Statistics 110: Introduction to Probability	Fall 2022

Midterm Review

Skyler Wu (skylerwu@college.harvard.edu) *Catherine Huang (catherinehuang@college.harvard.edu)*

Midterm Preparation Tips

The aim of this Midterm Review Guide is to provide general problem-solving strategies and help you recognize common problem types that will be on the midterm. Feel free to reach out to us if you have any questions or concerns - we're here for you, and our contact information is below! Most importantly, don't stress too much - y'all are supergeniuses and will do amazing! Below are some useful study action items, listed in a suggested chronological order (your specific workflow may vary):

1. Read chapters 1-4 of the textbook, *Introduction to Probability*.
2. Create personal review sheets (writing down formulas helps you better understand and remember them!).
3. Work through as many practice midterms as possible and *carefully* review the solutions *after*. You will generally be able to identify common threads and tricks as you do more and more practice problems.
4. For extra support: work through some past section and homework problems. If you have extra time, strategic practice problems (in Canvas) and other practice problems in the textbook are also helpful!
5. For more in-depth, extra-explanation-filled material review, please reference Matt DiSorbo's (AB '17) excellent virtual textbook, *Probability*, a companion guide to Stat 110:
https://bookdown.org/probability/beta/counting.html
6. Attend other Teaching Fellows' review sessions (see Ed and Google Calendar for details).

During the course of today's midterm review session, if you have *any or all questions* on specific concept clarification, midterm logistics, or would like us to work out a particular practice problem or past HW problem, please send them to our Google Form here:
https://bit.ly/CatSkyler110MidtermReview.

Section: Tuesday, 10/11, 3:00-4:15 PM, Science Center 222. We will convert our section into last-minute midterm review office hours. Please bring any and all questions! Happy to walk through any midterm review problems, too.

Office Hours: Please come with any questions!

1. Catherine: Tuesdays, 9:00 - 11:00 PM, Winthrop Dining Hall
2. Skyler: **Canceled**, because it is right after the midterm.

Emails: skylerwu@college.harvard.edu and catherinehuang@college.harvard.edu

Phone numbers: Feel free to reach out to us!

1. Skyler: (858) 205-9095
2. Catherine: (408) 707-7196

Contents

1 Concept Review

1.1 Counting

Multiplication Rule - Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment.

- **Sampling with replacement:** Suppose we have a jar with n distinguishable balls, numbered 1 to n, and suppose that each time we take a ball out of the jar, we return it to the jar. There are n possibilities for each sampled ball, and n^k ways to obtain an ordered sample of size k.

- **Sampling without replacement:** Suppose now that each time we take a ball out of the jar, we do NOT return it to the jar. There are $n(n-1)\dots(n-k+1)$ possibilities for k ordered samples from n balls total, for $1 \leq k \leq n$.

Factorial $n!$ (pronounced n "factorial") is the number of ways to order n people in line, by the multiplication rule.

$$n! = 1 \cdot 2 \cdot 3 \dots n \qquad\qquad 0! = 1$$

The Binomial Coefficient is used often in combinatorics. $\binom{n}{k}$ (read n choose k) is the number of subsets of size k of a set of size n. This is the same as choosing k people out of n without replacement where order doesn't matter.

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

Sampling Table - The sampling table describes the different ways to take a sample of size k out of a population of size n.

	Order Does Matter	**Order Doesn't Matter**
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

1.2 Set Theory

Sets and Subsets - A set is a collection of distinct objects. A is a subset of B (denoted as $A \subseteq B$) if every element of A is also included in B.

Empty Set - The empty set, denoted $\emptyset$, is the set that contains nothing.

Set Notation - Note that $A \cup B$, $A \cap B$, and A^c are all sets too.

- **Union** - $A \cup B$ (read A *union* B) means A *or* B
- **Intersection** - $A \cap B$ (read A *intersect* B) means A *and* B
- **Complement** - A^c (read A *complement*) occurs whenever A does not occur

Disjoint Sets - Two sets are disjoint if their intersection is the empty set (i.e. they don't overlap).

Partition - A set of subsets $A_1, A_2, A_3, ...A_n$ partition a space if they are disjoint and cover all possible outcomes (e.g. their union is the entire set). A simple case of a partitioning set of subsets is A, A^c

De Morgan's Laws - Expressing unions and intersections in terms of each other, via complements.

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

1.3 Properties and Axioms of Probability

Axioms of Probability (quite useful)

1. $P(\emptyset) = 0$, $P(S) = 1$
2. If A_1, A_2, A_3, ..., A_n are disjoint events then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Useful Properties

1. $P(A^c) = 1 - P(A)$
2. If $A \subseteq B$ then $P(A) \leq P(B)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

1.4 The Naive Definition of Probability (not always applicable!)

The naive definition of probability states that for an event A with a finite sample space S, the naive probability of A is:

$$P_{naive}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes in S}}$$

Often, we use counting to derive the naive probability of A.

Note: The naive definition requires that S be finite and that all outcomes be equally likely. It is usually only applicable in situations where there is symmetry in a problem (e.g. Heads or Tails on a fair coin), where outcomes are equally likely by design (e.g. a survey with a simple random sample), or where the naive definition is useful as a null model (e.g. assume the naive definition. Then, look at your observed data to see if that assumption is reasonable).

1.5 Birthday Problem

Be familiar with this famous problem from class. If you see the words "at least" in the problem, try **taking the complement** to simplify the problem!

1. Setup: There are k people in a room. What is the probability that at least one pair of people in the group have the same birthday?

