Lecture 1 - **Probability Theory Foundations**

BIOENG-210 Course Notes
Prof. Gioele La Manno

February 2024

1 Introduction to Random Variables

In the realm of probability theory, we encounter a fundamental shift in how we think about variables. While traditional mathematics deals with deterministic relationships, probability theory introduces us to the concept of random variables—a cornerstone of statistical analysis and modeling.

1.1 From Deterministic to Random

Traditional mathematical variables follow deterministic relationships: given an input x, a function $f(x)$ will always produce the same output y. For instance, in the equation $y = 2x + 1$, if we input $x = 2$, we will always get $y = 5$. This predictability characterizes deterministic relationships.

Random variables, however, introduce uncertainty into this framework. Instead of providing a single, deterministic output, a random variable is essentially a function that associates probabilities with possible outcomes. When we work with random variables, we shift our focus from asking ”What is the value?” to ”What is the probability of obtaining a particular value?”

Consider rolling a fair six-sided die. If we want to predict the outcome:

Deterministic approach: If we knew all the physics involved—the exact force applied, air resistance, surface friction, and initial position—we could theoretically calculate the exact number that would appear. Random variable approach: We can only say that each number (1 through 6) has a probability of 1/6 of occurring. Even with the same person rolling the same die in seemingly the same way, the outcome varies randomly according to these probabilities.

1.2 Formal Definition

Definition 1.1 (Measurable Set)**.** A set A in a measure space $(\Omega, \mathcal{F})$ is measurable if $A \in \mathcal{F}$, where $\mathcal{F}$ is a σ-algebra: a collection of subsets of Ω that is closed under complementation and countable unions.

Definition 1.2 (Borel Sets)**.** The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals in $\mathbb{R}$. Sets in $\mathcal{B}(\mathbb{R})$ are called Borel sets.

Definition 1.3 (Random Variable)**.** A random variable X is a function that maps outcomes from a sample space Ω to the real numbers $\mathbb{R}$, such that for any Borel set $B \subseteq \mathbb{R}$, the set $\{w \in \Omega : X(w) \in B\}$ is measurable.

While this definition might seem abstract, its practical implications are profound. A random variable allows us to:

- Assign numerical values to outcomes of random experiments
- Calculate probabilities of events using these numerical values

- Apply mathematical operations to uncertain quantities

This measurability condition is crucial as it ensures that we can assign probabilities to events involving the random variable. The Borel sets include all intervals and other ”well-behaved” subsets of the real line, allowing us to work with probabilities of events like $P(a \leq X \leq b)$ or $P(X \in S)$ for any reasonable set S. To put it simply, measurability ensures that we can meaningfully assign probabilities to sets of outcomes, while Borel sets give us a concrete way to work with subsets of real numbers while maintaining the mathematical properties we need for probability theory.

2 Support and Types of Random Variables

2.1 Support of a Random Variable

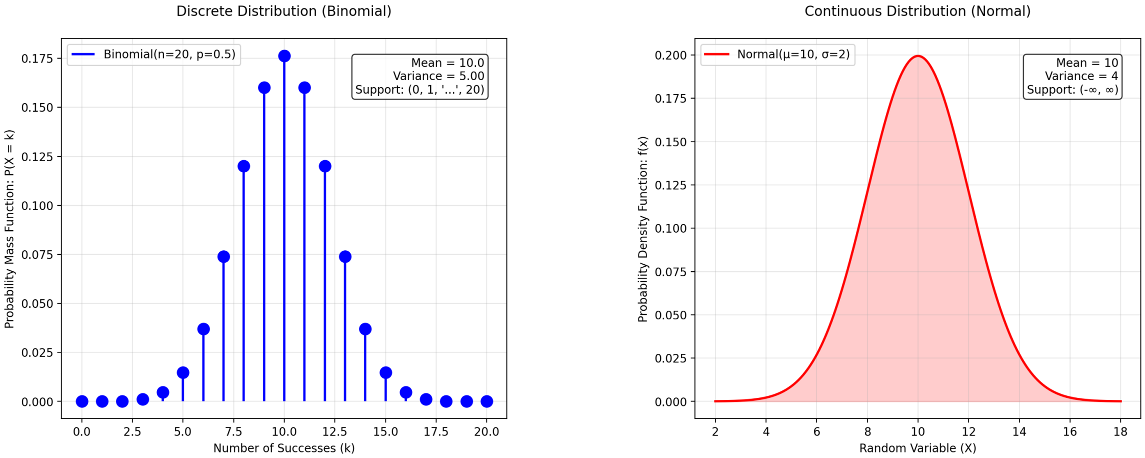
Definition 2.1 (Support)**.** The support of a random variable X, denoted $\text{Supp}(X)$, is the set of all possible values that X can take with non-zero probability (for discrete variables) or positive density (for continuous variables).

Understanding the support is crucial because it:

- Defines the domain where the random variable is meaningful
- Helps in identifying appropriate probability distributions
- Is essential for calculating probabilities and expectations

2.2 Types of Random Variables

Random variables can be categorized into two main types: discrete and continuous.



(a) Example of a Discrete RV

(b) Example of a Continous RV

Figure 1: **Discrete vs. Continous Random Variables**

2.2.1 Discrete Random Variables

Definition 2.2 (Discrete Random Variable)**.** A random variable X is discrete if it takes values from a countable set. Its probability distribution is described by a probability mass function (PMF) $p_X(x)$, where:

$$p_X(x) = P(X = x)$$

Properties of PMFs:

Property 2.3. For any PMF $p_X(x)$:

1. $p_X(x) \geq 0$ for all x
2. $\sum_{x \in \text{Supp}(X)} p_X(x) = 1$
3. $P(X \in A) = \sum_{x \in A} p_X(x)$ for any subset A of the support

Example 2.4 (Discrete Random Variable)**.** Consider rolling a fair six-sided die. The random variable X representing the outcome has:

- Support: $\text{Supp}(X) = \{1, 2, 3, 4, 5, 6\}$
- PMF: $p_X(x) = \frac{1}{6}$ for $x \in \text{Supp}(X)$

2.2.2 Continuous Random Variables

Definition 2.5 (Continuous Random Variable)**.** A random variable X is continuous if it takes values from an uncountable set (typically an interval). Its probability distribution is described by a probability density function (PDF) $f_X(x)$, where:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Properties of PDFs:

Property 2.6. For any PDF $f_X(x)$:

1. $f_X(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. $P(X \in A) = \int_A f_X(x)dx$ for any measurable set A
4. $P(X = x) = 0$ for any single point x

Example 2.7 (Continuous Random Variable)**.** The height of adults in a population is often modeled as a continuous random variable with a normal distribution. If X represents height:

- Support: $\text{Supp}(X) = (0, \infty)$
- PDF: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

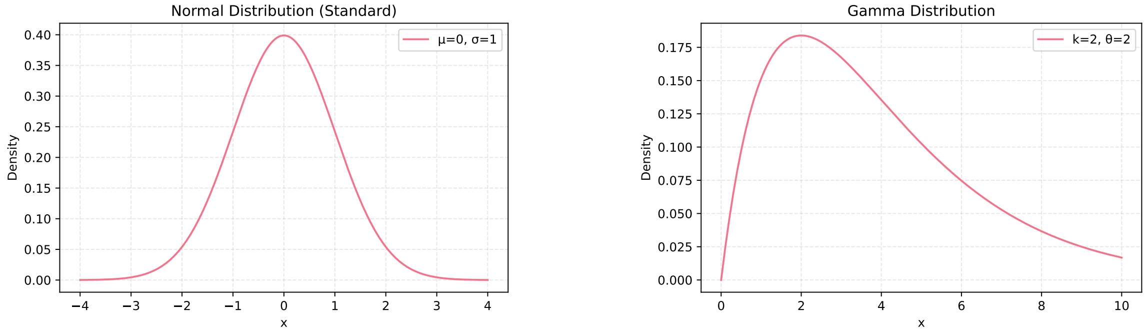
3 Functions Describing Random Variables

To fully characterize a random variable, we need several functions that provide different perspectives on its probability distribution.

