Math 55 - Spring 2004 - Lecture notes #18 - March 30 (Tuesday)

Goals for today:
- Continue discrete probability theory
- Conditional probability
- Independence
- Bernoulli trials

Now we start discussing conditional probability.
Here is an example that we would like to understand:

> A pharmaceutical company is marketing a new test for a certain medical condition. According to clinical trials, the test has the following properties:
>
> 1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% ("False negatives")
> 2. When applied to a healthy person, the test comes up negative in 80% of cases and positive in 20% ("False positives")
>
> Suppose that 5% of the US population has the condition. In other words, a random person has a 5% chance of being affected. When a random person is tested and comes up positive, what is the probability that the person actually has the condition?

This is an example of conditional probability: what is the probability of event A (person is affected) given that we know event B occurs (the person tests positive). We write this P(A|B), the probability of A given B.

Def: P(A|B) = P(A inter B)/P(B)

Justification: Let S be the original sample space, and P() the original probabilty function on S. Since we know B occurs, we have a new sample space, namely B subset S. What is the new probability function? If x in B, then P(x|B) must satisfy

```
  1 = sum_{x in B} P(x|B), so
```

the obvious choice is P(x|B) = P(x)/P(B).
So if A subset B is any event in the new sample space B,

```
then P(A|B) = sum_{x in A} P(x|B) = sum_{x in A} P(x)/P(B)
            = P(A)/P(B)
```

What if A is not a subset of B? If x in A but x not in B, then clearly P(x|B) = 0; if B occurs then x cannot occur.
Thus we finally get P(A|B) = P(A inter B)/P(B).

```
Let N = US population.
Returning to medical testing, the population consists of 4 groups:
  1) TP (true positives)  |TP|=90% of  5% of N = ( 9/200)*N, P(TP)=9/200
  2) FP (false positives) |FP|=20% of 95% of N = (19/100)*N, P(FP)=19/100
  3) TN (true negatives)  |TN|=80% of 95% of N = (76/100)*N, P(TN)=76/100
  4) FN (false negatives) |FN|=10% of  5% of N = ( 1/200)*N, P(FN)=1/200
Now let A = {person is affected} = TP U FN
        B = {person tests positive} = TP U FP
        A inter B = TP
  and finally P(A|B) = P(TP)/P(TP U FP)
                     = (9/200)/(9/200 + 19/100) = 9/47 ~ .19
So if a random person tests positive, there is only a 19% chance
that they really have it.

ASK&WAIT: What is P(B|A) = P(person tests positive | person is affected)?

ASK&WAIT: What is P(test correct when given to random person)?

ASK&WAIT: Let a "phony test" simply declare everyone healthy
          what is P(phony test correct when given to a random person)?

Ex: Suppose we toss 3 balls into 3 bins
ASK&WAIT: What is P(first bin empty)?
ASK&WAIT: What is P(second bin empty | first bin empty)?

Ex: Roll two fair dice, what is P(rolling a 6 | sum of dice is 10)?

Ex: Roll two fair coins, what is P(second is head | first is head)

Def: Two events A and B are independent if P(A inter B) = P(A)*P(B)

EX:  flip two coins, A = {HH, TH}, B = {HH, HT}, A inter B = {HH}
     P(A) = 1/2 = P(B), P(A inter B) = 1/4

Prop: If A and B are independent, then P(A|B) = P(A) and P(B|A) = P(B)
Proof: P(A|B) = P(A inter B)/P(B) = P(A)*P(B)/P(B) = P(A)
       P(B|A) = P(A inter B)/P(A) = P(A)*P(B)/P(A) = P(B)

ASK&WAIT: Throw 3 balls into 3 bins, are
          A = {first bin empty} and B = {second bin empty} independent?
ASK&WAIT: Throw 2 dice, are
          A = {rolling a 6} and B ={sum=10} independent?
```

