STEVENS
INSTITUTE OF TECHNOLOGY
1870

MA 540: Introduction to Probability Theory

Lecture 1: Intro and Combinatorics

Advice for this class

- In addition to using the slides, you should be taking notes to clarify your understanding
- When I present examples, you should attempt to solve them instead of just waiting for me to show you the answer. You will understand it better even if you try and fail, then understand the right way
- Questions are meant to be interactive. I would rather you answer a question wrong, or talk through the first part of an answer than not answer at all
- I don’t expect you to understand every single detail during the lectures. Sometimes you will be lost or confused about certain concepts.
  - It’s a long lecture at night, and concepts build on one another, both within and between lectures
  - You are expected to review the material later, spending at least as much time as the lecture itself
  - The purpose of HW is to solidify your understanding and focus your review
- We will derive many theoretical results in class, but homework and tests are focused more on applying these concepts

Advice for this class

- The first couple of lectures are relatively easier. The concepts in lectures 3 – 6 are more difficult, but very important
- If you'd like to discuss HW or class problems in greater depth, please come to office hours
- If you're having trouble with certain concepts, please come to office hours
- Videos are available about many of these topics. If you find it easier to gain intuition from these videos rather than the textbook, you can google the general topic (the section headings in the textbook or the slide headlines), but you'll need to know how to use the techniques we teach

Sample Space – Denoted Ω or S

- The sample space consists of all possible outcomes of an experiment

- Examples
  - Flipping one coin – sample space consists of {H, T}
  - Rolling 2 dice – sample space consists of {(1,1), (1,2), ..., (1,6), ..., (6,5), (6,6)}
  - Conditions
    - Outcomes must be mutually exclusive, that is, if the sample space consists of A – Z, if A happens, none of B – Z can happen
      - For 2 dice, we can't have A = at least one die shows 1, B = at most one die shows 6
    - Outcomes must be collectively exhaustive – Again, with sample space A – Z, the outcome must be one of A – Z, not some unenumerated outcome, e.g. Ψ
      - In other words, sample space must include all possible outcomes

Events

- An event is a collection of outcomes, or a subset $A \subseteq \Omega$

- Sometimes we don't care about the exact outcome, per se:
  - Rolling a 7 with 2 dice: We don't necessarily care whether its (1,6) or (4,3) etc.
  - Thresholds: Need above a certain score to get an A, or below a certain temperature to cancel classes

- We assign probabilities to events, not outcomes
  - An event could consist of exactly one outcome, but note that the singleton outcome ω is not the same as the event $\{\omega\}$

Probability

- The proportion of times an event will occur if running an experiment many times
  - May be interpreted as subjective belief, or betting preferences, as often an experiment simply can't be run many times
  - X% chance of developing a disease – Ultimately it will only be true or false

- Conditions:
  - Non-negativity: $P(A) \geq 0$
  - Normalization : $P(\Omega) = 1$
  - Additivity: for disjoint sets $A_1, A_2, \ldots$ $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

- From these axioms, other properties follow

- Why care? We can use probability to make choices, evaluate luck, etc.

Counting

- Intuitive way to explore probability is if the potential outcomes are known in advance, we can count all the potential outcomes, as well as all outcomes in an event
  - Example: Rolling 2 dice and getting at least one 6.
    - Total potential outcomes: 36. Total outcomes in event: 11

- It's tedious to enumerate all possibilities, so we'd like to use formulas to count faster
  - Permutations – number of ways to choose *k* out of *n* items, order matters
  - Combinations – like permutations, but order doesn't matter

Random variable

- Some variable whose value or realization is the outcome of a random experiment, and thus depends on chance

- We try to distinguish between the random variable, which is denoted as a capital (e.g. X) and represents the possible outcomes of the experiment, from the value of particular outcome, which is denoted as a small letter (e.g. x)
  - The term P(X = x) translates to " The probability that random variable X takes on the value x"
  - If X is a dice roll, P(X = 5) = 1/6
  - Can think of X as the daily high temperature in July in Hoboken, P(X > 100) means the probability the temperature exceeds 100