   **Assume that each person's birthday is equally likely to be any of the 365 days of the year, that leap years don't exist, and that people's birthdays are independent.*

2. Strategies:

(a) Complementary counting - sometimes, the number of / probability of the things / events we *don't* want is easier to compute:

$$\text{\# of favorable options} = \text{Total number of options} - \text{\# of undesirable options}$$

$$P(\text{favorable events}) = 1 - P(\text{undesirable events})$$

(b) Visualization: Imagine asking people's birthdays one-by-one in a line.

(c) Naive definition of probability:

$$P(\text{favorable event}) = \frac{\text{\# of favorable options}}{\text{total \# of options}}$$

3. Solution:

(a) Total number of options? In other words, with no restrictions, how many ways can we assign each of our k people a birthday? Answer: $365 \times 365 \times 365 \times ... \times 365 = 365^k$.

(b) Number of unfavorable options? In other words, how many ways can we assign birthdays to our k people such that *no pair of people* has the same birthday? Hint: how many birthday options do we have for the first person in line? What about the second person? The third?
Answer: $365 \times (365-1) \times (365-2) \times ... \times (365-k+1) = \frac{365!}{k!}$

(c) Let's put it all together:

$$P(\text{at least one match}) = 1 - P(\text{no matches}) = 1 - \frac{365 \times (365-1) \times (365-2) \times ... \times (365-k+1)}{365^k}.$$

4. Remarks:

(a) Be careful with the last term! It's $(365-k+1)$.

(b) Many problems in STAT 110 (especially on midterm) are "isomorphic" to the Birthday Problem. The setup is literally almost identical. Just changing the nouns, verbs, and setting. Oftentimes, the trick is to recognize that a problem is just a repackaging of the Birthday Problem.

*In general, if we have k agents, each of whom is assigned one of n possible values independently and with equal probability, the probability of a match is

$$1 - \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}, \text{ for } n \geq k. \text{ If } k > n, \text{ then } P(\text{match}) = 1.$$

1.6 Inclusion-Exclusion

The Principle of Inclusion-Exclusion is useful for finding the probabilities of unions of events that are not necessarily disjoint. For the two and three-event cases, we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Intuitively, we have:

$$P(\text{Union of many events}) = \text{Singles} - \text{Doubles} + \text{Triples} - \text{Quadruples} \ldots$$

Formally, for n events, we have:

$$P(A_1 \cup A_2 \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i,j} P(A_i \cap A_j) + \sum_{i,j,k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots A_n).$$

Example Problem: (De Montmort's Matching Problem) Consider a well-shuffled deck of n cards, labeled 1 through n. We flip over the cards one by one, saying the numbers 1 through n as we do so. What is the probability that at some point, the number we say aloud is the same as the number on the card being flipped?

De Montmort's Matching Problem (as well as other matching problems we have seen in section and homework problems) are common examples of Inclusion-Exclusion.

When solving Inclusion-Exclusion problems involving multiple items (e.g. multiple students, multiple playing cards, etc.) subject to two different mappings or orderings (e.g. the original ordering and a new ordering), keep the following tips in mind:

- The most common strategy is to define an event A_i as the event that item i or person i gets a "match." In De Montmort's Matching Problem, this means the event that the ith card flipped over has the number i written on it.
- $P(A_i) = \frac{1}{n}, P(A_i \cap A_j) = \frac{1}{n(n-1)}, P(A_i \cap \cdots \cap A_k) = \frac{1}{n(n-1)\cdots(n-k+1)}$
- In the inclusion-exclusion formula, there are n terms involving one event, $\binom{n}{2}$ terms involving two events, $\binom{n}{3}$ terms involving three events, and so forth.

1.7 Conditional Probability

Definitions:

1. **Joint Probability:** $P(A \cap B)$ or $P(A, B)$ - Probability of A and B. These two notations are interchangeable!
2. **Marginal (Unconditional) Probability:** $P(A)$ - Probability of A, before updating based on any evidence. We also call this the "prior" probability of A.
3. **Conditional Probability:** $P(A|B)$ - Probability of A given B occurred.

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

Here is a very useful equation that comes directly from rearranging the formula for conditional probability:

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

Here is a longer version that we can generalize to any number of events. It is important to remember that *the order in which events A, B, C, D occur is* ***not*** *important!* Hence, we can choose whichever order is easiest to work with.

$$P(A, B, C, D) = P(A) \cdot P(B|A) \cdot P(C|B, A) \cdot P(D|C, B, A)$$

*note: $P(A|A) = 1$ always holds true!

Conditional Probability is Probability: $P(A|B)$ is a probability as well, restricting the sample space to B instead of Ω. Any theorem that holds for probability also holds for conditional probability.
Here are some useful facts to remember:

1. Conditional probabilities (i.e., $P(A|B)$, or maybe with more events to conditon on) are *always* between 0 and 1!
2. $P(S|B) = 1$ (where S is the entire state space), and $P(\emptyset|B) = 0$ (where $\emptyset$ is the empty set).
3. $P(\cup_{j=1}^{n} A_j|E) = \sum_{j=1}^{n} P(A_j|E)$, *only if* $A_1, \ldots A_n$ are *disjoint*!
4. $P(A^c|E) = 1 - P(A|E)$.
5. PIE still holds! For example, $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$.

1.7.1 Disjointness and Independence

Disjoint Events: A and B are disjoint when they cannot happen simultaneously, or

$$P(A \cap B) = 0$$
$$A \cap B = \emptyset$$

Independent Events: A and B are independent if knowing one gives you no information about the other. A and B are independent if and only if one of the following equivalent statements hold:

$$P(A \cap B) = P(A)P(B)$$
$$P(A|B) = P(A)$$

*Note that $P(B|A) = P(B)$ is also equivalent to the above two statements!

