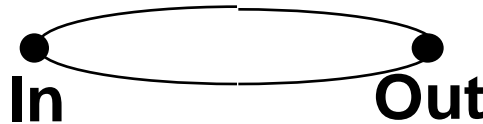
- $P(AB \ BC \ AC) = P(AB) + P(BC) + P(AC) - 2P(ABC)$
- $P(AB \ BC \ AC) = P(AB)+P(BC)+P(AC)-2P(ABC)$
 $= P(A)P(B) + P(B)P(C) + P(A)P(C) - 2P(A)P(B)P(C) = 3p^2-2p^3 \approx 3 \times 10^{-8}$ if $p = 10^{-4}$
- Thus, TMR has improved reliability considerably
- **Example:** (continued) The majority gate also can fail and produce the wrong output. If this failure (event D) is independent of A, B, C, and $P(D) = q$, what is the probability that the majority gate output is incorrect?
- Failure of majority gate means output = majority of inputs
- Incorrect majority gate output when majority gate is working correctly but two or more logic circuits fail OR majority gate fails while two or more logic circuits produce correct outputs
- Suppose that two or more circuits have failed and have output 0 when it should be 1. *If the majority gate has also failed*, then the majority gate output (which *should be* a 0 when two or more inputs are 0) is 1 instead, i.e. it is right!
- Moral: two wrongs do make a right every now and then!
- $\{A,B,C,D\}$ is an independent collection of events
- Let $E = AB \ BC \ AC$. Then, E and D are independent events
 $P(\text{output is incorrect}) = P(ED^c) + P(E^cD) = P(E)P(D^c) + P(E^c)P(D)$
- $P(\text{output is incorrect}) = (3p^2 - 2p^3)(1-q) + (1 - (3p^2 - 2p^3))q = q(1-2(3p^2-2p^3)) + (3p^2-2p^3)$
- The majority gate is a simple device compared to the logic circuits. It is reasonable to assume that $q \ll p$
- $P(\text{output is incorrect}) = q(1-2(3p^2-2p^3)) + (3p^2-2p^3)$
 If $p = 10^{-4}$ and $q = 10^{-6}$, then $P(\text{output is incorrect}) \approx q$!!
- Moral: the reliability of the majority gates determines the reliability of the TMR system
- Can a system be more reliable than its most reliable component?
- How do we model the failures of complicated systems?
- How do we calculate failure probability for system from the failure probabilities of the subsystems (parts?)
- Careful justification is necessary of assumptions of independence of failures
- **Example:** Telephone calls from City A to City B are routed through City C and thus traverse two links A–C and C–B. Calls successfully get through if *both* the links are in working condition, and fail otherwise



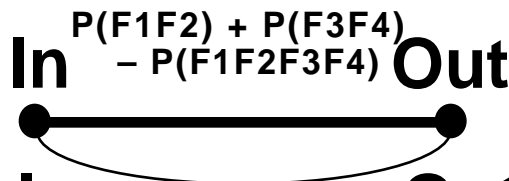
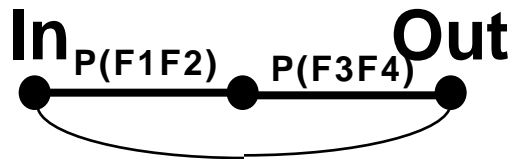
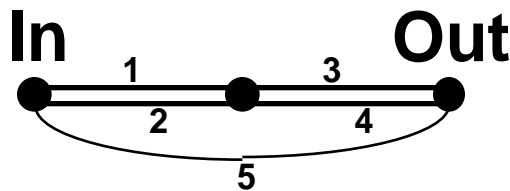
- This is a special case of a general system that works only if both subsystems work
- Graph model (circuit model) for reliability calculations
- Subsystems are represented by links (edges in graph)
- Special vertices called In and Out
- System works if there is a path from In to Out through components that are working



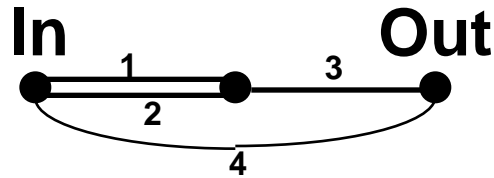
- Note that the links only model how the failures of various parts affect system failure. In the telephone call problem, the telephone lines are actually connected as in the graph model, but in general, the *actual* system might be quite different (a mechanical system, for example)
- Let A and B denote the events that the individual links are working
- $P(\text{system works}) = P(AB) = P(A)P(B)$ if A and B are assumed to be independent events
- More generally, if F_1, F_2, \dots, F_n denote failure events for n serial links, then $P(\text{system works}) = P(F_1^c F_2^c \dots F_n^c) = P(F_1^c)P(F_2^c) \dots P(F_n^c)$ if the events are assumed to be independent
- System works if there is a path from In to Out through components that are working



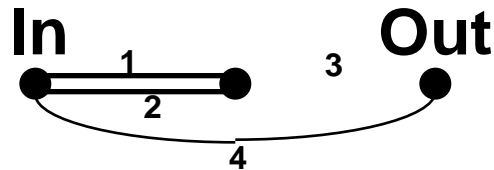
- Let A and B denote the events that the individual links are working
- $P(\text{system works}) = P(A \cup B) = P(A) + P(B) - P(AB) = P(A) + P(B) - P(A)P(B)$ if A and B are independent events
- More easily, $P(\text{system fails}) = P(A^c B^c) = P(A^c)P(B^c)$ if independent $= (1 - P(A))(1 - P(B)) = 1 - (P(A) + P(B) - P(A)P(B))$
- More generally, if F_1, F_2, \dots, F_n denote failure events for n parallel links, then $P(\text{system fails}) = P(F_1 F_2 \dots F_n) = P(F_1)P(F_2) \dots P(F_n)$ if the events are assumed to be independent
- What about more complicated systems?
- Analysis of more complicated graphs
- Replace parallel links with a single link with failure probability $P(F_1 F_2 \dots F_n)$
- Replace serial links with a single link with failure probability $1 - P(F_1^c F_2^c \dots F_n^c)$



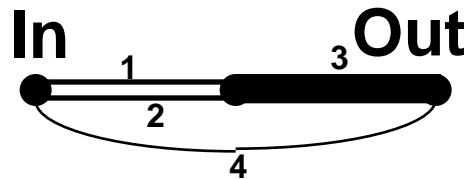
- Use conditional probabilities



- If link #3 has failed, system fails if link #4 fails



- If link #3 is good, system fails only if all of #1, #2, #4 fail



- $P(\text{system fails}) = P(F_4|F_3)P(F_3) + P(F_1F_2F_4|F_3^c)P(F_3^c)$
Independence of failures implies that $P(F_4|F_3) = P(F_4)$ and $P(F_1F_2F_4|F_3^c) = P(F_1F_2F_4) = P(F_1)P(F_2)P(F_4)$
- Hence, $P(\text{system fails}) = P(F_4)P(F_3) + P(F_1)P(F_2)P(F_4)P(F_3^c)$

Independent Trials

- Consider two experiments with probability spaces (Ω_1, F_1, P_1) and (Ω_2, F_2, P_2) respectively
- Ω_i = collection of outcomes of experiment #i
- F_i = family of events defined for experiment #i
- P_i = probability measure for experiment #i
- A *compound* experiment consists of performing both these *subexperiments*
- What is (Ω, F, P) for this compound experiment?
- Sample space Ω is the Cartesian product $\Omega_1 \times \Omega_2$, i.e., $\Omega = \{(x, y): x \in \Omega_1, y \in \Omega_2\}$
- **Example:** If the compound experiment consists of tossing a coin and rolling a die, then $\Omega = \{H, T\} \times \{1, 2, \dots, 6\} = \{(H, 1), (H, 2), \dots, (H, 6), (T, 1), (T, 2), \dots, (T, 6)\}$
- Similarly, $F = F_1 \times F_2 = \{(A, B): A \in F_1, B \in F_2\}$
- If compound outcome (x, y) occurs, then the compound event (A, B) is said to have occurred if $x \in A$ and $y \in B$. In other words, A occurred on first subexperiment and B on the second
- Example: (H, even) occurs if outcome = $(H, 2), (H, 4)$ or $(H, 6)$
- The statement “compound events (A, B) and (C, D) both occurred” means that AC occurred on the first subexperiment and BD on the second. Thus, $(A, B) \cap (C, D) = (A \cap C, B \cap D)$
- **Example:** If (H, even) occurred and so did (H, prime) , then $(H \cap H, \text{even} \cap \text{prime}) = (H, 2)$ occurred
- Unions are more complicated

