Statistics 110: Introduction to Probability Fall 2022

Final Review

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

Final Preparation Tips

Just like our midterm review guide, the aim of this final review guide is to provide general problem-solving strategies and help you recognize common problem types that will be on the final exam. 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, please do not stress too much. We've all seen how hard you've worked this semester. Hard work *will* get paid off!**

Below are some useful study action items, listed in a suggested order (your specific workflow may vary):

1. Read chapters 1-12 of *Introduction to Probability* (chapters 5-12 cover post-midterm material).
2. Create personal review sheets (yes, there are some cheat sheets already on Canvas, but writing concepts down yourself is another effective sweep of the information).
3. Work through as many practice finals 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

During the course of today's final review session, if you have *any or all questions* on specific concept clarification, logistics, or would like us to work out a particular practice problem or past HW problem, please send them to our Google Form here: http://bit.ly/CatSkyler110FinalReview.

Our Contact Information

This review session will be our last formal teaching activity of the semester. However, please do not hesitate to reach out to us via email or text for anything (legal):

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

Phone numbers: Feel free to text us!

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

Final Remarks

Thank you everyone for such an amazing semester! TFing was undoubtedly one of the highlights of our semester, and we've truly learned and grown so much from serving as your TFs. We hope you found our teaching at least somewhat marginally helpful towards your statistical studies. With that said, it was an honor to serve as your TFs. Godspeed and best of luck on the final exam and your educational journey. Y'all are gonna do fire.

Skyler and Cat, signing out.

Contents

1 Concept Review

1.1 Introduction to Continuous Random Variables

1.1.1 General Principles of Continuous Random Variables

We present this subsection in a Q&A format for readability and understanding:

1. **What is a Continuous Random Variable?** A continuous random variable can take on any possible value within a certain interval (for example, [0, 1]), whereas a discrete random variable can only take on variables in a list of countable values (for example, all the integers, or the values 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, etc.). Formally, a random variable is continuous if its CDF if *differentiable*. However, it's OK if the CDF is continuous but not differentiable at a finite number of points, like endpoints!

2. **Do Continuous Random Variables have PMFs?** No. **If X is a continuous random variable, then $P(X = x) = 0$ for any value x! Continuous random variables are continuous intervals, meaning the probability that the r.v. crystallizes to, say, exactly 1.00000...is *infinitesimally* small.** You may be asked to find whether a complex-looking random variable is discrete or continuous, and it might not seem obvious at first glance. This fact may be useful!

   Instead of PMFs (**bad!**), we can describe a continuous r.v. using its **PDF** (probability density function) — see below.

3. **How to I find the probability that a continuous random variable takes on a value within an interval?** Use the CDF (or the PDF, see below). To find the probability that a continuous random variable takes on a value in the interval $[a, b]$, subtract the respective CDFs.

   $$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

   What is the Cumulative Density Function (CDF)? It is the following function of x.

   $$F(x) = P(X \leq x)$$

   With the following properties.

   1) F is increasing.
   2) F is right-continuous.
   3) $F(x) \to 1$ as $x \to \infty$, $F(x) \to 0$ as $x \to -\infty$.
   4) F is differentiable. (mentioned earlier)

4. **What is the Probability Density Function (PDF)?** For continuous r.v. X with CDF F, the PDF, $f(x)$, is the derivative of the CDF, given by

   $$f(x) = F'(x).$$

   We can go from PDF to CDF by integrating. Notice the limits of integration!

   $$F(x) = \int_{-\infty}^{x} f(t)dt.$$

   Note that by the fundamental theorem of calculus,

   $$F(b) - F(a) = \int_{a}^{b} f(x)dx.$$

   Thus to find the probability that a continuous random variable takes on a value in an interval, you can integrate the PDF, thus finding the area under the density curve.

   The **support** of X is the set of all $x \in \mathbb{R}$ where $f(x) > 0$. You might be wondering, *didn't see just say earlier that the probability that X takes on any specific value is 0?* That is still true — the quantity $f(x)$ is *not* a probability, and we can even have $f(x) > 1$ for some values of x. **Watch out for category errors here!**

   Two additional properties of a PDF:

1) It must integrate to 1: $\int_{-\infty}^{\infty} f(x) = dx = 1$.

2) The PDF must always be nonnegative. $f(x) \geq 0$.

5. **How do I find the expected value of a continuous random variable?** Where in discrete cases you sum over the probabilities, in continuous cases you integrate over the densities.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The integral is taken over the entire real line, but if the support of X is not the entire real line, we can just integrate over the support.

6. **Review**: Expected value is *linear.* This means that for *any* random variables X and Y and any constants a, b, c, the following is true:

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

1.1.2 Discrete versus Continuous

	Discrete	**Continuous**
$P(X \leq x) =$	$F(x)$ (CDF)	$F(x)$ (CDF)
To find probabilities,	Add over PMF P(X = x)	Integrate over PDF f(x) = F'(x)
$E(X) =$	$\sum_x xP(X = x)$	$\int_{-\infty}^{\infty} xf(x)dx$
$E(g(X)) =$	$\sum_x g(x)P(X = x)$ (LOTUS Discrete)	$\int_{-\infty}^{\infty} g(x)f(x)dx$ (LOTUS Continuous)

1.1.3 Law of the Unconscious Statistician (LOTUS)

1. **How do I find the expected value of a function of a random variable?** Normally, you would find the expected value of X this way:

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

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

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

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

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

2. **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.

3. **What's the point?** 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.

1.1.4 Variance, Expectation and Independence, and e^x Taylor Series

1. The following Taylor series will be very helpful when working with Poisson distributions (introduced soon below):
$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$
2. Recall that variance is defined in the following manner:
$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2$$
3. If X and Y are independent, then
$$E(XY) = E(X)E(Y)$$

1.2 Descriptions of Named Continuous Random Variables (and Poisson)

1.2.1 Poisson Distribution (Discrete)

We include Poisson in this section because it was only touched upon right before the midterm exam, and because of its connections to continuous random variables such as the Gamma and Exponential distributions. Let us say that X is distributed $\mathrm{Pois}(\lambda)$. We know the following:

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. **PMF:** The PMF of $X \sim \mathrm{Pois}(\lambda)$ is:
$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$
The support of X is $\{0, 1, \ldots\}$.
4. **Expectation:**
$$E(X) = \lambda.$$
5. **Variance:**
$$\mathrm{Var}(X) = \lambda.$$
6. **Chicken-Egg Story:** Say that a chicken lays $N \sim \mathrm{Pois}(\lambda)$ 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:
   (a) $X \sim \mathrm{Pois}(\lambda p)$, and $Y \sim \mathrm{Pois}(\lambda q)$
   (b) X and Y are independent.
7. **Useful Properties:**
   (a) **Sum of Independent Poissons:** If $X \sim \mathrm{Pois}(\lambda_1)$ and $Y \sim \mathrm{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$, then $X + Y \sim \mathrm{Pois}(\lambda_1 + \lambda_2)$.
   (b) **Poisson given a sum of Poissons is Binomial** If $X \sim \mathrm{Pois}(\lambda_1)$ and $Y \sim \mathrm{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$, then the conditional distribution of X given $X + Y = n$ is
$$\mathrm{Bin}\Big(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\Big).$$
   (c) **Binomial taken to its limits is approximately Poisson:** Let $X \sim \mathrm{Bin}(n, p)$. Let $n \to \infty$ and $p \to 0$, with $\lambda = np$ staying fixed. Then, the PMF of X converges to the $\mathrm{Pois}(\lambda)$ PMF.

1.2.2 Uniform Distribution (Continuous)

Let us say that U is distributed $\text{Unif}(a,b)$. We know the following:

1. **Intuitive Definition:** U is just a completely random number between a and b: mathematically, this means that the PDF of U is *constant* over the interval (a,b). So, when you integrate over the PDF, you will get an area proportional to the length of the interval.

2. **Probability Proportional to Length:** For a uniform distribution, the probability of an draw from any interval on the uniform is *proportion to the length of the uniform.*

   Mathematically, let $U \sim \text{Unif}(a,b)$. Let (c,d) be a subinterval of (a,b) with $d-c=l$. Then, we have
   $$P(c < U < d) \propto d-c = l.$$
   Specifically, we also know that
   $$P(c < U < d) = \frac{d-c}{b-a} = \frac{l}{b-a}.$$

3. **Conditional Distribution:** Let $U \sim \text{Unif}(a,b)$ and let (c,d) be a subinterval of (a,b). Then, the conditional distribution (more on that later) of U given $U \in (c,d)$ is simply $\text{Unif}(c,d)$.

4. **"Standard Uniform":** If U is "standard uniform," this just means that $U \sim \text{Unif}(0,1)$.

5. **Example:** Bob throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. Bob's darts have a uniform distribution on the surface of the room. The uniform is the only distribution where the probably of hitting in any specific region is proportion to the area/length/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

6. **PDF and CDF:**

   $\text{Unif}(0,1)$ $$f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & x \notin [0,1] \end{cases} \qquad F(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x > 1 \end{cases}$$

   $\text{Unif}(a,b)$ $$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases} \qquad F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

7. **Expectation:**
   If $U \sim \text{Unif}(a,b)$, then
   $$E(U) = \frac{a+b}{2}$$

8. **Variance:**
   $$\text{Var}(U) = \frac{(b-a)^2}{12}$$

9. **Universality of Uniform** In words, when you plug any random variable into its own CDF, F, you get a $\text{Unif}(0,1)$ random variable. *For any continuous random variable X, we can transform it into a uniform random variable and back by using its CDF.*

   Let X be a continuous random variable, with CDF F. Because F is continuous and strictly increasing on X's support, $F^{-1} : (0,1) \to \mathbb{R}$ exists.

   We have the following results, which we pack into a property called the *Universality of the Uniform*:

   1) Let $U \sim \text{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X has CDF F.

- This says that if we start with $U \sim \text{Unif}(0,1)$ and a CDF F, then we can create a random variable X whose CDF is F, by plugging U into the inverse function F^{-1}.
- Fundamentally, $F^{-1}(U)$ is a function of a random variable, and recall that functions of random variables are random variables too!
- Universality of the Uniform just says this $F^{-1}(U)$ random variable has CDF F.