If events A and B are independent (often written as $A \perp\!\!\!\perp B$), then:

1. $A^c \perp\!\!\!\perp B$: the complement of A is independent of B.
2. $A^C \perp\!\!\!\perp B^c$: the complement of A is independent of the complement of B.
3. $A \perp\!\!\!\perp B^c$: A is independent of the complement of B.

Independence of 3 Events: Note that with three or more events, we introduce the concept of "pairwise" independence. Suppose we have three events A, B, C with the following statements:

1. $P(A \cap B) = P(A)P(B)$
2. $P(A \cap C) = P(A)P(C)$
3. $P(B \cap C) = P(B)P(C)$
4. $P(A \cap B \cap C) = P(A)P(B)P(C)$

If the first three equations hold, then A, B, C are *pairwise* independent. This is ***not*** (no, nada, nope) the same thing as *independence*. However, if **all four** equations hold, then A, B, C are *independent*.

Repeat after me. *Pairwise independence does not imply independence!*

Disjointness versus Independence: Note that disjointness and independence are **completely different**! In fact, if A and B are disjoint, then $P(A \cap B) = 0$ (why?), so disjoint events can be independent only if $P(A) = 0$ or $P(B) = 0$. (Knowing that A occurs tells us that B definitely did not occur, and vice versa.)

Conditional Independence: A and B are *conditionally* independent given C if

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Additionally, if A and B are conditionally independent given C, then:

$$P(A|C) = P(A|B, C)$$

$$P(B|C) = P(B|A, C)$$

Note that the three facts from earlier that result from A and B being independent also apply analogously for conditional independence! Just tag on a " $|C$" to each of the three statements.

**Note that conditional independence does not imply independence, and independence does not imply conditional independence! Furthermore, two events can be conditionally independent given an event E but not conditionally independent given E^c.*

1.7.2 Bayes' Rule and Law of Total Probability (LOTP)

Bayes' Rule - Bayes' Rule unites marginal, joint, and conditional probabilities. This is *one of the most important concepts of the course*, and one of the backbones of statistics.

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.}$$

The **Law of Total Probability (LOTP)** relates conditional probability to unconditional probability. This is a useful way to break up a harder problem into simpler pieces, conditioning on what we wish we knew. For any event B and set of events $A_1, A_2, A_3, ...A_n$ that partition a space, the following are true:

$$\boxed{P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + ... + P(B|A_n)P(A_n).}$$

The proof of LOTP arises from the fact that

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n).$$

In the simplest case where A is just one event, we have

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

We often use Bayes' Rule and LOTP together to solve problems involving updating our beliefs based on observed evidence. A well-chosen partition of the sample space into disjoint slices A_i will reduce a complicated problem into simpler (often given) pieces.

Tips and Tricks:

1. Condition on *what you want to know!*
2. Make sure that at least one of the events you condition on must occur! In other words, make sure that all of your cases truly cover the entire realm of possibilities.

1.7.3 Extra Conditioning

The key point here is that *conditional probabilities are still probabilities!*

1. **Bayes' Rule with Extra Conditioning**

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}$$

In the above equation, we are literally just adding a C to every term in the original Bayes' Rule! Observe that C *always* appears after the | bar!

$$P(A|B, C) = \frac{P(B, C|A)P(A)}{P(B, C)}$$

In the second equation (which *is also legit*), we are treating $B \cap C$ as *one* event, and just applying the normal form of Bayes' Rule!

*Again, both of the above formulas for $P(A \mid B, C)$ are legit! Intuitively, please just treat the commas as equivalent to the word "and."

2. **Law of Total Probability with Extra Conditioning**

   Let $A_1, A_2, \dots A_n$ be a partition of S. Then, we have:

$$P(B|C) = \sum_{i=1}^{n} P(B|A_i, C)P(A_i|C) = P(B|A_1, C)P(A_1, C) + \cdots + P(B|A_n, C)P(A_n|C)$$

1.8 Famous Applications of Conditional Probability

1.8.1 Simpson's Paradox

Simpson's Paradox says that X can be more successful than Y when compared within every group (Surgery Types, etc.), but still Y can be more successful than X when compared in the aggregate. This result is caused by a lurking variable, the groups, who (1) have a large impact on the success rate and (2) whose relative sizes are very different between X and Y. In statistical language, we have that:

$$P(A|B, C) < P(A|B^c, C) \qquad \text{and} \qquad P(A|B, C^c) < P(A|B^c, C^c) \qquad \textbf{but} \qquad P(A|B) > P(A|B^c)$$

where A is success, B is one of the things that you are comparing (e.g. Doctors), and C is one of the groups (e.g. Surgery Types).

	Heart	**Band-Aid**
Success	70	10
Failure	20	0

Dr. Hibbert

	Heart	**Band-Aid**
Success	2	81
Failure	8	9

Dr. Nick

Figure 1: The respective records of two doctors, Dr. Hibbert and Dr. Nick. Note that Dr. Hibbert had a higher success rate within each individual surgery type, but Dr. Nick had a higher *aggregate* success rate. (*Dr. Hibbert is performing the harder type of surgery much more often than Dr. Nick is.*)

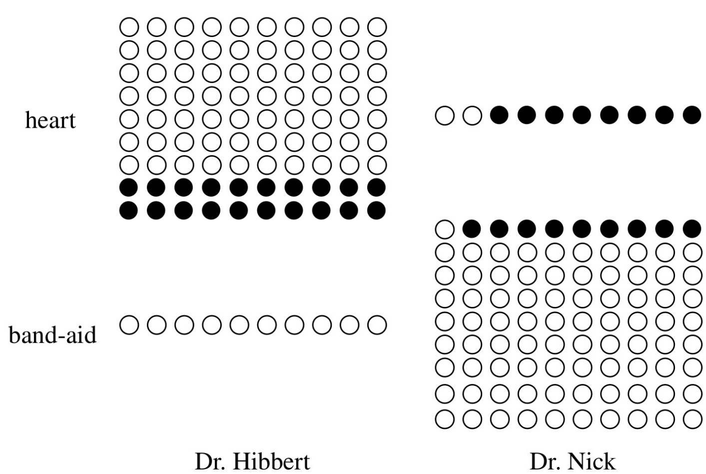

Figure 2: White dots represent successful surgeries, and black dots represent failed surgeries.