- If $(A,B) \cap (C,D)$ occurred, it is *not true* that $(A \cap C, B \cap D)$ occurred. The latter includes (x,y) with $x \in A$ and $y \in D$; the former does not
- **Example:** $(H,\text{even}) = \{(H,2), (H,4), (H,6)\}$ $(T,\text{prime}) = \{(T,2), (T,3), (T,5)\}$
Clearly, $(H,\text{even}) \cap (T,\text{prime})$ contains just the 6 compound outcomes listed above whereas $(H \cap T, \text{even} \cap \text{prime})$ includes outcomes such as $(T,4) \in (H,\text{even}) \cap (T,\text{prime})$
- Special compound events: $(A, _2)$ is just another way of saying that event A occurred on the first subexperiment
- We don't care *which* outcome occurred on the second; any will do, i.e., $_2$ occurs on second subexperiment. Similarly for $(_1, B)$
- What probability should be assigned to the compound events (A,B) ?
- Obviously, $P(A, _2) = P_1(A)$ since $P(A, _2)$ should just be the probability that A occurred; and $P(_1, B) = P_2(B)$ since $P(_1, B)$ should be the probability that B occurred.
- In other words, the compound probability measure should include the probability measures of the individual experiments as special cases.
- **Example:** The probability assignment shown satisfies $P_1(H) = P_1(T) = 1/2$, $P_2(i) = 1/6$ which are the probabilities for a fair coin and a fair die

H	1/24	3/24	1/24	3/24	1/24	3/24
T	3/24	1/24	3/24	1/24	3/24	1/24
	1	2	3	4	5	6

- The subexperiments are said to be *independent experiments* if for all $A \in F_1$, $B \in F_2$,
 $P(A,B) = P_1(A)P_2(B)$.
- The assumption is that the subexperiments are *physically independent*
- Physical independence implies stochastic independence
- In the above example, we do not have independent experiments even though the probabilities of the subexperiments are as they ought to be.
- **Example:** The probability assignment shown satisfies $P_1(H) = P_1(T) = 1/2$, $P_2(i) = 1/6$ which are the probabilities for a fair coin and a fair die

H	1/12	1/12	1/12	1/12	1/12	1/12
T	1/12	1/12	1/12	1/12	1/12	1/12
	1	2	3	4	5	6

- With this assignment, the assumed physical independence of the coin toss and die roll is also reflected in the stochastic independence of the events.
- Thus, just having the compound probability measure include the probability measures of the individual experiments as special cases is not sufficient to guarantee independence of the subexperiments. That is, just having

$$P(A, _2) = P_1(A)P_2(_2) = P_1(A) \text{ and } P(_1, B) = P_1(_1)P_2(B) = P_2(B)$$

is not enough. We also need to have

$$P(A,B) = P_1(A)P_2(B) \text{ for all events A and B}$$

in order to say that the subexperiments are *independent experiments*

- The case of two independent subexperiments is readily generalized to the case of many subexperiments

- **Generalization:** Compound outcome is (x_1, x_2, \dots, x_N) , x_i \in Ω_i
- Compound events are (A_1, A_2, \dots, A_N) , $A_i \in \mathcal{F}_i$ and are said to have occurred if $x_i \in A_i$ for $1 \leq i \leq N$
- The compound probability measure is $P(A_1, A_2, \dots, A_N) = \prod_{i=1}^N P_i(A_i)$
- The most important special case of this is when the subexperiments are *identical*, or *repetitions* of a single experiment. These are referred to as *repeated independent trials* of the (sub)experiment. An example of this that we studied earlier is sampling with replacement.

Repeated Independent Trials

- Simple experiment with probability space (Ω, \mathcal{F}, P) is repeated independently N times
- These are N independent trials, also called *Bernoulli trials*, of the same experiment
- We consider the compound experiment
- We have already encountered examples of this formulation
- Sampling from a set with replacement consists of independent repetitions of the experiment; sampling without replacement does not.
- Notions of relative frequency as a measure of probability are based on the assumption of repeated independent trials of the experiment
- If the repeated trials are not independent, the relative frequency does not reflect the probability of an event
- **Important difference:** with repeated independent trials, the *same event* can occur on *different trials*, so that the compound event (A_1, A_2, \dots, A_N) could very well be of the form $(A, A, A^c, A^c, A, \dots)$
- $(A, A, A^c, A^c, A, \dots)$ means that event A occurred on first, second, fifth, ... trials and A^c occurred on third, fourth, ... trials
- For a *single experiment*, A and A^c cannot occur on the same trial
- A is not independent of itself
- Yet it is perfectly valid to assign probability $P(A)P(A)P(A^c)P(A^c)P(A) \dots$ to the event $(A, A, A^c, A^c, A, \dots)$ because the first and second A 's are on *different* trials, and the A 's and A^c 's are occurring on *different* trials
- Independence results from the independent *trials*
- Suppose that A and B are *dependent* events defined on the simple experiment, that is, $P(AB) \neq P(A)P(B)$
- Nonetheless, the probability that A occurs on the *first* trial and B on the *second* trial is $P(A, B, \dots) = P(A)P(B)P(\dots) \dots = P(A)P(B)$
- Events that are dependent when we consider if they can occur simultaneously on any given trial are nonetheless independent if we ask whether they can occur on *different* (independent) trials
- The independence arises from the independent *trials*
- A most important question for problems involving N repeated independent trials is as follows: Let $B(k;N)$ denote the compound event that the event A occurred exactly k times on N trials, $0 \leq k \leq N$
- $P(A) = p$. What is $P(B(k;N))$?
- The event A can occur 0 or 1 or 2, ... , or N times on N trials
- *One and only one* of the events $B(0;N), B(1;N), \dots, B(N;N)$ must have occurred
- $B(0;N), B(1;N), \dots, B(N;N)$ is a *partition* of the (compound) sample space

- $\sum_{i=0}^N P(B(i;N)) = 1$
- If $B(k;N)$ has occurred, then we know that A occurred on exactly k of the N trials, and hence A^c must have occurred on the remaining $N - k$ trials
- Using independence of trials, can we set $P(B(k;N)) = [P(A)]^k [1 - P(A)]^{N-k} = p^k (1-p)^{N-k}$??
- Unfortunately, this is not quite right. So, where did we go wrong?
- We did not account for *where the events A occurred*
- **Example:** $N = 4$ and $k = 2$ $B(2;4)$ occurs if these occur

