How is probability defined?

Classical: $\frac{\text{number of ways an event can occur}}{\text{total number of outcomes}}$, assuming that all outcomes are equally likely
Relative Frequency: limiting proportion of times an event occurs in a long series of repetitions of an experiment
Subjective Probability: how sure a rational person is that an event will happen

What is probability?

- The **theory of probability** makes it possible to describe and quantify uncertainty

Sample spaces and Probability

- Probability is defined on an **event** resulting from a single **trial** of an **experiment**
- We denote events A and the probability of events $P(A) \in [0, 1]$
- The **sample space** S is the set of all possible outcomes of the experiment
  - The sample space can be discrete (bijective with $\mathbb{N}$) or non-discrete (continuous)
- An event $A \subseteq S$ is a subset of the sample space
  - A simple event is a single element of the sample space
  - Compound events contain more elements

Set notation

- $x \in A$ if the outcome x is in the event A
- Union: $A \cup B = \{x \mid x \in A \vee x \in B\}$
  - This corresponds to `or` (event A or event B) as long as the union is disjoint
- Intersection $A \cap B = \{x \mid x \in A \wedge x \in B\}$
  - This corresponds to `and`
- Complement $\bar{A} = \{x \in S, x \mid x \notin A\}$
  - This corresponds to `not`
- The empty event corresponds to the empty set $\emptyset$

Probability Definitions

- If S is discrete, we can assign probabilities to each outcome $P(a_i)$ such that
  - $0 \leq P(a_i) \leq 1$
  - $\sum_i P(a_i) = 1$
- In this case, the set of all $P(a_i)$ is a **probability distribution on** S
  - This doesn't say anything about what the values actually are, but questions will use values we are familiar with (i.e. the probability of a coin toss landing on heads is 0.5)
- The probability of an event is the sum of the probabilities of the individual outcomes that make it up

Elementary facts about probability functions

- $P(\emptyset) = 0$
- If $A_1, A_2, \ldots A_n$ are disjoint events, then the probability of their union is the sum of their probabilities

- $P(\bar{A}) = 1 - P(A)$

Odds

- The odds in favor of an event is the probability that it occurs over the probability that it doesn't: $\frac{P(A)}{1 - P(A)}$
  - The odds against is the reciprocal of this expression

- It is useful to think of probability in terms of number of ways an event can happen over the number of possible outcomes
  - This assumes that S is **equally likely**
  - Note that $P(S) = 1$

Addition rule

- If A and B are disjoint events, $|A \cup B| = |A| + |B|$
- Otherwise, $|A \cup B| = |A| + |B| - |A \cap B|$ (inclusion-exclusion again)
  - These are really the same rule; A and B being disjoint implies $|A \cap B| = 0$

Multiplication Rule

- If there are p ways to do thing 1 and q ways to do thing 2, there are pq ways to do things 1 and 2 in succession

Counting arrangements and permutations

- A **permutation** of size k of n objects is an ordered subset of the k objects
  - This is equal to $\dfrac{n!}{(n-k)!}$
  - Notation: $n^{(k)}$
- Applies when we select objects without replacement (i.e. pick k objects from n objects)
- If we select with replacement (repetition), there are actually n^k choices
- There are $n!$ ways to arrange n unique objects

Arrangements where symbols are repeated

- Here, we must compensate for identical symbols by dividing by the number of ways the symbols can be picked
  - If there are k identical symbols of a certain type, there are $k!$ ways to pick a particular arrangement
- In general, if there are n objects with k types, where there are c_1 objects of type 1, c_2 objects of type 2 (etc.), there are $\dfrac{n!}{c_1!c_2!\ldots c_k!}$ distinguishable arrangements of the n objects

Counting combinations

- $\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n!}{(n-k)!k!}$
- Combinations are the non-ordered equivalent of permutations; order does not matter
- As such, we compensate for the number of ways the same set of objects can be ordered

Useful series and sums

- Finite Geometric Series: $\sum_{i=0}^{n-1} t^i = 1 + t + t^2 + \cdots + t^{n-1} = \frac{1-t^n}{1-t}$ where $t \neq 1$
- Infinite Geometric Series: $\sum_{x=0}^{\infty} t^x = 1 + t + t^2 + \cdots = \frac{1}{1-t}$
- Binomial theorem: $(1+t)^n = \sum_{x=0}^{\infty} \binom{n}{x} t^x$
- Hypergeometric identity: $\binom{a+b}{n} = \sum_{x=0}^{\infty} \binom{a}{x} \binom{b}{n-x}$

Strategies for counting (general)

- If a restriction applies to one member of a group being selected, select that one first and work backwards from there, even if this is not the order they are really picked in
- Check for double counting: counting something twice will lead to a wrong answer
- Make sure that every possible event is counted
- If identical objects are being arranged, compensate for any possible ordering
- See if the question is asking for an ordered or unordered subset
- Do not compensate for ordering twice
- If we are picking unordered groups, use a combination to calculate the sample space instead of calculating the permutation and compensating for repetition
- Split into cases where $\leq$ and $\geq$ are involved
- Remove unnecessary details: instead of thinking of multiple groups (colored marbles, etc.), replace the question with the set of numbers $\{1 \ldots n\}$, then calculate the probability with respect to picking the corresponding elements