1.8.2 Monty Hall

Problem: Monty Hall is a person who really likes goats and cars. Suppose you're on a game show with Monty, and you're given the choice of three doors: behind one door is a car, behind the other two are goats. You pick a door, say door 1, and Monty, who knows what's behind the doors, opens another door which will always have a goat, let's say door 3. He then asks you, "Do you want to switch to door 2?" *Is it to your advantage to switch your choice?*

Solution: The Monty Hall problem is a classic example of how LOTP can be really useful: we utilize *wishful thinking*, **conditioning on what we wish we knew**: which door the car is behind! Let C_1, C_2, and C_3 be the events that the car is behind door 1, 2, and 3, respectively. Let W be the event that we choose the car (and win the game). Then we can write

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3).$$

Suppose without loss of generality that we initially choose door 1. Let's analyze our choices:

1. Switch. If the car is behind door 1, then switching to another door means we lose. If the car is behind either door 2 or door 3, we win, since Monty opens the door that has the goat, and we are guaranteed to switch to the door with the car. Recall that the car is behind each door with probability 1/3. We have
$$P(W) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \boxed{\frac{2}{3}}.$$

2. Stay. We will only win if the car is behind door 1.
$$P(W) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \boxed{\frac{1}{3}}.$$

Hence, we would prefer to switch doors. To help build further intuition for why switching is advantageous, let's consider the extreme case where we have 1,000 doors, 999 of which contain goats and 1 of which has a car. After you pick door 1, Monty opens 998 other doors with goats behind them. Here, it intuitively makes the most sense to switch to the other unopened door.

1.8.3 Gambler's Ruin and First-Step Analysis

Problem: Gamblers A and B have a series of bets, and they bet \$1 with each other, and the game ends when one of them is bankrupt. A starts with \$$i$ and B starts with \$$(N-i)$. p is the probability that A wins a bet, and $q = 1-p$ is the probability that B wins the bet. We wish to find P_i, which is the probability that A wins this game given that A starts with \$$i$.

We start off this problem by conditioning on the first step. If A wins the first bet, then the probability that A wins is P_{i+1}, and if A loses the first bet, then the probability that A wins is P_{i-1}. We know the boundary conditions are $P_0 = 0$ and $P_N = 1$. Thus we have the following recursive difference equation found using the Law of Total Probability (or first step analysis):

$$P_i = pP_{i+1} + qP_{i-1}$$

Solution: The above equation is known as a **difference equation**, and the general solution is as follows: starting at point i, say that the state of ruin is at 0 and the state of success is at N. Say also that the gambler has a probability p of moving up one step, and a corresponding probability $q = 1-p$ of moving down one step. The probability P_i that the gambler reaches N before 0 starting at state i is:

$$P_i = \begin{cases} \dfrac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N}, & \text{if } p \neq q \\ \dfrac{i}{N}, & \text{if } p = q = \dfrac{1}{2} \end{cases}$$

In homeworks and section, we have already seen problems that are either isomorphic to Gambler's Ruin or rely on first-step analysis in some other way. **Some signs that first-step analysis might be helpful include:**

- There are clearly defined boundary conditions in the problem. For example, in Hw 3 (the drunk man wandering on an integer number line), the boundary conditions are that the drunk man reaches k before ever reaching $-j$, or he does not.
- The game or scenario outlined in the problem is iterative, with one round or action happening after another. In the drunk man problem, a round involves the drunk man randomly moving either left or right on the integer line.
- The game or scenario outlined in the problem can be "resetted" in some way.
  - Consider the problem of two people playing a game over and over again, where person A has probability p of winning each game independently. Say we want to find the probability that the first person wins two games more than his opponent. This problem can be solved with first-step analysis, where we condition on how many of the first two games person A wins. (Resetting applies here because if person A wins one of the two games, then person A's net win count is back to 0, and the match is essentially reset.)

1.9 Random Variables: The Bread and Butter of Statistics

Formal Definition: A random variable X is a *function* mapping the sample space S into the real line. Think sum of two dice rolls example (drawn).

"Support" of a Random Variable: The "support" of a random variable is the set of values that a random variable X can take on with *strictly positive* probability. **IMPORTANT: Please always specify the supports of your random variables!** If a value k is outside the support of some random variable X, $P(X = k) = 0$!

Descriptive Definition: A random variable takes on a numerical summary of an experiment. The randomness comes from the the randomness of what outcome occurs. Each outcome has a certain probability. A discrete random variable may only take on a finite (or countably infinite) number of values. Random variables are often denoted by capital letters, usually X and Y.