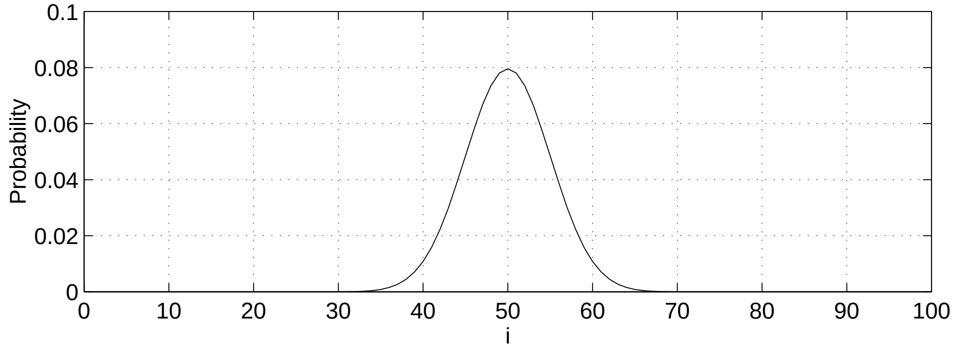
```
ASK&WAIT: Throw 2 dice, are
          A = {sum even}, B = {first die even} independent?

Def: Events A1, A2, ... , An are mutually independent if
     for every i and every subset J of {1,2,...,n} - {i} then
     P(Ai | inter_{j in J} Aj) = Pr(Ai)
     i.e. Ai does not depend on any combination of the other events

Thm: P(B inter A) = P(B)*P(A|B)
Proof: follows from definition of P(A|B)

Thm: P(A1 inter A2 inter ... inter An) =
     P(A1) * P(A2|A1) * P(A3|A1 inter A2) * P(A4| A1 inter A2 inter A3)
           * ... * P(An | A1 inter A2 inter ... inter An-1 )
Proof: induction on n:
       Base case: n=1: P(A1)=P(A1)
       Induction step: Assume
          P(A1 inter ... inter An-1)
                = P(A1) * ... * P(An-1 | A1 inter ... inter An-2)
       Then P(A1 inter ... inter An)
          = P(A1 inter ... inter An-1) * P(An | A1 inter ... inter An-1)
          = P(A1) * ... * P(An-1 | A1 inter ... inter An-2) *
            P(An | A1 inter ... inter An-1)     (by induction, as desired)

Corollary: Suppose A1, A2, ... , An are mutually independent. Then
           P(A1 inter A2 inter ... inter An) = P(A1)*P(A2)*...*P(An)
  Proof: in above proof, each
           P(Ai | A1 inter ... inter Ai-1) = P(Ai) by mutual independence

EX: Toss a fair coin 3 times. Let A={HHH}, A1={Hxx}, A2={xHx}, A3={xxH}
    A = A1 inter A2 inter A3
    P(A) = P(A1) * P(A2|A1) * P(A3|A1 inter A2)
         = P(A1) * P(A2)    * P(A3)
         = 1/2   *  1/2     * 1/2
         = 1/8 as expected
EX: Toss a biased coin 3 times, with P(H) = p
ASK&WAIT: what is P(A)?

Def: a Bernoulli trial is a (sequence) of (independent, identical)
     experiments, each of which has two outcomes

EX: Suppose we flip a fair coin 100 times. What is P(50 Heads)?
```

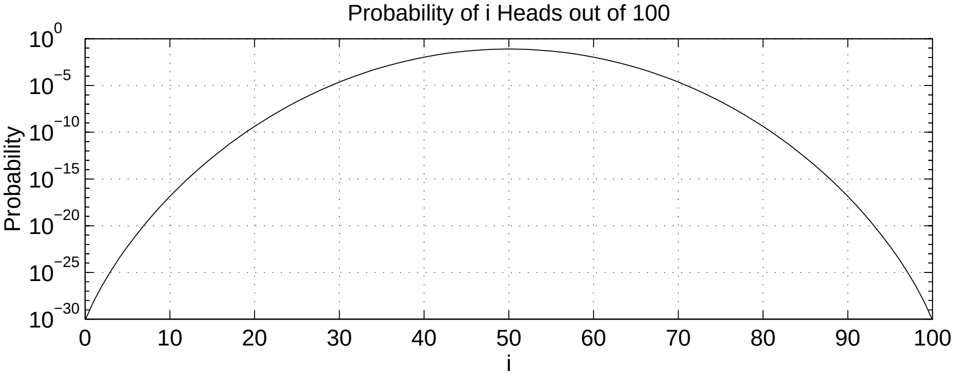
sample space S = {all sequences of 100 H's and T's},
each with P(x)=1/2^100 because
P(HTH...) = P(1st = H)*P(2nd = T)*P(3rd = H)* ... = 1/2^100
(or because it's a uniform distribution over 2^100 possibilities)
E = {all sequences with 50 heads, 50 tails}
ASK&WAIT: What is |E|? P(E)?
ASK&WAIT: Let E(i) = {i Heads out of n flips} What is |E(i)|? P(E(i))?
Note that E(i) and E(j) are disjoint, and
S = E(0) U E(1) U ... E(n), so P(S) = P(E(0)) + ... + P(E(n)) = 1
Check this: sum_{i=0 to n} P(E(i)) = sum_{i=0 to n} C(n,i)/2^n
= 2^(-n) * sum_{i=0 to n} C(n,i)
= 2^(-n) * (1+1)^n ... by the Binomial Theorem
= 1 as desired