Some basic principles of probability

1. $P(S) = 1$, since S is the set of all possible outcomes
2. For any event, A, $0 \leq P(A) \leq 1$
3. If for events A and B we have $A \subseteq B$, then $P(A) \leq P(B)$

De Morgan's Laws

- These are the same De Morgan's Laws from logic, but they are phrased in the language of sets instead of logic

> **De Morgan's Laws**
> $(A \cup B)^c = A^c \cap B^c$
> $(A \cap B)^c = A^c \cup B^c$

- These are easily illustrated on a Venn diagram

Inclusion exclusion rule

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Again, if A and B are disjoint, then $P(A \cap B) = \emptyset$, so the rule becomes $P(A \cup B) = P(A) + P(B)$ since $P(\emptyset) = 0$ by definition
- For three logical events, we must consider which intersections get double counted
  - We must subtract each time two sets intersect since they get counted twice
  - Doing this double-removes the intersection of all three sets, so we must add it back in
  - Thus, we get $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
- For more than three sets, this alternating pattern continues; we must add the intersections of 4 sets, subtract the intersection of 5, add the intersection of 6, etc.

Independence

- Two events are **independent** $\iff P(A \cap B) = P(A)P(B)$
  - I.e. the probability of one event does not change based on the outcome of another
- This definition extends to more than two events if every combination of the events adheres to the formula (possibly with more terms)

Conditional Probability

- The probability of a given event may "change" if we learn more information
  - Ex. the probability that it will rain today vs. the probability that it will rain today given that it is currently cloudy
- The probability of A given B (where $P(B) > 0$) is $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
  - Essentially, the probability of A and B happening given that B has happened
- Conditional probability leads to another, more semantic definition of independence: two events are independent if either $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$
  - This can be verified using the formula for conditional probability

Manipulating Conditional Probabilities

- Conditional probability behaves the same way as regular probability:
  - $P(A^C \mid B) = 1 - P(A \mid B)$
  - If A_1 and A_2 are disjoint, $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B)$
  - Etc.

Product rule

- $P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$
  - Probability of A and B happening is the probability of A times the probability of B happening given A
  - This formula is very useful because it allows us to find $P(B \mid A)$ if we know $P(A \mid B)$

Bayes Theorem

- The famous Bayes theorem gives us a way to find $P(A \mid B)$ if we know $P(B \mid A)$
- $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- This can be derived from the product rule

Law of Total Probability

- A series of sets $A_1 \dots A_k$ are a **partition** of the sample space if every combination of these sets is disjoint and the union of all the sets is the sample space, i.e. $\bigcup_{i=0}^{k} A_k = A$
  - I.e. these sets are a division of the sample space
- **Law of total probability**: $P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_k) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)$
- This is commonly used with complements, since a set and its complement are a partition of S: $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$

Strategies

- Draw a Venn Diagram
  - These can be used to visualize sets and their unions, intersects, and complements, especially when these are combined to form a complicated expression
  - Makes it easier to solve inclusion/exclusion problems

- So far, we've described sets in sample spaces to describe outcomes
- Now, we might want to describe the number of times a particular outcome happens using a variable
  - Ex. Let A describe the event that a coin tossed 3 times comes up heads 3 times, and let X denote "the number of times heads comes up when a coin is tossed three times"
  - The, we have $P(A) = P(X = 3)$
  - However, by using a variable, we open the door to looking at $P(X = 2)$, etc.

Random Variables

- If we let X denote "the number of heads when a coin is tossed three times", we can denote a value for X for each element of the sample space (i.e. how many heads occur)
  - We say that X is a **random variable**
  - We say that $\{0, 1, 2, 3\}$ is the **range** of X
- **Random variable**: a function that assigns a real number to each point in a sample space S
  - Often abbreviated RV or rv
  - The range of a random variable is often denoted $X(S)$ if X is the random variable

Discrete vs. continuous variables

- Variables are **discrete** if their range is finite or countably infinite
- Variables are **continuous** if their range is uncountably infinite (i.e. a non-0 range of $\mathbb{R}$)
  - In real life, we can only measure things discretely, but it still makes sense to think of some measurements as being continuous

Probability Functions

- Let X be a random variable with range A.
- The probability function of X *in the discrete case* is $f(x) = P(X = x)$
  - Defined for all $X \in A$
- The set of pairs $\{(x, f(x)) \mid x \in A\}$ is the **probability distribution** of X.

Properties of discrete probability functions

- For all x, $0 \leq f(x) \leq 1$
- $\sum\limits_{x \in A} f(x) = 1$
  - Note that this is not necessary true in the continuous case

Cumulative distribution function

- Sometimes, we want to know the probability of a compound event
  - The probability that a die roll is less than 4, for example
- The **cumulative distribution function** is defined as $F(x) = P(X \leq x)$
  - Tells us the probability that a variable is less than or equal to the inputted value
- For discrete variables, this is a step function; for continuous probability distributions, it may be a continuous function

Properties of CDF

1. $0 \leq F(x) \leq 1$
2. $F(x) \leq F(y)$ for $x < y$ ($F(x)$ is a non-decreasing function)
3. $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$
4. (If X only takes integer values) $f(x) = F(x) - F(x-1)$

Distributions

- Two random variables X and Y have the same distribution if their cumulative distribution function is the same for every input
  - This is denoted $X \sim Y$
  - Example: coin is heads and die roll is even