3.1 Probability Mass/Density Functions

As introduced earlier, PMFs and PDFs are the most basic descriptions of probability distributions. They tell us:

- For discrete RVs: The exact probability of each outcome
- For continuous RVs: The relative likelihood of values in any region



(a) PDF of a Normal Distribution

(b) PDF of a Gamma Distribution

Figure 2: **Examples of PDFs**

3.2 Cumulative Distribution Function (CDF)

Definition 3.1 (CDF)**.** The cumulative distribution function $F_X(x)$ of a random variable X is defined as:

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

Properties of CDFs:

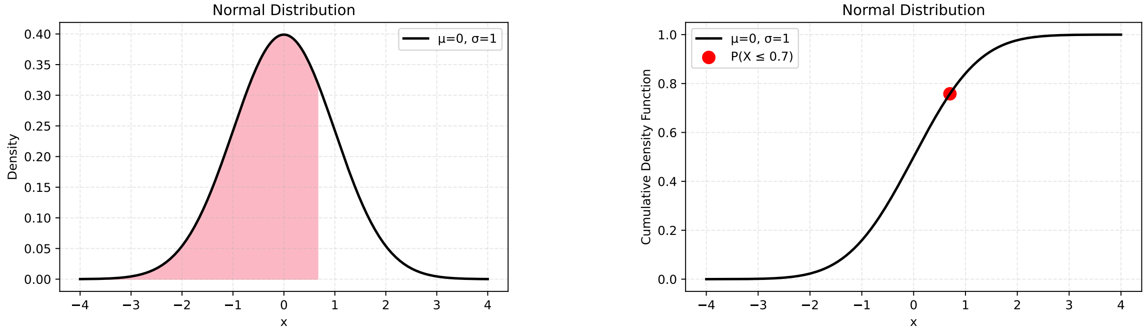
Property 3.2. For any CDF $F_X(x)$:

1. $F_X(x)$ is non-decreasing
2. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
3. $F_X(x)$ is right-continuous
4. For continuous RVs: $F_X(x)$ is continuous

Theorem 3.3 (Relationship between PDF and CDF)**.** For a continuous random variable with PDF $f_X(x)$:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

$$f_X(x) = \frac{d}{dx} F_X(x) \text{ (where the derivative exists)}$$



(a) Area corresponding to the CDF

(b) The CDF of a Normal distribtuion

Figure 3: **Representation of the CDF as an integral**

3.3 Quantile Function

Definition 3.4 (Quantile Function)**.** The quantile function $Q_X(p)$, also known as the inverse CDF, is defined for $0 < p < 1$ as:

$$Q_X(p) = \inf\{x : F_X(x) \geq p\}$$

The quantile function is particularly useful for:

- Finding percentiles of a distribution
- Generating random samples from a distribution
- Comparing different distributions

Example 3.5 (Quantile Function)**.** For a standard normal distribution:

- The median is $Q_X(0.5) = 0$
- The 95th percentile is $Q_X(0.95) \approx 1.645$

3.4 Additional Characterizing Functions

Several other functions are useful for describing random variables:

- **Survival Function:** $S_X(x) = P(X > x) = 1 - F_X(x)$
- **Hazard Function:** $h_X(x) = \frac{f_X(x)}{1-F_X(x)}$ (for continuous RVs)
- **Moment Generating Function:** $M_X(t) = E[e^{tX}]$
- **Characteristic Function:** $\phi_X(t) = E[e^{itX}]$

Each of these functions provides unique insights into the properties and behavior of the random variable.

4 Understanding Average Behavior: The Expectation

When we deal with uncertainty, one of our first questions is often ”What should we expect?” This seemingly simple question leads us to one of probability theory’s most powerful concepts: expectation.

4.1 The Intuition Behind Expectation

Imagine you’re a biologist studying the length of bacterial cells. If you could measure an infinite number of cells and take their average, what value would you get? This hypothetical ”infinite average” is essentially what expectation represents. It’s not just any average—it’s the theoretical average that emerges when random chance has played out infinitely many times.

4.1.1 A Historical Perspective

The concept of expectation has fascinating origins in gambling mathematics. In the 17th century, Blaise Pascal and Pierre de Fermat exchanged letters discussing the ”fair value” of unfinished games of chance. Their insights led to the formal concept of mathematical expectation—a way to quantify what we ”expect” to happen in the long run.

4.2 Formal Definition and Interpretation

Let's build the concept step by step:

Definition 4.1 (Expectation)**.** For a random variable X, the expectation $E[X]$ is:

$$E[X] = \begin{cases} \sum_x x p_X(x) & \text{for discrete X} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{for continuous X} \end{cases}$$

This formula might look intimidating, but it tells a simple story:

- For discrete variables: We multiply each possible value by its probability and sum them up
- For continuous variables: We do the same thing, but with infinitely many values using calculus

Example 4.2 (The Loaded Die)**.** Consider a loaded six-sided die where:

- The number 6 appears with probability 1/3
- All other numbers appear with probability 2/15 each

The expectation is:

$$E[X] = 1(\frac{2}{15}) + 2(\frac{2}{15}) + 3(\frac{2}{15}) + 4(\frac{2}{15}) + 5(\frac{2}{15}) + 6(\frac{1}{3}) = 4$$

Interestingly, even though 4 isn't the most likely outcome, it's our ”expected” value!

4.3 Properties That Make Life Easier

The true power of expectation lies in its properties. These aren't just mathematical rules—they're tools that help us understand complex situations by breaking them down into simpler parts.

4.3.1 The Linearity Property

Perhaps the most beautiful property of expectation is its linearity:

$$E[aX + b] = aE[X] + b$$

This means that expectations ”play nice” with linear operations. If you're measuring bacterial lengths in micrometers and want to convert to millimeters, you can:

- Either: Convert each measurement and then take the expectation
- Or: Take the expectation first and then convert
- Both give the same result!

