Selected Topics in Probability:
Course Notes and Announcements

Compiled and Expanded from Lecture Transcripts

An Undergraduate Mathematics Educator

Current Term

Contents

Course Announcements & Information

- **Upcoming Quiz Information:**
  - **Content Coverage:** The quiz will cover all material presented in lectures up to and including the topics of the current week. This means it will also include material corresponding to the *next* problem set, as one problem set cycle was missed earlier.
  - **Structure:** The quiz is expected to consist of approximately three main questions, with the possibility of an additional bonus question.
  - **Duration:** The allotted time for the quiz will be 1.5 hours (90 minutes).
  - **Formula Sheet:** A formula sheet, prepared by Hagit, will be made available on Moodle in the coming days. Please familiarize yourself with it once it is posted to understand which formulas are provided and which you are expected to know or derive.
- **Guidance for Quiz Preparation:**
  - **Primary Focus:** It is highly recommended to concentrate your study efforts on the problem sets assigned during this term and the style of questions and examples discussed in lectures and review sessions.
  - **Regarding Past Exams:** While past exam papers can sometimes be useful, please be aware that their style or emphasis might differ from the current course structure or your specific lecturer's approach. Therefore, they may not be fully representative of the upcoming quiz.
- **Notes on Problem Sets and Review Sessions:**
  - **Problem Set Scope:** Some problem sets are quite comprehensive. It's possible that not every single question from every problem set will be covered in exhaustive detail during the scheduled review sessions.
  - **Example of Review Coverage (from a recent session):** One review session might cover certain questions from a problem set (e.g., Q1, Q2), while another session might focus on other questions from the same set (e.g., Q3, Q4, Q5).
  - **Sampling and Code Questions:** For questions involving sampling algorithms and code, the review sessions will primarily focus on the theoretical underpinnings of the sampling methods. The specific code implementation details will generally be kept straightforward in the provided solutions.
- **Seeking Clarifications:**
  - Should any concepts, problem solutions, or administrative details remain unclear after attending lectures and review sessions, please do not hesitate to send an email for clarification.

1 Moment and Cumulant Generating Functions

Generating functions are remarkably powerful tools in probability theory. The Moment Generating Function (MGF), when it exists in a neighborhood around zero, offers a unique "fingerprint" for a probability distribution. Moreover, as its name aptly suggests, it can be used to systematically derive the moments of a random variable. Closely related is the Cumulant Generating Function (CGF), obtained from the MGF, which simplifies the calculation of cumulants—quantities related to moments that possess valuable statistical properties, particularly concerning sums of independent random variables.

1.1 The Moment Generating Function (MGF) of a Poisson Distribution

Let X be a random variable following a Poisson distribution with parameter $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$. Its Probability Mass Function (PMF) is given by:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots$$

The MGF of X, $M_X(t)$, is defined as $M_X(t) = \mathbb{E}[e^{tX}]$, provided this expectation exists for t in some neighborhood of 0.

Example 1.1 (Deriving the MGF of a Poisson(λ) RV)**.** We calculate $M_X(t)$ directly from its definition:

$$\begin{aligned}
M_X(t) = \mathbb{E}[e^{tX}] &= \sum_{k=0}^{\infty} e^{tk} P(X = k) \quad \text{(by definition of expectation for discrete RVs)} \\
&= \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t)^k \lambda^k}{k!} \quad \text{(factoring out } e^{-\lambda}\text{, constant w.r.t. } k) \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \quad \text{(combining terms with exponent } k)
\end{aligned}$$

We now recall the Taylor series expansion for the exponential function $e^y = \sum_{k=0}^{\infty} \frac{y^k}{k!}$. In our sum, we can identify $y = \lambda e^t$. Therefore, the sum becomes:

$$\sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{\lambda e^t}$$

Substituting this back into our expression for $M_X(t)$, we get:

$$M_X(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda e^t - \lambda} = e^{\lambda(e^t - 1)}$$

This MGF exists for all $t \in \mathbb{R}$.

Remark 1.2. The key step here was recognizing the familiar Taylor series. If this is not immediately apparent, one might try to manipulate the sum to resemble the PMF of another Poisson distribution that sums to 1, but for this standard MGF, the Taylor series approach is most direct and elegant.

1.2 Cumulant Generating Functions (CGFs) and Cumulants

Definition 1.3 (Cumulant Generating Function (CGF))**.** The Cumulant Generating Function (CGF) of a random variable X, denoted $K_X(t)$ or $\text{CGF}_X(t)$, is defined as the natural logarithm of its MGF:

$$K_X(t) = \log M_X(t)$$

provided $M_X(t)$ exists and is positive (which it will be if it exists, as $M_X(t) = \mathbb{E}[e^{tX}]$ and $e^{tX} > 0$).