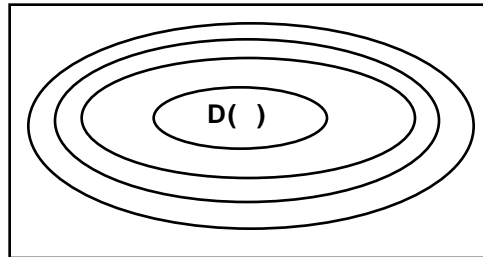
(A, A, A^c, A^c)	A occurred on trials $\#\{1,2\}$ with probability $p^2(1-p)^2$
(A, A^c, A, A^c)	A occurred on trials $\#\{1,3\}$ with probability $p^2(1-p)^2$
(A, A^c, A^c, A)	A occurred on trials $\#\{1,4\}$ with probability $p^2(1-p)^2$
(A^c, A, A, A^c)	A occurred on trials $\#\{2,3\}$ with probability $p^2(1-p)^2$
(A^c, A, A^c, A)	A occurred on trials $\#\{2,4\}$ with probability $p^2(1-p)^2$
(A^c, A^c, A, A)	A occurred on trials $\#\{3,4\}$ with probability $p^2(1-p)^2$
- Each of the 6 events has the same probability $p^2(1-p)^2$
- Why can't both (A, A, A^c, A^c) and (A, A^c, A, A^c) have occurred at the same time?
- The 6 events are mutually exclusive
- $P(C \cup D) = P(C) + P(D)$ if $CD = \emptyset$
- $P(B(2;4)) = 6p^2(1-p)^2 = 6p^2q^2$ where $q = 1-p$
- Note: the notation $q = 1-p$ is quite common, but is not universal. Be careful!
- The probability that A occurs on a *specified* set of k trials out of N (e.g., 1st, 3rd, 9th, ...) is $p^k(1-p)^{N-k} = p^kq^{N-k}$
- The probability that A occurs on some *other* specified set of k trials out of N (e.g., 1st, 2nd, 8th, ...) is also $p^k(1-p)^{N-k} = p^kq^{N-k}$
- *Both* these events could not have occurred at the same time. Both events have the same probability p^kq^{N-k}
- Each such event corresponds to a different subset of size k drawn from $\{1,2,3,\dots,N\}$. The sets are listed above for the case $N = 4, k = 2$.
- There are $\binom{N}{k}$ such subsets, and each corresponding event has probability p^kq^{N-k}
- These events are mutually exclusive
- The event $B(k;N)$ contains $\binom{N}{k}$ different mutually exclusive compound events of the form $(A, A, A^c, A^c, A, \dots)$, each with k A 's and $(N-k)$ A^c 's, each having probability p^kq^{N-k}
- $P(B(k;N)) = \binom{N}{k} p^k q^{N-k} = \binom{N}{k} p^k (1-p)^{N-k}$ is called a *binomial* probability
- $\sum_{k=0}^N P(B(k;N)) = \sum_{k=0}^N \binom{N}{k} p^k q^{N-k} = (p+q)^N = (p+1-p)^N = 1$ by the binomial theorem
- Binomial probabilities have been tabulated (see, for example, Abramowitz and Stegun, *Handbook of Mathematical Functions*)
- $P(A \text{ occurs no more than } k \text{ times on } N \text{ trials}) = \sum_{i=0}^k P(B(i;N)) = \sum_{i=0}^k \binom{N}{i} p^i q^{N-i}$

- $P(A \text{ occurs more than } k \text{ times on } N \text{ trials}) = \sum_{i=k+1}^N P(B(i;N)) = \sum_{i=k+1}^N \binom{N}{i} p^i q^{N-i}$
- $P(\text{at most } k) + P(\text{more than } k) = \sum_{i=0}^k \binom{N}{i} p^i q^{N-i} + \sum_{i=k+1}^N \binom{N}{i} p^i q^{N-i} = 1$
- The first sum contains $k+1$ terms, the second $N-k$ terms
- If $N = 100$, and you need to find $P(\text{at most } 95)$, just find $P(\text{more than } 95)$ and subtract from 1!
- Given $B(k;N)$ occurred, find the conditional probability of a specific sequence of k A's and $N-k$ A^cs, e.g. (A^c, A, A, \dots)
- The sequence is a subset of $B(k;N)$, so conditional probability is $p^k q^{N-k} / P(B(k;N)) = 1 / \binom{N}{k}$
- Given $B(k;N)$ occurred, find the conditional probability that A occurred on 3rd trial
- If A occurred on 3rd trial and $B(k;N)$ also occurred, then A must have occurred $k-1$ times on the remaining $N-1$ trials
- Independence of trials
- $P(A \text{ on 3rd and } B(k-1;N-1)) = p \times P(B(k-1;N-1))$
- $P(A \text{ on 3rd} | B(k;N)) = \frac{p \times P(B(k-1;N-1))}{P(B(k;N))} = p \times \frac{(N-1)! p^{k-1} q^{N-1-(k-1)}}{(k-1)!(N-k)!} \times \frac{k!(N-k)!}{N! p^k q^{N-k}} = \frac{k}{N}$
- In fact, $P(A \text{ on } i\text{-th} | B(k;N)) = k/N$
- Given $B(k;N)$ occurred, that is, event A occurred k times on N trials, the conditional probability that A occurred on the i -th trial is just k/N
- Given $B(k;N)$ occurred, the conditional probability that A occurred on the i -th as well as the j -th trial is $\frac{k(k-1)}{N(N-1)}$
- Independence of i -th and j -th trials and other $N-2$ trials
- $P(A \text{ occurred on } k \text{ of } N \text{ trials including the } i\text{-th and } j\text{-th trials}) = p^2 \times P(B(k-2;N-2))$
- $P(A \text{ on } i\text{-th and } j\text{-th} | B(k;N)) = \frac{p^2 \times P(B(k-2;N-2))}{P(B(k;N))} = p^2 \times \frac{(N-2)! p^{k-2} q^{N-2-(k-2)}}{(k-2)!(N-k)!} \times \frac{k!(N-k)!}{N! p^k q^{N-k}} = \frac{k(k-1)}{N(N-1)}$
- $P(A \text{ on } i\text{-th, } j\text{-th, } m\text{-th} | B(k;N)) = \frac{k(k-1)(k-2)}{N(N-1)(N-2)}$
- Conditional probability given $B(k;N)$ that A occurred on the i_1 -th, i_2 -th, ..., i_k -th trials?
- Here, we have specified the k trials on which A occurred!

$$P(A \text{ on } i_1\text{-th, } i_2\text{-th, } \dots, i_k\text{-th} | B(k;N)) = \frac{k(k-1)(k-2)\dots 1}{N(N-1)(N-2)\dots(N-k+1)} = 1 / \binom{N}{k}$$
- If A occurred k times on N trials, i.e., $B(k;N)$ occurred, the relative frequency of A is k/N . This is the *maximum-likelihood estimate* of $P(A) = p$
- Assume that $0 < p < 1$. What value of p maximizes the likelihood (probability) of the observation that A occurred k times on N trials?
- What value of p maximizes the likelihood (probability) $p^k(1-p)^{N-k}$ of the observation $(A, A, A^c, A^c, A, \dots)$ which consists of k A's and $N-k$ A^cs?
- Since $p^k(1-p)^{N-k}$ is a continuous function of p , we can use calculus to maximize
- $\frac{d}{dp} p^k(1-p)^{N-k} = k p^{k-1} (1-p)^{N-k} - p^k (N-k) (1-p)^{N-k-1}$
 $= p^{k-1} (1-p)^{N-k-1} [k(1-p) - p(N-k)] = 0$ if $p = \frac{k}{N}$