Discrete Uniform Distribution

- A random variable X takes values in $A = \{a, a+1, \dots b\}$ where each value is equally likely
  - Notation: $X \sim U(a, b)$
  - Ex. die roll is $U(1, 6)$, coin is $U(0, 1)$
- Since each value is equally likely, the probability of a single event happening is $\frac{1}{|A|} = \frac{1}{b-a+1}$
  - I.e. the probability function is constant at $\frac{1}{|A|}$

Hypergeometric Distribution

- A population consists of N objects of which r are "successful" and the remaining $N-r$ are "failures". Let a subset of n objects be chosen at random without replacement. The number of successes in that subset follows the **hypergeometric distribution**.
  - Notation: $X \sim hyp(N, r, n)$
  - Ex. the number of aces when 5 cards are drawn from a deck

- We have $f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, where $max\{0, n-(N-r)\} \leq x \leq \min\{r, n\}$

Binomial Distribution

- **Bernoulli trial**: experiment with probability of success p
- **Binomial distribution**: the resulting distribution of successes if n Bernoulli trials with success p
  - I.e. how many successes are observed
  - Notation: $A \sim Binomial(n, p)$
- There are two important assumptions about the binomial distribution
  - All trials are independent
  - All members of the sample share the same probability of success p
- Examples:
  - The number of heads if a coin is flipped 10 times
- We have $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, where $x \in \{0 \ldots n\}$

Binomial approximation of hypergeometric distribution

- If r and N are large, n is not too large, and $\frac{r}{N} = p \in [0, 1]$, then $X \sim hyp(N, r, n) \approx Binomial(n, p)$
  - This occurs because, for really large r and N compared to n, the draws are pretty close to independent (since the lack of replacement won't have a large effect)

Negative Binomial Distribution

- Consider an experiment in which Bernoulli trials are independently performed, each with probability of success p, until exactly k successes are observed. The number of failures before observing a success follows a **negative binomial distribution**
  - Notation: $X \sim NB(k, p)$
- Here, the number of trials is not fixed (compared to binomial distribution)
- We have $f(x) = \binom{x+k-1}{k-1} p^k (1-p)^x$ where $x \in \{0 \ldots n\}$

Geometric Distribution

- Special case of the negative binomial distribution where we stop after the first success
  - $X \sim NB(1, p)$
- We have $f(x) = (1-p)^x p$
- The number of trials is not fixed (unlike the binomial distribution)

- This distribution is **memoryless**:

Poisson Distribution

- Identity: $e^{-\mu}\frac{\mu^x}{x!} \approx \binom{n}{x}p^x(1-p)^{n-x}$ for large n and small p
- Poisson distribution: $f(x) = e^{-\mu}\dfrac{\mu^x}{x!}$ with parameter $\mu = np$
    - Note that this is a legit distribution since $\displaystyle\sum_{x=0}^{\infty} e^{-\mu}\frac{\mu^x}{x!} = 1$

Motivation

- Binomial distributions can be hard to estimate (before computers) because the $\binom{n}{x}$ terms could get really big with large sample spaces
- It was noted that as $n \to \infty$ and $p \to 0$ or $p \to 1$, $f(x) \to e^{-\mu}\dfrac{\mu^x}{x!}$ where $\mu = np$
    - Reminder, n is the number of trials and p is the probability of success

Poisson process

- Counting the number of occurrences of an event that happens at random points in time or space
- Three assumptions must be met
    - The events are independent
    - Events do not occur in clusters (i.e. chance of two events happening in small time step is near 0)
    - Events occur at uniform rate λ
        - $P(\text{event in } (t, t+\Delta t)) = \lambda\Delta t + o(\Delta t)$ as $\Delta t \to 0$
- If these are met, the setup is a poisson process

Order Notation (little o)

- A function $g(\Delta t)$ is $o(\Delta t)$ as $\Delta t \to 0$ if $\displaystyle\lim_{\Delta t \to 0} \frac{g(\Delta t)}{\Delta t} = 0$
- This is an error term that is negligible compared to Δt as it approaches 0
- Let X_t denote the number of events observed up to time t. If the above conditions are met, we have
    - $X \sim Poi(\lambda)$

- $F_t(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$

- Chapter 6 covers R, which is not taught in STAT 230

Summarizing Data on Random Variables

- Doesn't always make sense to present every datapoint that has been gathered since the amount of data can obfuscate any existing trends
- There are statistical strategies we can use to extract "key points"

Averages

- **Mean**: add all the data together and divide by the total number
- **Median**: sort the data in ascending order and pick the middle element
  - This is also known as the **expected value**
- **Mode**: the most commonly occurring item in the dataset

Expectation of a Random Variable

- Let X be a discrete random variable with range A and probability function $f(x)$ then $E(X)$ is the **expected value** of X, and is defined as $E[X] = \sum_{x \in A} x f(x)$
- This is the same as the mean of X since we have $\frac{x_1 + \cdots + x_n}{n} = \frac{k_1 \times x_1 + \cdots + k_r \times x_k}{n} = x_1 \frac{k_1}{n} + \cdots + x_r \frac{k_r}{n} = x_1 f(1) + \cdots + x_r f(r)$ where all $r \in A$
- We use the variable μ to refer to the mean

Winnings and net winnings

- Imagine you must pay x dollars to play a game with a winning of y dollars
  - Your **net winning** is $y - x$ dollars