Definition 1.4 (Cumulants)**.** The n-th cumulant of X, denoted κ_n, is defined as the n-th derivative of the CGF evaluated at $t = 0$:

$$\kappa_n = K_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n} K_X(t)\Big|_{t=0}$$

Example 1.5 (Cumulants of the Poisson(λ) Distribution)**.** From Example 1.1, we have $M_X(t) = e^{\lambda(e^t-1)}$ for $X \sim \text{Poisson}(\lambda)$. The CGF is:

$$K_X(t) = \log(e^{\lambda(e^t-1)}) = \lambda(e^t - 1) = \lambda e^t - \lambda$$

Now, let's find its derivatives with respect to t:

$$\begin{aligned} K_X'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}(\lambda e^t - \lambda) = \lambda e^t \\ K_X''(t) &= \frac{\mathrm{d}}{\mathrm{d}t}(\lambda e^t) = \lambda e^t \\ &\vdots \\ K_X^{(n)}(t) &= \lambda e^t \quad \text{for any integer } n \geq 1 \end{aligned}$$

Evaluating these derivatives at $t = 0$ gives us the cumulants:

$$\kappa_n = K_X^{(n)}(0) = \lambda e^0 = \lambda$$

Thus, for a Poisson(λ) distribution, all cumulants κ_n (for $n \geq 1$) are equal to λ. This is a distinctive and elegant property of the Poisson distribution!

1.3 Relationship Between Moments and Cumulants

Moments $m_n = \mathbb{E}[X^n] = M_X^{(n)}(0)$ (the n-th derivative of MGF at $t = 0$) and cumulants $\kappa_n = K_X^{(n)}(0)$ are intrinsically linked. Let's derive the first few relationships. A crucial fact we'll use is $M_X(0) = \mathbb{E}[e^{0 \cdot X}] = \mathbb{E}[1] = 1$.

1.3.1 First Cumulant: $\kappa_1 = m_1$

Proof. The CGF is $K_X(t) = \log M_X(t)$. Its first derivative is:

$$K_X'(t) = \frac{M_X'(t)}{M_X(t)}$$

Evaluating at $t = 0$:

$$\kappa_1 = K_X'(0) = \frac{M_X'(0)}{M_X(0)}$$

Since $M'_X(0) = m_1$ (the first moment, i.e., the mean $\mathbb{E}[X]$) and $M_X(0) = 1$, we have:

$$\kappa_1 = \frac{m_1}{1} = m_1$$

So, the first cumulant is identical to the mean of the distribution. □

1.3.2 Second Cumulant: $\kappa_2 = m_2 - m_1^2$

Proof. We differentiate $K'_X(t) = M'_X(t)/M_X(t)$ using the quotient rule:

$$\begin{aligned} K''_X(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{M'_X(t)}{M_X(t)}\right) \\ &= \frac{M''_X(t)M_X(t) - M'_X(t)M'_X(t)}{(M_X(t))^2} \\ &= \frac{M''_X(t)M_X(t) - (M'_X(t))^2}{(M_X(t))^2} \end{aligned}$$

Evaluating at $t = 0$:

$$\kappa_2 = K''_X(0) = \frac{M''_X(0)M_X(0) - (M'_X(0))^2}{(M_X(0))^2}$$

Substituting $M_X(0) = 1$, $M'_X(0) = m_1$, and $M''_X(0) = m_2$ (the second moment $\mathbb{E}[X^2]$):

$$\kappa_2 = \frac{m_2 \cdot 1 - (m_1)^2}{1^2} = m_2 - m_1^2$$

This expression $m_2 - m_1^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is precisely the variance of X, $\mathrm{Var}(X)$. Thus, the second cumulant is equal to the variance of the distribution. □

1.3.3 Third Cumulant: $\kappa_3 = m_3 - 3m_1m_2 + 2m_1^3$

Proof. Differentiating $K''_X(t)$ to find $K'''_X(t)$ requires careful application of the quotient rule again, or chain rule if expressed differently. After differentiation and evaluation at $t = 0$ (substituting $M_X(0) = 1, M'_X(0) = m_1, M''_X(0) = m_2, M'''_X(0) = m_3$), the result is:

$$\kappa_3 = m_3 - 3m_2m_1 + 2m_1^3$$

The third cumulant is related to the skewness of the distribution, providing a measure of its asymmetry. □

Remark 1.6. While you are generally not expected to memorize the formulas relating higher-order cumulants to moments (especially for $n \geq 3$), understanding their origin from derivatives of the CGF and being able to use them if provided is key. The process of deriving these relationships, as shown, is a good exercise in calculus and understanding of generating functions.