- A random variable can be generalized to a random vector, with multiple random values
  - For example, if I throw a dart and let the bullseye be the origin, the (x,y) location where the dart lands is a random variable

Discrete vs Continuous RVs

- Discrete random variables only take on a finite or countably infinite set of values
  - Examples: dice roll, number of heads from flipping n coins

- Continuous random variables take on an uncountably infinite set of values
  - Examples: Length, temperature, duration

Probability mass function

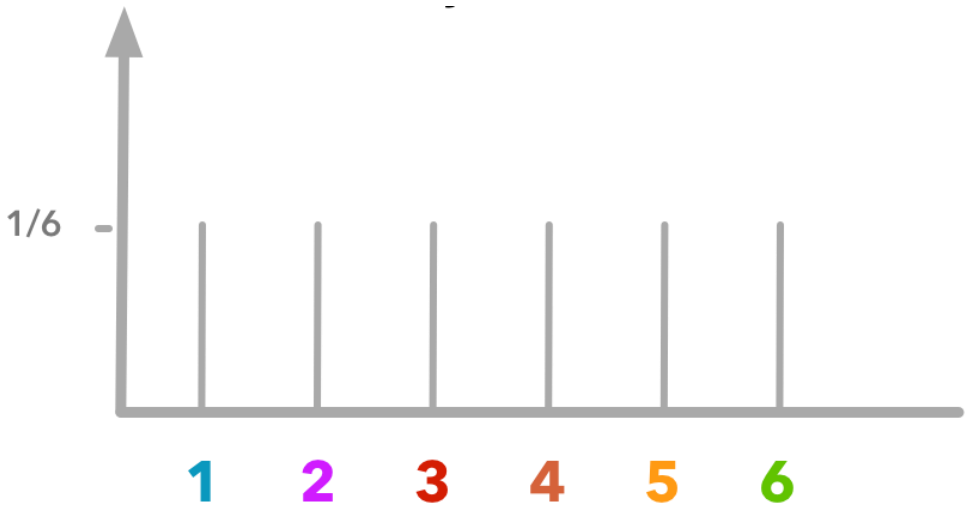
- For a discrete random variable, we can represent the probability any particular value will occur
  - For a die

    $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

    $P(\text{any other value}) = 0$

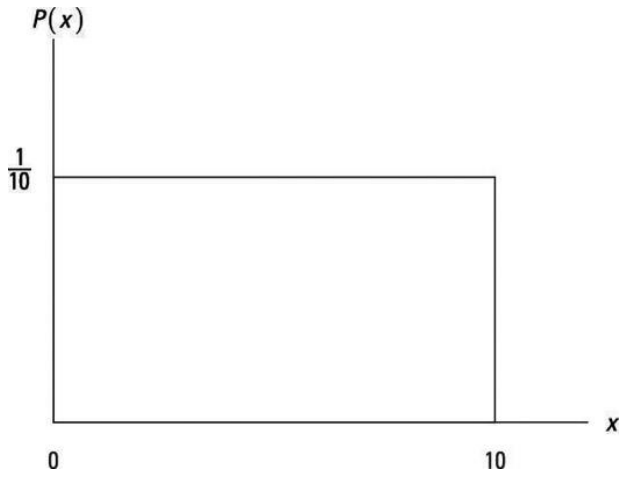
- Must obey probability axioms

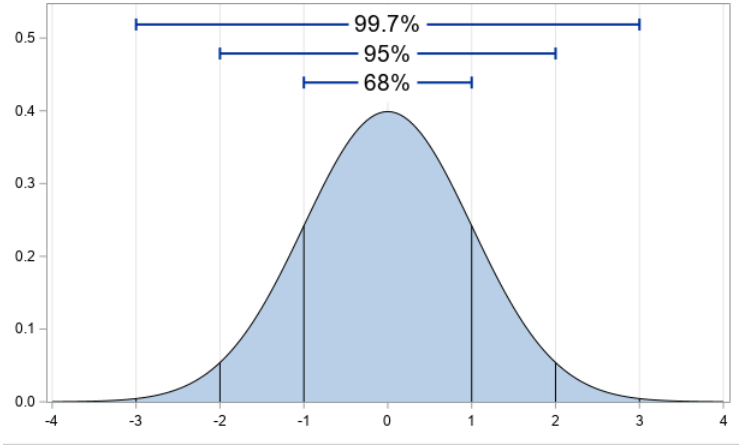
- The plot on the right shows the actual probability of each point