Some expected value problems

- You play a game where you win y dollars by rolling y on a 6-sided die. If it costs 3 dollars to play, what are your expected net winnings?
  - We have $E(Y) = (-2)\frac{1}{6} + (-1)\frac{1}{6} + (0)\frac{1}{6} + (1)\frac{1}{6} + (2)\frac{1}{6} + (3)\frac{1}{6} = \frac{1}{2}$
  - So, the expected value of the game is 50 cents
  - If you play the game a large number of times (n), you can expect to net $\frac{n}{2}$ dollars
  - If the cost of the game were $3.50, the expect value would be 0 and you couldn't

expect to make any money in the long term

"Expected value theorem"

- Let X be a discrete random variable with $range(X) = A$ and the probability function $f(x)$
- The expected value of some function $G(x)$ on X is given by $E[g(X)] = \sum_{x \in A} f(x)g(x)$
    - Here, $g(x)$ is "what we get" when the event x occurs with respect to X (i.e. a particular value)
- Note that $E[g(x)] \neq g(E[x])$
    - Ex. the expected value of the square of a random dice roll is not the same as the square of the expected value of a random dice roll
- We have $E[ag(x) + b] = a \times E[g(x)] + b$ where a and b are constants
    - This can be shown with properties of sums and the expected value function
    - This means that the expected value function $E[X]$ is a *linear operator*
- We also have $E[g(x) + h(x)] = E[g(x)] + E[h(x)]$
    - The sum of expected values is the same as the expected value of the sum
    - This still holds if $g(x)$ is a constant function: $E[g + h(x)] = E[g] + E[h(x)] = g + E[h(x)]$
        - The expected value of a constant is simply that constant: $E[g] = g$

Means of Distributions

- Since distributions often describe real-world events, there is value in knowing the mean/expected value of various distributions

Binomial

- If $X \sim Binomial(n, p)$, then X has the probability function $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$
- So, the expected value is $E[X] = \sum_{n}^{x=0} xf(x) = \cdots = np$
    - So, for a binomial distribution, $E[X] = \mu = np$

Poisson

- If $X \sim Poi(\mu)$, then X has the probability function $f(x) = e^{-\mu} \frac{u^x}{x!}$, so $E[X] = \sum_{x=0}^{\infty} xe^{-\mu} \frac{u^x}{x!} = \mu$
- So, for a poisson distribution, $E[X] = \mu$

Hypergeometric

- If $X \sim hyp(N, r, n)$, then $E[X] = n\frac{r}{N}$

Negative Binomial

- If $X \sim NB(k, p)$, then $E[X] = \frac{k(1-p)}{p}$

Geometric

- If $X \sim Geo(p)$, then $E[X] = \frac{1-p}{p}$

Variance of distributions

- The expected value is our "best guess" of what the value will be, but there are many values around it that are also fairly likely
- How do we quantify how likely our expected value is?
  - I.e. the deviation from the mean

Absolute deviation

- The **mean absolute deviation** is equal to the sum of the absolute differences between each datapoint and μ
  - Formula: $\sum_{x \in S} |\mu - x|$
  - However, absolute values are hard to work with, especially when calculus is involved
- The **mean squared deviation** is equal to the sum over $(\mu - x)^2$
  - This makes all the terms positive while doing away with the absolute value
- Expected squared deviation: $E[(x-\mu)^2]$

Variance

- The **variance** of X is denoted $Var(x) = E[(X - E[X])^2] = E[(X-\mu)^2]$
  - A useful formulation: $Var(x) = E[X^2] - E[X]^2$
- Some properties of variance
  - For all random variables X, $Var(X) \geq 0$ (variance is never negative)
  - $Var(X) = 0 \iff P(X = E[X]) = 1$
  - $E[X^2] \geq E[X]^2$
  - Larger values of $Var(X)$ mean that the data is more spread out around the mean

Standard Deviation

- The **standard deviation** of a random variable X, denoted $SD(X)$, is defined by $SD(X) = \sqrt{Var(X)}$
  - This is often used instead of variance to measure variability

Variance of Linear Transformations

- Let $Y = aX + b$, where a and b are constants.
- We have $Var(Y) = a^2 Var(X)$
  - Adding a constant b does not affect how "spread out" the dataset is
  - a is squared because variance is measured in square units

Variance of common distributions

- Binomial: $X \sim Bin(n, p)$ has variance $Var(X) = np(1-p)$
- Poisson: $X \sim Poi(\mu)$ has variance $Var(x) = \lambda$
- Hypergeometric: $X \sim hyp(N, r, n)$ has variance $Var(X) = n\frac{r}{N}(1 - \frac{r}{N})\binom{N-n}{N-1}$
- Negative binomial: $X \sim NB(k, p)$ has variance $Var(X) = \frac{k(1-p)}{p^2}$
- Geometric: If $X \sim Geo(p)$ has variance $Var(X) = \frac{1-p}{p^2} = \frac{1}{p^2} - \frac{1}{p}$

Continuity

- So far, we have discussed discrete variables
- We will now be looking at continuous ranges
  - Ex. pick a random number in the range $[0, 1]$
- With continuous variables, we have a theoretically infinite degree of accuracy
- Because of this, the probability of any elementary event is 0