1.4 Calculating Moments of Poisson(λ) using Cumulants

We've established two important facts:

1. For $X \sim \text{Poisson}(\lambda)$, all cumulants $\kappa_n = \lambda$ for $n \geq 1$ (Example 1.5).
2. We have general formulas relating κ_n to moments m_i (Section 1.3).

We can combine these to easily find the first few moments of a Poisson(λ) random variable.

Example 1.7 (First three moments of Poisson(λ) via cumulants)**.** Let $X \sim \text{Poisson}(\lambda)$. We know $\kappa_1 = \lambda, \kappa_2 = \lambda, \kappa_3 = \lambda$.

(a) **First moment ($m_1 = \mathbb{E}[X]$):** Using $\kappa_1 = m_1$, and $\kappa_1 = \lambda$:

$$m_1 = \lambda$$

(b) **Second moment ($m_2 = \mathbb{E}[X^2]$):** Using $\kappa_2 = m_2 - m_1^2$. We have $\kappa_2 = \lambda$ and $m_1 = \lambda$:

$$\lambda = m_2 - (\lambda)^2$$

Solving for m_2:

$$m_2 = \lambda + \lambda^2$$

(As a check: $\text{Var}(X) = \kappa_2 = \lambda$. Also, $\text{Var}(X) = m_2 - m_1^2 = (\lambda + \lambda^2) - \lambda^2 = \lambda$. This is consistent.)

(c) **Third moment ($m_3 = \mathbb{E}[X^3]$):** Using $\kappa_3 = m_3 - 3m_1m_2 + 2m_1^3$. We have $\kappa_3 = \lambda$, $m_1 = \lambda$, and $m_2 = \lambda + \lambda^2$:

$$\begin{aligned}
\lambda &= m_3 - 3(\lambda)(\lambda + \lambda^2) + 2(\lambda)^3 \\
\lambda &= m_3 - (3\lambda^2 + 3\lambda^3) + 2\lambda^3 \\
\lambda &= m_3 - 3\lambda^2 - 3\lambda^3 + 2\lambda^3 \\
\lambda &= m_3 - 3\lambda^2 - \lambda^3
\end{aligned}$$

Solving for m_3:

$$m_3 = \lambda + 3\lambda^2 + \lambda^3$$

This method elegantly demonstrates how the simple structure of Poisson cumulants can simplify moment calculations.

2 Identifying Distributions from MGFs

One of the most powerful aspects of MGFs is their uniqueness property: if an MGF $M_X(t)$ exists in a neighborhood of $t = 0$, it uniquely determines the probability distribution of X. This means if we are given an MGF and can recognize its form as belonging to a known distribution, we have effectively identified the distribution of X.

Example 2.1 (Degenerate Distribution from MGF: $M_X(t) = e^{ct}$)**.** Suppose a random variable X has an MGF given by $M_X(t) = e^{ct}$ for some real constant c. What is the distribution of X?

We can find the moments of X by differentiating $M_X(t)$: The first moment (mean): $m_1 = M'_X(0) = \frac{\mathrm{d}}{\mathrm{d}t}(e^{ct})\Big|_{t=0} = ce^{ct}\big|_{t=0} = ce^0 = c$. So, $\mathbb{E}[X] = c$.

The second moment: $m_2 = M''_X(0) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}(e^{ct})\Big|_{t=0} = c^2e^{ct}\big|_{t=0} = c^2e^0 = c^2$. So, $\mathbb{E}[X^2] = c^2$.

In general, the k-th moment is $m_k = M_X^{(k)}(0) = c^k$.

Now, let's calculate the variance of X: $\mathrm{Var}(X) = m_2 - m_1^2 = c^2 - (c)^2 = 0$. A random variable with zero variance must be a constant (almost surely). Since its mean $\mathbb{E}[X] = c$, it must be that $X = c$ with probability 1. This is known as a **degenerate distribution**, where all the probability mass is concentrated at a single point c. Its PMF is $P(X = c) = 1$ and $P(X = x) = 0$ for $x \neq c$. Its CDF is $F_X(x) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases}$.

Remark 2.2. It's crucial to remember that in $M_X(t)$, t is the argument of the MGF, which is a real variable. The MGF itself is a function that encodes information about the random variable X. Our goal is to deduce properties or the distribution of X, not t.

Example 2.3 (General Discrete Distribution from MGF: $M_X(t) = \sum_i p_i e^{tx_i}$)**.** Suppose an MGF is given in the form $M_X(t) = \sum_{i=1}^k p_i e^{tx_i}$, where $x_1, x_2, \ldots, x_k$ are distinct real numbers, $p_i > 0$ for all i, and $\sum_{i=1}^k p_i = 1$.