4.3.2 The Additivity Property

Another powerful property is additivity:

$$E[X + Y] = E[X] + E[Y]$$

This holds regardless of whether X and Y are independent! Think about its implications:

- Measuring total protein concentration? Add the expected concentrations of individual proteins
- Calculating total waiting time? Add the expected times for each stage

4.3.3 Expectation of a Product

A particularly subtle and important property concerns the expectation of a product. When we multiply random variables, the behavior of expectation becomes more complex:

Theorem 4.3 (Expectation of a Product)**.** For random variables X and Y:

$$E[XY] = E[X]E[Y] \text{ if and only if } X \text{ and } Y \text{ are independent}$$

This leads to an important observation about squared random variables:

$$E[X^2] = E[X \cdot X] \neq E[X]^2$$

This inequality is fundamental and captures a deep truth about random variables: if X has any variability at all, $E[X^2]$ will be strictly greater than $E[X]^2$. The difference between these quantities, in fact, defines the variance of X, which measures the spread of its distribution.

Example 4.4 (Understanding why)**.** Consider rolling a fair die where X is the outcome:

- $E[X] = 3.5$, so $E[X]^2 = 12.25$
- $E[X^2] = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)/6 = 91/6 \approx 15.17$
- The difference $E[X^2] - E[X]^2 \approx 2.92$ is positive and represents the variance, measuring how spread out the distribution is from its mean
- This difference must be positive for any non-constant random variable, reflecting the fundamental presence of variability in the distribution

Definition 4.5 (Conditional Expectation)**.** The conditional expectation $E[X|Y]$ represents the expected value of X given knowledge of Y. It can be thought of as:

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$$

for continuous variables, with an analogous sum for discrete variables.

A fundamental property is the Law of Total Expectation:

$$E[X] = E[E[X|Y]]$$

This seemingly circular statement has profound implications for probability theory and statistics.

Definition 4.6 (Marginal Expectation)**.** The marginal expectation of a random variable X can be computed from the marginal distribution $f_X(x)$:

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

for continuous variables, or

$$E[X] = \sum_x x\, P(X = x)$$

for discrete variables.

This relates to the conditional expectation through marginalization:

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

Example 4.7 (Marginal vs Conditional Expectation)**.** Consider a joint distribution of student grades X and study hours Y:

- The marginal expectation $E[X]$ gives the average grade across all students, regardless of study time.
- The conditional expectation $E[X|Y = y]$ gives the average grade for students who studied y hours.
- We can recover $E[X]$ by averaging over all possible study hours:

$$E[X] = \int E[X|Y = y]\, f_Y(y)\, dy.$$

This connection between marginal and conditional expectations provides a bridge between overall population-level statistics and more detailed conditional analyses. It is particularly useful in statistical modeling, where we often need to move between these different levels of analysis.

5 Moments of a Random Variable

The concept of moments provides a systematic way to describe the shape and properties of a probability distribution. Think of moments as increasingly detailed ”measurements” of a distribution's characteristics.

Moments provide information about the shape and spread of the distribution of X. The k-th moment of a random variable X is defined as the expectation of X^k, denoted $E[X^k]$:

$$E[X^k] = \int_{-\infty}^{\infty} x^k p(x)\, dx$$

- **First Moment (Mean):** $E[X]$
- **Second Moment:** $E[X^2]$. (captures both the mean and the spread of X around the mean, related to variance).

  Do not confuse with Variance (see below): Defined as $\mathrm{Var}(X) = E[X^2] - (E[X])^2$
- **Higher-Order Moments:**
  - Third Moment (related to Skewness): relates to the asymmetry of the distribution around the mean.
  - Fourth Moment (related to Kurtosis): relates to the “tailedness” of the distribution.

5.1 Central Moments

Central moments are particularly important as they describe the distribution's shape relative to its mean.

Definition 5.1 (Central Moments)**.** The kth central moment of a random variable X is defined as:

$$\mu_k = E[(X - E[X])^k]$$

The first few central moments have special interpretations:

- First central moment ($k = 1$): Always zero (by definition)
- Second central moment ($k = 2$): Variance
- Third central moment ($k = 3$): Related to skewness (asymmetry)
- Fourth central moment ($k = 4$): Related to kurtosis (tail behavior)

5.2 Variance of a Random Variable

Now we can properly understand variance as the second central moment:

Definition 5.2 (Variance)**.** The variance of a random variable X is:

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

This definition connects our earlier discussion of product expectations with the concept of spread in a distribution. The equality $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ is particularly useful in calculations and helps explain why $E[X^2] \neq E[X]^2$ is so important.

Key properties of variance include:

Property 5.3. For random variables X and Y:

1. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ for any constant a
2. $\mathrm{Var}(X + b) = \mathrm{Var}(X)$ for any constant b
3. If X and Y are independent: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Property 5.4 (Zoom-in: Variance of the Sum of Independent Variables)**.** For any two independent random variables X and Y, the variance of their sum satisfies:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Theoretical Explanation: In general, the variance of the sum of two random variables is given by:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y),$$

where the covariance $\mathrm{Cov}(X, Y) = E\big[(X - E[X])(Y - E[Y])\big]$ measures the degree to which X and Y change together.

When X and Y are independent, their joint behavior is completely unrelated. This independence implies that the covariance is zero:

$$\mathrm{Cov}(X, Y) = 0.$$

Therefore, the formula simplifies to:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

This property is particularly useful because it allows us to compute the overall variability (or uncertainty) of the sum of independent random variables by simply adding their individual variances. This principle finds applications in many areas such as error analysis, finance, and signal processing.

Example: Tossing Two Fair Coins Consider two independent coin tosses. Let:

- X be an indicator variable for the outcome of the first coin, where $X = 1$ if the coin shows heads and $X = 0$ if tails.
- Y be similarly defined for the second coin.

Since the coins are fair, the probability of heads is $P(X = 1) = P(Y = 1) = 0.5$, and the probability of tails is $P(X = 0) = P(Y = 0) = 0.5$. For a Bernoulli random variable (like X or Y) with parameter $p = 0.5$, the variance is:

$$\mathrm{Var}(X) = p(1 - p) = 0.5 \times 0.5 = 0.25.$$

Similarly, $\mathrm{Var}(Y) = 0.25$.

Since the coin tosses are independent, the variance of the sum $X + Y$ is:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 0.25 + 0.25 = 0.5.$$

Thus, the sum $X + Y$ (which can take on the values 0, 1, or 2) has a variance of 0.5. This example clearly illustrates how the property simplifies the calculation of the variance when dealing with independent random variables.

Practical Implication: In practice, this property is extremely valuable. For instance, if multiple independent measurements are taken in an experiment, the overall uncertainty in the total measurement is just the sum of the uncertainties (variances) of the individual measurements. This insight helps in designing experiments and in understanding the behavior of combined random processes.

6 The Art and Science of Parametrizing Distributions

In the study of natural phenomena, we often encounter patterns in how random variables are distributed. These patterns, far from being arbitrary, can be captured through mathematical functions with specific parameters. The concept of parametrization lies at the heart of how we model and understand randomness in the real world.

6.1 The Philosophy of Parametric Distributions

When we say a random variable follows a particular distribution with certain parameters, what do we really mean? This seemingly simple statement carries deep implications about how we understand and model random phenomena.

Definition 6.1 (Parametric Distribution)**.** A parametric distribution is a family of probability distributions indexed by a set of parameters θ. We denote the probability density function as $p_\theta(x)$ to emphasize its dependence on these parameters.