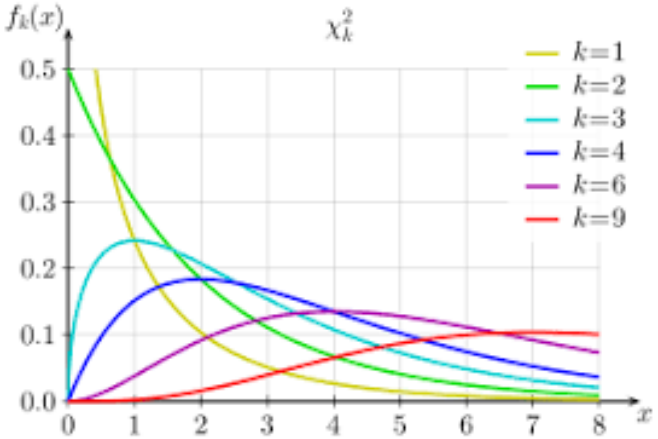


Probability density function

- For continuous random variables, we can't really examine individual values the same way
  - Imagine a random variable takes on all values in [0, 10] with equal probability (called a uniform distribution). The probability of picking any exact value, like 1, 2, π, $\sqrt{7}$, etc. is exactly 0.
  - Instead, we can look at the integral of the distribution over an interval to calculate probabilities





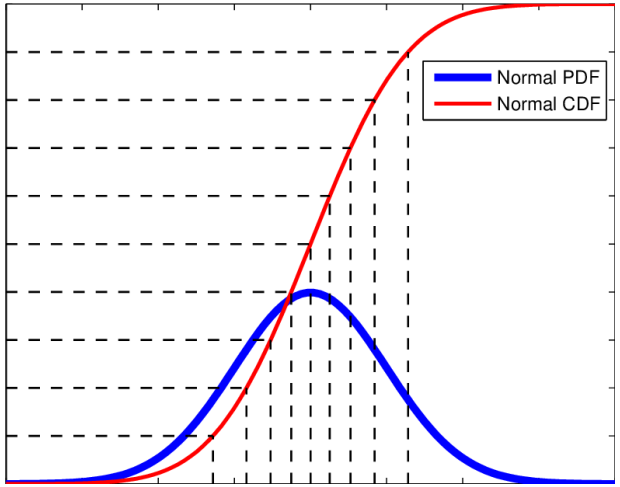


Cumulative distribution functions

- Sometimes it may be useful to look not at the probability at a given point x, but rather the cumulative probability up until that point
  - P(random variable ≤ value) usually written $P(X \leq x)$
  - Can be understood as the integral of the pdf from $-\infty$ to x

- The actual value x is greater than what proportion of the values from the distribution of the random variable

- Conveniently on the [0, 1] scale



Expectation

- In the simplest understanding, this is merely the average of a random variable, often denoted as μ
  - Still approximately, but more accurately, it is the weighted average value of the random variable where the weights correspond to the probability of the RV taking on that value
  - In discrete random variables: $\sum_x xP(X = x)$
  - In continuous random variables: $\int_{-\infty}^{\infty} x f(x)\, dx$
- The expectation in some sense captures the “center of mass” of a distribution
- Nice properties like linearity, monotonicity

Variance

- Expectation captures the center of a distribution, Variance characterizes its spread
- Different formulas to achieve same outcome, but here's one simple one:

$$\mathrm{E}\left[(X-\mu)^2\right]$$

  - One way to see this is the "average squared distance from the average"
- Equivalently

$$\mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$$

- Variance isn't on the scale of the data, taking the square root gives standard deviation
  - Each has some advantages over the other