This is the MGF of a discrete random variable X that takes on the values $\{x_1, x_2, \ldots, x_k\}$ with corresponding probabilities $P(X = x_i) = p_i$. Why? Because by definition, $\mathbb{E}[e^{tX}] = \sum_{i=1}^k e^{tx_i} P(X = x_i)$. Comparing this with the given $M_X(t)$, we can identify the values and their probabilities. The CDF of such a random variable would then be $F_X(x) = \sum_{x_j \leq x} p_j$. To identify a specific distribution, one would look at the x_i values (the support) and the probabilities p_i.

3 Inequalities in Probability Theory

Inequalities are indispensable tools in probability and statistics. They allow us to:

- Provide bounds on probabilities or expectations that might be difficult or impossible to calculate exactly.
- Understand the limiting behavior of sequences of random variables.
- Make robust statements that hold under general conditions.

3.1 Jensen's Inequality

Jensen's inequality provides a fundamental relationship between the expectation of a convex (or concave) function of a random variable and the function of its expectation.

Definition 3.1 (Convex Function)**.** A function $g : I \to \mathbb{R}$, where I is an interval in $\mathbb{R}$, is said to be **convex** if for all $x_1, x_2 \in I$ and for all $\lambda \in [0, 1]$:

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2)$$

Geometrically, this means the line segment connecting any two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ on the graph of g lies on or above the graph of g between these points. If g is twice differentiable on I, then g is convex if and only if $g''(x) \geq 0$ for all $x \in I$. A function g is **concave** if $-g$ is convex. For a twice-differentiable function, this means $g''(x) \leq 0$ for all $x \in I$, and the inequality in the definition is reversed: $g(\lambda x_1 + (1-\lambda)x_2) \geq \lambda g(x_1) + (1-\lambda)g(x_2)$.

Theorem 3.2 (Jensen's Inequality)**.** *Let X be a random variable such that its expectation $\mathbb{E}[X]$ is finite and lies in the domain of g.*

(a) If g is a convex function, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.

(b) If g is a concave function, then $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$.

These inequalities hold provided $\mathbb{E}[g(X)]$ also exists and is finite.

3.1.1 Application: Reciprocal of a Positive Random Variable

Problem 3.1 (Jensen's Inequality Example 1)**.** Prove that for a positive random variable X (i.e., $P(X > 0) = 1$) with finite expectation, $E\left[\frac{1}{X}\right] \geq \frac{1}{E[X]}$.

Proof. We aim to apply Jensen's inequality. Let the function be $g(x) = \frac{1}{x}$. The random variable X takes values in $(0, \infty)$. We need to check the convexity of $g(x)$ on this interval. We examine the second derivative of $g(x)$: $g'(x) = -\frac{1}{x^2} = -x^{-2}$ $g''(x) = -(-2)x^{-3} = \frac{2}{x^3}$ For $x \in (0, \infty)$, $x^3 > 0$, so $g''(x) = \frac{2}{x^3} > 0$. Since $g''(x) > 0$ for all $x \in (0, \infty)$, the function $g(x) = \frac{1}{x}$ is convex on $(0, \infty)$.

Remark 3.3 (Importance of Domain for Convexity)*.* The second derivative test is often the most convenient way to establish convexity for differentiable functions. It's crucial to confirm that the condition ($g''(x) \geq 0$ for convexity) holds over the entire support of the random variable X. If X could take values where $g(x)$ is undefined or not convex, Jensen's inequality (in this form) would not be applicable.

Since $X > 0$, $\mathbb{E}[X]$ will also be positive. The function $g(x) = 1/x$ is convex on the domain of X. Applying Jensen's inequality for convex functions (part (a)):

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

Substituting $g(x) = 1/x$:

$$\mathbb{E}\left[\frac{1}{X}\right] \geq \frac{1}{\mathbb{E}[X]}$$

This completes the proof. This inequality is quite useful, for example, showing that the harmonic mean is less than or equal to the arithmetic mean. □

3.1.2 Application: The Arithmetic Mean-Geometric Mean (AM-GM) Inequality

Problem 3.2 (Jensen's Inequality Example 2)**.** Prove the AM-GM inequality for n positive numbers $x_1, x_2, \ldots, x_n > 0$:

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} x_i$$

(This states that the Geometric Mean is less than or equal to the Arithmetic Mean).

Proof. We follow the insightful hint to consider a random variable X that takes each value x_i with probability $1/n$. Then, we apply Jensen's inequality to the function $g(x) = \log(x)$.