2) Let X be a random variable with CDF F. Then $F(X) \sim \text{Unif}(0,1)$.

- For this part, we start with a random variable X with CDF F, and our goal is to use these "resources" to *construct* a Unif$(0,1)$ random variable.
- Universality of the Uniform says we can do this by plugging random variable X into its *own* CDF F.

Example Let's say that a random variable X has a CDF

$$F(x) = 1 - e^{-x}.$$

By the Universality of the the Uniform, if we plug in X into this function F, then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0,1).$$

Similarly, since $F(X) \sim U$ then $X \sim F^{-1}(U)$. The key point is that **for any continuous random variable X, we can transform it into a uniform random variable and back by using its CDF.**

1.2.3 Normal Distribution (Continuous) (a.k.a. Gaussian)

The Normal distribution is a famous continuous distribution with a bell-shaped PDF. Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

1. **PDF:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

2. **CDF:** It's too difficult to write this one out. We express the CDF of the Standard Normal Z as the function $\Phi(x)$.

3. **Standard Normal:** The Standard Normal, denoted Z, is $Z \sim \mathcal{N}(0,1)$.

4. **Short-Hand Expressions:** The $N(0,1)$ PDF and CDF are typically quite unwieldy, so we define the following:

   (a) Let $\varphi(z)$ represent the PDF of a Standard Normal (i.e., $\mathcal{N}(0,1)$) random variable.

   (b) Let $\Phi(z)$ represent the CDF of a Standard Normal random variable.

5. **Expectation:** For $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$E(X) = \mu$$

6. **Variance:**

$$\text{Var}(X) = \sigma^2$$

   Note that the first and second parameters of a Normal random variable are the expectation and variance of that variable, respectively!

7. **Symmetry:** Here are some helpful symmetry properties of the Standard Normal PDF and CDF. Again, these only apply to the PDF and CDF of the *Standard Normal* distribution, Did we mention that these properties only apply for *Standard Normal* random variables!

   (a) Symmetry of PDF: $\varphi(z) = \varphi(-z)$ (i.e., φ is an "even" function)

(b) Symmetry of Tail Areas: $\Phi(z) = 1 - \Phi(-z)$

(c) Symmetry of $Z \sim \mathcal{N}(0,1)$: $P(-Z \leq z) = P(Z \geq -z) = 1 - \Phi(-z)$

8. **Scaling and Transformation Properties** In general, intuitively, every time we stretch or scale a Normal random variable, we end up changing it into another Normal random variable:

   (a) If we add c to a Normal random variable, then its mean increases additively by c.

   (b) If we multiply a Normal random variable by c, then its variance increases multiplicatively by c^2.

   Formally, we present the following concepts:

   (a) **Definition:** Let $Z \sim \mathcal{N}(0,1)$. If $X = \mu + \sigma Z$, then we say that X is Normal, with mean μ and variance σ^2. Note that σ must be > 0, else you would have a degenerate random variable. Symbolically, we write:

   $$X \sim \mathcal{N}(\mu, \sigma^2)$$

   (b) **Standardization:** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. The *random variable* $\frac{X-\mu}{\sigma}$ is called the "standardized version" of X.

   (c) **PDF and CDF of a Normal Random Variable:** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $f(x)$ and $F(x)$ be the PDF and CDF of X, respectively. Then, we have the following:

   $$f(x) = \varphi(\frac{x-\mu}{\sigma})\frac{1}{\sigma}$$

   $$F(x) = \Phi(\frac{x-\mu}{\sigma})$$

9. **Sum of independent Normals is Normal:** If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent Normal r.v.'s, then means and variances are both additive. Formally,

   $$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

   Furthermore,

   $$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2).$$

1.2.4 Exponential Distribution (Continuous)

Let us say that X is distributed Expo(λ). We know the following:

1. **Story:** You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but it's never true that a shooting star is ever "due" to come because you've waited so long. Your waiting time is memorylessness, which means that the time until the next shooting star comes does not depend on how long you've waited already.

2. **Example:** The waiting time until the next shooting star is distributed Expo(4). The 4 here is λ, or the rate parameter, or how many shooting stars we expect to see in a unit of time. The expected time until the next shooting star is $\frac{1}{\lambda}$, or $\frac{1}{4}$ of an hour. You can expect to wait 15 minutes until the next shooting star.

3. **All Exponentials are Scaled Versions of Each Other:** We can use scaling to get from the simple Expo(1) to the general Expo(λ) : if $X \sim$ Expo(1), then

   $$Y = \frac{X}{\lambda} \sim \text{Expo}(\lambda).$$

   Conversely,

   $$Y \sim \text{Expo}(\lambda) \implies X = \lambda Y \sim \text{Expo}(1)$$

4. **PDF and CDF:** The PDF and CDF of a Exponential is:

$$f(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty) \qquad\qquad F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \in [0, \infty)$$

5. **Expectation and Variance:** If $X \sim \text{Expo}(1)$, we can obtain $E(X)$ and $\text{Var}(X)$ through standard integration: $E(X) = 1$, and $\text{Var}(X) = 1$. For general $Y = \frac{X}{\lambda} \sim \text{Expo}(\lambda)$, we have

$$E(Y) = \frac{1}{\lambda}E(X) = \frac{1}{\lambda},$$

$$\text{Var}(Y) = \frac{1}{\lambda^2}\text{Var}(X) = \frac{1}{\lambda^2}.$$

6. **Memorylessness:** The Exponential Distribution is the sole continuous memoryless distribution. This means that it's always "as good as new", which means that the probability of it failing in the next infinitesimal time period is the same as any infinitesimal time period. This means that for an exponentially distributed X and any real numbers t and s,

$$P(X > s + t | X > s) = P(X > t).$$

Given that you've waited already at least s minutes, the probability of having to wait an additional t minutes is the same as the probability that you have to wait more than t minutes to begin with. Here's another formulation.

$$X - s | X > s \sim \text{Expo}(\lambda).$$

Conditional on $X > s$, the additional waiting time $X - s$ is still distributed $\text{Expo}(\lambda)$. This further implies that

$$E(X|X > s) = s + E(X) = s + \frac{1}{\lambda}.$$

1.2.5 Beta Distribution (Continuous)

1. **Story:** The Beta distribution is continuous on the internal $(0, 1)$ and is a generalization of the $\text{Unif}(0, 1)$ distribution (allowing the PDF to be non-constant on that interval). Let us say that X is distributed $\text{Beta}(a, b)$, where $a > 0$ and $b > 0$.

2. **PDF:** The PDF of a $\text{Beta}(a, b)$ r.v. is

$$\frac{1}{\beta(a, b)} x^{a-1}(1 - x)^{b-1},$$

where the constant $\dfrac{1}{\beta(a, b)} = \dfrac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}$ is the Beta normalizing constant, chosen to make the PDF integrate to 1. (Γ is the Gamma function. We'll learn about some cool properties of the Gamma function later!)

3. **Properties of the Beta Distribution:**

   (a) If $X \sim \text{Beta}(a, b)$, then $\mu = E(X) = \dfrac{a}{a + b}$.

   (b) $\text{Var}(X) = \dfrac{\mu(1 - \mu)}{a + b + 1}$.

   (c) If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upward. If $a > 1$ and $b > 1$, then the PDF opens down.

   (d) If $a = b$, the PDF is symmetric about 1/2. If $a > b$, the PDF favors values larger than 1/2, and if $a < b$, the PDF favors values smaller than 1/2.

   (e) **$\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$.** (Plug in $a = 1$ and $b = 1$ to the PDF to verify!)

4. **Bayes' Billiards** We include this fact mainly because the integral presented below closely resembles integrating a Beta PDF. For any integers k and n with $0 \leq k \leq n$,

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}.$$

We show that left hand side (LHS) equals right hand side (RHS) through the same scenario: Start with $n+1$ balls, n white and 1 gray. Let's randomly throw each ball onto the unit interval $[0,1]$, and let X be the number of white balls tot he left of the gray ball; $X \in \{0, 1, \dots, n\}$.

LHS: To get $P(X = k)$, we can use LOTP, conditioning on the position of the gray ball. Conditioning on the gray ball being at position p in the $[0,1]$ interval, the number of white balls landing to the left of p is distributed $\text{Bin}(n, p)$. We get the left hand side of Bayes' Billiards from LOTP, conditioning on the position of the gray ball (which is distributed $\text{Unif}(0,1)$).

RHS: Note that it doesn't matter what we do first – assign the colors of the balls or throw them. We can first randomly throw each ball onto the interval and only *then* choose one ball at random to paint gray. By symmetry, any one of the $n+1$ balls is equally likely to be painted gray, so $P(X = k) = \dfrac{1}{n+1}$.

Since both sides of the equation equate to $P(X = k)$, they are equal. Bayes' Billiards is both a useful result in and of itself and a way for us to solve for the Beta normalizing constant without using calculus: Because PDFs must integrate to 1, the normalizing constant $\beta(a,b)$ satisfies $\beta(a,b) = \int\limits_0^1 x^{a-1}(1-x)^{b-1}dx$. If we substitute $a-1$ for k and $b-1$ for $n-k$, it follows that $\binom{a+b-2}{a-1}\beta(a,b) = \dfrac{1}{a+b-1}$, so $\beta(a,b) = \dfrac{1}{(a+b-1)\binom{a+b-2}{a-1}} = \dfrac{(a-1)!(b-1)!}{(a+b-1)!}$.

1.2.6 Gamma Distribution (Continuous)

1. **PDF:** A random variable $Y \sim \text{Gamma}(a, \lambda)$, where $a > 0$ and $\lambda > 0$, has the following PDF, for support $y > 0$. The "Γ" symbol represents the Gamma function, discussed below.

$$f(y) = \frac{1}{\Gamma(a)}(\lambda y)^a e^{-\lambda y}\frac{1}{y},$$

2. **Gamma Function:** The Gamma function Γ is an extension of the factorial function to all real (and complex) numbers, with the argument shifted down by 1. In closed form, the Gamma function is defined as:

$$\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt.$$

For the purposes of this class, you *do not* need to know or use the closed-form definition. Rather, pay attention to the below handy properties of the Gamma function:

(a) Recursive definition: $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$.