Example: Let's have X take on the number from a dice roll. So $X = 1$ (this is an event!) with probability 1/6 and $X = 2$ (this is also an event) with probability 1/6 and so on. In this case, X is our *discrete random variable* and $X = 1$, $X = 2$, and so on are our *events*. In notational form, we can write $P(X = 1) = 1/6$, which means that the probability that X takes on the value of 1 is 1/6. Here is the **distribution** of X:

$$X = \begin{cases} 1 & \text{with probability } \frac{1}{6} \\ 2 & \text{with probability } \frac{1}{6} \\ ... & \\ 6 & \text{with probability } \frac{1}{6} \end{cases}$$

Random Variables Expressing Events: Let's say that X is the random variable expressing the sum of two dice rolls. $\{X = 7\}$ is the event that your dice summed to 7, and $\{X \leq 7\}$ is the event that your dice summed to at most 7. $\{X = 2, X = 3, X = 5, X = 7, X = 11\}$ is also an event (the event that X is prime), and so is $\{X = 2, X = 9\}$.

Functions of Random Variables: A function of a random variable is also a random variable! A function of multiple random variables is also a random variable!

Let X be a discrete random variable, and let g be a function from $\mathbb{R}$ to $\mathbb{R}$. Then, the PMF of the random variable $g(X)$ is defined as

$$P\big(g(X) = y\big) = \sum_{x:g(x)=y} P(X = x)$$

Independence of Random Variables: If X and Y are independent, then the following hold if *for all possible values of* $x, y \in \mathbb{R}$,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$

and

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

If X and Y are independent, then any function of X is independent of any function of Y.

Identically distributed: Two random variables are identically distributed if they have the same distribution. This says *nothing* about whether they are independent or equal.

"i.i.d": This means "independent and identically distributed." If we have i.i.d. random variables, then they are all identically distributed and independent — knowing the value of one does not tell us anything new about the others.

1.10 Expectation

Expected Value (also know as *mean*, *expectation*, or *average*) can be thought of as the "weighted average" (wink, wink) of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the possible values that X can take, the expected value of X can be calculated as follows:

$$E(X) = \sum_i x_i P(X = x_i)$$

EXPECTATION IS A FIXED VALUE! IT IS NOT A RANDOM VARIABLE!

Note that if two random variables have the same distribution (even if they are dependent), their marginal expected values are equal! For example, the expected value of the number of heads you flip using a fair coin is equal to the expected value of the number of tails you flip using that same fair coin. Of course, the two quantities are very dependent, but we don't care.

Conditional Expected Value is calculated like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x xP(X = x|A)$$

1.10.1 Linearity of Expectation

Linearity of Expectation allows us to find the expected value of combinations of random variables.

1. Please note that the following properties below apply to *any* random variables X and Y, and constants a and b.

2. To emphasize, X and Y *do not* have to be independent. Again, repeat after me: X and Y *do not* have to be independent.

3. **Linearity of Expectation Properties**

   (a) Expectations can be summed:
   $$E(X + Y) = E(X) + E(Y)$$

   (b) Expectation respects scalar multiplication:
   $$E(cX) = cE(X)$$

(c) We can put these two properties together:

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

4. Did we mention that X and Y *do not* have to be independent?

1.10.2 Indicator Random Variables, Symmetry, Fundamental Bridge

Indicator Random Variables is random variable that takes on either 1 or 0. The indicator is always an indicator of some event. If the event occurs, the indicator is 1, otherwise it is 0. They are useful for many problems that involve counting and expected value. We usually denote the indicator random variable corresponding to the event A as "I_A."

Distribution: $I_A \sim \text{Bern}(p)$ where $p = P(A)$. Yes, indicator random variables are just a fancy way of saying *Bernoulli random variables!*

Fundamental Bridge: The expectation of an indicator for A is the probability of the event A:

$$E(I_A) = P(A).$$

Notation:

$$I_A = \begin{cases} 1 & \text{A occurs} \\ 0 & \text{A does not occur} \end{cases}$$

Here are some useful properties of indicator random variables. Let A and B be events. Then, we can use the following properties:

1. $I_A^k = I_A$, for any positive integer k. Think: $1^k = 1$ and $0^k = 0$.
2. $I_{A^c} = 1 - I_A$.
3. $I_{A \cap B} = I_A I_B$.
4. $I_{A \cup B} = I_A + I_B - I_A I_B$ (analogous to PIE)

1.10.3 Using Linearity, Fundamental Bridge, and Symmetry Together

One useful strategy when trying to find the expected value of some complex random variable (especially with multiple trials, each not necessarily independent) is to write it as the sum of indicator random variables. Because linearity of expectation *does not* require the random variables to be independent, we can oftentimes simply problems significantly. **Some of your new favorite phrases in Stat 110 will be "by symmetry," "by linearity of expectation," and "by the fundamental bridge." Wink. Wink. Nudge. Nudge.**

Again, the steps are as follows for finding the expected value of some random variable X:

1. Define an indicator r.v. I_j $(1 \leq j \leq n)$ such that the n indicator random variables sum up to X
2. Use the fundamental bridge to find $E(I_1) = P(I_1 = 1)$
3. Apply symmetry if applicable: $E(I_j) = E(I_1)$
4. Apply linearity to the expected value of this sum of indicators to get the sum of expected values. $E(X) = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n) = nE(I_1)$

1.10.4 LOTUS

How do we find the expected value of a function of a random variable? Normally, we would find the expected value of X this way:

$$E(X) = \sum_x xP(X = x)$$

LOTUS states that you can find the expected value of a *function of a random variable* g(X) this way:

$$E(g(X)) = \sum_x g(x)P(X = x)$$

What's a function of a random variable? A function of a random variable is also a random variable! For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ could be the number of bike wheels you see in an hour. Both are random variables.

Why is LOTUS important? You don't need to know the PDF/PMF of $g(X)$ to find its expected value. All you need is the PDF/PMF of X.