The notation $p_\theta(x)$ tells us something profound: the same basic "shape" or "family" of distributions can generate different specific distributions based on the values we choose for θ. Think of it as having a basic recipe (the distribution family) that can be adjusted (through parameters) to create different variations, each suited to a particular situation.

Consider modeling the distribution of cell sizes across different tissue samples. We might assume that cell sizes follow a normal distribution, where $p_\theta(x)$ represents this distribution with parameters $\theta = (\mu, \sigma^2)$. In one tissue, the cells might have a larger average size and more variability, while in another, they might be smaller and more uniform. By simply adjusting the values of μ and σ^2, the same normal distribution family can be tailored to fit the unique characteristics of each tissue type. This illustrates the power of parametric distributions: they provide a flexible framework that can capture diverse biological phenomena through appropriate parameter choices.

6.2 Understanding Parameters: Beyond Location and Scale

A common misconception in probability theory is that parameters always correspond to intuitive properties like location (where the distribution is centered) or scale (how spread out it is). While this simple interpretation works for some distributions, like the normal distribution where μ represents location and σ^2 represents scale, reality is often more complex.

Consider the gamma distribution with parameters k and θ. Here, the parameters interact in subtle ways - k affects both the shape and location of the distribution, while θ acts as a scaling factor. Even more intricate is the beta distribution, where both parameters α and β simultaneously influence the shape and location of the distribution in ways that defy simple categorization.

6.3 The General Form of Parametric Distributions

When we write $p_\theta(x)$, we're expressing that our distribution is a function of the form $f(x, \theta)$. This notation emphasizes the dual nature of probability distributions: they depend both on the random variable x (what we observe) and the parameters θ (what we need to specify). Different values of θ give us different members of the same family of distributions, like different variations on a theme in music.

Consider the exponential distribution, a cornerstone of survival analysis and reliability theory. Its probability density function is given by:

$$p_\lambda(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Here, λ acts as our parameter, controlling how quickly the distribution decays. While all exponential distributions maintain their characteristic "decay" shape, the parameter λ determines whether this decay happens rapidly or gradually. This single parameter captures the essential behavior of phenomena as diverse as radioactive decay and customer service times.

6.4 The Impact on Biological Data Science

In biological data science, the choice of parametrization can profoundly affect our ability to model and understand natural phenomena. Consider gene expression levels, which often follow a log-normal distribution:

$$p_{\mu,\sigma^2}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

This parametrization naturally captures the multiplicative nature of biological processes. When we observe gene expression, the parameters μ and σ^2 tell us not just about the location and spread of the data, but about the underlying biological mechanisms that generate this pattern.

Even more interesting is the case of RNA-seq count data, where the negative binomial distribution has become the standard model. Its parametrization in terms of mean μ and dispersion θ directly connects to the biological reality of gene expression:

$$p_{\mu,\theta}(k) = \binom{k+\theta-1}{k} \left(\frac{\theta}{\theta+\mu}\right)^\theta \left(\frac{\mu}{\theta+\mu}\right)^k$$

This formulation emerged not just from mathematical convenience, but from the need to model both the average expression level and the biological variability independently.

6.5 The Art of Choosing Parametrizations

Choosing the right parametrization is as much an art as it is a science. Good parametrizations should serve multiple masters: they must be interpretable, allowing scientists to connect parameters to real-world phenomena; they must be statistically tractable, enabling efficient estimation and inference; and they must be computationally convenient, facilitating numerical calculations and stability.

Sometimes these goals conflict. A parametrization that makes perfect biological sense might be difficult to estimate from data, or one that's computationally efficient might obscure the underlying biology. The skill lies in finding the right balance for the problem at hand.

7 Looking Forward

As we move forward to study specific distributions and their applications, the importance of parametrization will become increasingly clear. We'll see how different parametrizations of the same distribution can illuminate different aspects of the phenomena we study, and how choosing the right parametrization can make the difference between insight and confusion. The concepts we've explored here will prove essential as we delve into parameter estimation, hypothesis testing, and the broader landscape of statistical inference in biological data science.

8 Common Distributions in Biological Data Science

When we study biological systems, certain patterns emerge again and again. These patterns, captured by specific probability distributions, aren't just mathematical conveniences—they often reflect fundamental biological processes. Let's explore these distributions and understand why they appear so frequently in biological data.

8.1 The Normal and Log-Normal Duo: A Tale of Two Distributions

The normal and log-normal distributions form a fascinating pair, each telling us something different about biological processes. Their relationship offers deep insights into how biological variables behave.

8.2 The Normal Distribution: A Mathematical Marvel

The story of the normal distribution is one of the most fascinating in the history of mathematics and science. It begins in the 18th century with Abraham de Moivre, who was studying the behavior of coin flips and gambling games. He discovered something remarkable: as you sum more and more random variables, their distribution tends toward a specific bell-shaped curve. This observation would later become known as the Central Limit Theorem, one of the most profound results in probability theory.

However, it was Carl Friedrich Gauss who would cement this distribution's place in scientific history. While working on the method of least squares for astronomical calculations, Gauss needed to understand the distribution of measurement errors. He sought a function that would represent the "typical" pattern of random deviations from a true value. Let's follow a path similar to how Gauss might have reasoned about this distribution.

Start by imagining a histogram of measurement errors. Common sense and observation tell us several things:

- Small errors should be more common than large ones
- Positive and negative errors should be equally likely (symmetry)
- Very large errors should be increasingly rare
- The total probability should sum to one

How might we construct such a function mathematically? Let's build it step by step:

- The simplest function that creates a "hill" shape is the negative square:

  $$-x^2$$

  This gives us symmetry and decreasing values as we move away from zero.

- To ensure our function stays positive and approaches zero for large values (rather than negative infinity), we can use the exponential function:

  $$e^{-x^2}$$

  This is our basic "bell shape."

- To control the width of the bell, we need a scale parameter σ. Dividing by σ^2 (squaring as guarantee to have it positive) gives us this control:

$$e^{-x^2/\sigma^2}$$

- For flexibility in location, we want to center our bell anywhere on the real line. Subtracting a location parameter μ from x accomplishes this:

$$e^{-(x-\mu)^2/\sigma^2}$$

- For our function to be a proper probability density, we need:

$$\int_{-\infty}^{\infty} Ae^{-\frac{(x-\mu)^2}{\sigma^2}} dx = 1$$

The solution follows in three key steps:

a) Make the substitution $z = \frac{x-\mu}{\sigma}$, giving:

$$A\sigma \int_{-\infty}^{\infty} e^{-z^2} dz = 1$$

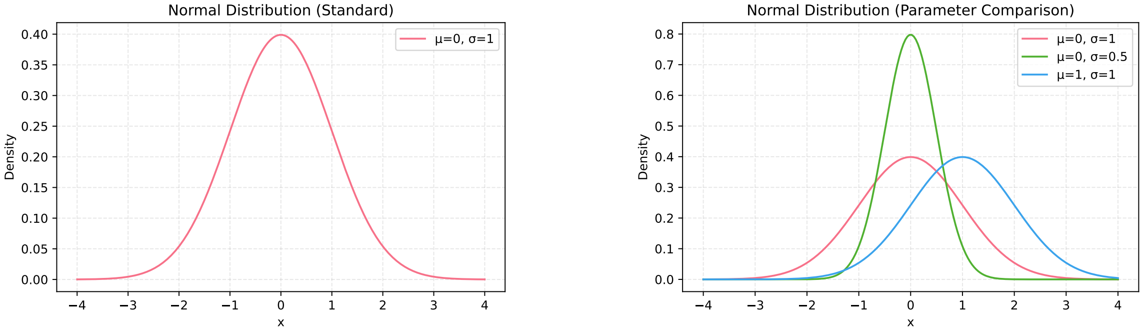
b) Remind the integral $\int_{-\infty}^{\infty} e^{-z^2} dz$ equals $\sqrt{\pi}$ (easy to see in polar coordinates)

c) Therefore:

$$A\sigma\sqrt{\pi} = 1 \quad \Rightarrow \quad A = \frac{1}{\sigma\sqrt{\pi}}$$

And this could be it! For mathematical convenience (e.g. simplify moments) and elegance we introduce the factor of $\frac{1}{2}$ in the exponent and $\frac{1}{\sqrt{2\pi}}$ to sum to 1:

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


Normal Distribution (Standard)

Normal Distribution (Parameter Comparison)

(a) Standard normal distribution

(b) Normal - varying parameters

Figure 4: **The Log-Normal Distribution: Multiplicative effects in biological processes**

8.2.1 The Log-Normal Distribution: Multiplicative Biology

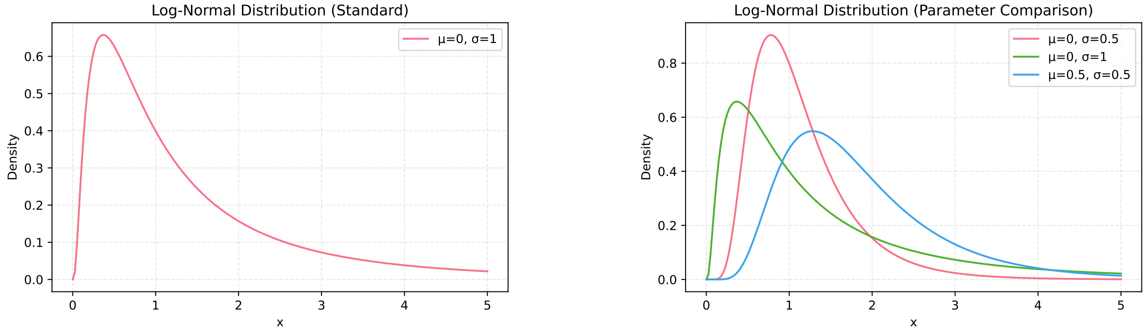
Enter the log-normal distribution. When we say a variable X is log-normally distributed, we mean that ln(X) follows a normal distribution. Its density function,

$$f_{\mu,\sigma^2}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

might look similar to the normal, but it describes fundamentally different phenomena.

The log-normal distribution emerges when processes are multiplicative rather than additive. Think about cell growth: each division multiplies the population by some factor. Or consider gene expression, where each step in the process (transcription, translation, protein folding) has a multiplicative effect on the final outcome.

This is why gene expression data is often log-transformed before analysis—it converts these multiplicative effects into additive ones, making the data approximately normal on the log scale. This isn't just a mathematical trick; it reflects the underlying biology of gene regulation.



(a) Log-normal distribution

(b) Log-normal - varying parameters

Figure 5: **The Log-Normal Distribution: Multiplicative effects in biological processes**

8.3 The Gamma Distribution: Waiting for Biology to Happen

The gamma distribution often appears when we're dealing with waiting times or accumulations in biological processes. Its probability density function,

$$f_{k,\theta}(x) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)}$$

is governed by two parameters: k (shape) and θ (scale).

A notable aspect of the gamma distribution is its clear mean-variance relationship:

$$E[X] = k\theta, \quad \mathrm{Var}(X) = k\theta^2.$$

This relationship matters because it ties the expected waiting time directly to the variability in that waiting time. In biological contexts, understanding both the central tendency and the spread is crucial. For example, when modeling the time required for a cell to divide, the mean informs us about the typical cycle duration, while the variance reveals the degree of biological variability in that process.

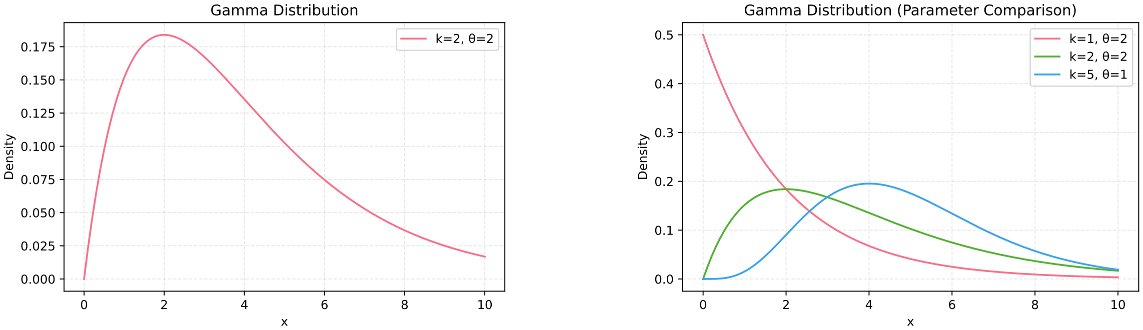
The gamma distribution is especially flexible due to its two parameters. By adjusting k and θ, it can capture a wide range of shapes:

- For $k < 1$, the distribution is highly skewed with a pronounced peak near zero, reflecting processes where short waiting times are common but long delays are possible.

- As k increases, the distribution becomes more symmetric and bell-shaped, approaching a normal-like form for large k.
- The scale parameter θ stretches or compresses the distribution, allowing it to model different time scales or magnitudes.

This dual parameter flexibility makes the gamma distribution particularly useful for modeling diverse biological processes, where both the average behavior and the variability around that average are of interest.

Consider the time it takes for a cell to divide. This process requires multiple sequential steps: DNA replication, chromatin condensation, nuclear envelope breakdown, and so on. The gamma distribution naturally models such sequences of events, where we're waiting for k things to happen, each taking an exponentially distributed time.



(a) An example of gamma distribution

(b) Gamma - different shape and scale parameters

Figure 6: **The Gamma Distribution: Modeling waiting times and positive continuous data**

But the gamma distribution's utility extends beyond waiting times. It's also excellent for modeling continuous, positive quantities that show right-skewed distributions, such as:

- Protein concentrations in cells
- Gene expression burst sizes
- Duration of biological processes

8.4 The Beta Family: Proportions and Compositions

When we deal with proportions or compositional data in biology, the beta and Dirichlet distributions become invaluable. The beta distribution, with density

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

is constrained to the interval [0,1], making it perfect for modeling proportions.

Think about allele frequencies in a population, or the percentage of cells expressing a particular marker. The beta distribution's flexibility allows it to model various shapes of proportional data, from U-shaped (where most values are near 0 or 1) to bell-shaped or skewed distributions.