(b) Factorial generalization: $\Gamma(n) = (n-1)!$ if n is a positive integer.

We can use these properties, along with pattern-matching, to find the mean, variance, and other moments of the Gamma distribution (not shown).

3. **Mean and Variance:** If $X \sim \text{Gamma}(a, \lambda)$,

1. $E(X) = \dfrac{a}{\lambda}$
2. $\text{Var}(X) = \dfrac{a}{\lambda^2}$

3. $E(X^c) = \frac{1}{\lambda^c} \cdot \frac{\Gamma(a+c)}{\Gamma(a)}, c \geq -a.$

4. **Properties:**

   1. If $X \sim \text{Gamma}(a, \lambda)$, then $cX \sim \text{Gamma}(a, \frac{\lambda}{c})$.
   2. If $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$, and X and Y are independent, then $X + Y \sim \text{Gamma}(a + b, \lambda)$.

5. **Connections to Exponential:** The Gamma distribution is a generalization of the Exponential distribution. In fact, the Gamma(a, λ) random variable is the sum of a i.i.d Expo(λ) random variables. Again, if $X_1, \ldots, X_n \sim^{i.i.d}$ Expo(λ), then

$$X_1 + \ldots + X_n \sim \text{Gamma}(a, \lambda).$$

6. **Beta-Gamma Connection: The Bank-Post Office Story** The Bank-Post Office story gives us a really neat way to connect the Beta and Gamma distributions. While running errands, you decide to go to the bank and then to the post office. Let $X \sim \text{Gamma}(a, \lambda)$ be your waiting time at the bank and $Y \sim \text{Gamma}(b, \lambda)$ to be your waiting time at the post office, with X and Y independent. Let $T = X + Y$ be the total wait time and $W = \frac{X}{X+Y}$ be the fraction of the waiting time you spent at the bank. The Bank-Post Office story tells us the following:

   - $T \sim \text{Gamma}(a + b, \lambda)$
   - $W \sim \text{Beta}(a, b)$
   - T (the total) and W (the fraction) are independent!
   - We can also use these findings to find Beta expectation and the Beta normalizing constant – without any calculus!

1.2.7 χ^2 Distribution (Continuous)

We now introduce the second-to-last continuous distribution in Stat 110: the χ^2 distribution. We present the definition and a few useful properties:

1. **Definition:** Let $V = Z_1^2 + \cdots + Z_n^2$, where $Z_1 \ldots Z_n$ are i.i.d $N(0, 1)$ random variables. Then, we say that V has a "Chi-Square distribution with n degrees of freedom." Symbolically, we write:

$$V \sim \chi_n^2$$

2. **Theorem (Special Gamma):** χ_n^2 is a special case of the Gamma distribution. Specifically, we have:

$$\chi_n^2 \sim Gamma\Big(\frac{n}{2}, \frac{1}{2}\Big)$$

   .

3. **Theorem (Mean and Variance):** Let $V \sim \chi_n^2$. Then, we have the following:

$$E(V) = n$$

$$Var(V) = 2n$$

   We can use linearity of expectation and/or properties of Normal to derive the two results below.

4. **Useful Fact (Sample Variance of Normals):** Let $Z_1 \ldots Z_n$ be i.i.d $N(\mu, \sigma^2)$ random variables. We define S_n^2, the "sample variance" of $Z_1 \ldots Z_n$, as:

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (Z_j - \bar{Z}_n)^2$$

It turns out that S_n^2 (after appropriate scaling) is Chi-Square. Specifically, we have:

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that the sample variance S_n^2 is "unbiased" for estimating the true variance, σ^2. Mathematically, this means that $E(S_n^2) = \sigma^2$.

1.2.8 Student-t Distribution (Continuous)

We now introduce the very last continuous distribution in Stat 110: the Student-t distribution. We present its definition and a few useful properties:

1. **Definition:** Let $Z \sim N(0,1)$, $V \sim \chi^2_n$, and let Z be independent of V. Define T as the following:

$$T = \frac{Z}{\sqrt{\frac{V}{n}}}$$

   Then, we say that T has a "Student-t distribution with n degrees of freedom." Symbolically, we write that $T \sim t_n$.

2. **Properties of Student-t:** We present three useful properties of the Student-t distribution:

   (a) *Symmetry:* If $T \sim t_n$, then $-T \sim t_n$.

   (b) *Cauchy (t_1):* Recall that the Cauchy distribution can be thought of as the ratio of two i.i.d. $N(0,1)$ random variables. As such the t_1 distribution is the same thing as the Cauchy distribution.

   (c) *Convergence to Normal:* As $n \to \infty$, the t_n distribution approaches the standard Normal distribution. For finite n, the t-distribution PDF has fatter tails than that of the standard Normal distribution.

1.3 Conjugacy Relationships

1.3.1 Beta-Binomial Conjugacy

Let p be a random variable $\in [0,1]$, and suppose our goal is to estimate p after observing some Binomially-distributed data with parameter p. As an example, suppose we have a coin that lands Heads with probability p, but we don't know what p is. We can only *infer* what p is after observing coin tosses — if we observe n tosses, the number that come up as Heads is distributed $\text{Bin}(n,p)$. Formally, our setup is

$$p \sim \text{Beta}(a,b),$$
$$X|p \sim \text{Bin}(n,p).$$

Note that X is *not* marginally Binomial; it is only *conditionally* Binomial, given p. (The marginal distribution of X is called the *Beta-Binomial distribution.*

Beta-Binomial conjugacy tells us that the posterior distribution of p after observing $X = k$, is
$\mathbf{p}|(\mathbf{X}=\mathbf{k}) \sim \text{Beta}(\mathbf{a}+\mathbf{k}, \mathbf{b}+\mathbf{n}-\mathbf{k})$.

This relationship is fascinating: when going from prior to posterior distributions of p, we don't leave the family of Beta distributions! We just add the number of observed successes, k, to the first parameter of the prior Beta (a), and the number of observed failures, $n-k$, to the second parameter b.

We say that **Beta is the *conjugate prior* of the Binomial** because if p has a *prior* distribution that is Beta-distributed, then the *posterior* distribution of p given observed data is also Beta-distributed.
Given that we never leave the Beta family when going from prior to posterior, we can *sequentially* update our beliefs as we get more and more evidence.

1.3.2 Gamma-Poisson Conjugacy

Consider a Poisson process with a rate of arrival of λ (for example, buses per hour), where λ is unknown. Let us place a prior of $\lambda \sim Gamma(r_0, b_0)$ on this unknown rate, λ: r_0 and b_0 are known, positive constants and r_0 is an integer. Let Y be the number of buses that arrive in a period of t hours. Mathematically, by our understanding of the Poisson process, our story has the following setup:

$$\lambda \sim \text{Gamma}(r_0, b_0)$$

$$Y|\lambda \sim \text{Pois}(\lambda t)$$

We obtain the following results:

1. Marginally, we have the following distribution for the number of arrivals in t hours:

$$Y \sim \text{NBin}(r_0, \frac{b_0}{b_0 + t})$$

2. Conditional on our observation that $Y = y$, we have the following posterior distribution for λ:

$$\lambda|Y = y \sim \text{Gamma}(r_0 + y, b_0 + t)$$

3. Intuitively, we can imagine that there had been r_0 buses in b_0 hours. We are just incrementing the number of bus arrival counts, and of course, our total time.

1.4 Poisson Processes

The Exponential, Poisson, and Gamma distributions are linked by a common story, which is the story of the *Poisson process.* A Poisson process is a sequence of arrivals occurring at different points in continuous time, such that the number of arrivals in a particular interval of time has a Poisson distribution. A **Poisson process with rate** λ has the following properties:

1. The number of arrivals that occur in a **continuous** interval of length t is a Pois(λt) random variable.
2. The numbers of arrivals that occur in disjoint time intervals are independent of each other. For example, the numbers of arrivals in the time intervals $(0, 10)$, $[10, 12)$, and $[15, \infty)$ are independent.

Theorem (Count-time duality). In a poisson process, if T_n is the time until the nth arrival (continuous) and N_t is the number of arrivals before or at time t (discrete), then

$$P(T_n > t) = P(N_t < n).$$

Count-time duality connects a discrete random variable N_t, which counts the number of arrivals, with a continuous random variable T_n, which marks the time of the nth arrival. The event that the nth arrival has not happened yet as of time t is equivalent the event that up until time t, there have been fewer than n arrivals.

Inter-arrival times: Since $N_t \sim \text{Pois}(\lambda t)$, we have that

$$P(T_1 > t) = P(N_t = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t},$$

by count-time duality.

$P(T_1 \leq t) = 1 - e^{-\lambda t}$, and this is the Expo($\lambda$) CDF! (Great teaching moment here: *pattern-matching expressions to known CDFs and PDFs is an effective problem solving strategy.*)

Hence, by count-time duality, the times between arrivals are i.i.d Expo(λ)!

Independence of inter-arrival times: Combining count-time duality, part (2) of the definition (disjoint time intervals of Poisson processes are independent), and memorylessness of the Exponential, we have that the inter-arrival times are i.i.d Expo(λ) random variables. $T_1 \perp\!\!\!\perp T_2 - T_1$: the time intervals in question are disjoint, so whatever happened before the first arrival is irrelevant once the first arrival occurs. As such, $T_2 - T_1$ also has an Expo(λ) distribution. The same goes with $T_3 - T_2$: $T_3 - T_2 \sim$ Expo(λ), independently of *both* T_1 and $T_2 - T_1$.

Total time until nth arrival: What is the distribution of T_n, the total time until the nth arrival? Let's start with $n = 2$. T_2 is the sum of two independent Expo(λ) random variables, T_1 and $T_2 - T_1$. **This sum is NOT distributed Expo. Do not make this category error!** This sum follows another named distribution that we have not yet learned, so (1) don't worry about it for now, but (2) **don't erroneously say $T_n \sim$ Expo**.

1.5 Moment Generating Functions (MGFs)

Before we talk MGFs, we should define what in the world a "moment" is (and some other important terms). Let X be a random variable with mean μ and variance σ^2. And, let n be a positive integer.