Terminology and Notation

- A random variable X is said to be continuous if its range $X(S)$ is an interval $(a, b) \in \mathbb{R}$
  - X can take any value between a and b
- Examples
  - Time
  - Distance
  - (We don't consider the Planck length or time)
- We can no longer use the axiom of probability $\sum_{x \in [0,1]} f(x) = 1$; must update it to a continuous x by using the integral: $\int_0^1 f(x)\,dx = 1$
- Again, we cannot have a probability function (since $\forall x \in \mathbb{R}, f(x) = 0$), we must use a probability density function

Probability density function

- A probability density function has the following properties
  1. $f(x) \geq 0$
  2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
  3. $P(a \leq X \leq b) = \int_a^b f(x)\,dx$
     - I.e. the probability that the function is in the range $[a, b]$
- The probability density function (PDF) is *not* a probability function, but it can be used to gain information about probabilities
- Ex. A spinner is spun and lands at an angle θ. We can define its PDF as $f(x) = \begin{cases} 0.25 & 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$. Note that this satisfies $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- **Support** of a PDF $f(x)$ is defined as $supp(f) = \{x \in \mathbb{R} : f(x) \neq 0\}$
  - All values of x such that $f(x)$ is not 0

    - This is essentially a lower and upper bound on $f(x)$ that lets you avoid integrating between ∞ and $-\infty$ each time an end is unbounded
- Note that our previous assertion about $f(x)$ being 0 is correct: $P(x=a) = P(a \leq X \leq a) = \int_a^a f(x)\,dx = 0$ by the properties of integrals

Cumulative density function (continuous version)

- For discrete functions, we have $F(x) = P(X \leq x)$
- For continuous functions, the CDF is defined as $F(x) = \int_{-\infty}^{x} f(u)\,du$
- By the fundamental theorem of calculus, we have $\frac{d}{dx}F(x) = f(x)$, where $f(x)$ is continuous
- The CDF more useful and less difficult to work with because it lets us easily find the probabilities of ranges: $P(a \leq X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_a^b f(x)\,dx$
    - This integration is done anyway when calculating the same probability using the PDF

Approaches for finding $F(X)$ from $f(x)$

- Treat each piece of $f(x)$ separately (i.e. use a step function)
- Note that $F(X) = 0$ for every $X < \text{minimum value in the support of } f(x)$
- Note that $F(X) = 1$ for every $X > \text{maximum value in the support of } f(x)$
- For the middle, find $F(x) = \int_{-\infty}^{x} f(u)\,du$

Percentiles and Quantiles

- Let X be a continuous random variable with CDF $F(x)$
- The p^{th} **quantile** of the of X is the value $q(p)$ such that $P(X \leq q(p)) = p$
    - I.e. the probability that X is less than or equal to $q(p)$
    - Ex. If p is 0.9, $q(p)$ is the value where 90% of the possible values of X are below it
    - The value $q(p)$ is the 100th percentile of the distribution
    - If p is 0.5, then $q(0.5)$ is the median of X
        - In fact, it makes sense to think about quantiles as an extension of the median, where we look at "partitions" other than the lower half and other half
- We can find a quantile by solving $F(X) = p$, which leads to $x = q(p)$

Change of Variables

- What if we wanted to find the CDF or PDF of some function $g(x)$ of x
  - We "solve" the whole equation in order to write it in terms of x
  - Ex. $P(\frac{1}{x} \leq y) \rightarrow P(x \geq \frac{1}{y}) \rightarrow 1 - P(x < \frac{1}{y}) \rightarrow 1 - F_X(\frac{1}{y})$
- Algorithm for change of variables (where $y = g(x)$)
  1. Write the CDF of Y as a function of X
     - $F_Y(y) \rightarrow P(Y \leq y) \rightarrow P(g(X) \leq y)$
  2. Use $F_X(x)$ to find $F_Y(y)$. We can differentiate this if we want to find the PDF of Y, $F_Y(y)$
  3. Find the range of values of y

Expectation and Variance

- For discrete random variables, the expectation is $E[g(x)] = \sum_{x \in \mathbb{Z}} g(x)f(x)$
- Similarly, for continuous random variables, we have $E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$
  - When $g(x) = x$, we have $E[x] = \int_{-\infty}^{\infty} xf(x)\,dx$
- Thus, we have $Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\,dx$
- We still have the shortcut for computing variance: $Var(x) = E[X^2] - E[X]^2$
- Sometimes, σ^2 is used to represent $Var(x)$, so we have $\sigma^2 = E[X^2] - \mu$

Distributions for continuous variables

Continuous uniform distribution

- X has a **continuous uniform distribution** on (a, b) if any interval of the same fixed length has the same probability
  - Since it is a continuous distribution, any specific *number* still has probability 0
  - X has the PDF $f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{otherwise} \end{cases}$
  - X has the CDF $f(x) = \begin{cases} 0 & x < a \\ \int_a^x \frac{1}{b-a}\,du = \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$
- This is denoted $X \sim U(a, b)$
- We have $E[X] = \frac{a+b}{2}$ and $Var(x) = \frac{(b-a)^2}{12}$

Exponential Distribution (aka. power law)