- Exercise: show that $p^k(1-p)^{N-k}$ has a maximum, not a minimum, at $p = \frac{k}{N}$
- $p^k(1-p)^{N-k}$, the probability of the observed sequence of k A's and $N-k$ A^c's is maximum at $p = k/N$, the relative frequency of A on the N trials
- This is the justification for using the observed relative frequency of an event A as an estimate of $P(A)$
- We now consider a sequence of repeated independent trials of an experiment
- On each trial, either A or A^c occurs. Assume $0 < p < 1$
- Let $C(k)$ denote the event that A occurred *for the first time* on the k -th trial
- What is $P(C(k))$?
- Occurrence of $C(k)$ implies that A *did not occur* on the 1st, 2nd, ... , $(k-1)$ -th trials
- $C(k)$ occurs if and only if $(A^c, A^c, A^c, \dots, A^c, A)$
 $\qquad\qquad\qquad 1 \qquad 2 \qquad 3 \qquad \dots \qquad k-1 \qquad k$
- By independence of the trials, $P(C(k)) = (1-P(A))^{k-1}P(A) = (1-p)^{k-1}p = q^{k-1}p$ ($q = 1-p$)
- Are we ignoring lots of things here?
- $C(1) = A$ occurred on first trial
- $C(2) = A$ occurred for the first time of second trial, i.e. (A^c, A) occurred
- In computing $P(C(1)) = p$, we ignored what happened on succeeding trials. Is this OK?
- Think of an *infinitely long* sequence of independent trials
- $C(1) = (A, \quad , \quad , \quad , \dots)$
 $C(2) = (A^c, \quad A, \quad , \quad , \dots)$
 $C(3) = (A^c, \quad A^c, \quad A, \quad , \dots)$
- $P(C(1)) = P(A) \times 1 \times 1 \times \dots = P(A)$
- $P(C(2)) = P(A^c) \times P(A) \times 1 \times 1 \dots$
- Thus, in computing $P(C(k))$, we can *ignore* what happened on further trials without affecting the probability calculations
- In fact, it is not even necessary to assume that further trials were performed!
- Note that $C(i) \cap C(j) = \emptyset$ for $i \neq j$
- $P(A \text{ occurs at least once on the first } k \text{ trials}) = P(C(1) \cup C(2) \cup C(3) \cup \dots \cup C(k))$
 $= P(C(1)) + P(C(2)) + \dots + P(C(k))$ because the events are mutually exclusive
 $= p + qp + q^2p + \dots + q^{k-1}p = p \frac{1 - q^k}{1 - q} = 1 - q^k$ ($p = 1 - q$)
- $P(A \text{ occurs at least once on the first } k \text{ trials}) = 1 - q^k$
- Instead of summing geometric series, it is *much easier* to calculate that the *complementary event* $D(k)$ (that A *did not occur at all* on the first k trials) has probability q^k
- $D(k)$ denotes the event that A *did not occur at all* on the first k trials
 $D(k) = (A^c, A^c, \dots, A^c, \dots)$
 $\qquad\qquad\qquad 1 \qquad 2 \qquad \dots \qquad k \qquad k+1$
- $D(k) = (C(1) \cup C(2) \cup \dots \cup C(k))^c$
- $P(D(k)) = q^k$
- $D(1) \cup D(2) \cup \dots \cup D(k) \cup \dots$
- Isn't $C(1) \cup C(2) \cup C(3) \cup \dots \cup C(k)$ the event that A occurs *exactly once* on the first k trials?

- $C(2)$ means that A occurred for the first time on the 2nd trial. There is no information as to what happened on the remaining trials; A could have occurred again on those trials
- $C(1), C(2), \dots$ etc form a countable sequence of mutually exclusive events
- Axiom III gives $P(C(1) \cup C(2) \cup \dots \cup C(k) \cup \dots) = P(C(1)) + P(C(2)) + \dots$
 $= p + qp + q^2p + \dots + q^{k-1}p + \dots = p(1 + q + q^2 + \dots + q^{k-1} + \dots) = p \times 1/(1-q) = 1 \quad (q < 1)$
- Provided that $q < 1$ (i.e. $p > 0$), the probability that A occurs at least once (i.e. *some* time) on an infinite sequence of trials is 1
- The complementary event that A *never occurs* has probability 0
- Probability measures are set-continuous functions
- For a *telescoping* sequence of sets such as $D(1) \supset D(2) \supset \dots \supset D(k) \supset \dots$, let $D(\infty)$ denote $\lim_k D(k)$



- The set-continuity property says that $P(D(\infty)) = P(\lim_k D(k)) = \lim_k P(D(k))$
 - In other words, the \lim can be interchanged with the $P(\bullet)$
 $P(\lim_k D(k)) = \lim_k P(D(k)) = \lim_k q^k = 0$ provided $q < 1$
 - $P(D(\infty)) = 0$
 - However, $D(\infty)$ is *not* the empty set, and we are *not asserting* that $D(\infty)$ will never occur
 - Remember: events with probability 0 are not impossible events (in the sense of \emptyset)
 - **Example:** Fred and Wilma take turns tossing a biased coin with $P(\text{Head}) = p$. The first to toss Head wins. Fred starts the game (gets to toss first)
 - What is $P(F) = P(\text{Fred wins})$? $P(W) = P(\text{Wilma wins})$?
 $P(\text{game goes on forever with neither one winning})$?
 - Fred wins if Head occurs for the first time on the 1st, or the 3rd, or the 5th, or ... toss, that is, if $C(1) \cup C(3) \cup C(5) \cup \dots$ occurs
 - $P(F) = P(C(1)) + P(C(3)) + \dots = p + q^2p + q^4p + \dots = p/(1-q^2) = p/(1-q)(1+q)$
 $= 1/(1+q)$ since $1-q = p$
 - Wilma wins if Head occurs for the first time on the 2nd, or the 4th, or the 6th ... toss, that is, if $C(2) \cup C(4) \cup C(6) \cup \dots$ occurs
 - $P(W) = P(C(2)) + P(C(4)) + \dots = qp + q^3p + q^5p + \dots = qp/(1-q^2) = qp/(1-q)(1+q)$
 $= q/(1+q)$ since $1-q = p$
 - $P(F) = \frac{1}{1+q}$ • $P(W) = \frac{q}{1+q}$
- Hence, $P(\text{game goes on forever with neither winning}) = P(F^c \cap W^c) = P((F \cup W)^c)$
 $= 1 - P(F \cup W) = 1 - (P(F) + P(W)) = 0$
- The odds of Fred winning are $1/(1+q):q/(1+q) = 1:q$ which are always better than 1:1

- More generally, if there are m players taking turns, the win probabilities are in the ratio $1:q:q^2:q^3: \dots :q^{m-1}$ and the probability of the i -th player winning is thus

$$\frac{q^{i-1}}{1+q+q^2+q^3+ \dots +q^{m-1}}$$
- Let $m > n$. Given that $D(n)$ occurred, what is the conditional probability of $C(m)$?
- Given that the first n trials resulted in A^c , what is the probability that A occurred for the first time on the m -th trial?
- $P(C(m)|D(n)) = \frac{P(C(m) \cap D(n))}{P(D(n))} = \frac{P(C(m))}{P(D(n))}$ because $C(m) \cap D(n) = \frac{q^{m-1}p}{q^n} = q^{m-n-1}p$
 $= P(C(m-n))$
- Since $m > n$, let $m = n + r$
- $P(C(n+r)|D(n)) = P(C(r))$
- Given that the first n trials resulted in A^c , what is the probability that a *further* r trials are needed to observe A for the first time?
 $q^{r-1}p = P(C(r)) = P(A \text{ occurs for the first time on the } r\text{-th trial})$
- In other words, if the desired event did not occur on the first n trials, *forget* that the first n trials ever occurred. Just count the $(n+1)$ -th as the first trial, $(n+2)$ -th as second, etc and compute probability that A occurs for the first time on the r -th trial
- This is called the memoryless property of independent trials
- **Example:** $P(F)$ and $P(W)$ via the memoryless property and theorem of total probability
- We condition on $D(1) = \text{Tail on first toss}$ and $D(1)^c = \text{Head on first toss}$
- $D(1)^c = \text{Fred wins! } P(F|D(1)^c) = 1$
- $P(F|D(1)) = P(W)$. Having tossed Tail on first try, Fred plays second fiddle to Wilma
- $P(F|D(1)^c) = 1$
- $P(F) = P(F|D(1)^c)P(D(1)^c) + P(F|D(1))P(D(1)) = p + P(W) \times q$
- $P(W) = P(W|D(1)^c)P(D(1)^c) + P(W|D(1))P(D(1)) = 0 + q \times P(F) = q \times P(F)$ since Wilma is in Fred's shoes if Tail occurs on Fred's first toss
- $P(F) = p + q^2P(F)$ giving $P(F) = p/(1-q^2) = 1/(1+q)$
- Also, $P(W) = qP(F) = \frac{q}{1+q}$
- Note that $P(F) + P(W) = 1$ as was found earlier also
- **Example:** Let A and B denote mutually exclusive events. In a series of repeated independent trials, what is the probability that A occurs before B does?
- Let $C = (A \cap B)^c$, i.e. neither A nor B occurs
- $P(C) = 1 - P(A) - P(B)$
- If A occurs (for the first time) on the k -th trial and B has not occurred as yet, then C must have occurred on the 1st, 2nd, ... $(k-1)$ -th trial
- Probability of this event is $(P(C))^{k-1}P(A)$
- $P(A \text{ occurs before } B) = \text{sum of such probabilities over } k$
- $P(A \text{ occurs before } B) = \sum_{k=1}^{\infty} (P(C))^{k-1}P(A) = \frac{P(A)}{1 - P(C)} = \frac{P(A)}{P(A) + P(B)}$
- $P(B \text{ occurs before } A) = \frac{P(B)}{P(A) + P(B)} = 1 - P(A \text{ occurs before } B)$