1. **Definition:** The "n^{th} moment of X" is defined as $E(X^n)$.
2. **Definition:** The "n^{th} central moment of X" is defined as $E((X - \mu)^n)$.
3. **Definition:** The "n^{th} standardized moment of X" is defined as $E((\frac{X-\mu}{\sigma})^n)$.

Definition: For any random variable X, the "moment generating function" (MGF) of X is defined as the following, where t is a dummy variable.

$$M_X(t) = E(e^{tX})$$

The purpose of the t is simply to serve as a dummy variable / placeholder and to keep the MGF as an actual function (of t), rather than just some fixed constant. We say "the MGF exists" if $M_X(t)$ is finite on some open interval $(-a, a)$ containing 0. Otherwise, we say that "the MGF of X does not exist."

Below are some fundamental properties of the MGF:

1. The MGF "determines the distribution": If two random variables X and Y have the same MGF, then they must have the same distribution! Note that even if $M_X(t) = M_Y(t)$ only on a super tiny interval $(-a, a)$ containing 0, then $X \sim Y$. Warning: the MGF says *nothing* about whether X and Y are *equal* or *independent*.
2. Extracting Moments: The n^{th} moment of a random variable X, denoted as $E(X^n)$, can be found by evaluating the n^{th} derivative with respect to tr of its MGF, $M_X(t)$, at $t = 0$. Mathematically, we write the following:
$$E(X^n) = M_X^{(n)}(0)$$
3. Taylor Series and Pattern-Matching: used to derivative the above moment-extracting formula.
   - The Taylor series expansion of $M_X(t)$ about 0 is written below:
   $$M_X(t) = \sum_{n=0}^{\infty} M_X^{(n)}(0)\frac{t^n}{n!}.$$
   - By LOTUS, we also have
   $$M(t) = E(e^{tX}) = E\left(\sum_{n=0}^{\infty} X^n \frac{t^n}{n!}\right).$$
   - We are allowed to move the expectation to *inside* the sum. The math behind this is gnarly, but we can write
   $$M(t) = \sum_{n=0}^{\infty} E(X^n)\frac{t^n}{n!}.$$

- By pattern-matching the coefficients of the two expansions, we get $E(X^n) = M^{(n)}(0)$.

4. MGF of the Sum of Two *Independent* Random Variables: Let X and Y be two *independent* random variables with MGFs $M_X(t)$ and $M_Y(t)$, respectively. Then, the MGF of the random variable $X+Y$, denoted as $M_{X+Y}(t)$, is as follows:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

5. MGF of a Location-Scale Transformation: Let X be a random variable with MGF $M_X(t)$. Let $Y = a + bX$, for constants a, b. Then the MGF of Y, denoted as $M_Y(t)$, is as follows:

$$M_Y(t) = E(e^{tY}) = E(e^{t(a+bX)}) = E(e^{at}e^{tbX}) = e^{at}E(e^{(tb)X}) = e^{at}M_X(bt)$$

*a note on notation: $M_X^{(k)}(t)$ refers to the k^{th} derivative of the MGF of X, $M_X(t)$, with respect to t.

1.6 Joint, Marginal, and Conditional Distributions

Let X and Y be two random variables (possibly dependent). We define the following:

1. **Joint Distribution:** The joint distribution of X and Y describes the probability of the vector (X, Y) falling into any subset of the $\mathbb{R}^2$ plane.

2. **Marginal Distribution:** The marginal distribution of X is the individual distribution of X, ignoring whatever behavior or influence Y might have.

3. **Conditional Distribution:** The conditional distribution of X given Y is the updated distribution for X after observing $Y = y$ (i.e., Y crystallizing to some specific value).

A few remarks:

1. Joint distributions typically contain more information than marginal distributions. Intuitively, this is because we can always derive the marginal distribution of a random variable X *from* its joint distribution with another random variable Y.

2. If X and Y are independent random variables, then their joint PMF/PDF (depending on whichever one is relevant) is *just the product of their marginal PMF/PDFs!* As such, we do not get any additional information from studying X and Y jointly.

1.6.1 Joint Distributions

We begin by defining the following terms:

1. **Joint CDF:** For *any* 2 random variables X and Y, the joint CDF is defined as the following:

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

Furthermore, using the CDF $F(x, y)$, the probability that (X, Y) falls into the 2D rectangle $[x_1, x_2] \times [y_1, y_2]$ is

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - \bigg(F(x_2, y_1) - F(x_1, y_1)\bigg).$$

Note that if X and Y are discrete, then the joint CDF may have many jumps and flat-regions, making it hard to work with.

2. **Joint PMF:** For 2 *discrete* random variables X and Y, the joint PMF is defined as the following:

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

Just like regular univariate PMFs, joint PMFs should also sum up to 1. Specifically, we mean that we are summing over *all possible values* of X and Y. Mathematically, we write:

$$\sum_x \sum_y P(X = x, Y = y) = 1$$

3. **Joint PDF** For 2 *continuous* random variables X and Y, the joint PDF is defined as the following:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y), \text{ where } F_{X,Y} \text{ is the joint CDF of } X \text{ and } Y.$$

   With less fancy math symbols, the joint PDF is obtained by taking the partial derivative of the joint CDF with respect to x, and then taking the partial derivative again with respect to y.

   As with all PDFs in general, the joint PDF must integrate to 1. Mathematically, we mean

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

   Of course, the joint PDF must also be *nonnegative* for all values of X and Y.

   We can use the joint PDF to find the probability that X and Y will be in a particular region of values. Let's look at the following examples:

$$P(1 < X < 4, 2 < Y < 5) = \int_{2}^{5} \int_{1}^{4} f_{X,Y}(x,y) dx dy$$

$$P(X < 4, 2 < Y < 5) = \int_{2}^{5} \int_{-\infty}^{4} f_{X,Y}(x,y) dx dy$$

$$P(X < Y) = \int_{x}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy$$

1.6.2 Marginal Distributions

Discrete Random Variables

Let us start by taking X and Y to be 2 discrete random variables. To recap, we've seen the following definitions:

1. **Marginal PMF:** The marginal PMF of X is defined as the following (with no concern for Y):

$$P(X = x)$$

2. **Marginal CDF:** The marginal CDF of X is defined as the following (with no concern for Y):

$$P(X \leq x)$$

Now, we can derive the **marginal PMF** and **marginal CDF** of X from their joint distribution counterparts:

1. **Marginal PMF from Joint PMF:** We can derive the marginal PMF of X from the joint PMF of X and Y by summing over all possible values of Y:

$$P(X = x) = \sum_{y} P(Y = y, X = x)$$

2. **Marginal CDF from Joint CDF:** We can derive the marginal CDF of X from the joint CDF of X and Y by taking the limit as y approaches infinity:

$$P(X \leq x) = \lim_{y \to \infty} P(X \leq x, Y \leq y)$$

Continuous Random Variables

Now, let us take X and Y to be 2 continuous random variables. Recall the following definitions:

1. **Marginal PDF:** The marginal PDF of X is defined as the following (with no concern for Y):

$$f_X(x) = \frac{d}{dx}F_X(x), \text{ where } F_X \text{ is the marginal CDF of } X.$$

2. **Marginal CDF:** The marginal CDF of X is defined as the following (with no concern for Y). Note this is the same definition from the discrete case.

$$P(X \leq x)$$

We can derive the **marginal PDF** and **marginal CDF** from their joint counterparts:

1. **Marginal PDF from Joint PDF:** We can derive the marginal PDF of X from the joint PDF of X and Y by integrating over all possible values of Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

2. **Marginal CDF from Joint CDF:** See discrete section. Literally, copy-paste.

1.6.3 Conditional Distributions

Discrete Random Variables

Again, let us start by taking X and Y to be 2 discrete random variables. We explore and define the following:

1. **Conditional PMF:** The conditional PMF of Y given $X = x$ is defined as the following:

$$P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

We should state here that there could potentially be a different conditional PMF for every possible value of X!

2. **Conditional PMFs are still PMFs:** Like all PMFs, a conditional PMF should sum up to 1 over all possible values of Y (while holding X as fixed). Mathematically, we mean:

$$\sum_y P(Y = y \mid X = x) = 1$$

3. **Connections between Conditional PMFs:** Notice how the conditional PMF looks kind of like Bayes' Rule? Well, that's not a coincidence! We present the following relationship:

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

Continuous Random Variables

Now, let us take X and Y to be 2 continuous random variables. We explore and define the following:

1. **Conditional PDF:** The conditional PDF of Y given $X = x$ is defined as the following, for all x with $f_X(x) > 0$. We typically view the conditional PDF of Y as a function of y, with x being fixed:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Of course, conditional PDFs are PDFs, so they must integrate to 1:

$$\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)dx = 1$$

2. **Connections between Joint PDF and Conditional PDF:** From our expression above, we can derive the following:
$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) \cdot f_X(x)$$
$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y) \cdot f_Y(y)$$

3. **Continuous Form of Bayes' Rule:** Analogous to the discrete case, we have the following that follows from our above statements:
$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)}, \text{ for } f_X(x) > 0.$$

4. **Continuous Form of LOTP:** Analogous to the discrete case, we obtain the following:
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)dy$$

1.6.4 Independence

Any 2 random variables X and Y are independent if the following condition holds, where F_X and F_Y are their marginal CDFs, and $F_{X,Y}$ is their joint CDF:
$$F_{X,Y}(x,y) = F_x(x)F_Y(y)$$

Discrete Random Variables
Now, if X and Y are both discrete, three equivalent conditions for X and Y being independent are the following:

1. $P(X = x, Y = y) = P(X = x)P(Y = y)$
2. $P(Y = y \mid X = x) = P(Y = y)$
3. $P(X = x \mid Y = y) = P(x = x)$

Note that these three equivalent conditions must hold for *all* possible values of X and Y! If there exists any values of X, Y for which any of these conditions fail, then X, Y are **not** independent!

Continuous Random Variables
Now, if X and Y are both continuous, three equivalent conditions for X and Y being independent are the following:

1. $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
2. $f_{Y|X}(y \mid x) = f_Y(y)$
3. $f_{X|Y}(x \mid y) = f_X(x)$

As in the discrete case, note that these three equivalent conditions must hold for *all* possible values of X and Y! If there exists any values of X, Y for which any of these conditions fail, then X, Y are **not** independent!