When we deal with proportions or compositional data in biology, the beta and Dirichlet distributions become invaluable. The beta distribution, with density

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

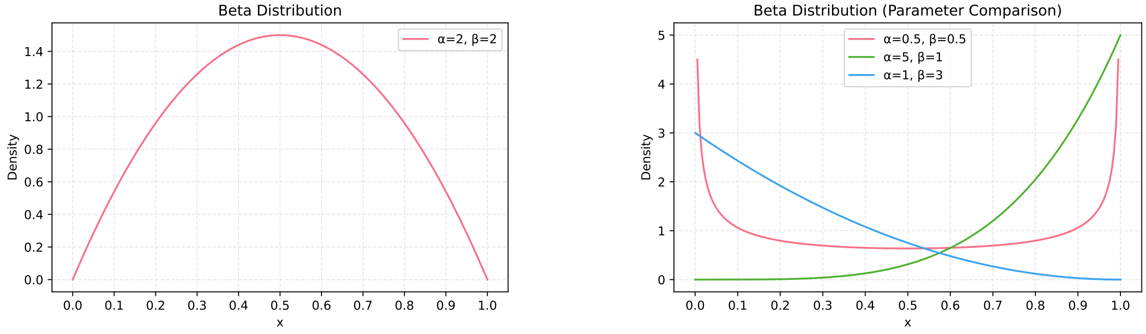
is constrained to the interval [0,1], making it perfect for modeling proportions.

Think about allele frequencies in a population, or the percentage of cells expressing a particular marker. The beta distribution's flexibility allows it to model various shapes of proportional data, from U-shaped (where most values are near 0 or 1) to bell-shaped or skewed distributions.

Example 8.1 (Allele Mutation Frequency)**.** Consider a gene with two alleles, A and a, in a population. The frequency of allele A, denoted by p, represents the probability that a randomly chosen allele is A. Rather than being a fixed number, this frequency can vary across populations due to evolutionary forces like genetic drift, mutation, and selection.

By treating p as a random variable, we acknowledge that it, too, can have a probability distribution. The beta distribution is a natural choice for modeling p because its support is $[0, 1]$. For example, if prior data suggests that allele A is relatively common, one might choose beta parameters with $\alpha > \beta$, resulting in a distribution that concentrates more mass near 1. Conversely, if allele A is typically rare, parameters with $\alpha < \beta$ would yield a distribution with more mass near 0.

This approach is especially useful in Bayesian analyses, where the beta distribution can serve as a prior for allele frequency, encapsulating our uncertainty and variability about this biological parameter.



(a) An example of beta distribution

(b) Betas - varying shape parameters

Figure 7: **The Beta Distribution: Modeling proportions and probabilities**

8.5 The Dirichlet Distribution: A Multivariate Generalization

The Dirichlet distribution serves as the multivariate generalization of the beta distribution, providing a sophisticated tool for modeling proportions that must sum to 1 across multiple categories. Its probability density function is given by:

$$p_\alpha(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\sum x_i = 1$ and $B(\alpha)$ is the multivariate beta function.

8.5.1 Key Properties

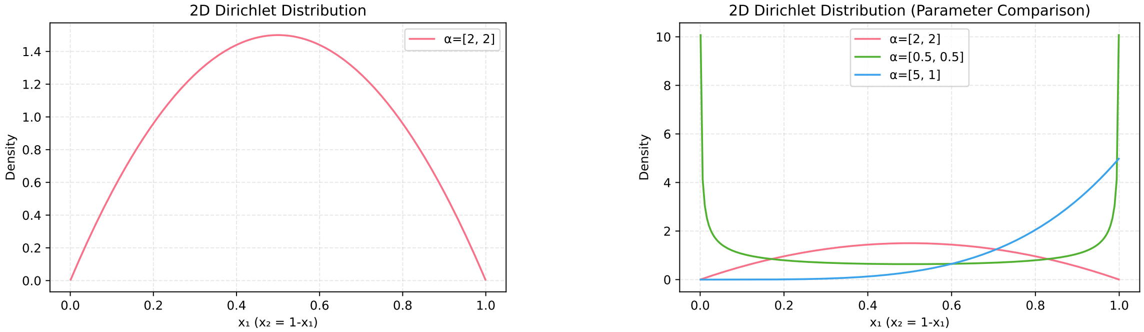
The Dirichlet distribution has several properties that make it particularly useful in biological data analysis:

- Each component is constrained to $(0, 1)$
- The sum of all components equals 1
- The parameters α_i can be interpreted as pseudo-counts or prior observations
- It serves as the conjugate prior for the multinomial distribution

8.5.2 Applications in Biology

The Dirichlet distribution naturally arises in various biological contexts:

- **Cell Type Composition:** In tissue analysis, modeling the proportions of different cell types, where each x_i represents the fraction of cells of type i
- **Microbiome Analysis:** Describing species abundance distributions, where each component represents the relative abundance of a particular species
- **Gene Expression:** Modeling the relative expression levels of different isoforms of a gene
- **Population Genetics:** Representing allele frequencies in multi-allelic systems



(a) An example of Dirichlet distribution

(b) Dirichlet - different parameters

Figure 8: **The Dirichlet Distribution: A Multivariate Generalization**

8.6 The Count Distributions: Discrete Events in Biology

When we move from continuous measurements to counting discrete events in biology, we enter the realm of discrete distributions. These distributions form a fascinating hierarchy, each building on the previous one to capture increasingly complex biological phenomena.

8.6.1 The Bernoulli Distribution: The Quantum of Biology

At its most fundamental level, many biological events are binary: a gene is either expressed or not, a mutation either occurs or doesn't, a cell either divides or doesn't. The Bernoulli distribution models these simplest of random events with a single parameter p:

$$P(X = k) = p^k(1-p)^{1-k}, \quad k \in \{0, 1\}$$

Despite its simplicity, the Bernoulli distribution forms the building block for more complex discrete distributions. It's like the atom of random counting—more complex distributions are built by combining these simple binary events.

8.6.2 The Binomial Distribution: Counting Successes

When we observe n independent Bernoulli trials, we enter the domain of the binomial distribution. Its probability mass function:

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

perfectly describes situations like:

- Counting mutated cells in a fixed-size population
- Number of successful PCR amplifications in a set of reactions
- Proportion of affected offspring in genetic crosses

The binomial distribution teaches us something profound about biological sampling: even when the underlying process is simple (like a coin flip), the act of counting these events over multiple trials introduces predictable patterns of variation.

Consider a PCR experiment designed to amplify a specific gene from a DNA sample. In this setup, multiple reaction wells are prepared, and each well represents an independent trial with a certain probability p of successful amplification. Factors such as primer specificity, enzyme efficiency, and sample purity all influence p. Even under optimized conditions, not every well will yield a detectable product.

For instance, if the success probability is $p = 0.9$ and we perform $n = 20$ PCR reactions, the number of successful amplifications is modeled by a binomial random variable. The expected number of successes is

$$E[X] = np = 20 \times 0.9 = 18,$$

and the variance is

$$\mathrm{Var}(X) = np(1 - p) = 20 \times 0.9 \times 0.1 = 1.8.$$

This quantification of variability is essential: if the observed number of successes deviates significantly from these predictions, it may indicate technical issues or sample variability. Thus, the binomial model not only describes the expected outcomes but also provides a framework for assessing the reliability and consistency of PCR experiments.