1.11 Variance

Variance and its Properties

1. Variance is the "expected squared distance away from the mean".
   VARIANCE IS A FIXED VALUE! IT IS NOT A RANDOM VARIABLE!
2. Mathematically, variance is calculated as follows, in two equivalent forms. The second form may be easier to work with, as we can use LOTUS to find $E(X^2)$.
$$\text{Var}(X) = E\left[\left(X - E(X)\right)^2\right] = E(X^2) - E(X)E(X)$$
3. The *standard deviation* of a random variable is simply the square root of the variance. Note that standard deviation has the same units as the random variable itself.
$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$
4. These properties will be very helpful, and hold for *all* random variables:
   (a)
$$\text{Var}(cX) = c^2\text{Var}(X)$$
   (b)
$$\text{Var}(X) = \text{Var}(-X)$$
   (c)
$$\text{Var}(X + b) = \text{Var}(X)$$
   (d)
$$\text{Var}(aX + b) = a^2\text{Var}(X) \text{ for any constants } a, b.$$
   (e) Variance is *ALWAYS nonnegative!*
$$\text{Var}(X) = E(X^2) - E(X)E(X) \geq 0 \longrightarrow E(X^2) \geq E(X)E(X).$$
5. The following **only holds if X and Y are independent**:
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$
6. For general X and Y, we have the following: you are not expected to know this for the midterm, but we like sneak peeks.
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

1.11.1 2 Useful Ways To Calculate $E(X^2)$:

1. Using LOTUS:
$$E(X^2) = \sum_{x \in S} x^2 P(X = x), \text{ where } S \text{ is the support of } X.$$

2. Using the definition of variance:
$$E(X^2) = \text{Var}(X) + E(X)E(X)$$

1.12 PMFs, CDFs, Independence

A **distribution** describes the probability that a random variable takes on certain values. Some distributions are commonly used in statistics because they can help model real life phenomena. In other words, think of a "distribution" as "rules" that govern what values a random variable can take on, and with what probability.

Probability Mass Function (PMF): (Discrete Only) gives the probability that a random variable takes on the value X.

$$P_X(x) = P(X = x)$$

Sanity check: PMFs are probabilities, so PMFs can NEVER be negaitve! PMFs must also add up to 1! Put another way, $X = k$ is an *event*!

Cumulative Distribution Function (CDF): gives the probability that a random variable takes on the value x or less. CDFs exist for *all random variables, regardless of whether they are discrete or continuous!*

$$F_X(x_0) = P(X \leq x_0)$$

Note that $P(X \leq x_0) = 1 - P(X > x_0)$. This may prove to be quite useful at times!

For Discrete Distributions ONLY: We can calculate the CDF by adding up the values of PDF. For example, if $X \sim Bin(n, p)$,

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

Properties of CDFs

1. CDFs are strictly increasing! If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
2. CDFs are right continuous. CDFs may have some jumps, but the CDF must always be continuous from the right. Formally, for any a,
$$F(a) = \lim_{x \to a^+} F(x)$$
3. CDFs must converge to 0 and 1 in the limits:
$$\lim_{x \to -\infty} F(x) = 0$$
$$\lim_{x \to +\infty} F(x) = 1$$

Sanity check: CDFs are probabilities, so CDFs can NEVER be greater than 1, nor can they be less than 0!

Independence: Intuitively, two random variables are independent if knowing one gives you no information about the other. X and Y are independent if for ALL values of x and y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

1.13 Named Distributions

1.13.1 Bernoulli

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial, or $n = 1$. Let us say that X is distributed $\text{Bern}(p)$. We know the following:

1. **Story:** X is either 1 or 0. X "succeeds" (is 1) with probability p, and X "fails" (is 0) with probability $1 - p$.

2. **Example:** A fair coin flip is distributed $\text{Bern}(\frac{1}{2})$.

3. **Support:** The support of X is $\{0, 1\}$.

4. **PMF:** The probability mass function of a Bernoulli is:

$$P(X = x) = p^x(1-p)^{1-x}$$

or simply

$$P(X = x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

5. **Expectation:**

$$E(X) = p$$

1.13.2 Binomial

Let us say that X is distributed $\text{Bin}(n, p)$. We know the following:

1. **Story:** X is the number of "successes" that we will achieve in n *independent* trials, where each trial can be either a success or a failure, each with the *same* probability p of success. We can also say that X is a sum of multiple independent $Bern(p)$ random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. We can express the following:

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

2. **Example:** If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed $\text{Bin}(10, \frac{3}{4})$, or, letting X be the number of free throws that he makes, X is a Binomial Random Variable distributed $\text{Bin}(10, \frac{3}{4})$.

3. **Support:** If $X \sim \text{Bin}(n, p)$, the support of X is $\{0, 1, \ldots, n\}$.

4. **PMF:** The probability mass function of a Binomial is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

5. **Expectation:**

$$E(X) = np$$

6. **Useful Properties:**

   (a) If $X \sim \text{Bin}(n, p)$, then we can write $X = X_1 + \ldots + X_n$, where each X_i are i.i.d $\text{Bern}(p)$.

   (b) If $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ (note the equal probabilities of success), and X is independent of Y, then $X + Y \sim \text{Bin}(n + m, p)$. We can think of this as combining two sets of independent trials together.

   (c) Let $X \sim \text{Bin}(n, p)$. Define the probability of failure $q = 1 - p$. Then, $n - X \sim \text{Bin}(n, q)$.