Another test for independence of two continuous random variables X and Y is the following: suppose that the joint PDF of $f_{X,Y}$ of X and Y can be factored as
$$f_{X,Y}(x,y) = g(x)h(y)$$
for all x and y, with g and h nonnegative functions. Then, we can conclude that X and Y are independent.

1.6.5 2D LOTUS

Oftentimes, we may want to find the expectation of a random variable that is itself a function of two random variables X and Y. In this situation, we use the 2D version of LOTUS:

1. **Discrete Case:** Let g be a function from $\mathbb{R}^2$ to $\mathbb{R}$. If X and Y are both discrete, we have the following:
$$E(g(X,Y)) = \sum_x \sum_y g(x,y) \cdot P(X=x, Y=y)$$
2. **Continuous Case:** Let g be a function from $\mathbb{R}^2$ to $\mathbb{R}$. If X and Y are both continuous, we have the following:
$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y) dx dy$$

1.6.6 Hybrid Forms

It seems that we have not covered the cases for when X is discrete and Y is continuous, or vice versa. These cases tend to be rarer and more nuanced than their purely continuous or discrete counterparts. We include the following tables to account for these cases, relatively painlessly:

	Y **discrete**	Y **continuous**
X **discrete**	$\sum\limits_y P(X=x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$\sum\limits_y f_X(x\|Y=y)P(Y=y)$	$\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

Figure 1: Hybrid Forms of LOTP

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x) = \frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x) = \frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x) = \frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x) = \frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

Figure 2: Hybrid Forms of Bayes' Rule

1.7 Covariance and Correlation

Intuitively, covariance and correlation are two numerical summaries that measure the tendency for two random variables to go up or down together relative to their means. Positive covariance (and by extension, correlation) between X and Y suggests that as X increases, Y also tends to increase. Negative covariance between X and Y suggests that as X increases, Y generally tends to decrease. We provide the following formal definitions:

1. **Covariance:** The covariance between *any* two random variables X and Y is defined as the following:
$$Cov(X,Y) = E((X-EX)(Y-EY))$$
Equivalently, we have a (usually) simpler definition:
$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

2. **Correlation:** The correlation between *any* two random variables X and Y is defined as the following:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

   Intuitively, correlation is just a unit-less version of covariance in the sense that we are dividing by the square root of the product of the variances. This way, we obtain a useful property of correlation: correlation is *always* between -1 and 1.

An important theorem is that if *any* two random variables are *independent*, then they are *uncorrelated* (i.e., $Corr(X,Y) = 0$). The reverse direction of this statement is **not necessarily true!**

Now, we present the following key properties of covariance:

1. $Cov(X,X) = Var(X)$
2. $Cov(X,Y) = Cov(Y,X)$
3. $Cov(X,c) = 0$, for any constant c.
4. $Cov(aX,Y) = aCov(X,Y)$, for any constant a.
5. $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$
6. $Cov(X+Y,Z+W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W)$
7. $Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y)$. **No, this is *not* a typo!** The “2” is very important!
8. In general, $Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) + 2\sum_{i<j} Cov(X_i, X_j)$

1.8 Multivariate Distributions

1.8.1 Multinomial Distribution

The Multinomial distribution is a generalization of the Binomial distribution that accounts for multiple possible outcomes, not just two (success or failure).

Let's say that the vector $\vec{X} = (X_1, X_2, \ldots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \ldots, p_k)$, $\sum_{j=1}^{k} p_j = 1$.

We can say the following about $\vec{X}$:

1. **Story:** We have n items, and each item can fall into any of k buckets, independently of other items. p_j is the probability of the item falling into item j, for $j \in \{1, \ldots, k\}$. X_1 is the number of items in bucket 1, X_2 is the number of items in bucket 2, etc., so that $X_1 + \ldots + X_k = n$.

2. **Example:** Every spring, assume that 2000 Harvard first-years are randomly and independently sorted into one of 12 upperclassman houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_{12}(2000, \vec{p})$, where $\vec{p} = (1/12, 1/12, \ldots, 1/12)$. Note that $X_1 + X_2 + \ldots + X_{12} = 2000$, and they are dependent.

3. **Joint PMF:** For $n_1, \ldots, n_k$ satisfying $n_1 + \ldots n_k = n$, the joint PMF of $\vec{X} \sim \text{Mult}_k(n, \vec{p})$ is

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot {p_1}^{n_1} {p_2}^{n_2} \cdots {p_k}^{n_k}.$$

   You might be asking, *where does the* $\frac{n!}{n_1! n_2! \ldots n_k!}$ *term come from?* This term is called the *multinomial coefficient* and represents the number of permutations of n objects where you have counts $n_1, n_2, \ldots, n_k$ of each object.

4. **Marginal PMF of each X_j:** The marginals of a Multinomial are Binomial: if $\vec{X} \sim \text{Mult}_k(n, \vec{p})$, then $X_j \sim \text{Bin}(n, p_j)$. In the marginal case, we only care whether items fall into bucket j or not – we disregard the other buckets.

5. **Lumping:** For any distinct buckets i and j, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$. Think about this as merging buckets i and j into one larger bucket and only caring about whether items fall into this larger bucket. The random vector obtained from merging two buckets is still Multinomial, for example:

$$(X_1, X_2, X_3) \sim \text{Mult}_3(n, (p_1, p_2, p_3)) \implies (X_1, X_2 + X_3) \sim \text{Mult}_2(n, (p_1, p_2 + p_3)).$$

6. **Conditioning:** Suppose we know X_1, the number of objects in bucket 1. Our distribution for the rest of the random vector $(X_2, \ldots, X_k)$ still follows a Multinomial distribution (with updated probabilities):

$$(X_2, \ldots, X_k) | X_1 = n_1 \sim \text{Mult}_{k-1}(n - n_1, (p'_2, \ldots, p'_k)),$$

   where $p'_j = \frac{p_j}{p_2 + \ldots + p_k}$.

   We have to update our probabilities because we are in a new conditional "world," and probabilities must always add up to 1. This is just normalization.

7. **Covariance:** For $i \neq j$,

$$\text{Cov}(X_i, X_j) = -np_i p_j.$$

   This makes intuitive sense: since we have a fixed total number of items, if one bucket has more items, some other bucket is more likely to have fewer items.

1.8.2 Multivariate Normal Distribution

The Multivariate Normal distribution is a generalization of the Normal distribution into higher dimensions.

A vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ is Multivariate Normal (MVN) if ***any linear combination* of the** X_j is Normally distributed. That is,

$$t_1X_1 + t_2X_2 + \ldots + t_kX_k$$

is Normal for any set of constants $t_1, t_2, \ldots, t_k$.

The parameters of the Multivariate Normal are the **mean vector** $\vec{\mu} = (\mu_1, \ldots, \mu_k)$ and the **covariance matrix** where the $(i,j)^{th}$ entry is $\text{Cov}(X_i, X_j)$. You might be asking, *why do we not need the variances of each component to fully specify a Multivariate Normal?* The variances of each component are already encoded in the covariance matrix on the diagonal: $\text{Cov}(X_j, X_j) = \text{Var}(X_j)$.

We present the following properties of Multivariate Normal random vectors:

1. **Sub-vectors:** Any sub-vector of a Multivariate Normal is also Multivariate Normal. For example, if (X_1, X_2, X_3) is Multivariate Normal, then so is the subvector (X_1, X_2). This comes from the fact that any linear combination of Multivariate Normal is also Multivariate Normal → we can just set t_3, the coefficient corresponding to X_3, to 0.

2. **Concatenation:** If $\vec{X} = (X_1, \ldots, X_n)$ and $\vec{Y} = (Y_1, \ldots, Y_m)$ are Multivariate Normal, and $\vec{X}$ and $\vec{Y}$ are independent, then the concatenated random vector $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is Multivariate Normal. Note that $\vec{X}$ and $\vec{Y}$ *must be independent*!

3. **MGF:** The joint MGF of a Multivariate Normal $(X_1, \ldots, X_k)$ is

$$E(e^{t_1X_1+\ldots+t_kX_k}) = \exp\left(t_1E(X_1) + \ldots + t_kE(X_k) + \frac{1}{2}\text{Var}(t_1X_1 + \ldots + t_kX_k)\right).$$

   Notes about this:

   - Here, "exp" stands for "e to the power of."
   - In general, the *joint MGF* of a random vector $\vec{X} = (X_1, \ldots, X_k)$ is defined as

   $$M_{\vec{X}}(\vec{t}) = E(e^{\vec{t}'\vec{X}}) = E(e^{t_1X_1+\ldots+t_kX_k}),$$

   for input $\vec{t} = (t_1, \ldots, t_k) \in \mathbb{R}^k$. Think about this as a multivariable generalization of the single-variable MGF we learned a few weeks back.

   *note: $\vec{t}'$ means the transpose of $\vec{t}$.

4. **Uncorrelated implies independent:** If any two components of a Multivariate Normal random vector are uncorrelated (i.e. correlation = 0), then they are independent.

   Note that this does not generally apply to random variables and vectors! In general, we only have guarantees that independence implies uncorrelated.

1.9 Transformations

Often, it is very useful to consider transformations of random variables, i.e. $g(X)$. With LOTUS, we can find $E(g(X))$ given the distribution of X. But how can we get the *distribution* of $Y = g(X)$? (Note that Y is a random variable as well.)

1.9.1 Discrete

Let X be a discrete r.v. If g is one-to-one, then we know that

$$P(g(X) = y) = P(X = g^{-1}(y)),$$

and we can use this substitution to easily derive the PMF for $g(X)$.

1.9.2 Continuous

One Variable Transformations
Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in $Y = g(X)$. If g is differentiable and one-to-one (every value of X gets mapped to a unique value of Y), then the following **change-of-variables formula** holds:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

We can then differentiate to get the PDF. We differentiate the left side with respect to y and right size with respect to x, and the extra $\left|\frac{dx}{dy}\right|$ term on the right comes from using the chain rule.