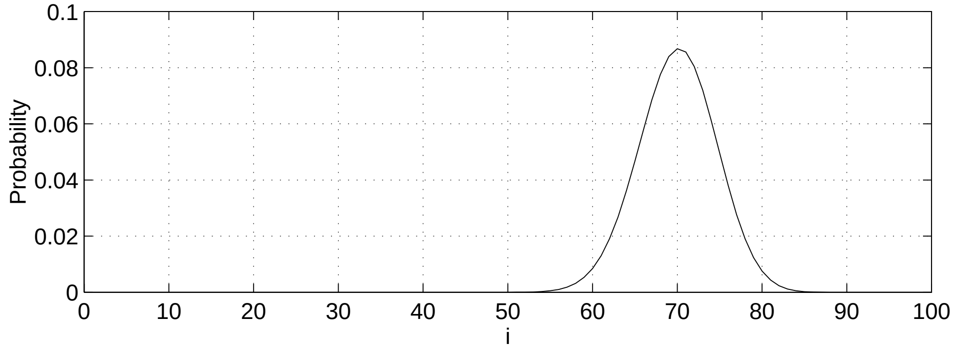
EX; Now flip a biased coin, with P(H) = p and P(T) = 1-p, 100 times
The sample space is the same as above.
But not all P(x) are the same
ASK&WAIT: What is P(50 Hs followed by 50 Ts)?
ASK&WAIT: What is P(50 Hs and 50 Ts, in some fixed order)?
ASK&WAIT: What is P(50 Hs and 50 Ts, in any order)?
Now flip a biased coin n times
ASK&WAIT: What is P(i Hs and n-i Ts, in any order)?
ASK&WAIT: What is sum_{i=0 to n} P(i Hs and n-i Ts, in any order)?

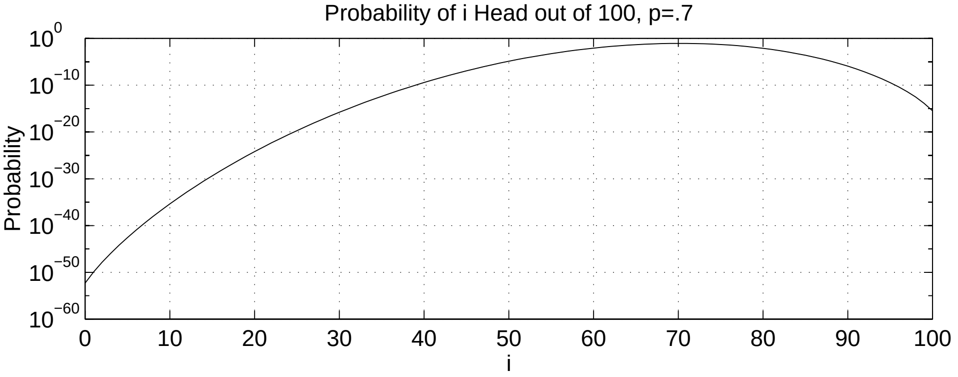
Theorem: If you flip a biased coin n times, with P(H) = p,
the probability of getting i Heads is C(n,i)*p^i*(1-p)^(n-i)

What does P(getting i heads out of n flips) look like as a function of i?
Let's look for n=100, p = .5 , and for n=100, p=.7
Comments on the plots:
when p=.5, the probability is largest at i=50 (equal numbers of heads and tails), and quickly gets smaller for larger or smaller i.
It gets so small that it is easier to look at a logarithmic scale (second plot), where the probability of getting 30 Hs and 70 Ts (or 70 Hs and 30 Ts), is about 10^(-5),
and the probability of getting 10 Hs (or 10Ts) is down to 10^(-17).
when p = .7, then most noticeable feature of the 3rd plot is that it look very much like the first plot, except slid over to have its peak at 70 Hs instead of 50 Hs. This makes sense because with P(H) = .7, one expects close to 70 Hs out of 100. We will return later to explain the remarkable resemblance of these two plots when we discusse the Central Limit Theorem.


Probability of i Heads out of 100


Probability of i Heads out of 100


Probability of i Head out of 100, p=.7


Probability of i Head out of 100, p=.7

```
EX: Probability of {a flush in poker} (5 cards of same suit)
    P(flush) = 4*P(A), A = {flush in hearts}
    A = A1 inter A2 inter ... inter A5
    Ai = {ith card is a heart}
    By Theorem: P(A) = P(A1) * P(A2|A1) * P(A3|A1 inter A2) * ...
ASK&WAIT: What is P(flush)?

EX: You go to a casino, which advertises the following game:
    You pick a number from 1 to 6. Then they role 3 die, and
    you win if your number comes up at least once.
ASK&WAIT: The casino claims that your chance of winning is 50%,
          since it is 1/6 for each die, each die is independent,
          so the probability is 3*(1/6)=1/2. Is this argument reasonable?
    Let's figure out the real probability of winning at this game.
    Let Ai = {your number comes up on die i}, and A = A1 U A2 U A3.
    We want P(A). The casino said P(A) = P(A1) + P(A2) + P(A3) = 3*(1/6)=1/2
    But this is only true if the Ai are disjoint, which they are not
    (your number can come up twice). So we need inclusion/exclusion:

    Recall: P(A1 U A2) = P(A1) U P(A2) - P(A1 inter A2)
ASK&WAIT: what is P(A1 U A2 U A3)?
ASK&WAIT: What is P(Ai inter Aj)?
ASK&WAIT: What is P(A1 inter A2 inter A3)?
ASK&WAIT: What is P(A) = P(winning)? Should you play an even bet?
```