- Probability is $\sum_{k=r}^{r+m-1} P(C(k;r))$
- $P(\text{no more than } N \text{ trials needed to observe } A \text{ } r \text{ times}) = \sum_{k=r}^N P(C(k;r)) = 1 - \sum_{k=N+1}^{\infty} P(C(k;r))$
- $P(\text{more than } N \text{ trials needed to see } A \text{ occur } r \text{ times}) = \sum_{k=N+1}^{\infty} P(C(k;r))$
- $= P(A \text{ occurred fewer than } r \text{ times on } N \text{ trials}) = \sum_{k=0}^{r-1} P(B(k;N))$
- Negative binomial probabilities arise in studies of queues and waiting times
- $C(k;r)$ is the event that you have to wait for k trials to observe A r times
- Generalizations to waiting for multiple events can be done in straightforward manner
- **Example:** Boxes of Soggies cereal contain either a picture of Luke or a picture of Darth. $P(L) = 2/3$ and $P(D) = 1/3$. Over in Cedar Rapids IA, Mrs Kirk's little Jimmy demands that she buy cereal boxes until he has acquired one picture of each. What is $P(\text{Mrs Kirk buys } N \text{ boxes})$?
- Obviously $N \geq 2$
- $P(\text{Mrs Kirk buys } N \text{ boxes}) = P(L, L, L, \dots, L, D) + P(D, D, D, \dots, D, L)$
 $= \left(\frac{2}{3}\right)^{N-1} \left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^{N-1} \left(\frac{2}{3}\right)$
- Exercise: what is the probability that the last box she buys has a picture of Darth? has a picture of Luke?
- What is the probability that she has to buy no more than N boxes?
- $P(2 \text{ boxes}) + P(3 \text{ boxes}) + \dots + P(N \text{ boxes}) = \sum_{k=2}^N \left(\frac{2}{3}\right)^{k-1} \left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^{k-1} \left(\frac{2}{3}\right)$
- These geometric series can be summed
- Alternatively, more than N boxes are needed if all N boxes have picture of Luke only (or Darth only)
- $P(\text{more than } N \text{ boxes}) = \left(\frac{2}{3}\right)^N + \left(\frac{1}{3}\right)^N$ • $P(\text{no more than } N \text{ boxes}) = 1 - \left(\frac{2}{3}\right)^N - \left(\frac{1}{3}\right)^N$
- Further generalizations are possible
- Suppose that boxes of Soggies contain a picture of Han or Luke or Darth. $P(H) = P(L) = 0.4$, $P(D) = 0.2$
- Little Jimmy wants a picture of each. $P(\text{Mrs Kirk buys } N \text{ boxes})$?
- Suppose that Han's picture is in the N -th box
- The first $N-1$ boxes can contain any arbitrary sequence of D and L *except* DDD...D and LLL...L
- $P(N \text{ boxes and last is Han}) = 0.4 \times [(0.6)^{N-1} - (0.4)^{N-1} - (0.2)^{N-1}]$
- Similar argument holds if last picture is Luke (same prob. as Han) or Darth (different probability)
- $P(N \text{ boxes and last is Han}) = 0.4 \times [(0.6)^{N-1} - (0.4)^{N-1} - (0.2)^{N-1}]$
- $P(N \text{ boxes and last is Luke}) = 0.4 \times [(0.6)^{N-1} - (0.4)^{N-1} - (0.2)^{N-1}]$
- $P(N \text{ boxes and last is Darth}) = 0.2 \times [(0.8)^{N-1} - 2 \times (0.4)^{N-1}]$
- $P(N \text{ boxes}) = \text{sum of these three probabilities}$

Decision Making under Uncertainty – Revisited

- Consider an event A that has probability either p_0 or p_1 depending on the state of nature. We previously considered the case when the experiment is performed once, and we decide what the state of nature is based on the outcome of the trial. The *decision rule* specifies our decision when we observe that A has occurred as well when A^c has occurred. In Chapter 2, we studied the following example.
- Example (re-visited):** We have a coin that is either a fair coin (the null hypothesis H_0) or has probability $2/3$ of turning up Heads (the alternative hypothesis H_1). We toss the coin once and observe the outcome. The likelihood matrix is

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

The maximum-likelihood decision rule is indicated by the shading. If the coin toss results in Heads, we choose hypothesis H_1 while if the result is Tails, we choose hypothesis H_0 . Suppose that H_0 happens to be the true hypothesis. The maximum-likelihood decision rule says that we declare H_1 to be true whenever we observe Heads. The probability of Heads when H_0 happens to be the true hypothesis is $1/2$. Thus, the Type I error probability or false alarm probability is $1/2$. Similarly, suppose instead that H_1 happens to be the true hypothesis. Our decision rule says that we declare H_0 to be true whenever we observe Tails. The probability of Tails when H_1 happens to be the true hypothesis is $1/3$. Thus, the Type II error probability or missed detection probability is $1/3$.

- Likelihood ratio:** The likelihood ratio is defined as

$$(\text{observation}) = \frac{P_1(\text{observation})}{P_0(\text{observation})}$$

where $P_i(\text{observation})$ denotes the likelihood (that is, the probability) of the observation under hypothesis H_i . Note that $(\text{Heads}) = (2/3)/(1/2) = 4/3$ while $(\text{Tails}) = (1/3)/(1/2) = 2/3$ in our example.

- Maximum-likelihood (ML) decision rule:** This rule can be stated in the following notation:

$$(\text{observation}) = \frac{P_1(\text{observation})}{P_0(\text{observation})} \begin{matrix} > \\ < \end{matrix} \begin{matrix} H_1 \\ H_0 \end{matrix} \quad 1$$

which means that the likelihood ratio for the observation is compared to the *threshold* of 1, and the decision is H_1 if $(\text{observation}) > 1$ while the decision is H_0 if $(\text{observation}) < 1$. Thus, in our example, when we observe Heads, we compute $(\text{Heads}) = 4/3$ which is greater than 1, so we decide that H_1 is the state of nature (that is, we have a biased coin). On the other hand, if we observe Tails, we compute $(\text{Tails}) = 2/3$ which is smaller than 1, so we decide that H_0 is the state of nature (that is, we have a fair coin).

- Maximum-a-posteriori-probability (MAP) or Bayes' decision rule:** This can also be stated in terms of the likelihood ratio as

$$(\text{observation}) = \frac{P_1(\text{observation})}{P_0(\text{observation})} \begin{matrix} > \\ < \end{matrix} \begin{matrix} H_1 \\ H_0 \end{matrix} = \frac{P(H_0)}{P(H_1)} = \frac{p_0}{p_1}$$

where $P(H_0) = p_0$ and $P(H_1) = p_1$ are the probabilities of the hypotheses. Notice that we are comparing the likelihood ratio to a different threshold than the one used by the ML decision rule. Also, if $p_0 = p_1$, the MAP rule reduces to the ML rule.

- **Minimum cost Bayes' decision rule:** Here, the likelihood ratio is compared to yet another threshold:

$$(\text{observation}) = \frac{P_1(\text{observation})}{P_0(\text{observation})} > \frac{H_1}{H_0} \frac{(C_{10} - C_{00})P(H_0)}{(C_{01} - C_{11})P(H_1)}$$

where C_{ij} is the cost incurred for deciding that hypothesis H_i is true when in fact hypothesis H_j happens to be the true state of nature.

- In our example, the false alarm probability was 1/2 and the missed detection probability was 1/3. These are quite large because of our rush to judgement — we only tossed the coin once before making the decision. What if we tossed the coin many times before deciding?
- **Example:** We have a coin that is either a fair coin (the null hypothesis H_0) or has probability 2/3 of turning up Heads (the alternative hypothesis H_1). We toss the coin three times and observe the outcome. The likelihood matrix is

	TTT	TTH	THT	THH	HTT	HTH	HHT	HHH
H_0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
H_1	1/27	2/27	2/27	4/27	2/27	4/27	4/27	8/27

and the maximum-likelihood decision rule is as shown by the shading. Note that H_1 is accepted if and only if two or more heads occur. Now, the missed detection probability is only $7/27 < 1/3$ (though the false alarm probability remains at 1/2.)