1.13.3 Geometric

Let us say that X is distributed Geom(p). We know the following:

1. **Story:** X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

2. **Example:** If each pokeball we throw has a $\frac{1}{10}$ probability to catch Mew, the number of failed pokeballs before we catch Mew will be distributed Geom($\frac{1}{10}$).

3. **Support:** The support of X is $\{0, 1, \ldots\}$.

4. **PMF:** With $q = 1 - p$, the probability mass function of a Geometric is:

$$P(X = k) = q^k p$$

5. **Expectation:**

$$E(X) = \frac{q}{p}$$

6. **First Success (a directly related distribution):** Let $Y \sim \mathcal{FS}(p)$ and $X \sim \text{Geom}(p)$. Y represents the number of trials before the first success, *including* the first success. Since $Y - 1 = X$, the PMF and expectation of a First Success are:

$$P(Y = k) = q^{k-1} p$$

$$E(Y) = \frac{q}{p} + 1 = \frac{1}{p}$$

1.13.4 Negative Binomial

Let us say that X is distributed NBin(r, p). We know the following:

1. **Story:** X is the number of "failures" that we will achieve before we achieve our rth success. Our successes have probability p.

2. **Example:** Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed NBin($3, .6$).

3. **Support:** The support of X is $\{0, 1, \ldots\}$.

4. **PMF:** With $q = 1 - p$, the probability mass function of a Negative Binomial is:

$$P(X = n) = \binom{n + r - 1}{r - 1} p^r q^n$$

5. **Expectation:**

$$E(X) = r\frac{q}{p}$$

   Since the $X \sim \text{NBin}(r, p)$ can be interpreted as a sum of r Geom(p) r.v.'s, this is true by linearity of expectation.

1.13.5 Hypergeometric

Let us say that X is distributed HGeom(w, b, n). We know the following:

1. **Story:** In a population of b undesired objects and w desired objects, X is the number of "successes" we will have in a draw of n objects, without replacement.

2. **Examples:** The classic application of the Hypergeometric distribution involves sampling n balls from a urn with w white and b black balls. X is the number of white balls you will draw in your sample of size n.

   Other examples of the Hypergeometric distribution in action are as follows:

   (a) Elk Capture-Recapture Problem - You have N elk, you capture n of them, tag them, and release them. Then you recollect a new sample of size m. How many tagged elk are now in the new sample?

   (b) Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n of the Pokemon in the forest, and X is the number of Pikachus in our encounters.

   (c) The number of aces that you draw in 5 cards (without replacement).

3. **Support:** The support of X is $\{0, 1, \ldots, \min(n, w)\}$.

4. **PMF:** Let $X \sim \text{Hypergeometric}(w, b, n)$. The probability mass function of X is:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

5. **Expectation:** Let $X \sim \text{Hypergeometric}(w, b, n)$. We can write X as the sum of Bernoulli indicator r.v.'s, $X = I_1 + \cdots + I_n$, where (going with the black and white balls example) $I_j = 1$ if the jth ball in the sample is white, 0 otherwise. By symmetry, $I_j \sim \text{Bern}(p), p = \dfrac{w}{w+b}$. By linearity of expectation,

$$E(X) = \frac{nw}{w+b}$$

1.13.6 Some useful ideas/connections between Binomial and Hypergeometric

1. **Independent vs. Dependent** the Binomial distribution models *independent* trials, while the Hypergeometric distribution models inherently *dependent* sampling – we are sampling without replacement.

2. **Link between Binomial and Hypergeometric:** If $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, with $X \perp\!\!\!\perp Y$ (this means "independent of"), then the conditional distribution of X given $X + Y = r$ is $\text{Hypergeometric}(n, m, r)$. (see pg. 135 in *Introduction to Probability* (2e)).

3. **Hypergeometric with Super Large Population:** If $X \sim \text{Hypergeometric}(w, b, n)$, and $N = w + b \to \infty$ (i.e., there are a LOT of individuals in the population), then the PMF of X converges to the $\text{Bin}(n, p)$ PMF, where $p = \frac{w}{w+b}$. One way to think about why this makes sense is that if you have an effectively infinite population, sampling without replacement is effectively negligibly different from sampling with replacement. Specifically, for super large values of $N = w + b$, we can approximate the $\text{Hypergeometric}(w, b, n)$ PMF with the $\text{Bin}(n, \frac{w}{w+b})$ PMF.

1.13.7 Poisson

1. **Story:** There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

2. **Example:** A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, the number of accidents in a month at that intersection is distributed Pois(2). The number of accidents that happen in two months at that intersection is distributed Pois(4).

3. **Support:** Let $X \sim \text{Pois}(\lambda)$. The support of X is $\{0, 1, \ldots\}$.

4. **PMF:** Let $X \sim \text{Pois}(\lambda)$. The PMF of X is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

5. **Expectation:**

$$E(X) = \lambda$$

6. **Useful Properties:**

   (a) **Sum of Independent Poissons:** If $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$, then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

   (b) **Poisson given a sum of Poissons is Binomial** If $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$, then the conditional distribution of X given $X + Y = n$ is

   $$\text{Bin}\Big(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\Big).$$

   (c) **Binomial taken to its limits is approximately Poisson:** Let $X \sim \text{Bin}(n, p)$. Let $n \to \infty$ and $p \to 0$, with $\lambda = np$ staying fixed. Then, the PMF of X converges to the $\text{Pois}(\lambda)$ PMF.

1.13.8 Chicken-Egg

Say that a chicken lays $N \sim \text{Pois}(\lambda)$ eggs and that each one has a probability p of hatching, *independently of other eggs*. Let $q = 1 - p$. Let X be the number of eggs that hatch and Y be the number of eggs that do not. Two elegant facts arise from this:

1. $X \sim \text{Pois}(\lambda p)$, and $Y \sim \text{Pois}(\lambda q)$
2. X and Y are independent.