We can alternatively write the change-of-variables formula as

$$f_Y(y)dy = f_X(x)dx,$$

where the term $f_Y(y)dy$ can be interpreted as the probability that Y is in a tiny interval of length dy (with an analogous interpretation for $f_X(x)dx$). It makes intuitive sense to set the quantities $f_X(x)dx$ and $f_Y(y)dy$ equal to one another: the probability that $g(X)$ is in a tiny interval of $g(x)$ values should equal the original probability that X is in a tiny interval of x values, since $Y = g(X)$ is a deterministic function of X.

It may seem kind of odd to have $\frac{dx}{dy}$. Don't panic! Just remember that

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}.$$

Tips for applying the change-of-variables formula:

1. Make sure g is differentiable and one-to-one. A good sign of this is if g is strictly increasing or strictly decreasing.
2. Express your final answer for the PDF of Y as a function of y.
3. Specify the support of Y.
4. We can choose whether to compute $\frac{dx}{dy}$ or $\frac{dy}{dx}$ (and take the reciprocal) - these give the same result.

Two-Variable Transformations
Similarly, let's say we know the joint distribution of U and V but are also interested in the random vector (X, Y) found by $(X, Y) = g(U, V)$. If g is differentiable and one-to-one, then the following is true:

$$f_{X,Y}(x, y) = f_{U,V}(u, v)\left|\left|\frac{\delta(u, v)}{\delta(x, y)}\right|\right| = f_{U,V}(u, v)\left|\left|\begin{array}{cc} \frac{\delta u}{\delta x} & \frac{\delta u}{\delta y} \\ \frac{\delta v}{\delta x} & \frac{\delta v}{\delta y} \end{array}\right|\right|.$$

The outer || signs around our matrix tells us to take the absolute value. The inner || signs tells us to the matrix's determinant. Thus the two pairs of || signs tell us to take the absolute value of the determinant matrix of partial derivatives. In a 2x2 matrix,

$$\left\| \begin{matrix} a & b \\ c & d \end{matrix} \right\| = |ad - bc|$$

The determinant of the matrix of partial derivatives is referred to the **Jacobian**, denoted as J.

$$\left| \begin{matrix} \frac{\delta u}{\delta x} & \frac{\delta u}{\delta y} \\ \frac{\delta v}{\delta x} & \frac{\delta v}{\delta y} \end{matrix} \right| = J$$

We can generalize this to multi-variable transformations:

Multi-Variable Transformations
Let $\vec{X} = (x_1, \ldots, X_n)$ be a continuous random vector with joint PDF $f_{\vec{x}}$. Let $\vec{Y} = g(\vec{X})$ be an invertible function, and mirror this by letting $\vec{y} = g(\vec{x})$. Since g is invertible, we also have that $\vec{X} = g^{-1}(\vec{Y})$.

If all the partial derivatives $\dfrac{\partial x_i}{\partial y_j}$ exist and are continuous, we can form the **Jacobian matrix** $J = \dfrac{\partial \vec{x}}{\partial \vec{y}} =$

$$J = \frac{\partial \vec{x}}{\partial \vec{y}} = \begin{pmatrix} \frac{\delta x_1}{\delta y_1} & \frac{\delta x_1}{\delta y_2} & \cdots & \frac{\delta x_1}{\delta y_n} \\ \vdots & & & \vdots \\ \frac{\delta x_n}{\delta y_1} & \frac{\delta x_n}{\delta y_2} & \cdots & \frac{\delta x_n}{\delta y_n} \end{pmatrix}.$$

We can check whether g is invertible by verifying that the determinant of J is not 0. In that case, the joint PDF of $\vec{Y}$ is

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(g^{-1}(\vec{y}) \cdot |\left|\frac{\partial \vec{x}}{\partial \vec{y}}\right||,$$

where $|\left|\dfrac{\partial \vec{x}}{\partial \vec{y}}\right||$ is the absolute value of the determinant of $\dfrac{\partial \vec{x}}{\partial \vec{y}}$.

1.9.3 Convolutions

A **convolution** is the sum of independent random variables. We have already seen a few examples of this:

- **Binomial:** If $X_1, \ldots, X_n$ are i.i.d. Bern(p), then $X_1 + \ldots + X_n \sim \text{Bin}(n, p)$.
- **Negative binomial:** If $G_1, \ldots, G_r$ are i.i.d. Geom(p), then $G_1 + \ldots + G_r \sim \text{NBin}(r, p)$.
- **Poisson:** If $X_1, \ldots, X_n$ are i.i.d. Pois(λ), then $X_1, + \ldots + X_n \sim \text{Pois}(n\lambda)$.
- **Normal:** If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and X and Y are independent, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

In the general case where X and Y can have any distribution (they just need to be independent), we can find the distribution of $T = X + Y$ by using convolution sums and integrals.

Convolution Sum: used for *discrete* X and Y.

$$P(T = t) = \sum_x P(Y = t - x)P(X = x) = \sum_y P(X = t - y)P(Y = y).$$

Note: This is just LOTP, conditioning on the possible values of X (or the possible values of Y).

Convolution Integral: used for *continuous* X and Y.

$$f_T(t) = \int_{-\infty}^{\infty} f_Y(t - x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(t - y) f_Y(y) dy.$$

1.10 Order Statistics

1. **What is an order statistic?** Say we have n i.i.d random variables $X_1, X_2, \dots, X_n$. If we arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. $X_{(1)}$ is the smallest out of $X_1, \dots, X_n$, and $X_{(n)}$ is the largest. The order statistics are *dependent* random variables – for any value of $X_{(i)}$, $X_{(i+1)} \geq X_{(i)}$.

2. **Distribution:** Taking n i.i.d. random variables $X_1, X_2, X_3, \dots X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are as follows:

$$F_{X_{(i)}}(x) = P(X_{(j)} \leq x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n \binom{n-1}{i-1} F(x)^{i-1} (1 - F(X))^{n-i} f(x)$$

In general, the order statistics of $X_1, \dots, X_n$ will not follow a named distribution, but the order statistics of the standard Uniform distribution are an exception.

3. **Order Statistics of Standard Uniforms:** Let $U_1, \dots, U_n$ be i.i.d Unif$(0, 1)$. Then for $0 \leq x \leq 1, f(x) = 1$ and $F(x) = x$, so the PDF of $U_{(j)}$ is

$$f_{U_{(j)}}(x) = n \binom{n-1}{j-1} x^{j-1} (1-x)^{n-j}.$$

This is the Beta$(j, n-j+1)$ PDF! So $U_{(j)} \sim \text{Beta}(j, n-j+1)$, and $E(U_{(j)}) = \dfrac{j}{n+1}$.

Again, this is the key takeaway: $U_{(j)} \sim \text{Beta}(j, n-j+1)$.

4. **Universality of the Uniform:** We can also express the distribution of the order statistics of n i.i.d. random variables $X_1, X_2, X_3, \dots X_n$ in terms of the order statistics of n uniforms. We have that

$$F(X_{(j)}) \sim U_{(j)}.$$

1.11 Conditional Expectation and Variance

Intuitively, as we receive more and more information, our perceptions on how a random variable behaves will naturally change. As such, we introduce conditional expectation and conditional variance.

We provide the following definitions and theorems:

1. **Conditional Expectation (given an event):** Let A be an event with positive probability.

   i) If Y is a discrete random variable, then:

$$E(Y|A) = \sum_y y \cdot P(Y = y|A)$$

   ii) If Y is a continuous random variable, then:

$$E(Y|A) = \int_{-\infty}^{\infty} y \cdot f(y|A) dy$$

   Note that we can calculate the conditional PDF $f(y|A)$ in two ways:

   a)

$$f(y|A) = \frac{d}{dy} F(y|A) = \frac{d}{dy} P(Y \leq y|A)$$

b)

$$f(y|A) = \frac{P(A|Y = y)f(y)}{P(A)}$$

In both the discrete and continuous cases, $E(Y|A)$ is a fixed value!

2. **Law of Total Expectation (LOTE):** Let $A_1 \dots A_n$ be a set of events that partition the sample space. Intuitively, this means that at most one of $A_1 \dots A_n$ can occur at any given time, but also that *at least* one of $A_1 \dots A_n$ must occur at any given time. Succintly, one and only one of $A_1 \dots A_n$ can occur. Then, for any random variable Y, we have:

$$E(Y) = \sum_{i=1}^{n} E(Y|A_i) \cdot P(A_i)$$

Doesn't this look similar to and sound similar to LOTP? Well, that's not a coincidence! What if Y was an indicator variable?

3. **Conditional Expectation (given the *event* that another random variable takes on a *specific* value:** Yes, this is a super long title, but we will soon see why this level of specificity is necessary. Let X be a random variable, and suppose we know that X crystallizes to little x. Note that X crystallizing to little x *is an event!*

   i) If Y is a discrete random variable, then we have the following:

$$E(Y|X = x) = \sum_{y} y \cdot P(Y = y|X = x)$$

   ii) If Y is a continuous random variable, then we use the conditional PDF $f_{Y|X}(y|x)$:

$$E(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy$$

What is important to note is that $E(Y|X = x)$ **is *always* a function of little** x – in other words, a special $g(x)$. Now, to find $E(Y|X)$, we simply replace every little x we see with a big X (see below).

4. **Conditional Expectation (given a random variable, in general):** Let Y and X be random variables. Intuitively, the conditional expectation of Y given X, written as $E(Y|X)$, (notice how we *do not* specify what specific value X takes on) can be thought of as our **best prediction of** Y, assuming we get to know X.

To be clear, $E(Y|X)$, is itself a random variable. Again, $E(Y|X)$ *is* a random variable! In fact, $E(Y|X)$ is actually a random variable that is a function of X – in other words, think of $E(Y|X)$ as a really special $g(X)$.

As alluded to earlier, the general strategy for finding $E(Y|X)$ is to first find $E(Y|X = x)$ for some dummy variable x. Then, wherever we see little x, we plug in a big X.

5. Now, we present a few useful properties of conditional expectation:

   i) **Independence:** If X and Y are independent, then $E(Y|X) = E(Y)$.

   ii) **Taking out what's known:** For any function h, $E(h(X)Y|X) = h(X) \cdot E(Y|X)$.

   iii) **Linearity of Expectation:** Conditional expectations are still expectations!

$$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$$

   For any constant c, we have:

$$E(cY|X) = cE(Y|X)$$

6. **Adam's Law:** This theorem is typically used when we *want to find a marginal expectation.* Sometimes, a marginal expectation might be hard to find on its own, but a conditional expectation might be easier.

$$E(Y) = E(E(Y|X))$$

Again, $E(Y|X)$ is a function of the random variable X. Thus, the outer expectation is taken *with respect to* the random variable X! In other words, we can find the expectation of Y by first conditioning on a random variable that we wished we knew – in this case, X.