- $X \sim exp(\theta = \frac{1}{\lambda})$ is defined by its PDF
  - X has the CDF $F(x) = 1 - e^{-\lambda x} = 1 - e^{-\frac{x}{\theta}}$
  - X has the PDF $f(x) = \lambda e^{-\lambda x} = \frac{e^{-\frac{x}{\theta}}}{\theta} = \frac{1}{\theta} e^{-\frac{x}{\theta}}$ for $x > 0$, 0 otherwise
  - Here, λ is the same rate as in the poisson process
- We have $E[X] = \theta$ and $Var(x) = \theta^2$, which can be found using the gamma function
- Motivation
  - Imagine a situation where cars passing an intersection follow a poisson process. What is the distribution of the time until the first car passes?
  - We have the CDF as $F(x) = P(X \leq x) = 1 - P(X > x) = 1 - P(\text{no event occurs in } (0, x)) = 1 - P(Y_x = 0)$, where $Y_x \sim Poi(\lambda t)$. This leads to our CDF
- Alternate parameterization: $\theta = \frac{1}{\lambda}$ is the scale parameter
  - So the CDF is $F(x) = 1 - e^{-\frac{x}{\theta}}$ and the PDF is $F(x) = \frac{e^{-\frac{x}{\theta}}}{\theta}$
- **Survivor function**: The complement of the CDF $e^{-\frac{x}{\theta}}$ is an often used form
- **Memoryless property**: The amount of events that have already happened is not present
  - The geometric distribution is also memoryless
- Continuous analog to the geometric distribution

Gamma Function

- The **gamma function** $\Gamma(x)$ is defined as $\Gamma(\alpha) = \int_0^\infty y^{\alpha - 1} e^{-y} \, dy = (\alpha - 1)!$ as defined for $\mathbb{N}$ for all $\alpha \geq 0 \in \mathbb{R}$
- We can use its properties to solve some integrals related to probabilities
  - Example: expected value of exponential distribution

Example

- A battery range of a car is on average 15000km with an exponential distribution. What is the chance that a 1000km trip can be completed without needing a battery replacement
  - So, $P(X > 1000)$ where $X \sim exp(15000)$
  - $P(X > x) = 1 - e^{-\frac{x}{\theta}} \implies P(X \leq x) = e^{-\frac{x}{\theta}}$
  - So, answer is $e^{-\frac{1000}{15000}} = 0.936\ldots$

Computer-Generated Random Numbers

- Computers can generate pseudo-random numbers: $U \sim U(0, 1)$
  - We can simulate any distribution using this

- Let $F^{-1}(x)$ denote the inverse CDF defined on $(0,1)$, where F and F^{-1} are continuous
    - $F^{-1}(U=(0,1))$ has the same distribution (and thus CDF, PDF) as X, namely $F(x)$
- What if F is not continuous:
    - We can used a *generalized inverse*: $F^{-1}(u) = \inf\{x, F(x) \geq u\}$
    - Here, infimum $(\inf)$ is the highest lower bound of the set
- **Inverse Transform Sampling Theorem**: If $U \sim U(0,1)$, then the random variable X defined by the transformation $X = F^{-1}(U)$ has cumulative distribution function $F(x)$

Normal Distribution

- X follows a **normal distribution** (aka a **gaussian distribution**) with a mean μ and variance σ^2 if the PDF of X is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ where $x \in \mathbb{R}$
    - We say $X \sim N(\mu, \sigma^2)$ or $X \sim G(\mu, \sigma)$
- **Standard normal** distribution $N(0,1)$ is often used
    - This has PDF $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and CDF $\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}\, dy$
    - A distribution $X \sim N(\mu, \sigma^2)$ can be normalized as such: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- Properties of the normal distribution
    1. Symmetric about the mean μ: $P(X \leq \mu - t) = P(X \geq \mu + t)$ where $t \in \mathbb{R}$
    2. There is a single peak at μ
    3. Parameters are the mean μ and variance σ
- Describes many natural phenomena (and useful for generative art); possibly the most important distribution
- Continuous analog to the binomial distribution

- So far, we've only used **univariate distributions**: distributions that measure one variable
- **Multivariate distributions** are measurements of *multiple random variables* or *repeated measurements of the same quantity*

Joint Probability

- Let X and Y be discrete random variables with the same sample space
    - X and Y don't necessarily have to have the same range
- The **joint probability function** of X and Y is $f(x, y) = P(X = x, Y = y) = P(\{X = x\} \cap \{Y = y\})$, $x \in X(S), y \in Y(S)$
    - Ex: the joint probability of two die rolls in succession X and Y is $\frac{1}{36}$

Properties of Joint Probability

- Properties of multivariate probability functions (same as univariate ones)
    1. $0 \leq f(x, y) \leq 1$
    2. $\sum_{x,y} f(x, y) = 1$
- Like always, computing joint probability involves adding up all the possible outcomes
    - $P((X, Y) \in A) = \sum_{(x,y) \in A} f(x, y)$

Marginal Probability

- Let discrete X, Y have probability function $f(x, y)$. The **marginal probability function** of X is $f_X(x) = P(X = x) = \sum_{y \in Y(S)} f(x, y)$. We also have $f_Y(y) = P(Y = y) = \sum_{x \in X(S)} f(x, y)$

Independence of joint probabilities

- Discrete X and Y with probability function $f(x, y)$ and marginal probability functions $f_X(x)$ and $f_Y(y)$ are **independent** iff $f(x, y) = f_X(x) \times f_Y(y)$ for all $x \in X(S)$ and $y \in Y(S)$
    - Alternate formulation: $P(X = x, Y = y) = P(X = x) \times P(Y = y)$ for all x and y
- Extension to more variables: discrete $X_1, X_2, \ldots X_n$ with probability function $f(x_1, x_2, \ldots x_n)$ are independent iff $f(x_1, x_2, \ldots x_n) = f(x_1) \times f(x_2) \times \cdots \times f(x_n) = \prod_{k=1}^{n} f(x_k)$