1.14 Useful Math Tricks

1. Series representation for e:

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots$$

2. Series representation for e^x:

$$\lim_{n\to\infty} (1 + \frac{x}{n})^n = e^x$$

3. "Choosing k team members is the same action as choosing $n - k$ non-team members."

$$\binom{n}{k} = \binom{n}{n-k}$$

4. Hockey-Stick Identity

$$\sum_{t=k}^{n} \binom{t}{k} = \binom{k}{k} + \binom{k+1}{k} + \dots + \binom{n}{k} = \binom{n+1}{k+1}$$

5. Vandermonde's Identity: "Forming a team of Harvard and Yale students. How many students do you choose from each school?"

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

6. Combinatorics to Exponentiation: "When forming a subset from a set, each element is either in or out of the subset. A subset can contain anywhere from 0 to n elements."

$$\sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n$$

2 Problem Solving Tips

- **Practice and pattern match.** Does the structure of the problem resemble something you've seen before? For example, try matching the problem with the story of a known distribution.

  Once you understand the material, the best thing you can do is practice problems. The more you do, the better you'll get at pattern matching :)

- **Test simple and extreme cases.** If the problem gives you variables for values, make up numbers to make the problem more concrete. In general: start with easy examples. Try extreme cases. Draw pictures. Testing is especially helpful it you're struggling to pattern match or need more intuition.

- **Clearly define any events and random variables you introduce.** Clear notation means clear thinking. Write out your notation explicitly ("Let A be the event that the coin lands heads") and then write out what you're supposed to find in terms of that notation. For indicator random variables, this means clearly defining when $I = 1$ and when $I = 0$.

- **Look for symmetry.** Symmetry makes things easier and simpler. For example, if probabilities and/or expectations associated with person 1 also apply to any other person, then we can find information about person 1 and generalize to all people j.

- **Consider indicator random variables.** If you're trying to count something, break up what you're trying to count into one indicator random variable per item. Indicators break what you're counting into smaller pieces. Define one indicator r.v. per person or item, and use symmetry and the fundamental bridge to simplify the problem.

- **Check that you didn't commit category errors.** Category errors arise when applying the wrong tools to certain problems or incorrectly associating different ideas together. For example, if X is an r.v., then writing $P(X)$ is a category error − X is a random variable, not an event! Getting a probability greater than 1 is another example of a category error.

- **Be thorough.** Remember to fully explain anything you do. This not only helps whoever reads your work understand your intentions (and reward you for them!), but it also helps you more clearly navigate through your own thought processes.

- **Do sanity checks.** Before you move on from a problem, ask yourself if the answer makes sense - "is the number as small or as large as I would expect it to be?"

- **Check the support of a distribution.** Please, please, please explicitly state the support of a distribution/PMF, whenever applicable! You may likely encounter questions that ask "Is [insert distribution here] a good model for this situation?" For example, recall that a Bernoulli random variable can only take on the values of 0 or 1.

Binomial, Bernoulli, Neg. Binomial, Geometric, Hypergeometric

Use this table to see the differences between the Binomial, Bernoulli, Negative Binomial, Geometric, and Hypergeometric Distributions. All of them include drawing elements from a sample, the differences include the rule for how many you draw, and whether or not those draws are made with replacement. This table includes all of the discrete distributions that we have covered minus the Poisson.

	Draw With Replacement (prob of success constant)	**Draw Without Replacement** (prob success changes)
Fixed Number of Trials (n)	Binomial (Bernoulli - Special Case $n = 1$)	Hypergeometric
Draw until k successes	Negative Binomial (Geometric - Special Case $k = 1$)	Negative Hypergeometric (Not required for Stat 110)

Discrete Distributions

Distribution	PMF and Support	Expected Value	Variance	Equivalent To
Bernoulli Bern(p)	$P(X = 1) = p$ $P(X = 0) = q$	p	pq	$\text{Bin}(1, p)$
Binomial Bin(n, p)	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k \in \{0, 1, 2, \ldots n\}$	np	npq	Sum of n Bern(p)
Geometric Geom(p)	$P(X = k) = q^k p$ $k \in \{0,\ 1,\ 2,\ \ldots\}$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\text{NBin}(1, p)$
First Success $\mathcal{FS}(p)$	$P(X = k) = q^{k-1} p$ $k \in \{1,\ 2,\ 3,\ \ldots\}$	$\frac{1}{p}$	$\frac{q}{p^2}$	Geom(p) + 1
Negative Binomial NBin(r, p)	$P(X = n) = \binom{n+r-1}{r-1} p^r q^n$ $n \in \{0,\ 1,\ 2,\ \ldots\}$	$r\frac{q}{p}$	$r\frac{q}{p^2}$	Sum of r Geom(p)
Hypergeometric Hypergeometric(w, b, n)	$P(X = k) = \binom{w}{k}\binom{b}{n-k} / \binom{w+b}{n}$ $k \in \{0,\ 1,\ 2,\ \ldots\}$	$n\frac{w}{b+w}$	$\frac{w+b-n}{w+b-1} n \frac{\mu}{n}(1 - \frac{\mu}{n})$	
Poisson Pois(λ)	$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ $k \in \{0,\ 1,\ 2,\ \ldots\}$	λ	λ	

*NOTE: Negative Binomial is number of **failures** until r successes. Negative Binomial is **NOT** the total number of trials.*

Good luck!! Y'all are going to do amazing! Best wishes, godspeed, and salúd!
- Cat and Skyler