8.6.3 The Poisson Distribution: Events in Time and Space

Sometimes, the total number of trials is unknown, but we do know the average rate of events, λ. In such cases, the Poisson distribution provides a natural model.

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

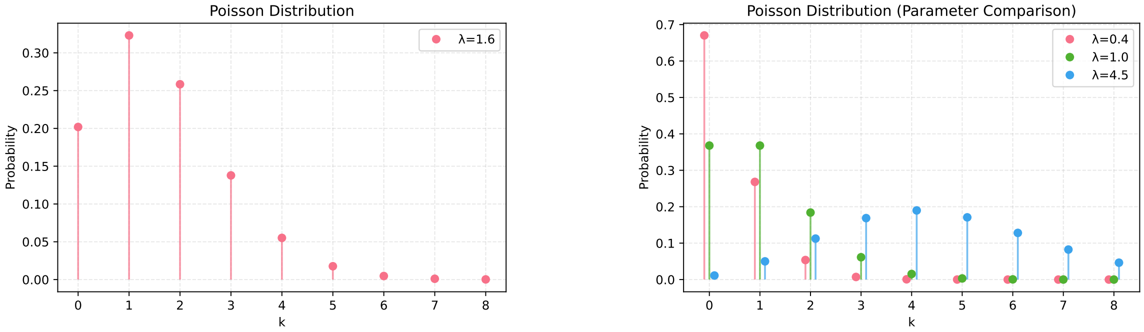
The Poisson distribution emerges naturally when we count rare events over fixed intervals of time or space. What's remarkable is that it has only one parameter, λ, which equals both its mean and variance. This distribution beautifully models phenomena like:

- Mutations occurring along a DNA sequence
- Cell division events in a time interval
- Number of binding sites in a genomic region

Imagine analyzing a long stretch of DNA where mutations occur randomly and independently along the sequence. Suppose the average mutation rate is λ mutations per kilobase. For a region spanning L kilobases, the expected number of mutations is $L\lambda$, and under the Poisson model, the variance is also $L\lambda$. For instance, if $\lambda = 0.2$ mutations per kilobase and we examine a 10-kilobase segment, we expect:

$$\text{Mean} = 10 \times 0.2 = 2 \quad \text{and} \quad \text{Variance} = 2.$$

This framework allows geneticists to determine whether the observed mutation counts are in line with the expected random variation or if they suggest the presence of additional biological factors affecting mutation rates.



(a) An example of Poisson distribution

(b) Poisson - different rate parameters

Figure 9: **The Poisson Distribution: Modeling rare events and counts**

8.6.4 The Negative Binomial: When Biology Is More Complex

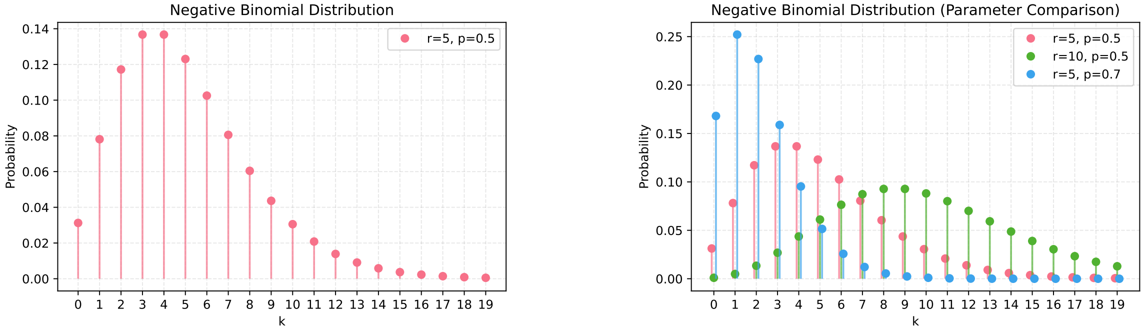
Real biological data often shows more variability than the Poisson distribution predicts. This extra variation, known as overdispersion, suggests that additional factors or complexities in the biological system and extra noise may be contributing to the observed outcomes. This overdispersion leads us to the negative binomial distribution, which can be parametrized in terms of a mean μ and dispersion parameter θ:

$$P(X = k) = \binom{k + \theta - 1}{k} \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \left(\frac{\mu}{\theta + \mu}\right)^{k}$$

This parametrization is particularly useful because:

- μ represents the expected count
- θ controls the amount of extra variability
- As $\theta \to \infty$, we recover the Poisson distribution

This makes the negative binomial distribution ideal for RNA-seq data, where biological variability often exceeds what a pure Poisson model would predict. It's become the standard model in differential expression analysis, capturing both technical and biological variability in gene expression counts.



(a) An example of negative binomial distribution

(b) Negative binomial - varying parameters

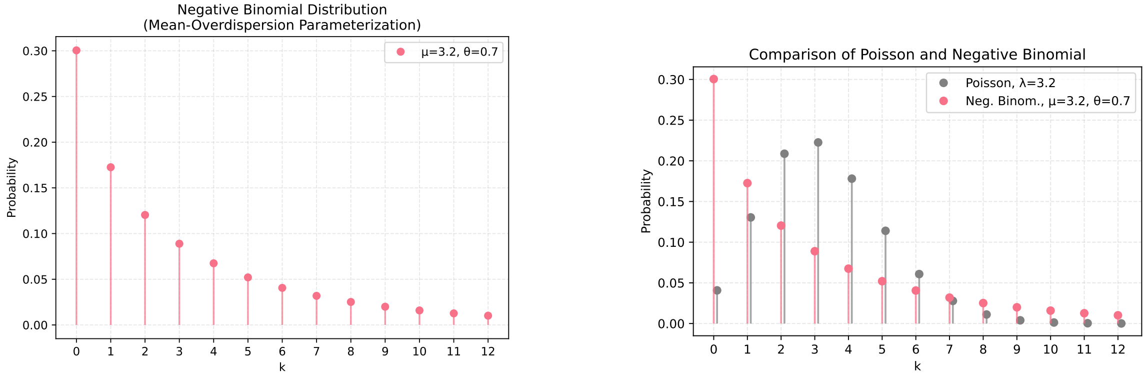
Figure 10: **The Negative Binomial Distribution: Modeling overdispersed count data**

8.6.5 Alternative Parametrization: A Practical Note

In many biological applications, we prefer to work with the negative binomial in terms of its mean μ and overdispersion parameter $r = 1/\theta$. This parametrization provides a direct link to the amount of extra variability in our data compared to a Poisson model. The variance under this parametrization becomes:

$$\text{Var}(X) = \mu + r\mu^2$$

When r approaches 0, we get closer to a Poisson distribution, while larger values of r indicate greater overdispersion. This interpretation is particularly intuitive in the context of gene expression data, where r can be thought of as a measure of biological variability.