Common distributions

- Normal – Bell shaped, so more likely to be near the mean than far from it. Used very frequently
- Uniform – Every possible outcome has equal probability. Discrete and continuous versions
- Bernoulli – Generalized coin flip
- Binomial – Number of successes out of n trials. Sum of n Bernoulli variables
- Poisson – Used to model the number of events that occur in a given timeframe given a certain rate
- Note that none of these are one single distribution, but a *family* of distributions, where any particular instance depends on choice of parameters.
  - In a Bernoulli, the probability of success alone will determine how the distribution looks

Moments

- Different distributions have different shapes, but there are an infinite number of these shapes
- We may wish to describe these shapes using only a few values
- First moment is mean, second central moment is variance
- n^{th} raw moment is given by $E(X^n)$
- n^{th} central moment is given by $E([X - E(X)]^n)$
- While you can calculate all moments of a distribution, we usually only see references to versions the first 4: mean (center), variance (spread), skew (symmetry), kurtosis (tail heaviness)

Jointly distributed random variables

- Some random variables may be related to one another
  - Height and weight, house price and square footage
- Instead of lying on a line, the density of a joint distribution lies on a plane
- This is still a probability distribution, and must obey the rules of probability
- Integrating the height of the curve on the plane can tell us the probability of getting a value in that region

Conditional distributions

- Think about two random variables which may have a relationship, e.g. how many hours someone studies and their grades

- We may wish to ask about one of the random variables if we know something about the other
  - E.g. If we know someone studied for X hours, what is the probability of scoring at least Y on the midterm?

- A distribution of X conditional on Y (or X given Y) is the distribution of X when we posit a particular value of Y
  - We can use information about the relationship between X and Y along with the value of Y to better perceive the distribution of Y
  - For example, the distribution of heights for men is different than the distribution of heights for women

Independence

- In other cases, the 2 distributions have no impact on one another
  - Think about someone's height vs how much they like chocolate. Knowing something about one of these doesn't tell you anything about the other

- Formally $f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$
  - Think about it this way: the probability of getting the pair (x, y) is exactly the product of the probability of getting the value x and the probability of getting the value y
  - The reason is these variables have no influence on one another, so if you know the value of y, the distribution of x doesn't change
  - If the probability of being 6 feet or taller is 30% and the probability of liking chocolate is 70%, we expect 21% of the population is over 6 feet and likes chocolate
  - If the probability of being male is 50%, we might expect ***more than 15%*** of the population to be over 6 feet tall males. That's because it's plausible that 50% of men are over 6 feet tall, while only 10% of women are. In that case, 25% of the population are 6 feet tall and male

Bivariate normal

- A particular joint distribution wherein both individual random variables are marginally (that is, unconditionally) normal

- This can be completely described using only 5 parameters:
  - Means for both random variables
  - Variances for both random variables
  - Covariance parameter between variables

- As usual, knowing something about X can give you information about Y

Transformations of variables

- Sometimes we care not about a random variable X itself, but some function of the random variable g(X)
  - Consider looking at the car market and changing the variable from miles per gallon to gallons per mile. This is a simple reciprocal, but how does it change the way a distribution looks?
  - Other times you may have such spread out data you may wish to take the log. Think about wealth in the population

- We will discuss formulas for calculating these transformed distributions

- Some named distributions are special cases or transformations of other distributions

Central Limit Theorem

- States that under certain conditions, the average of many i.i.d. random variables is itself normally distributed with the mean of the sampling distribution is μ, the mean of each of the RVs, and the standard deviation of the sampling distribution is $\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}$
- This holds in many cases even if the RVs aren't normally distributed themselves!
- Upshot: Most of the time, taking the empirical mean of a sample is a good way to approximate the theoretical mean of the underlying distribution, and collecting more data helps

Onto today's main topic: Combinatorics

But first, let's take a break

Basic principle of counting

- We run n experiments
- Each experiment *i* can result in k_i different outcomes
- The total number of possible outcomes is

$$\prod_{i=1}^{n} k_i \quad = \quad k_1 \times k_2 \times \cdots \times k_n$$

Breaking a Password

- Let’s say you must create a 10 character long password where you can use any letter in the alphabet in any slot (assume case doesn’t matter)

- That gives the following options:

AAAAAAAAAA

AAAAAAAAAB

AAAAAAAAAC

....