- Note that the *sequence* in which heads and tails occur does not make any difference as far as the decision rule is concerned. We could have equally easily just noted the total number of heads that occurred and based our decision on that. Thus, a simplified version of the likelihood matrix would look as follows:

	0 Heads	1 Head	2 Heads	3 Heads
H_0	1/8	3/8	3/8	1/8
H_1	1/27	6/27	12/27	4/27

In terms of likelihood ratios, we see that it does not matter very much whether we take the observation to be the *sequence of Heads and Tails*, or just the number of Heads and Tails.

Thus, $(\text{HTT}) = \frac{P_1(\text{HTT})}{P_0(\text{HTT})} = \frac{\frac{2}{3} \times \frac{1}{3} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}} = \frac{16}{27}$ while $(1 \text{ head}) = \frac{P_1(1 \text{ head})}{P_0(1 \text{ head})} = \frac{\binom{3}{1} \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3}}{\binom{3}{1} \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}} = \frac{16}{27}$ also

with the $\binom{3}{1}$ cancelling out from numerator and denominator.

- **Example:** We have a coin that either has $P(\text{Heads}) = p_0$ (the null hypothesis H_0) or has $P(\text{Heads}) = p_1$ (the alternative hypothesis H_1). We toss the coin N times and observe that k Heads have occurred. The likelihood ratio is given by

$$(k \text{ Heads}) = \frac{P_1(k \text{ Heads})}{P_0(k \text{ Heads})} = \frac{p_1^k (1 - p_1)^{N-k}}{p_0^k (1 - p_0)^{N-k}}$$

We could have included a $\binom{N}{k}$ in the numerator and denominator but the terms would have cancelled out anyway. The three decision rules that we described above are all *threshold tests* on the likelihood ratio: the likelihood ratio is compared to a threshold and the decision rule is to choose H_1 if and only if the likelihood ratio exceeds . What happens to be depends on whether we are using the maximum-likelihood rule or the MAP rule or the minimum cost rule. We now show that all these decision rules can be expressed in terms of threshold tests on k ,

the number of Heads that occurred in which we compare k to a (different) threshold.

If $p_1 > p_0$, the decision rule is to choose H_1 if and only if k exceeds this threshold, while if $p_1 < p_0$, the decision rule is to choose H_1 if and only if k is smaller than this threshold. This is eminently reasonable. If Heads are more likely when H_1 is the true state of nature, then we should accept H_1 upon seeing lots of Heads and reject H_1 if we see only a few Heads. On the other hand, if Heads are more likely when H_0 is the true state of nature, then we should choose H_1 upon seeing only few Heads and reject H_1 if we see lots of Heads.

- The three different decision rules discussed above are all of the form

$$(k \text{ Heads}) = \frac{P_1(k \text{ Heads})}{P_0(k \text{ Heads})} = \frac{p_1^k(1 - p_1)^{N-k}}{p_0^k(1 - p_0)^{N-k}} \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} .$$

Taking logarithms* on both sides, we can write the decision rule as

$$\ln (k \text{ Heads}) = k \ln \frac{p_1}{p_0} + (N-k) \ln \frac{1 - p_1}{1 - p_0} \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} \ln ,$$

that is,
$$k \ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1} \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} \ln + N \ln \frac{1 - p_0}{1 - p_1} ,$$

- If $p_1 > p_0$, then $1 - p_1 < 1 - p_0$. Consequently, both $\ln \frac{p_1}{p_0}$ and $\ln \frac{1 - p_0}{1 - p_1}$ are positive, and hence the quantities multiplying k and N are positive. Hence we can write the decision rule as

$$k \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} \frac{\ln + N \ln \frac{1 - p_0}{1 - p_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1}}$$

- On the other hand, if $p_1 < p_0$, then $1 - p_1 > 1 - p_0$. Consequently, both $\ln \frac{p_1}{p_0}$ and $\ln \frac{1 - p_0}{1 - p_1}$ are less than 0, and the quantities multiplying k and N are negative. Hence we can write the decision rule as

$$k \begin{matrix} H_0 \\ > \\ < \\ H_1 \end{matrix} \frac{\ln + N \ln \frac{1 - p_0}{1 - p_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1}}$$

- As noted above, if $p_1 < p_0$, the rule chooses H_0 for large values of k because Heads are more likely under H_0 than under H_1 . Note that at least three of the four logarithms on the right side of the above inequality are negative.
- In what follows, we shall assume for simplicity that $p_1 > p_0$. Analogous results hold for the case when $p_1 < p_0$, and the diligent reader can work these out as an exercise.

* Here, we are using the fact that the logarithmic function is a monotone increasing function, that is, if x and y are positive numbers, then $x > y$ if and only if $\log x > \log y$. Thus, to determine which of $(k \text{ Heads})$ and \dots is the larger number, we can compare their logarithms. The number which has the larger logarithm is the larger number. Note that the base of the logarithms is immaterial (as long as you use the same one on both sides!)

- In what follows, we shall assume for simplicity that $p_1 > p_0$ so that the decision rule is

$$k \begin{array}{l} H_1 \\ > \\ < \\ H_0 \end{array} \frac{\ln \quad + N \ln \frac{1 - p_0}{1 - p_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1}} =$$

where all the terms on the right (with the possible exception of $\ln \quad$) are positive. Of course, we compute the numerical value of the threshold ahead of time, and just compare k , the observed number of heads to \quad .

- As a special case, we get the ML decision rule if we set $\quad = 1$ and $\ln \quad = 0$. The ML decision rule is a threshold test on k , the number of Heads, which we can write as

$$k \begin{array}{l} H_1 \\ > \\ < \\ H_0 \end{array} \frac{N \ln \frac{1 - p_0}{1 - p_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1}} = \text{ML}$$

Note that all three terms on the right side are positive, and hence ML , the maximum-likelihood threshold, is always a positive quantity.

- Example:** If $N = 100$, $p_1 = 2/3$ and $p_0 = 1/2$, the ML decision rule is to compare the number of heads to $\text{ML} = 58.496\dots$. The ML decision rule thus chooses H_1 if 59 or more heads are observed on 100 tosses, and it chooses H_0 if 58 or fewer heads are observed on 100 tosses. This is quite reasonable — on 100 tosses of a fair coin, we should expect to see roughly 50 heads, whereas a coin with $P(\text{Heads}) = 2/3$ should come up Heads roughly 67 times. Our threshold is set between these two levels which is intuitively satisfying.
- The MAP decision rule uses a threshold $\quad = \theta_0 / \theta_1$. The corresponding threshold test on k can be expressed as

$$k \begin{array}{l} H_1 \\ > \\ < \\ H_0 \end{array} \text{ML} + \frac{\ln \frac{\theta_0}{\theta_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1 - p_0}{1 - p_1}} = \text{MAP}$$

- Suppose that $\theta_0 > \theta_1$. Then the term in the numerator is positive. Since the terms in the denominator are also positive, we get that $\text{MAP} > \text{ML}$. In the example above, if $\theta_0 = 0.8$ and $\theta_1 = 0.2$, then $\text{MAP} = \text{ML} + 2 = 60.496\dots$. Thus, the MAP decision rule chooses H_0 if 59 or 60 heads are observed on 100 tosses, whereas the ML decision rule would choose H_1 in such cases. (In all other cases, the MAP decision rule and ML decision rule would agree on the choice of hypothesis) This is in accordance with our intuition. Since H_0 is much more likely to be the true hypothesis than H_1 , we should accept H_1 only if the evidence in its favor is quite strong. MAP increases with θ_0 and exceeds 100 as θ_0 approaches 1, that is, if the null hypothesis H_0 is very much more likely than the alternative hypothesis H_1 , then the minimum average error probability is incurred by the stick-your-head-in-the-sand rule that ignores all the experimental evidence and always chooses H_0 .
- If $\theta_0 < \theta_1$, the whole argument above holds in reverse. Now, $\text{MAP} < \text{ML}$, so that we are more prepared to accept H_1 at the drop of a hat.