Let X be a discrete random variable such that $P(X = x_i) = \frac{1}{n}$ for $i = 1, \ldots, n$. Consider the function $g(x) = \log(x)$. Its domain is $(0, \infty)$, which is appropriate since all $x_i > 0$. Let's check for concavity/convexity by examining the second derivative: $g'(x) = \frac{1}{x}$ $g''(x) = -\frac{1}{x^2}$ For $x \in (0, \infty)$, $x^2 > 0$, so $g''(x) = -\frac{1}{x^2} < 0$. Since $g''(x) < 0$ for all $x \in (0, \infty)$, the function $g(x) = \log(x)$ is concave on $(0, \infty)$.

Next, we calculate the necessary expectations for our random variable X: The expectation of X is the arithmetic mean of the x_i's:

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The expectation of $g(X) = \log(X)$ is:

$$\mathbb{E}[\log(X)] = \sum_{i=1}^{n} \log(x_i) P(X = x_i) = \sum_{i=1}^{n} \log(x_i) \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} \log(x_i)$$

Now, we apply Jensen's inequality for concave functions (part (b): $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$):

$$\frac{1}{n} \sum_{i=1}^{n} \log(x_i) \leq \log\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)$$

To make this look like the AM-GM inequality, we use properties of logarithms on the left side: The sum of logarithms is the logarithm of the product: $\sum_{i=1}^{n} \log(x_i) = \log(x_1 x_2 \ldots x_n)$. So, $\frac{1}{n} \sum_{i=1}^{n} \log(x_i) = \frac{1}{n} \log\left(\prod_{i=1}^{n} x_i\right) = \log\left(\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right)$. The inequality thus becomes:

$$\log\left(\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right) \leq \log\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)$$

Since $\log(x)$ is a strictly monotonically increasing function, if $\log(A) \leq \log(B)$, then $A \leq B$. We can exponentiate both sides (i.e., apply the function e^y, which is also strictly increasing):

$$e^{\log\left(\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right)} \leq e^{\log\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)}$$

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i$$

This is precisely the AM-GM inequality.

Remark 3.4 (Why the logarithm function?). The choice of $g(x) = \log(x)$ is strategic. The AM-GM inequality relates a product (geometric mean) to a sum (arithmetic mean). The logarithm is the perfect tool for transforming products into sums (via $\log(ab) = \log a + \log b$) and roots/powers into products (via $\log(a^{1/n}) = \frac{1}{n} \log a$). This transformation allows Jensen's inequality, which is fundamentally about sums (expectations), to be elegantly applied to an inequality involving products.

□

3.2 Chernoff Bounds

Chernoff bounds are a class of inequalities that provide exponentially decreasing bounds on the tail probabilities of sums of independent random variables. They are typically much sharper (tighter) than bounds derived from Markov's or Chebyshev's inequality, especially for large deviations.

The derivation involves the MGF and Markov's inequality. A key quantity is the Cramér-Chernoff transform, also known as the rate function.

Definition 3.5 (Cramér-Chernoff Transform / Rate Function)**.** Let X be a random variable with MGF $M_X(t) = \mathbb{E}[e^{tX}]$. The Cramér-Chernoff transform (or rate function) associated with X is defined as:

$$\Lambda_X^*(a) = \sup_{t \in \mathbb{R}} (ta - \log M_X(t))$$

For bounding upper tail probabilities like $P(X \geq a)$ where $a > \mathbb{E}[X]$, the supremum is often taken over $t \geq 0$.

Theorem 3.6 (Chernoff Bound (Upper Tail))**.** *For a random variable X and any real number a:*

$$P(X \geq a) \leq e^{-\Lambda_X^*(a)}$$

where $\Lambda_X^*(a) = \sup_{t \geq 0}(ta - \log M_X(t))$. (A similar bound exists for lower tails, $P(X \leq a)$, typically involving $\sup_{t \leq 0}$.)

3.2.1 Derivation of the Chernoff Bound from Markov's Inequality

Proof. We wish to bound $P(X \geq a)$. For any $t > 0$, the event $\{X \geq a\}$ is equivalent to the event $\{tX \geq ta\}$. Since e^x is a strictly increasing function for real x, if $t > 0$, then $\{tX \geq ta\}$ is equivalent to $\{e^{tX} \geq e^{ta}\}$. Thus, for any $t > 0$:

$$P(X \geq a) = P(e^{tX} \geq e^{ta})$$

Let $Y = e^{tX}$. Since t is real, e^{tX} is always a non-negative random variable. We can apply Markov's inequality to Y. Markov's Inequality states: For a non-negative random variable Y and any $b > 0$, $P(Y \geq b) \leq \frac{\mathbb{E}[Y]}{b}$. Applying this with $Y = e^{tX}$ and $b = e^{ta}$:

$$P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

Recognizing $\mathbb{E}[e^{tX}]$ as the MGF $M_X(t)$, we have:

$$P(X \geq a) \leq \frac{M_X(t)}{e^{ta}} = e^{-ta} M_X(t)$$

This inequality holds for any $t > 0$. To obtain the tightest possible bound from this family of inequalities (parameterized by t), we should choose $t > 0$ to minimize the right-hand side expression $e^{-ta} M_X(t)$. So, $P(X \geq a) \leq \inf_{t>0}(e^{-ta} M_X(t))$. Let's rewrite the term being minimized: $e^{-ta} M_X(t) = e^{-ta} e^{\log M_X(t)} = e^{-(ta - \log M_X(t))}$. Minimizing e^{-Y} is equivalent to maximizing Y. Thus, minimizing $e^{-(ta - \log M_X(t))}$ with respect to $t > 0$ is equivalent to maximizing $ta - \log M_X(t)$ with respect to $t > 0$. Therefore,

$$\inf_{t>0}(e^{-(ta - \log M_X(t))}) = e^{-\sup_{t>0}(ta - \log M_X(t))}$$

The lecture presentation used $t \geq 0$ for the supremum. If $t = 0$, then $ta - \log M_X(0) = 0 - \log(1) = 0$, leading to the bound $P(X \geq a) \leq e^0 = 1$, which is a trivial but correct bound. The optimization is usually most effective for $t > 0$, particularly when $a > \mathbb{E}[X]$. Defining $\Lambda_X^*(a) = \sup_{t \geq 0}(ta - \log M_X(t))$, we arrive at the Chernoff bound:

$$P(X \geq a) \leq e^{-\Lambda_X^*(a)}$$

□

3.2.2 Chernoff Bound for the Sample Mean of I.I.D. Random Variables

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Let $S_n = \sum_{i=1}^n X_i = n\bar{X}_n$ be the sum. We are interested in finding a bound for $P(\bar{X}_n \geq a)$. This is equivalent to $P(S_n \geq na)$. We apply the Chernoff bound to the random variable S_n with the threshold na:

$$P(S_n \geq na) \leq e^{-\Lambda_{S_n}^*(na)}$$

where $\Lambda_{S_n}^*(na) = \sup_{t \geq 0}(t(na) - \log M_{S_n}(t))$. A key property for sums of independent random variables is that the MGF of the sum is the product of their MGFs. Since X_i are i.i.d., they all have the same MGF, say $M_X(t)$. So, $M_{S_n}(t) = M_{\sum X_i}(t) = \mathbb{E}[e^{t\sum X_i}] = \mathbb{E}[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}]$ (by independence) $= \prod_{i=1}^n M_X(t) = (M_X(t))^n$. Therefore, $\log M_{S_n}(t) = \log((M_X(t))^n) = n \log M_X(t)$. Substituting this into the expression for $\Lambda_{S_n}^*(na)$:

$$\begin{aligned}
\Lambda_{S_n}^*(na) &= \sup_{t \geq 0}(tna - n \log M_X(t)) \\
&= \sup_{t \geq 0} n(ta - \log M_X(t)) \\
&= n \sup_{t \geq 0}(ta - \log M_X(t)) \quad \text{(since } n > 0\text{, it can be factored out of the sup)} \\
&= n\Lambda_X^*(a)
\end{aligned}$$

Thus, the Chernoff bound for the sample mean of i.i.d. random variables is:

$$P(\bar{X}_n \geq a) \leq e^{-n\Lambda_X^*(a)}$$

This is a very powerful result, indicating that the probability of the sample mean deviating significantly from certain values decreases exponentially with the sample size n, provided $\Lambda_X^*(a) > 0$.

Example 3.7 (Chernoff Bound for I.I.D. Bernoulli Random Variables)**.** Let $X_i \sim \text{Bernoulli}(p)$ be i.i.d. Bernoulli random variables with parameter $p \in (0, 1)$. The MGF of a single Bernoulli(p) RV is $M_X(t) = (1-p)e^{t \cdot 0} + pe^{t \cdot 1} = 1 - p + pe^t$. We need to find $\Lambda_X^*(a) = \sup_{t \geq 0}(ta - \log(1 - p + pe^t))$. Let $f(t) = ta - \log(1 - p + pe^t)$. To find the supremum (for $a > \mathbb{E}[X] = p$), we differentiate with respect to t and set the derivative to zero:

$$f'(t) = a - \frac{pe^t}{1 - p + pe^t}$$

Setting $f'(t^*) = 0$ to find the optimal t^*:

$$a = \frac{pe^{t^*}}{1 - p + pe^{t^*}}$$

Solving for e^{t^*}: $a(1 - p + pe^{t^*}) = pe^{t^*}\ a(1-p) + ape^{t^*} = pe^{t^*}\ a(1-p) = pe^{t^*}(1-a)$