(a) Alternative parameterization

(b) Changing mean and dispersion

Figure 11: **Alternative Parameterization of the Negative Binomial: Mean and overdispersion**

8.7 Special-Purpose Distributions

Beyond these common distributions, biology often requires specialized distributions for specific phenomena. Your slides mention several other important cases that deserve attention:

8.7.1 Multinomial and Dirichlet-Multinomial

These distributions extend the binomial and negative binomial to cases with multiple categories. They're essential for modeling:

- Cell type proportions in single-cell RNA sequencing
- Species counts in microbiome data
- Multiple allele frequencies in populations

8.7.2 Zero-Inflated Distributions

Many biological processes produce more zeros than our standard distributions predict. Zero-inflated versions of the Poisson and negative binomial distributions help model phenomena like:

- Single-cell gene expression (where many genes show zero expression)
- Rare species counts in ecological samples
- Protein-protein interaction data

These distributions are particularly important in modern high-throughput biology, where the presence of excess zeros is often biologically meaningful rather than just technical noise.

9 Operations on Random Variables: Building Complex from Simple

The true elegance of probability theory emerges when we consider how random variables combine and transform. Nature rarely presents us with simple, isolated random processes. Instead, we observe the results of multiple processes interacting, combining, and transforming in complex ways. Understanding these operations isn't just a mathematical exercise—it's essential for modeling real biological phenomena.

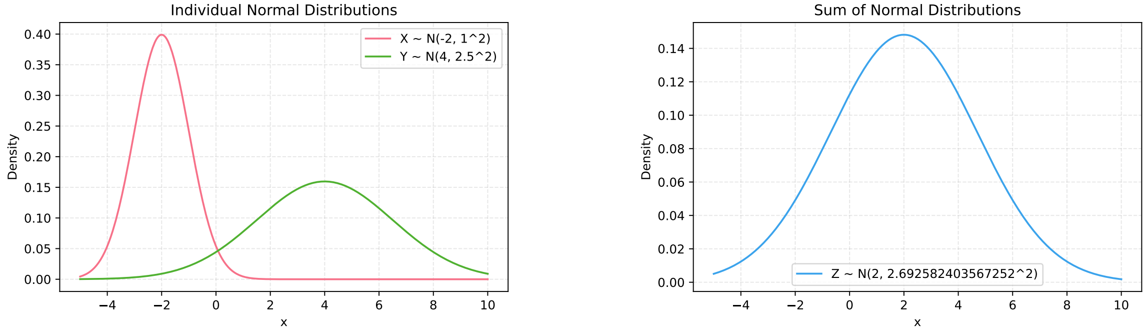
9.1 Convolution: The Addition of Randomness

Note: while this topic and those coming after it are part of lecture #1, they will be covered during the second class of this course. When we add independent random variables, we enter the realm of convolution. Consider a cell's total protein content: it results from countless individual production and degradation events, each contributing to the final amount. Or think about the time it takes for a cell to complete its cycle: it's the sum of multiple sequential phases, each with its own random duration. These are natural examples of convolution at work. In both cases, convolution illustrates how the aggregation of many small, random contributions produces a complex yet structured overall behavior. This process reveals the remarkable ability of biological systems to generate orderly outcomes from the interplay of numerous random events with their own distributions!

Definition 9.1 (Convolution)**.** For independent random variables X and Y with PDFs f_X and f_Y, the PDF of their sum $Z = X + Y$ is:

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)dx$$

This formula, though intimidating at first glance, tells a beautiful story about how randomness combines. For each possible value of the sum z, we consider all the ways it could arise from adding values of X and Y, weighing each combination by its probability. Nature offers us some remarkable patterns in how distributions combine. The sum of independent normal distributions is again normal—a property that helps explain the ubiquity of the normal distribution in biology. When we add independent Poisson distributions, we get another Poisson distribution, reflecting how random events accumulate over time. Perhaps most elegantly, when we sum independent exponential distributions with the same rate, we obtain a gamma distribution—a pattern often seen in waiting times for multiple biological events.



(a) Original distributions before convolution

(b) Resulting distribution after convolution

Figure 12: **Convolution of Random Variables: The sum of independent random variables often leads to more symmetric, bell-shaped distributions, a manifestation of the Central Limit Theorem**

9.2 Mixture Distributions: Nature's Heterogeneity

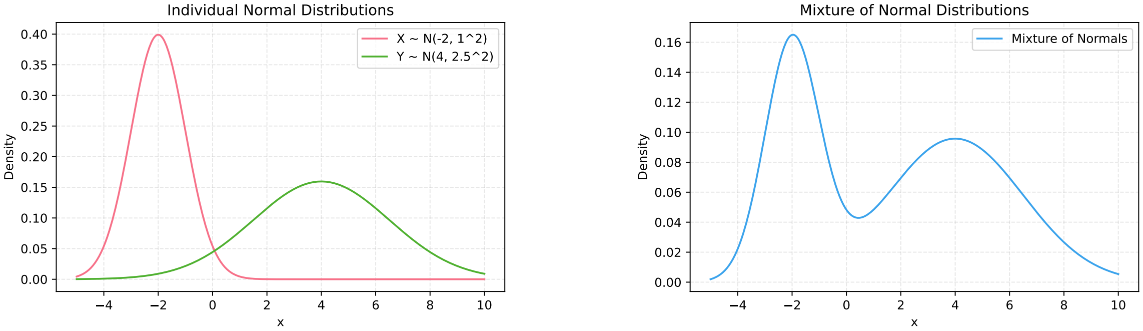
Sometimes, what we observe isn't a sum but rather a sampling from different populations. A tissue sample might contain multiple cell types, each with its own gene expression pattern. A bacterial culture might comprise cells in different growth phases. These scenarios lead us to mixture distributions—a different way that randomness combines.

Definition 9.2 (Mixture Distribution)**.** A mixture distribution has PDF:

$$f(x) = \sum_{i=1}^{k} w_i f_i(x), \quad \sum_{i=1}^{k} w_i = 1$$

where $f_i(x)$ are component densities and w_i are mixture weights.

The mathematics of mixtures reflects a fundamentally different process than convolution. Instead of adding random variables, we're randomly selecting which distribution to sample from. This leads to fascinating behaviors: the mean of a mixture is the weighted average of the component means, but the variance includes both the average of component variances and the variability between component means. This additional variance term quantifies the extra uncertainty introduced by not knowing which population we're sampling from.



(a) Component distributions with their weights

(b) Resulting mixture distribution

Figure 13: **Gaussian Mixture: The combination of two normal distributions can create complex, multimodal patterns commonly seen in heterogeneous biological samples**

9.3 Transformation: Reshaping Randomness

The third fundamental operation occurs when we apply a function to a random variable. This process, known as transformation, is particularly important in biological data analysis. When we take the logarithm of gene expression data, or calculate the ratio of two measurements, we're transforming random variables.

Theorem 9.3 (Change of Variables)**.** If $Y = g(X)$ where g is monotonic and differentiable, then:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

This formula captures how probability distributions reshape themselves under transformation. The logarithmic transformation, so common in biological data analysis, converts multiplicative effects into additive ones—not just mathematically, but in a way that often reveals underlying biological mechanisms. Power transformations can stabilize variance, making data more amenable to statistical analysis. Exponential transformations model growth processes where the rate of change depends on the current value.