Conditional joint probability

- The **conditional probability function** of X given $Y = y$ is denoted $f_{x|y}(x \mid y) = P(X = x \mid Y = y) = \dfrac{P(X = x, Y = y)}{P(Y = y)} = \dfrac{f(x, y)}{f_Y(y)}$
  - $f_{y|x}(y \mid x)$ can be similarly defined

Multinomial Distribution

- Sometimes experiments have more than 2 outcomes
  - Ex. roulette game ($\frac{9}{19}$ of red winning, $\frac{9}{19}$ of black winning, $\frac{1}{19}$ of house winning)
- Example question: what is the chance of a sequence of roulette games being `RRRRBBBBHH` ?
  - Answer: $\frac{10!}{4!4!2!}\left(\frac{18}{38}\right)^4\left(\frac{18}{38}\right)^4\left(\frac{2}{38}\right)^2$
- Properties of multinomial distribution with parameters k and $p_1 \dots p_k$
  1. Individual trials are independent and have k possible outcomes where the sum of each trial's probability is 1 ($p_1 + \cdots + p_k = 1$)
  2. There are n trials and the total number of outcomes is $n = X_1 + \dots X_k$, where each X_i a possible outcome with probability p_i
- We have $f(x_1, \dots x_k) = \dfrac{n!}{x_1! \times x_2! \times \cdots \times x_k!}(p_1^{x_1} \times p_2^{x_2} \times \cdots \times p_k^{x_k}) = \dfrac{n!}{x_1! \times x_2! \times \cdots \times x_k!}\prod_{i=0}^{k} p_i^{x_i}$
- We also have $X_i \sim Binomial(n, p_i)$ and $X_i + X_k \sim Binomial(n, p_i + p_k)$
  - I.e. their marginal distribution is a binomial distribution
- Example: A bag contains 5 red, green marbles and 10 blue marbles. 6 are drawn from the bag with replacement. What is the probability that we draw two of each marble?

Expected Value

- Let X and Y be jointly distributed random variables with probability function $f(x, y)$. Then for some $g : \mathbb{R}^2 \to \mathbb{R}$, $E[g(X, Y)] = \sum_{(x,y)} g(x, y)f(x, y)$
  - This applies to the general case: $E[g(X_1, \dots, X_n)] = \sum_{(x_1,\dots,x_n)} g(x_1, \dots, x_n)f(x_1, \dots, x_n)$
- $E[X]$ is still linear: $E[X + Y] = E[X] + E[Y]$

Covariance

- If X and Y are jointly distributed, then $Cov(X, Y)$ denotes the **covariance** between X and Y.
- It is defined by $Cov(X, Y) = E[(X - E[X]) \times (Y - E[Y])]$

    - Easier formula: $Cov(X,Y) = E[XY] - E[X] \times E[Y]$
- Covariance is
    - *Positive* if Y increases as X increases
    - *Negative* if Y decreases as X increases
- The larger the number, the stronger the relationship between the two variables is
- If X and Y are independent, then $Cov(X,Y) = 0$ since in this case $E[XY] = E[X] \times E[Y]$
    - However, a covariance of 0 does not necessarily imply independence

Variance and Covariance Identities

- $Cov(X,X) = Var(x)$
- $Var(aX + bY) = a^2Var(X) + 2abCov(X,Y) + b^2Var(Y)$
- If X and Y are independent
    - $Cov(X,Y) = 0$, so
    - $Var(X+Y) = Var(X-Y) = Var(X) + Var(Y)$
    - $Var(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} {a_i}^2{\sigma_i}^2$

Correlation

- The **correlation** of X and Y is denoted $corr(X,Y)$ and is defined by $\rho = \dfrac{Cov(X,Y)}{SD(X) \times SD(Y)}$, where $-1 \leq \rho \leq 1$
- Correlation measures the strength of the *linear* relationship between X and Y
- The linear relationship is
    - *Positive* if $\rho \approx 1$
    - *Negative* if $\rho \approx -1$
    - *Nonexistent* if $\rho \approx 0$. This doesn't mean there is no relationship between X and Y, just not a *linear* one
- Essentially a normalized version of the covariance

Linear Combinations of Random Variables

- Let $X_1 \ldots X_n$ be jointly distributed random variables with joint probability function $f(x_1, \ldots, x_n)$.
- **Linear Combination**: $a_1X_1 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$ where $a_1, \ldots, a_n \in \mathbb{R}$

Common Linear Combinations

- Total: $T = \sum_{i=1}^{n} X_i$, where a_i is 1
- Sample mean: $\bar{X} = \sum_{i=1}^{n} \frac{1}{n} X_i$, where $a_i = \frac{1}{n}$
  - We also have $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$
  - So, variability decreases with the number of samples taken into account
  - Finally, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Expected value: $E\left[\sum_{i=1}^{n} a_i\right] = \sum_{i=1}^{n} a_i E[X_i]$

Linear Combinations of Normally Distributed Variables

- Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$. Then, $Y \sim N(a\mu + b, a^2\sigma^2)$
- If we have $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, then $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$
  - I.e. the linear combination of independent, normally distributed random variables is also a normal distribution
- General case: $\sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$
  - When $a_i = 1$, we have $\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$