Of course, we can also condition on more than one random variable. We present to you, **Adam's Law with Extra Conditioning:** Let X, Y, Z be random variables. Then, we have:

$$E(Y|Z) = E(E(Y|X, Z))$$

7. **Conditional Variance:** Similar to conditional expectation, we can also define construct conditional variance. The conditional variance of Y given X is defined as the following:

$$Var(Y|X) = E\Bigg((Y - E(Y|X))^2|X\Bigg)$$

Equivalently, we have:

$$Var(Y|X) = E(Y^2|X) - E(Y|X)E(Y|X)$$

Again, just like the conditional expectation $E(Y|X)$, the conditional variance $Var(Y|X)$ is also a random variable – namely, a function of X!

8. **Eve's Law:** Sometimes, we want to find the marginal variance of Y, but it might be a bit difficult. At times, it may be easier to work with the conditional distribution of Y given X:

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

Intuitively, think of a population where each person has a value of X (their age) and a value of Y (their height). $E(Var(Y|X))$ can be interpreted as the "within-group variation," or the average amount of variation in height within each age group. $Var(E(Y|X))$ can be interpreted as the "between-group variation," which can be construed as the variance of average heights *across* different age groups. Eve's Law tells us that the total variation within the population is derived from the sum of these two sources of variation.

1.12 Statistical Inequalities

We begin by presenting a collection of commonly-used inequalities in statistics:

1. **Cauchy-Schwarz:** Let X and Y be r.v.s with finite variances. Then, we have:

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$$

2. **Jensen's Inequality:** let X be a random variable.

   (a) Let g be a convex function (a.k.a. "concave up"). Then, we have the following:

$$E(g(X)) \geq g(E(X)).$$

   (b) Let h be a concave function (a.k.a. "concave down"). Then, we have the following:

$$E(h(X)) \leq h(E(X)).$$

   Remember the caveman and his cave! If you can fit a caveman underneath the function g to protect him from the rain, then g is concave! Else, it is convex! Also, remember that *equality only holds* if g has a second derivative of 0 (i.e., g is just a straight line).

   "Convex" means that the 2nd derivative ≥ 0. "Concave" means the 2nd derivative ≤ 0. A straight line is *both convex and concave!*

3. **Markov's Inequality:** Let X be a random variable, and let $a > 0$ be a constant. Then, we have:

$$P(|X| \geq a) \leq \frac{E(|X|)}{a}$$

4. **Chebyshev's Inequality:** Let X be a random variable with mean μ and variance σ^2. Let $a > 0$ be a constant. Then, we have:

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

5. **Chernoff's Inequality (est. Harvard University, Professor Emeritus Herman Chernoff):** Let X be a random variable, and let $a > 0$ and $t > 0$ be constants. Then, we have:

$$P(X \geq a) \leq \frac{E(e^{tX})}{e^{ta}}$$

   Note that the numerator is the MGF of X, if it exists.

1.13 Limit Laws

1.13.1 Law of Large Numbers

For this section, let $X_1, X_2, X_3, \ldots$ be i.i.d random variables with finite mean μ and finite variance σ^2. For any positive integer n, define the following:

$$\bar{X}_n = \frac{X_1 + X_2 \cdots + X_n}{n}$$

$\bar{X}_n$ is referred to as the "sample mean" of X_1 through X_n. Note that the "sample mean" is itself *still* a random variable with mean μ and variance $\frac{\sigma^2}{n}$ (you can verify this using linearity of expectation and properties of variance). Now, we present two critical results regarding how sample means behave as n increases. Intuitively, the "sample mean" $\bar{X}_n$ "converges" to the true mean μ.

1. **Strong Law of Large Numbers (SLLN):** Intuitively, the sample mean $\bar{X}_n$ converges to the true mean μ "pointwise." Mathematically, we say that $P(\bar{X}_n \to \mu) = 1$.

2. **Weak Law of Large Numbers (WLLN):** The WLLN states that the sample mean $\bar{X}_n$ "converges in probability" to the true mean μ. Mathematically, for any $\epsilon > 0$, we have:

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0, \text{ as } n \to \infty$$

1.13.2 Central Limit Theorem

For this section, too, let $X_1, X_2, X_3, \ldots$ be i.i.d random variables with finite mean μ and finite variance σ^2. Again, for any positive integer n, let $\bar{X}_n$ denote the sample mean of X_1 through X_n:

$$\bar{X}_n = \frac{X_1 + X_2 \cdots + X_n}{n}$$

The WLLN and SLLN tell us that as $n \to \infty$, the sample mean will *eventually* converge to the true mean μ. But we also may want to know what distribution the sample mean adopts as $n \to \infty$. The Central Limit Theorem answers this question:

1. **Central Limit Theorem:** As $n \to \infty$, we have the following "asymptotic distribution" for $\bar{X}_n$:

   $$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to N(0,1)$$

   The left-hand-side of this expression is known as the "standardized form" of $\bar{X}_n$: we standardize a random variable by subtracting its mean and dividing by its standard deviation. Intuitively, the above expression means that as n gets large, the CDF of $\bar{X}_n$ will become closer and closer to Φ, the CDF of the standard Normal distribution.

2. **Central Limit Theorem, Approximation Form** Let n be a large positive integer. Then, we have the following approximate distribution for $\bar{X}_n$:

   $$\bar{X}_n \dot{\sim} N(\mu, \frac{\sigma^2}{n})$$

   *The "$\dot{\sim}$" symbol means "approximate distribution."

   The beauty of the Central Limit Theorem is that it does not care about what the original distribution of each individual X_1 is – the X_j could be Poisson, Geometric, Cauchy, or some weird (but valid) distribution with finite mean and variance, but in the long run, the sample mean will always converge to a normal distribution (with some terms and conditions beyond the scope of this course).

   **Note that both CLT and LLN do not apply to the Cauchy distribution (because the mean is not defined).*

   Long story short, the Law of Large Numbers tells us what specific value $\bar{X}_n$ will converge to as $n \to \infty$. We're not concerned with the distribution at $n = \infty$ (yes, mathematically an abomination, but we'll let it slide): we know that $\bar{X}_n$ will converge to a *constant* at $n = \infty$.

   The Central Limit Theorem tells us that on our long, windy, flower-filled path towards $n = \infty$ (yes, I know it's mathematically an abuse of notation), our $\bar{X}_n$ will have an *approximately* Normal distribution. The mean and variance of this approximate Normal distribution can be directly calculated using the tools we have at hand (like linearity of expectation, variance of a sum, etc.)

1.14 Markov Chains

1.14.1 Introduction and Markov Property

Definition: A **Markov chain** is a sequence of random variables $X_0, X_1, X_2, \ldots$ taking values in the *state space* $\{1, 2, \ldots, M\}$, where for all $n \geq 0$,

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i).$$

We call the above expression the **Markov property**, and in words, it states that given the entire past history $X_0, X_1, \ldots, X_n$, only the *most recent* term, X_n, matters for predicting X_{n+1}.

- If we think of time n as the present, then another way to conceptualize the Markov Property is that given the present, the past and the future are conditionally independent.

The quantity $P(X_{n+1} = j | X_n = i)$ is called the **transition probability** from state i to state j.

The **state space** of a Markov chain is the set of possible values of the X_n. Since we deal with Markov chains with a finite state space in this class, we can express the state space as the finite set $\{1, 2, \ldots, M\}$.

1.14.2 Transition Matrix

To describe the dynamics of a Markov chain, we need to specify the probabilities of moving from any state to any other state, that is, $P(X_{n+1} = j | X_n = i)$ for all $i, j \in \{1, \ldots, M\}$. We present the following concepts and definitions:

1. The **transition matrix** $Q = (q_{ij})$ accomplishes this by storing all of the transition probabilities $q_{ij} = P(X_{n+1} = j | X_n = i)$. Remember that i always indexes *rows* and j always indexes *columns*!

   (a) The (i, j) entry of Q (i.e. row i, column j) is the probability of moving from state i to state j in *one* step of the Markov chain.

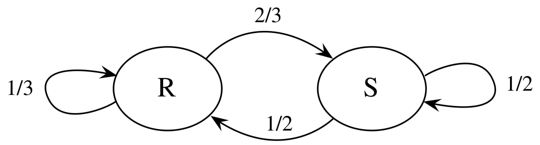
   (b) If our state space is $\{1, \ldots, M\}$, then Q will be an $M \times M$ matrix.

   (c) Each row of Q sums to 1. (Think about why this is so!)

   Example: Suppose we had the following transition matrix for a two-state Markov chain:

$$\begin{array}{c c} & \begin{array}{cc} S & R \end{array} \\ \begin{array}{c} S \\ R \end{array} & \begin{pmatrix} 1/2 & 1/2 \\ 2/3 & 1/3 \end{pmatrix} \end{array}$$

   These transition probabilities can also be represented with a diagram:



   Each state is represented by a circle, and each arrow indicates a possible one-step transition and the corresponding probability of that transition being realized.

2. **n-Step Transition Probabilities:** The n-step transition probability $q_{ij}^{(n)}$ is the probability of being at state j exactly n steps after being at state i. To find this, we can take the nth power of the transition matrix Q, through matrix multiplication:

$$q_{ij}^{(n)} \text{ is the } (i,j) \text{ entry of } Q^n.$$

3. **Marginal Distributions:** We can get the marginal distribution of X_n as follows:

$$P(X_n = j) \text{ is the } j\text{th component of } \mathbf{t}Q^n,$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_M)$ and $t_i = P(X_0 = i)$. Note that this means we need to know not only the transition matrix but also the *initial conditions* of the Markov chain. We obtain the above result through LOTP, conditioning on X_0. The probability that the chain is in state j after n steps is

$$P(X_n = j) = \sum_{i=1}^{M} P(X_0 = i)P(X_n = j|X_0 = i) = \sum_{i=1}^{M} t_i q_{ij}^{(n)},$$

which is the jth component of $\mathbf{t}Q^n$ by definition of matrix multiplication.