$$e^{t^*} = \frac{a(1-p)}{p(1-a)}$$

For this t^* to be positive (which is required for $a > p = \mathbb{E}[X]$ for a non-trivial bound), we need $e^{t^*} > 1$, which means $\frac{a(1-p)}{p(1-a)} > 1$. This typically holds if $p < a < 1$. If $a \leq p$, the supremum of $f(t)$ for $t \geq 0$ is achieved at $t = 0$, yielding $\Lambda_X^*(a) = f(0) = 0$. In this case, the Chernoff bound $P(\bar{X}_n \geq a) \leq e^0 = 1$ is trivial but correct. We are usually interested in the non-trivial case where $a > p$. Assuming $p < a < 1$, then $t^* = \log\left(\frac{a(1-p)}{p(1-a)}\right) > 0$. Now we substitute this t^* (or rather e^{t^*}) back into $\Lambda_X^*(a) = t^*a - \log M_X(t^*)$. First, find $M_X(t^*)$: $M_X(t^*) = 1 - p + pe^{t^*} = 1 - p + p\frac{a(1-p)}{p(1-a)} = 1 - p + \frac{a(1-p)}{1-a} = \frac{(1-p)(1-a)+a(1-p)}{1-a} = \frac{(1-p)(1-a+a)}{1-a} = \frac{1-p}{1-a}$. So, $\log M_X(t^*) = \log\left(\frac{1-p}{1-a}\right)$. Then,

$$\begin{aligned}
\Lambda_X^*(a) &= a \cdot t^* - \log M_X(t^*) \\
&= a \log\left(\frac{a(1-p)}{p(1-a)}\right) - \log\left(\frac{1-p}{1-a}\right) \\
&= a\left(\log a + \log(1-p) - \log p - \log(1-a)\right) - \left(\log(1-p) - \log(1-a)\right) \\
&= a \log a - a \log p + a \log(1-p) - a \log(1-a) - \log(1-p) + \log(1-a) \\
&= a \log\left(\frac{a}{p}\right) + (a-1)\log(1-p) - (a-1)\log(1-a) \\
&= a \log\left(\frac{a}{p}\right) + (1-a)\left(\log(1-a) - \log(1-p)\right) \quad \text{(factoring out } -(a-1) = 1-a\text{)} \\
&= a \log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right)
\end{aligned}$$

This expression is known as the Kullback-Leibler (KL) divergence between two Bernoulli distributions with parameters a and p, often denoted $D_{\mathrm{KL}}(\text{Bernoulli}(a)||\text{Bernoulli}(p))$ or simply $D_{\mathrm{KL}}(a||p)$ in this context. So, for $X_i \sim \text{Bernoulli}(p)$ i.i.d., and $a > p$:

$$P(\bar{X}_n \geq a) \leq e^{-nD_{\mathrm{KL}}(a||p)}$$

This is a classic and very useful form of the Chernoff bound for sums of Bernoulli trials (e.g., bounding the probability that the observed frequency $a = \bar{X}_n$ deviates significantly from the true probability p).

4 Generating Random Samples from Distributions

The ability to generate random samples from various probability distributions is fundamental in many areas, including statistical simulation, Monte Carlo methods, testing statistical hypotheses, and more. While modern software often provides built-in functions for common distributions, understanding the underlying principles of how such samples can be generated is crucial. One of the most foundational methods is Inverse Transform Sampling.

4.1 Inverse Transform Sampling

The inverse transform sampling method (also known as the inversion method or Smirnov transform) is a technique for generating samples from a random variable X if its Cumulative Distribution Function (CDF), $F_X(x)$, is known and invertible. The method relies on our ability to generate samples from a standard uniform distribution, $U \sim \text{Uniform}(0, 1)$.

Lemma 4.1 (Inverse Transform Sampling)**.** *Let $F_X(x)$ be the CDF of a random variable X. Define the (generalized) inverse CDF as:*

$$F_X^{-1}(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}, \quad \text{for } u \in (0, 1)$$

If U is a random variable uniformly distributed on $(0, 1)$ (i.e., $U \sim \text{Uniform}(0, 1)$), then the random variable $Y = F_X^{-1}(U)$ has the same distribution as X. That is, Y follows the distribution with CDF $F_X(y)$.