....

ZZZZZZZZZY

ZZZZZZZZZZ

10 spots with 26 characters

_ _ _ _ _ _ _ _ _ _

In each spot, we can put 26 characters, so there are

26 * 26 * 26 * 26 * 26 * 26 * 26 * 26 * 26 * 26 = 26^10 possible passwords

What if it had to be 9 characters long?

What if we had to use numbers instead of letters?

- Options:

000000000

000000001

000000002

……

……

9999999998

9999999999

_ _ _ _ _ _ _ _ _ _

In each spot, we can put any of 10 digits, so there are

10*10*10*10*10*10*10*10*10*10 = 10^10

Number of different samples for sampling with replacement

- Using the previous slides, if there are k different options to fill n slots, and we can recycle the options to fill multiple slots, there are

  k^n different possible passwords

- This is akin to sampling with replacement. It's equivalent to picking one of k balls out of a hat, writing down which ball, putting the ball back, and repeating this process a total of n times
- One upshot: k^n grows much faster in n than in k
  - That is, to make passwords harder to crack, usually better to add one more slot instead of allowing one more acceptable character

Quick example

- Which rule is harder to crack?
  - Password must be exactly 4 characters long, choose from the letters where case doesn’t matter
  - Password must be exactly 6 characters long, choose from the numbers

Quick question

- Which rule is harder to crack?
  - Password must be exactly 4 characters long, choose from the letters where case doesn’t matter
  - Password must be exactly 6 characters long, choose from the numbers
- Essentially asking which is larger, 26^4 or 10^6
  - 26^4 = 456,976
  - 10^6 = 1 million

What if we can’t replace anything?

- What if Stevens implements a new strange rule for passwords:
  - Must be 10 digits long
  - Only numbers are acceptable
  - No numbers may repeat
- For one thing, it’s more tedious to enumerate the possibilities
- The formula is perhaps harder, too

0123456789

0123456798

0123456879

0123456897

……

9876543201

9876543210

What if we can’t replace anything?

- What if Stevens says we have a strange rule for passwords:
  - Must be 10 digits long
  - Only numbers are acceptable
  - No numbers may repeat
- For one thing, it’s more tedious to enumerate the possibilities
- The formula is perhaps harder, too

1^{st} slot: 10 options remaining

2^{nd} slot: 9 options remaining

3^{rd} slot: 8 options remaining

4^{th} slot: 7 options remaining

….

i^{th} slot: $n - i + 1$ options remaining

….

10^{th} slot: 1 option remaining

Factorials

- In other words, if we have to find the number of unique orderings for n distinct items, the formula is:

n*(n-1)*(n-2)…*3*2*1

- Compactly written as n!
- This is fairly simple, but pretty interesting:

If we have n students in this class, and we run a race, there are n! different orders you could finish

There are 20 horses running in the Kentucky Derby, the probability of guessing the exact order without knowing anything about horses is 1/20! = 1 in 2,432,902,008,176,640,000

or about one in 2.4 quintillion. **Good luck!**

- Note this is much smaller than n which would be the case with replacement, 20^20

~2.4 x 10^18 vs ~10^26, or about 40 million times smaller

What if we only want to know who finishes on the podium?

- What if we don't care about who gets 19^{th} place in the Kentucky Derby, but we do care about who comes in 1^{st}, 2^{nd} and 3^{rd}
  - Frankly I'd be surprised if anyone here cares who comes in first place, but I digress

- 20 horses can come in first
- Contingent on first place, 19 horses can come in second
- Contingent on first and second, 18 horses can come in third

- This looks kind of familiar, so can find a way to tie in factorials?