1.14.3 Classification of States (Transience and Periodicity)

Here, we introduce terminology for describing the various characteristics of a Markov chain.

Properties of States

1. **Recurrent:** A state is recurrent if it can be visited over and over again in the long run.
2. **Transient:** A state is transient if there is a positive probability of never returning to it.
3. **Period:** The period of a state i is the greatest common divisor of all the possible numbers of steps it can take to return to state i when starting at i.
4. **Periodic:** A state is periodic if its period does not equal 1.
5. **Aperiodic:** A state is aperiodic if its period equals 1.

Properties of Markov Chains as a Whole

1. **Irreducible:** A Markov chain is irreducible if for any two states i and j, it is possible to go from i to j in a finite number of steps. This means that we can get to any state from any other state. Irreducible implies that **all states are recurrent** and that **all states have the same period**.
2. **Reducible:** A Markov chain that is not irreducible is considered reducible.
3. **Aperiodic:** A Markov chain is aperiodic if all of its states are aperiodic.
4. **Periodic:** A Markov chain that is not aperiodic is considered periodic.

*note: "Aperiodic" and "periodic" are used to describe both individual states and Markov Chains as a whole! Double-check the context in which these terms are used!

1.14.4 Stationary Distribution

In the long run, Markov chains will eventually spend all its time in recurrent states – but what fraction of the time will it spend in each of the recurrent states? This question is answered by the *stationary distribution* of the chain. For irreducible and aperiodic Markov chains, the stationary distribution tells us both the long-run probability of being in any state, and the long-run proportion of time the chain spends in each state, *regardless of initial conditions.* We provide the following definitions and properties:

1. **Definition:** A row vector $\mathbf{s} = (s_1, \ldots, s_M)$ (where $s_i \geq 0$ and $\sum_i s_i = 1$ is a **stationary distribution** for a Markov chain with transition matrix Q if

$$\sum_i s_i q_{ij} = s_j$$

for all j. We can also write this as one matrix equation:

$$\mathbf{s}Q = \mathbf{S}.$$

Intuitively, this states that if we make move while in s, we will stay in s forever, so s is the stationary distribution.

To solve for the stationary distribution, you can solve the following matrix equation for $\mathbf{s}$: $(Q' - I)(\mathbf{s}') = 0$. The $'$ symbol denotes taking the transpose, and I is the identity matrix with the same dimensions as Q.

2. **Eigenvector Connection:** In linear algebra terminology, $\mathbf{s}$ is a left eigenvector of Q with eigenvalue 1. To work with *right eigenvectors* (the calculation of which is more often taught in linear algebra classes), note that $\mathbf{s}$ is a right eigenvector of Q^T, where the T operator means taking the transpose of matrix Q.

3. **Stationary Distribution is Marginal:** Note that $\mathbf{s}$ is the *marginal* distribution of X_n. The conditional PMF of X_n given $X_{n-1} = i$ is still given by the ith row of the transition matrix Q.

4. **Existence and Uniqueness:** For every *irreducible* Markov chain, there exists a unique stationary distribution.

5. **Convergence:** Let $X_0, X_1, \ldots$ be an irreducible, aperiodic Markov chain with stationary distribution $\mathbf{s}$ and transition matrix Q. If we run this chain for a long time, the marginal distribution of X_n will converge to $\mathbf{s}$ regardless of initial conditions. We write this as $P(X_n = i) \to s_i$ as $n \to \infty$. In terms of Q, Q^n converges to a matrix in which each row is $\mathbf{s}$.

6. **Expected Time to Return:** Say we had an irreducible Markov chain with stationary distribution $\mathbf{s}$, and let r_i be the expected time it takes the chain to return to state i, given it starts at i. Then

$$s_i = \frac{1}{r_i}.$$

1.14.5 Reversibility

The reversibility condition for a Markov chain helps us find the stationary distribution in the scenario that we have a large state space and it is difficult to compute. We provide the following definitions and concepts:

1. **Definition:** Let Q be the transition matrix of a Markov chain. Suppose there is $\mathbf{s} = (s_1, \ldots, s_M)$ with $s_i \geq 0, \sum_i s_i = 1$. If

$$s_i q_{ij} = s_j q_{ji}$$

for all states i and j, then the chain is **reversible** with respect to $\mathbf{s}$. Intuitively, reversibility means that the chain behaves the same way regardless of whether time runs forwards or backwards. If you record a video of a reversible chain, and then show the video to a friend, either in the normal way or with time reversed, your friend will not be able to tell whether time is running forwards or backwards in the video.

2. **Reversible implies Stationary:** If a chain is reversible with respect to some row vector $\mathbf{s}$, then we know that $\mathbf{s}$ is the stationary distribution. Proof is by the following equality:

$$\sum_i s_i q_{ij} = \sum_i s_j q_{ji} = s_j \sum_i q_{ji} = s_j,$$

where the last equality holds because each row of Q sums to 1.

1.14.6 Three Specific Types of Reversible Markov Chains, and their Stationary Distributions

In this section, we introduce three types of reversible Markov chains with stationary distributions that are quick and easy to find:

1. **Symmetric Matrix**

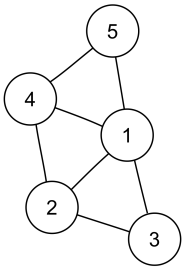
   (a) If Q is a symmetric matrix (i.e. $q_{ij} = q_{ji}$), then the reversibility condition is satisfied, and the stationary distribution is uniform over the state space:
   $$\mathbf{s} = (1/M, 1/M, \ldots, 1/M).$$

   (b) A generalization of (1): if each *column* of the transition matrix Q sums to 1, then
   $$\mathbf{s} = (1/M, 1/M, \ldots, 1/M)$$
   is a stationary distribution.

2. **Random Walk on an Undirected Network** Think of a network as a collection of *nodes* joined by *edges*; the network is *undirected* if direction does not matter, i.e. edges can be traversed in any direction. The *degree* of a node is the number of edges attached to it, and the *degree sequence* of a network with nodes $1, 2, \ldots, n$ is the vector $\mathbf{d} = (d_1, \ldots, d_n)$, where d_j is the degree of node j.

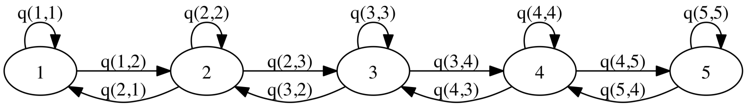
   (A self-loop – i.e. edge from node to itself – counts 1 towards the degree of a node.)

   Example: the following undirected network has degree sequence $\mathbf{d} = (4, 3, 2, 3, 2)$.



   If a Markov chain is a walk on an undirected network, then
   $$d_i q_{ij} = d_j q_{ji},$$
   and the stationary distribution is proportional to the degree sequence. In the example above, then,
   $$\mathbf{s} = \left(\frac{4}{14}, \frac{3}{14}, \frac{2}{14}, \frac{3}{14}, \frac{2}{14} \right).$$

3. **Birth-Death Chain:** A *birth-death chain* is a type of Markov chain where it is only possible to go one step to the left or one step to the right (except at boundaries), but it is impossible to jump farther than one step. Formally, $q_{ij} > 0$ if $|i - j| = 1$, and $q_{ij} = 0$ if $|i - j| \geq 2$. Below is an example of a birth-death chain (loops from a state to itself are allowed to have 0 probability):



   A birth-death chain is reversible! We can construct the stationary distribution by letting
   $$s_j = \frac{s_1 q_{12} q_{23} \cdots q_{j-1,j}}{q_{j,j-1} q_{j-1,j-2} \cdots q_{21}}$$
   for all states $2 \leq j \leq M$. We then choose s_1 such that the s_j sum to 1.

1.14.7 Markov Chain Monte Carlo

A *Monte Carlo* method involves generating random values to approximate a quantity. We study *Markov Chain Monte Carlo*, a class of algorithms that allows us to, given some stationary distribution $\mathbf{s}$, essentially *build our own Markov chain* whose stationary distribution is s. Specifically, we look at the **Metropolis-Hastings** algorithm within the MCMC class:

Let $\mathbf{s} = (s_1, \dots, s_M)$ be our desired stationary distribution, and suppose $P = (p_{ij})$ is the transition matrix for the Markov chain we currently have. We can run P, but it does not have the desired stationary distribution. Our goal is to modify P and construct a *new* Markov chain $X_0, \dots, X_n, \dots$ so that it does have our desired stationary distribution $\mathbf{s}$.

In the algorithm, we start at any state X_0, and suppose we are currently at X_n. To make one move of the new chain, we do the following:

1. If $X_n = i$, propose a new state j.
2. Compute the *acceptance probability*:
$$a_{ij} = \min\left(\frac{s_j p_{ji}}{s_i p_{ij}}, 1\right).$$
3. With probability a_{ij}, accept the proposal (i.e., go to j), setting $X_{n+1} = j$. Otherwise, stay at i (i.e. $X_{n+1} = i$).

Table of Distributions

Note that discrete and continuous distributions are separated by double horizontal lines.

Distribution	PDF and Support	Expectation	Variance
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q$	p	pq
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq
Geometric Geom(p)	$P(X=k)=q^kp$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2
Negative Binom. NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^rq^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$		
Hypergeometric Hypergeometric(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\frac{w+b-n}{w+b-1}n\frac{\mu}{n}(1-\frac{\mu}{n})$
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ	λ
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x\,-\,\mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ	σ^2
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$^1/_\lambda$	$^1/_{\lambda^2}$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^ae^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	$^a/_\lambda$	$^a/_{\lambda^2}$
Beta Beta(a, b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$
Chi-Squared χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,1)$	n	$2n$
Multinomial Mult$_k(n,\vec{p})$	$P(\vec{X}=\vec{n})=\binom{n}{n_1\dots n_k}p_1^{n_1}\dots p_k^{n_k}$ $n=n_1+n_2+\dots+n_k$	$n\vec{p}$	$\mathrm{Var}(X_i)=np_i(1-p_i)$ $\mathrm{Cov}(X_i,X_j)=-np_ip_j$

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