Sketch of Proof and Intuition. We want to show that the CDF of $Y = F_X^{-1}(U)$ is $F_X(y)$. That is, $P(Y \leq y) = F_X(y)$. $P(Y \leq y) = P(F_X^{-1}(U) \leq y)$. If F_X is continuous and strictly increasing, then F_X^{-1} is its standard inverse, and the event $F_X^{-1}(U) \leq y$ is equivalent to the event $U \leq F_X(y)$ (by applying F_X to both sides, which preserves the inequality because F_X is increasing). So, $P(F_X^{-1}(U) \leq y) = P(U \leq F_X(y))$. Since $U \sim \text{Uniform}(0, 1)$, its CDF is $F_U(u) = u$ for $u \in (0, 1)$. Therefore, $P(U \leq F_X(y)) = F_X(y)$, because $F_X(y)$ is a value between 0 and 1. The use of inf in the definition of $F_X^{-1}(u)$ (the quantile function) correctly handles cases where F_X is not strictly increasing (i.e., has flat regions) or has jumps (as in discrete distributions), ensuring the result holds more generally. □

Algorithm for Inverse Transform Sampling: To generate a single sample x from a distribution with CDF F_X:

1. Generate a random number u from $\text{Uniform}(0, 1)$.
2. Compute $x = F_X^{-1}(u)$. This x is now a sample from the distribution characterized by F_X.

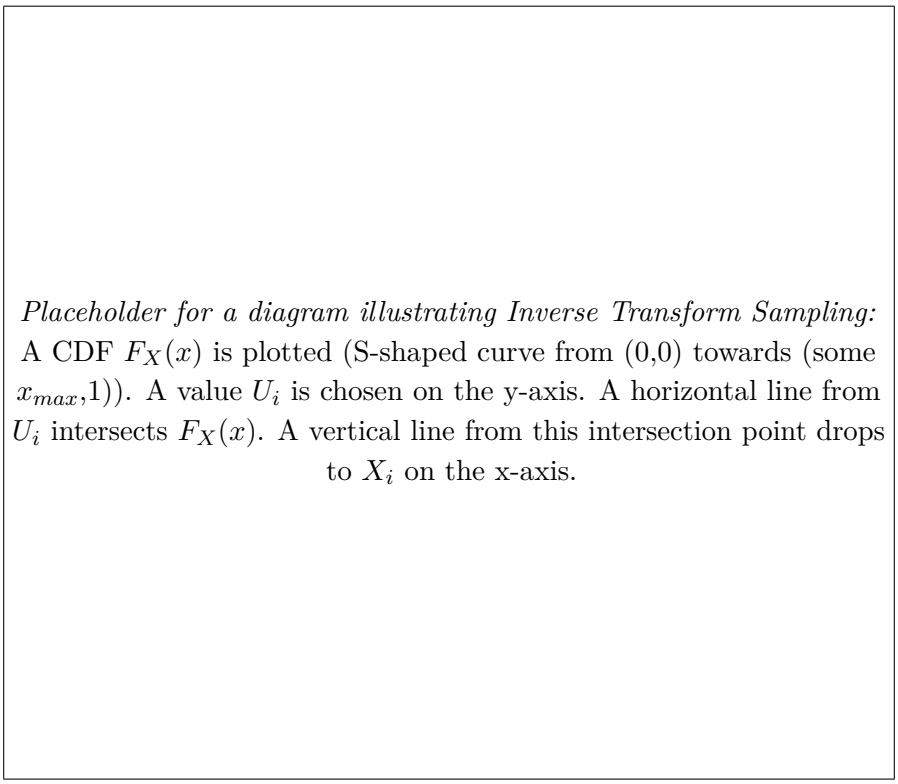
To generate n i.i.d. samples $X_1, \ldots, X_n$ from F_X:

1. Generate n i.i.d. samples $U_1, \ldots, U_n$ from $\text{Uniform}(0, 1)$.
2. For each $i = 1, \ldots, n$, compute $X_i = F_X^{-1}(U_i)$.

Pictorial Intuition: Imagine the graph of the CDF $F_X(x)$. The y-axis ranges from 0 to 1, and the x-axis represents the values of the random variable.

1. **Sample U:** Pick a random height u on the y-axis (this is your $U_i \sim \text{Uniform}(0, 1)$).
2. **Invert:** From this height u, draw a horizontal line to the right until it intersects the graph of $F_X(x)$.
3. **Find X:** From the intersection point, draw a vertical line down to the x-axis. The value x where this line meets the x-axis is your sample $X_i = F_X^{-1}(U_i)$.

Regions on the x-axis where $F_X(x)$ is steep (corresponding to high probability density) will have a larger range of u values mapping to them, so they will be "hit" more frequently. Conversely, regions where $F_X(x)$ is flat (low probability density) will be hit less often. This mechanism naturally generates samples according to the desired distribution.



Remark 4.2. The primary challenge in applying the inverse transform sampling method often lies in finding a closed-form expression for the inverse CDF, $F_X^{-1}(u)$, or in being able to compute it efficiently. For some common distributions (like Exponential), $F_X^{-1}(u)$ is simple. For others (like Normal), it's not available in closed form, and other methods are typically used.