MODULE 2: STATISTICAL DISTRIBUTIONS

Comprehensive Course Analysis for Statistics for Data Science

Course: Statistics for Data Science
Instructor: Bruno Damásio
Institution: NOVA IMS (Information Management School)
Academic Year: 2025/2026

TABLE OF CONTENTS

1. Executive Summary
2. Module Overview
3. Subject Analysis
    - 3.1 Motivation and Important Concepts
    - 3.2 Discrete Distributions
    - 3.3 Continuous Distributions
4. Knowledge Base Requirements
5. Techniques and Methods
6. Challenges and Common Pitfalls
7. Complexity Rankings
8. R Programming Reference
9. Distribution Relationships
10. Practice Problems and Solutions

---

1. EXECUTIVE SUMMARY

This module covers the foundational statistical distributions essential for data science applications. The content spans both discrete distributions (Bernoulli, Binomial, Poisson) and continuous distributions (Normal, Chi-squared, Student's t, Snedecor's F). Understanding these distributions is critical for statistical modeling, hypothesis testing, and building the mathematical framework for data analysis.

Key Learning Outcomes:

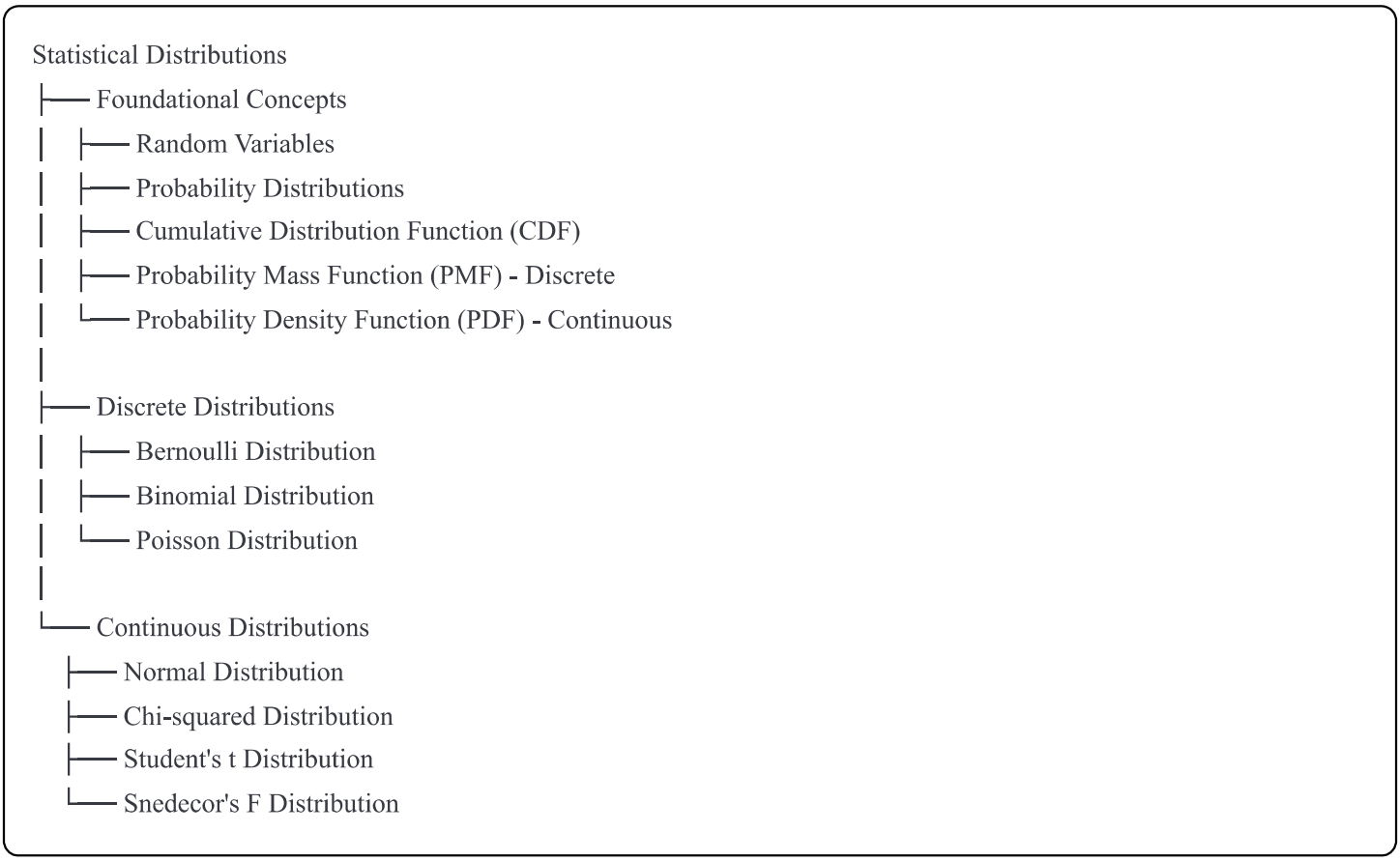
- Understand random variables and probability distributions
- Master discrete probability distributions and their applications