Using factorials to get permutations

- 20 options for 1st
- 19 options for 2nd
- 18 options for 3rd
- 17 options for 4th
- 16 options for 5th

.....

- 2 options for 19th
- 1 option for 20th

We need the information in black to figure out the number of possible top 3 finishes, but we'd also need the information red to calculate the number of possible orders for all horses

20*19*18 is kind of like 20!, but it's as though we ignore the last 17 slots, and so we don't multiply them

Permutation formula for number of possible top k finishers in a field of size n

$$\frac{\boldsymbol{n!}}{\boldsymbol{(n-k)!}} = \frac{20!}{(20-3)!} = \frac{20!}{17!} = 20 * 19 * 18 * \frac{17!}{17!}$$

Quick question

- Baskin Robbins famously has 31 flavors of ice cream (so you can try a different flavor each day of the month!). People are asked to rank their favorite 5 flavors in order. How many different lists of this type are possible?

Quick question

- Baskin Robbins famously has 31 flavors of ice cream (so you can try a different flavor each day of the month!). People are asked to rank their favorite 5 flavors in order. How many different lists of this type are possible? (You can leave it as a fraction)

- $\frac{n!}{(n-k)!} = \frac{31!}{(31-5)!} = \frac{31!}{26!} = 31 * 30 * 29 * 28 * 27 = 20{,}389{,}320$

Seems like it'd be hard for us all to agree on a list. Maybe we need to sample more ice cream

What if we're choosing groups where the order doesn't matter

- If a class of 15 people has a presentation component where 5 people will present each of the next 3 days, how many different sets of people can go on the first day?
  - From the students' perspective, they don't care when they go in the class, because they'll have to be ready before class starts anyway. All that matters is which day they have to present

- Using the formula $\frac{n!}{(n-k)!} = \frac{15!}{(15-5)!}$ will overstate the number of different sets, since it counts

A-B-C-D-E as different than A-B-C-E-D, but it's the same from the student perspective

  - This makes sense when finishing first is different than finishing second, but here, **after choosing the first 5 students, order doesn't matter. We're simply choosing a group, not individual positions**

What if we're choosing groups where the order doesn't matter

- If a class of 15 people has a presentation component where 5 people will present each of the next 3 days, how many different sets of people can go on the first day?
  - From the students' perspective, they don't care when they go in the class, because they'll have to be done before class starts anyway

- Do we know how many different orderings there are for each set of 5 students? Of course we do! Factorials!
  - For each unique group of 5 students, there are 5! different orderings

- We can take our permutation formula and simply divide the result by the factorial number of elements in the chosen set

What if we're choosing groups where the order doesn't matter

- If a class of 15 people has a presentation component where 5 people will present each of the next 3 days, how many different sets of people can go on the first day?
  - From the students' perspective, they don't care when they go in the class, because they'll have to be done before class starts anyway

- We can take our permutation formula and simply divide the result by the factorial number of elements in the chosen set

$$\binom{n}{k} = \frac{n!}{(n-k)!}/k! = \boxed{\frac{n!}{(n-k)!\,k!}} = \frac{15!}{(15-5)!\,5!} = \frac{15!}{10!\,5!}$$

Quick question

A team of mathletes has 6 players, but only 4 will compete in the competition, and the other 2 will serve as alternates

1. How many different possible sets of starters are there?

2. How many different possible sets of alternates are there?

3. What if we put the starters into 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} chairs, how many different possible sets of starters are there?

Quick question

A team of mathletes has 6 players, but only 4 will compete in the competition, and the other 2 will serve as alternates

1. How many different possible sets of starters are there?

2. How many different possible sets of alternates are there?

1 and 2 have the same answer. Intuitively, choosing the starters also implicitly chooses the alternates and vice versa. You can also look at the structure of the answer

$$\frac{6!}{2!\,4!} = 15$$

Quick question

A team of mathletes has 6 players, but only 4 will compete in the competition, and the other 2 will serve as alternates

3. What if we put the starters into 1st, 2nd, 3rd, and 4th chairs, how many different possible sets of starters are there?