- All of the above also applies to the more general form of Bayes' decision rules that minimize the average cost (or risk) in making a decision.
- **Exercise:** For what value of θ_0 does λ_{MAP} equal 100 (so that H_1 is never accepted?) For what value of θ_0 does λ_{MAP} become negative (so that H_1 is always accepted?)
- **Error probabilities**

Consider a general decision rule of the form

$$(k \text{ Heads}) = \frac{P_1(k \text{ Heads})}{P_0(k \text{ Heads})} = \frac{p_1^k (1-p_1)^{N-k}}{p_0^k (1-p_0)^{N-k}} \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} .$$

which can be expressed as

$$k \begin{matrix} H_1 \\ > \\ < \\ H_0 \end{matrix} \frac{\ln \frac{1-p_0}{1-p_1} + N \ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1}{p_0} + \ln \frac{1-p_0}{1-p_1}} =$$

when $p_1 > p_0$. Now, suppose that H_0 happens to be the true hypothesis. The decision rule correctly decides in favor of H_0 if the number of heads that occur on N tosses is smaller than k . The decision rule makes an error if the number of heads is greater than k . Hence, the false alarm probability is

$$P_{FA} = P(\text{false alarm}) = P(\text{more than } k \text{ heads when } H_0 \text{ is true}) = \sum_{k > k} \binom{N}{k} p_0^k (1-p_0)^{N-k}.$$

Similarly, if H_1 happens to be the true hypothesis, the decision rule makes an error if the number of heads is smaller than k . Hence, the missed detection probability is

$$P_{MD} = P(\text{missed detection}) = P(\text{fewer than } k \text{ heads when } H_1 \text{ is true}) = \sum_{k < k} \binom{N}{k} p_1^k (1-p_1)^{N-k}.$$

- The false alarm probability P_{FA} is sometimes called the *significance level* of the decision rule. A decision to accept H_1 is said to be "significant at the 5% level" if $P_{FA} = 0.05$ for the decision rule being used, that is, the probability that we have accepted H_1 by mistake is 0.05.
- $P(\text{detection}) = 1 - P(\text{missed detection}) = 1 - P_{MD}$ is called the *power* of the decision rule.
- If the hypotheses have probabilities $P(H_0) = \theta_0$ and $P(H_1) = \theta_1$, then P_{FA} and P_{MD} can be thought of as the conditional probabilities of error given H_0 and given H_1 respectively. Then, the theorem of total probability gives the average error probability as $\theta_0 P_{FA} + \theta_1 P_{MD}$. The MAP decision rule attempts to minimize the average error probability.
- The more general Bayesian decision rule that minimizes the average cost uses a different threshold than the MAP decision rule. If P_{FA} and P_{MD} are the error probabilities corresponding to this threshold, then the average cost is $\theta_0 C_{00} + \theta_1 C_{11} + \theta_0 (C_{10} - C_{00}) P_{FA} + \theta_1 (C_{01} - C_{11}) P_{MD}$ which reduces to $\theta_0 P_{FA} + \theta_1 P_{MD}$ if $C_{00} = C_{11} = 0$ (no cost for making a correct decision) and $C_{10} = C_{01}$ (unit cost for making a wrong decision either way).
- The threshold k can be written as $\ln \frac{1-p_0}{1-p_1} + N \ln \frac{1-p_0}{1-p_1}$ and is an increasing function of θ_0 . Notice that as θ_0 (and hence θ_1) increases, the false alarm probability P_{FA} decreases because fewer and fewer

terms are included in the first sum displayed above. Thus, by suitable choice of α or β , we can adjust P_{FA} to satisfy any design constraint (e.g., the decision rule must have $P_{FA} = 0.001$, say). Unfortunately, as α and β increase, so does P_{MD} because the second sum displayed above includes more and more terms. Thus, any decision rule can be viewed as a compromise between the desires to make each of these error probabilities as small as possible.

- The significance level P_{FA} can be made as small as we like by increasing α (equivalently β). Unfortunately, this has the effect of increasing P_{MD} and hence decreasing the power of the test. Thus, suppose that we compromise and set a constraint α on P_{FA} . We want our decision rule to be such that $P_{FA} = \alpha$. Of course, we could even make $P_{FA} = 0$ by choosing $\alpha = 100$ in our example (more generally, we can set $\alpha = M$) so that we never choose H_1 and hence never have a false alarm, but this would make the power of the test 0 since we would never detect H_1 either. Thus, of all possible tests that satisfy the constraints $P_{FA} = \alpha$, we would like to use the one which has the maximum power (equivalently, the smallest missed detection probability). Such a decision rule is called a *Neyman-Pearson decision rule*. It is the decision rule with the smallest missed-detection probability (that is, the maximum power) among all decision rules satisfying the constraint $P_{FA} = \alpha$.
- For our example, the construction of the Neyman-Pearson decision rule is intuitively obvious. We know that if $\alpha = 100$, then $P_{FA} = 0$. Now, consider the decision rules corresponding to $\alpha = 99, \alpha = 98, \dots$ and calculate P_{FA} for each. Since P_{FA} will increase as we decrease the threshold (as will the power), we will reach a point where the decision rule with $\alpha = M$ satisfies the design constraint $P_{FA} = \alpha$ but the decision rule with $\alpha = M-1$ does not. The Neyman-Pearson decision rule thus is a threshold test on k , the number of heads observed on N tosses, with threshold $k_{NP} = M$.
- The skeptical reader will have noticed that while the design procedure clearly produces a decision rule satisfying $P_{FA} = \alpha$, and the construction procedure is producing a series of decision rules with increasing power, there is no guarantee that the rule that we have found has the *maximum* power among all possible decision rules satisfying $P_{FA} = \alpha$. For example, a decision rule of the form “choose H_1 if you observe 12, 18, or 21 heads, and choose H_0 otherwise” may well satisfy $P_{FA} = \alpha$, and it may be that it has power greater than the one obtained by our construction. The reader may rest assured that, although the formal proof of the result is not included in these notes, the Neyman-Pearson decision criterion always results in a decision rule that is a threshold test of the likelihood ratio, that is, it is of the form

$$(\text{observation}) = \frac{P_1(\text{observation})}{P_0(\text{observation})} \begin{matrix} > \\ < \end{matrix} \begin{matrix} H_1 \\ H_0 \end{matrix}$$

where γ is an appropriately chosen threshold. Of course, the way to choose the threshold is to choose it such that the false alarm probability satisfies $P_{FA} = \alpha$.

- **Summary**

- Various design criteria and methodologies for decision rules result in threshold tests on the likelihood ratio in which the alternative hypothesis H_1 is accepted if and only if the likelihood ratio exceeds a threshold γ . The decision rules can also be expressed in terms of threshold tests on the observations themselves.