- Apply continuous distributions in real-world scenarios
- Compute probabilities using R programming
- Recognize relationships between different distributions

---

2. MODULE OVERVIEW

2.1 Core Concepts Hierarchy



2.2 Module Structure

Section	Topics Covered	Slide Range
Motivation & Concepts	Random variables, CDF, PMF, PDF	Pages 2-10
Discrete Distributions	Bernoulli, Binomial, Poisson	Pages 11-36
Continuous Distributions	Normal, χ^2, t, F distributions	Pages 37-72

---

3. SUBJECT ANALYSIS

3.1 MOTIVATION AND IMPORTANT CONCEPTS

3.1.1 Random Variables

Definition: A random variable is a variable that can take on certain numerical values with certain probabilities.

Key Properties:

- Associates numerical outcomes with probability values
- Foundation for all probability distributions
- Can be discrete (countable outcomes) or continuous (uncountable outcomes)

3.1.2 Probability Distribution

Definition: The collection of probabilities associated with a random variable specifying how the total probability (always equals 1) is distributed among various possible outcomes.

Importance in Data Science:

- Building blocks of statistical models
- Describes the data generation process
- Enables probability estimation for any specific observation

3.1.3 Cumulative Distribution Function (CDF)

Mathematical Definition:

$$F(x) = P(X \leq x), \text{ for all } x$$

Properties:

- Non-decreasing function
- Right-continuous
- $F(-\infty) = 0$, $F(\infty) = 1$
- Applies to both discrete and continuous random variables

Example (Three Fair Coins): For X = number of heads when tossing three fair coins:

$$
F(x) = \begin{cases} 0, & -\infty < x < 0 \\ 1/8, & 0 \le x < 1 \\ 1/2, & 1 \le x < 2 \\ 7/8, & 2 \le x < 3 \\ 1, & 3 \le x < \infty \end{cases}
$$

3.1.4 Probability Mass Function (PMF) - Discrete

Definition:

$$f(x) = P(X = x), \text{ for all } x$$

Application: Used for discrete random variables where outcomes are countable integers.

3.1.5 Probability Density Function (PDF) - Continuous

Definition:

$$F(x) = \int_{-\infty}^{x} f(t)dt, \text{ for all } x$$

Key Insight: For continuous distributions, probabilities are calculated as areas under the curve.

Example (Logistic Curve):

$$f(x) = e^{-x} / (1 + e^{-x})^2$$

$$P(1.5 \le X \le 2.5) = \int_{1.5}^{2.5} f(x)dx$$

3.2 DISCRETE DISTRIBUTIONS

3.2.1 Bernoulli Distribution

Definition: A discrete probability distribution of a random variable which takes value 1 with probability p and value 0 with probability q = 1 - p.

Notation: $X \sim Ber(p)$

PMF:

$$f(x) = p^x \times (1-p)^{1-x}, \text{ for } x \in \{0, 1\}$$

Properties:

Property	Value
Mean (Expected Value)	$E(X) = p$
Variance	$Var(X) = p(1-p) = pq$

Derivation of Mean:

$$E[X] = P(X=1)\times 1 + P(X=0)\times 0 = p\times 1 + q\times 0 = p$$

Derivation of Variance:

$$E[X^2] = P(X=1)\times 1^2 + P(X=0)\times 0^2 = p$$
$$Var[X] = E[X^2] - E[X]^2 = p - p^2 = p(1-p) = pq$$

Applications:

- Coin toss modeling (fair or biased)
- Yes/No questions
- Success/Failure experiments
- Binary classification outcomes

3.2.2 Binomial Distribution

Definition: The discrete probability distribution of the number of successes in a sequence of n independent Bernoulli trials.

Notation: $X \sim Bin(n, p)$

PMF:

$$f(x) = P(X = x) = C(n,x) \times p^x \times (1-p)^{(n-x)}, \text{ for } x = 0, 1, 2, ..., n$$

Where $C(n,x) = n! / [x!(n-x)!]$ is the binomial coefficient.

Properties:

Property	Value
Mean	$E[X] = np$
Variance	$Var[X] = np(1-p)$

Special Case: When n = 1, the binomial distribution reduces to a Bernoulli distribution.