3 is a permutation question because order matters

You can do this directly 6*5*4*3 = 360

Or using the permutation formula $\frac{6!}{(6-4)!} = \frac{720}{2}$ = 360

Recap: n slots to fill, k choices, n ≥ k

- Sampling with replacement $= k^n$ — Think: No restrictions
- Permutations $= \frac{n!}{(n-k)!}$ — Think: Can't recycle
- Combinations $= \frac{n!}{(n-k)!k!}$ — Think: Can't recycle and changing the order doesn't change set

As we move down, we get fewer different options, as we restrict ourselves from differentiating sets in some way

Harder question: 4B from textbook

- From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed?

- What if 2 of the men are feuding and refuse to serve on the committee together?

Harder question: 4B from textbook

- From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed?
- What if 2 of the men are feuding and refuse to serve on the committee together?
- First part – $\binom{5}{2} * \binom{7}{3} = 10 * 35 = 350$

Harder question: 4B from textbook

- From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed?

- What if 2 of the men are feuding and refuse to serve on the committee together?

- Second part – committees with neither man $\binom{5}{2}\binom{5}{3} = 100$

  committees with man A $\binom{5}{2}\binom{5}{2} = 100$

  committees with man B $\binom{5}{2}\binom{5}{2} = 100$

  total=300

Binomial Theorem

The binomial theorem

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Inductive proof of Binomial Theorem

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

- Base Case: when n = 1, the equation above simplifies to $x^0y^1 + x^1y^0 = x + y$, which is correct
- Inductive step: On the board

Backing into the binomial distribution

Seemingly basic question: If I flip a coin n times, what is the probability I get exactly k heads?

- For now, let's assume the probability of getting a heads or a tails is the same
  - That means the probability of getting any "string" of heads or tails is identical: $.5^n$
    - so we can ignore this coefficient throughout, for now

Backing into the binomial distribution

Seemingly basic question: If I flip a coin n times, what is the probability I get exactly k heads?
For now, let's assume the probability of getting a heads or a tails is the same

Simple thought:

Numerator: The number of ways we can get k heads out of n flips

Denominator: all different possible outcomes of n flips

Backing into the binomial distribution

Numerator: The number of ways we can get k heads out of n flips

Insight 1: once we know we have k heads, we also know we have n-k tails

Insight 2: What if we had exactly k blue coins and n-k red coins?

We can rearrange all the coins n! ways, but it won't matter how we rearrange the heads internally, or the tails internally

> Think: Swapping the heads in HTH keeps the sequence as HTH, but swapping the first heads with the tails gives THH, which is a different sequence

That means we can divide by all reorderings of k blue coins as well as all reorderings of n-k red coins

Backing into the binomial distribution

Insight 2: What if we had exactly k blue coins and n-k red coins?

We can rearrange all the coins n! ways, but it won't matter how we rearrange the heads internally, or the tails internally

That means we can divide by all reorderings of k blue coins as well as all reorderings of n-k red coins

This actually looks exactly the same as the combination formula:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Backing into the binomial distribution

Numerator: The number of ways we can get k heads out of n flips

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

Denominator: all different possible outcomes of n flips

Much easier: just 2^n

- Each flip has 2 possibilities, so it's like sampling with replacement

- Sidenote: this is also the total number of subsets in a set of size n
  - Line up every element in the set, and decide whether it's in (1) or out (0). We have n slots with 2 choices for each one. Hence, 2^n

Backing into the binomial distribution

- Seemingly basic question: If I flip a coin n times, what is the probability I get exactly k heads?
- We just derived this formula

$$\binom{n}{k} / 2^n$$

Backing into the binomial distribution

- Seemingly basic question: If I flip a coin n times, what is the probability I get exactly k heads?
- We just derived this formula

$$\binom{n}{k} / 2^n$$

Let's look at this another way:

- Flip n coins. In each slot: 50% chance of H, 50% chance of T
- We must get k heads, so we must also get n-k tails. Therefore, the probability of any string of length n with k heads has probability $(.5^k)*(.5^{n-k}) = .5^n$
- How many distinct arrangements? We just did this, it's $\binom{n}{k}$
- Full probability is thus $\binom{n}{k} * .5^n = \binom{n}{k} * (\frac{1}{2})^n = \binom{n}{k} / 2^n$ Phew! Same answer

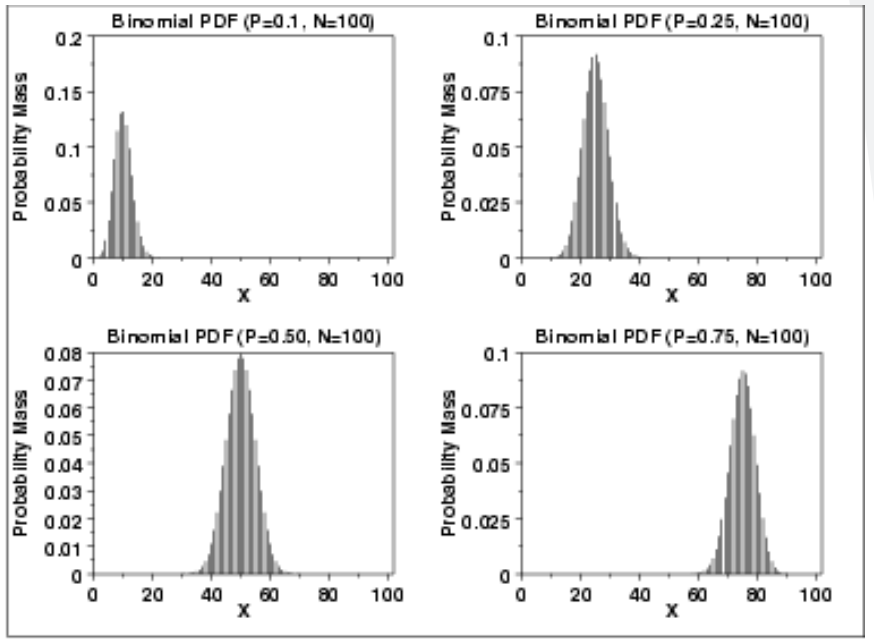
Generalizing to binomial distribution

- If we have n independent trials where each trial is independent and each has probability of success **p,** probability of failure is **q** = 1-p
  - Ex: guessing on a multiple choice test. 4 choices, 20 questions. n= 20, p =.25
- The probability of getting k successes in n tries is given by the formula

$$\binom{n}{k} * p^k q^{n-k}$$

  - Number of distinct ways to rearrange multiplied by the probability of getting a particular string with k successes in n tries

- Notice that if p = .5, q=.5, so $p^k q^{n-k}$ collapses to p^n, so the probability of getting k successes is the same as getting n-k successes, or k failures



Poll on Canvas due Thursday night by 11:59 PM

- How many letters in your first name when spelled out in English?
- If you're going to a party scheduled to start at 7 PM, how many minutes early or late do you usually show up? (6:45 PM = 15 minutes early = -15. 8 PM = 60 minutes late = 60)
- How many different people did you text/message/email etc. on Sunday?
- How many phone calls have you had in the past week?
- How many cups of coffee do you have each week?
- How long is your commute (in minutes) to Stevens?
- How many books do you read each year?
- How many hours do you sleep per night?
- On a scale from 1 – 10, how much did you enjoy taking this poll?

Homework 1

- Due Wednesday 9/20
- Problems from textbook (**9th edition!**) Chapter 1:
  - 1
  - 7
  - 9
  - 13
  - 20
- Theoretical problems from textbook (**9th edition!**) Chapter 1:
  - 10 a - c
- Calculate:
  - $\int_1^2 \frac{1}{x} dx$
  - $\iint\limits_{\substack{x+y<1,\\ x>0,\ y>0}} (x+y)^2 dxdy$