Automatic Repeat Request (ARQ) Communication Systems

- Data communication over a channel that changes 0's to 1's and 1's to 0's occasionally
- $P(\text{bit error}) = p \ll 1$
- The occurrences of errors in the bits are *independent* events
- $\underline{C} = (C_1, C_2, C_3, \dots, C_N)$ = transmitted bit sequence
 $\underline{R} = (R_1, R_2, R_3, \dots, R_N)$ = received bit sequence
- $\underline{R} = \underline{C} \oplus \underline{E}$ where \oplus denotes bit-by-bit Exclusive OR (XOR) and $\underline{E} = (E_1, E_2, E_3, \dots, E_N)$ is the *bit error* sequence
- $R_i = C_i \oplus E_i$ for $1 \leq i \leq N$
- If $E_i = 0$, $R_i = C_i$
- If $E_i = 1$, $R_i = \bar{C}_i = \begin{cases} 0 & \text{if } C_i = 1 \\ 1 & \text{if } C_i = 0 \end{cases}$
- Transmission errors exactly in those positions where $E_i = 1$
- \underline{E} is also known as the channel error pattern
- Suppose that \underline{C} is transmitted, and $\underline{R} = \underline{C} \oplus \underline{E}$ is received
- e = number of channel errors = number of 1's in \underline{E} = Hamming weight of \underline{E}
- $P(\underline{E}) = p^e(1-p)^{N-e} = p^e q^{N-e}$ where $q = 1-p$ as usual
- $E_i = 0 \implies R_i = C_i$
- If $E_i = 0$ for all i , $1 \leq i \leq N$, that is, if $e = 0$, then $\underline{R} = \underline{C}$ and the transmitted bit sequence is received correctly
- $P(\underline{R} = \underline{C}) = P(e = 0) = q^N$
- If $p = 10^{-4}$, $N = 500$, then $P(\underline{R} = \underline{C}) = 0.951\dots$
- N transmitted bits = k data bits + $(N-k)$ overhead bits
- Some of the overhead bits are for header, address, timestamp, etc.
- Some of the overhead bits are *parity check* bits for error protection
- A parity check bit tells whether the total number of 1's in some pre-specified set of bit positions is odd or even
- Multiple parity check bits are used to check different subsets
- VRC, LRC, CRC
- Example: The data bits are 01011. The 6th (parity check) bit is a 1 to indicate that there are an odd number of 1's in positions $\{1,2,3\}$. The 7th (parity check) bit is a 0 to indicate that there are an even number of 1's in positions $\{3,4,5\}$.
- $\underline{C} = 0101110$
- The transmitter computes the parity check bits and includes them as part of the $N-k$ overhead bits in \underline{C}
- The receiver *recomputes* the parity check bits from the *received* data bits in \underline{R} and compares the *computed* parity bits with the *received* parity check bits in \underline{R}
- If $e = 0$, then $\underline{R} = \underline{C}$ and the receiver's recomputed parity check bits match the received parity check bits exactly
- In such cases, the receiver accepts the received bits as correct in all respects
- Usually (but *not* always), the decision to accept is correct
- There is a *small* chance that the received bits are in error even in case of a match
- $\underline{C} = 0101110$ $6 = \{1,2,3\}, 7 = \{3,4,5\}$
- $\underline{R} = 1001110$ $e = 2$

- The receiver recomputes parity check bits and finds that they match the received parity check bits!
- What if the recomputed parity bits do not exactly match the received parity check bits?
- This is called a parity-check failure
- $\underline{C} = 0101110$ $6 = \{1,2,3\}, 7 = \{3,4,5\}$
- $\underline{R} = 0001110$ $e = 1$
- Recomputed bit 6 is 0 while the received bit 6 is 1
- If parity check failure occurs, the receiver *knows for sure* that transmission errors have occurred and that $\underline{R} \neq \underline{C}$
- In cases of parity check failure, the receiver requests that the message be *repeated* that is, be re-transmitted
- Receiver never requests re-transmission of a perfectly correct transmission
- Re-transmission requests may be unnecessary
- $\underline{C} = 0101110$ $6 = \{1,2,3\}, 7 = \{3,4,5\}$
- $\underline{R} = 0101111$ $e = 1$
- Recomputed bit 7 is 0 while the received bit 7 is 1
- In this example, the *data bits* are correct and it is the parity bit which is in error
- In the previous example, the data bits were in error while the parity bits were correct
- The receiver has no way of knowing which of these possibilities has occurred
- Receiver uses the sensible and conservative strategy of requesting re-transmissions in all cases of parity check failure
- Summary: If parity checks fail, receiver requests that the message be re-transmitted
- Receiver accepts message if parity checks are satisfied
- $P(\text{correct reception}) = q^N$
- $P(\text{acceptance})$ is *very slightly* larger than q^N
- $P(e > 0 | \text{parity checks fail}) = 1$
- $P(\text{parity checks fail} | e > 0) = 1$
- $P(e > 0 | \text{parity checks OK}) = 0$
- ARQ communication systems in common use correctly detect *all* occurrences of 0, 1, 2, or 3 bit errors (and ask for repeats as necessary)
- At least 4 bit errors must occur in order to fool the system into accepting an incorrect transmission
- However, the *overwhelming majority* of the $\binom{N}{4}$ patterns of 4 bit errors are successfully detected, as are most of the $\binom{N}{5}$ patterns of 5 bit errors, ... etc
- If $p = 10^{-4}$, $N = 500$, then $P(e = 0) = q^{500} = 0.951227\dots$
 $P(e = 1) = 500pq^{499} = 0.04757\dots$; $P(e = 2) = \binom{500}{2} p^2 q^{498} = 0.00186893\dots$
 $P(e = 3) = \binom{500}{3} p^3 q^{497} = 1.97 \times 10^{-5}$ Sum $1 - 2 \times 10^{-6}$
- $P(e=0) + P(e=1) + P(e=2) + P(e=3) = 1 - 2 \times 10^{-6}$
- $P(\text{parity checks satisfied}) = P(e = 0) + P(e > 0 \text{ checks satisfied}) < 0.951227\dots + 2 \times 10^{-6}$
- $P(e > 0 | \text{parity checks OK}) < 2 \times 10^{-6} / 0.951229 = 2 \times 10^{-6}$
- ARQ communication systems are used in RF systems as well as in computer networks

- Long messages are divided at the transmitter into numbered *packets*, each with its own header and parity bits
- Receiver re-assembles packets into message
- How many data bits should be in each packet?
- Here, $N - k = r$ is fixed regardless of how many data bits k are in each packet
- If $k \gg r$, we save on the overhead
- If a packet is transmitted M times, data rate = k/NM
- “Obvious answer #1”: Make k as large as possible so as to minimize overhead
- $P(\text{correct}) = q^N \rightarrow 0$ as N
- If k is very large, too many re-transmissions will be requested → data rate is low
- “Obvious answer #2”: Make k as small as possible so as to maximize $P(\text{correct}) = q^N$
- Although most packets get through successfully, only a few data bits are in each packet → data rate is low
- Few data bit/packet means large overhead per data bit
- Consider the transmission of a large number J of packets
- A packet will be transmitted just once (i.e., the receiver accept the packet on the very first transmission) with probability $Q = q^N$
- JQ packets require 1 transmission
- A packet will have to be transmitted twice with probability $(1-Q)Q = (1-q^N)q^N$
- $J(1-Q)Q$ packets will have to be transmitted twice
- A packet will have to be transmitted L times with probability $(1-Q)^{L-1}Q = (1-q^N)^{L-1}q^N$
- $J(1-Q)^{L-1}Q$ packets will have to be transmitted L times, $L = 1, 2, 3, \dots$
- The total number of packet transmissions (including repeats) to send J packets is thus $J(Q + 2(1-Q)Q + \dots + L(1-Q)^{L-1}Q) = JQ[1 + 2(1-Q) + \dots]$
- We need $JQ[1 + 2(1-Q) + \dots] = JQ \times \frac{1}{[1-(1-Q)]^2} = \frac{J}{Q}$ packet transmissions to send J packets
- Average data rate = $\frac{kJ}{NJ/Q} = \frac{kQ}{N} = \frac{kq^N}{N} = \frac{kq^{k+r}}{k+r}$
- What value of k maximizes this?
- Once again, we have a discrete maximization problem
- Look at ratio of two successive values
- Ratio = $\frac{(k+1)q^{k+r+1}}{k+1+r} \times \frac{k+r}{kq^{k+r}} = q \times \left(1 + \frac{r}{k(k+r+1)}\right) = 1$ if $k(k+r+1) = \frac{qr}{1-q}$,
that is, when $k \approx \sqrt{\frac{qr}{1-q}} - \frac{r}{2}$. This is because we can approximate $k(k+r+1) \approx (k + r/2)^2$
- **Example:** $p = 10^{-4}$ and $r = 30$
The approximate solution is 532.69 which can be rounded off to 533
- A simple calculation shows the maximum occurs at 533
- Typical values for k are 1024 bits (128 bytes) for XMODEM or YMODEM, and 1024 bytes for ZMODEM
- Ethernet uses a maximum packet size of 1526 bytes
- Summary: ARQ systems provide highly reliable (error-free) data communication
- Schemes are quite robust to variations in channel model
- Delays, caused by repetition of incorrectly received packets, can cause problems sometimes