Indicator Random Variables

- Let $A \subset S$ be an event. $\mathbb{1}_A$ is the **indicator random variable** of the event A and is defined by $\mathbb{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \in \bar{A} \iff \omega \notin A \end{cases}$
  - These are also called *Bernoulli Random Variables*
- We have $E[\mathbb{1}_A] = P(A)$ and $Var(\mathbb{1}_A) = P(A)(1 - P(A))$
- We have $Cov(\mathbb{1}_A, \mathbb{1}_B) = P(A \cap B) - P(A)P(B)$

Solving complex sounding problems

- The weights of male and female geese follow the normal distributions M and F respectively. What is the probability that the female goose is heavier if two geese are selected at random?
  - We want $P(F > M) = P(F - M > 0)$

- $F - M$ is a linear combination, so we can find its distribution, then calculate the probability

Central Limit Theorem

- If $X_1 \dots X_n$ are independent random variables from the same distribution, with mean μ and variance σ^2, then as $n \to \infty$, then the distribution of $S_n = \sum_{i=1}^{n} X_i$ approaches the shape of the probability density function of $N(n\mu, n\sigma^2)$
  - Consequence: $\bar{X}$ approaches $N(\mu, \sigma^2)$
  - If $X_1 \dots X_n$ are normal with $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

Central Limit Theorem

If $X_1 \dots X_n$ are independent random variables with mean μ and variance σ^2, then as $n \to \infty$, the CDF of the sum $X_1 + \dots + X_n$ approaches $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$

Steps in a central limit problem

1. Verify the required assumptions
   - Random variables are independent
   - Random variables have the same mean and variance
2. Identify the mean μ and variance σ^2
3. Apply the CLT and probability rules to obtain the solution

Guidelines for using the CLT

- Regular distribution: more than 30+ observations
- Close to unimodal, relatively symmetric, close to being continuous: 5-15+ observations
- Highly Skewed, very discrete: 50+ observations

Normal approximation to binomial

- If $X \sim Binomial(n, p)$, then for large n, the random variable $W = \frac{X - np}{\sqrt{np(1-p)}}$ has approximately a $N(0, 1)$ distribution.

Continuity Correction

- When using the normal approximation to the binomial, we are approximating a discrete distribution with a continuous one using the CLT
- This leads to an error factor because an extra discrete "bucket" is counted when using an inequality (for the continuous one, the range just stops there)
  - I.e. we have to account for the "buckets"
- To correct it, add and subtract 0.5 from both the right and left of an inequality respectively:
  - For $P(a \leq X \leq b)$, compute $P(a - \frac{1}{2} \leq X \leq b + \frac{1}{2})$
  - For $P(X < b)$, compute $P(X < b - \frac{1}{2})$
  - For $P(X = x)$, compute $P(x - \frac{1}{2} \leq X \leq x + \frac{1}{2})$

Normal Approximation to Poisson

- If $X \sim Poisson(\mu)$, then the CDF of $Z = \frac{X - \mu}{\sqrt{\mu}}$ approaches that of $N(0, 1)$ as $n \to \infty$
  - Motivation: $P(X > \mu)$ (the CDF) using the normal approximation is $P(X > \mu) = P(\frac{X - \mu}{\sqrt{\mu}} > \frac{\mu - \mu}{\sqrt{\mu}}) = P(Z > 0)$; since $Z \sim N(0, 1)$, this is the approximation
- We still have to remember to use the continuity correction

Moment Generating Functions

- In addition to the PDF and CDF of a distribution, there is a third function that uniquely determines a distribution: the **moment generating function**
- The MGF is given by $M_X(t) = E[e^{tX}]$, $t \in \mathbb{R}$
  - If X is discrete with PF $f(x)$, then we have $M_X(t) = \sum_{x=0}^{\infty} e^{tx} f(x)$, $t \in \mathbb{R}$
  - If X is continuous with PDF $f(x)$, then we have $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$, $t \in \mathbb{R}$

Properties of MGF

1. $M_X(t) = 1 + tE[X] + \frac{t^2 E[X^2]}{2!} + \frac{t^3 E[X^3]}{3!} + \ldots$
   - I.e. there is a Taylor-ish expansion
2. If $M_X(t)$ is defined in the neighborhood of $t = 0$, then $\frac{d}{dt^k} M_X(0) = E[X^k]$
   - I.e. we can find the kth moment of the distribution by taking the derivative

Uniqueness Theorem

- Let X and Y have MGFs $M_X(t)$ and $M_Y(t)$. If $M_X(t) = M_Y(t)$ for all t, then X and Y have the same distribution

MGFs of common distributions

Name	Dist.	Moment Generating Function
Normal	$X \sim N(\mu, \sigma^2)$	$M_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}$
Poisson	$X \sim Poi(\lambda)$	$M_X(t) = e^{\lambda(e^t - 1)}$
Binomial	$X \sim Binomial(n, p)$	$(1 - p + pe^t)^n$

Multivariate Moment Generating Functions

- Let X and Y be independent distributions with moment generating functions $M_X(t)$ and $M_Y(t)$. Then, the MGF of $X + Y$ is $M_X(t) \times M_Y(t)$
  - This follows from the properties of exponents
  - This generalizes to more than 2 distributions
- This result can be used to prove the central limit theorem!