Understanding the Formula:

- p^x: probability of x successes
- (1-p)^(n-x): probability of (n-x) failures
- C(n,x): number of ways to arrange x successes in n trials

Derivation of Mean (via Bernoulli sum):

If $X_1, X_2, ..., X_n$ are independent Bernoulli(p) random variables:
$X = X_1 + X_2 + ... + X_n$
$E[X] = E[X_1] + E[X_2] + ... + E[X_n] = p + p + ... + p = np$

R Implementation:

```r
# PMF: probability of exactly x successes
dbinom(x, size=n, prob=p)

# CDF: probability of x or fewer successes
pbinom(x, size=n, prob=p)
```

Example Problem: *12-question multiple choice quiz, 5 options per question, random guessing.*

Finding $P(X \leq 4)$:

```r
```

```
# Method 1: Sum individual probabilities
dbinom(0, size=12, prob=0.2) +
dbinom(1, size=12, prob=0.2) +
dbinom(2, size=12, prob=0.2) +
dbinom(3, size=12, prob=0.2) +
dbinom(4, size=12, prob=0.2)
# Result: 0.9274445

# Method 2: Use CDF directly
pbinom(4, size=12, prob=0.2)
# Result: 0.9274445
```

3.2.3 Poisson Distribution

Definition: A discrete probability distribution expressing the probability of a given number of events occurring in a fixed interval of time or space, given a known constant mean rate.

Notation: X ~ Poisson(λ)

PMF:

```
f(x) = P(X = x) = (λ^x × e^(-λ)) / x!, for x = 0, 1, 2, ...
```

Where:

- e = 2.71828... (Euler's number)
- λ = rate parameter (average number of events)
- x! = factorial of x

Properties:

Property	Value
Mean	$E(X) = \lambda$
Variance	$Var(X) = \lambda$

Key Feature: Mean equals variance (equidispersion).

Assumptions:

1. Events occur independently

2. Average rate is constant
3. Two events cannot occur at exactly the same instant

Applications:

- Phone calls per hour at a call center
- Radioactive decay events per second
- Daily mail volume
- Doctor/museum visits
- Website hits per minute

R Implementation:

```
r

# PMF: probability of exactly x events
dpois(x, lambda=λ)

# CDF: probability of x or fewer events (lower tail)
ppois(x, lambda=λ)

# Upper tail: probability of more than x events
ppois(x, lambda=λ, lower=FALSE)
```

Example Problem: *12 cars crossing a bridge per minute on average. Find $P(X \geq 17)$.*

```
r

# Lower tail (16 or fewer)
ppois(16, lambda=12)  # 0.898709

# Upper tail (17 or more)
ppois(16, lambda=12, lower=FALSE)  # 0.101291
```

3.3 CONTINUOUS DISTRIBUTIONS

3.3.1 Normal Distribution

Definition: A continuous probability distribution characterized by its bell-shaped curve, defined by mean μ and variance σ^2.

Notation: $X \sim N(\mu, \sigma^2)$

PDF:

$$\varphi(x) = (1 / \sqrt{(2\pi\sigma^2)}) \times \exp\{-(1/(2\sigma^2))(x - \mu)^2\}$$

Standard Normal Distribution: $Z \sim N(0, 1)$

- Mean = 0
- Standard deviation = 1
- Also known as the Z distribution

Properties:

Property	Value
Mean	μ
Variance	σ^2
Standard Deviation	σ
Symmetry	Symmetric around μ
Total Area	1

Effect of Parameters:

- Larger $\sigma^2 \rightarrow$ wider spread
- Smaller $\sigma^2 \rightarrow$ narrower spread
- μ shifts the distribution left or right

R Implementation:

```
r
```

```
# PDF: density at point x
dnorm(x, mean=μ, sd=σ)

# CDF: P(X ≤ x)
pnorm(x, mean=μ, sd=σ, lower.tail=TRUE)

# Upper tail: P(X > x)
pnorm(x, mean=μ, sd=σ, lower.tail=FALSE)

# Quantile: find x given probability
qnorm(p, mean=μ, sd=σ, lower.tail=TRUE)
```

Probability Calculations:

1. **"Less than" problems:**

```
r

# P(X ≤ 120) for N(102, 64)
pnorm(120, mean=102, sd=8, lower.tail=TRUE)  # 0.9877755
```

2. **"Greater than" problems:**

```
r

# P(X > 120) for N(102, 64)
pnorm(120, mean=102, sd=8, lower.tail=FALSE)  # 0.01222447
```

3. **"Between" problems:**

```
r

# P(0 ≤ Z ≤ 1.75) for standard normal
pnorm(1.75, mean=0, sd=1) - pnorm(0, mean=0, sd=1)  # 0.4599408
```

4. **"More extreme than" problems:**

```
r

# P(|Z| > 2) for standard normal
pnorm(2, mean=0, sd=1, lower.tail=FALSE) * 2  # 0.04550026
```

5. **Finding quantiles:**

```r
# What value separates top 10%?
qnorm(0.9, mean=102, sd=8, lower.tail=TRUE)  # 112.2524
qnorm(0.1, mean=102, sd=8, lower.tail=FALSE)  # 112.2524

# What z-scores bound the middle 90%?
qnorm(0.95, mean=0, sd=1, lower.tail=TRUE)  # 1.644854
```

3.3.2 Chi-Squared (χ^2) Distribution

Definition: If $X_1, X_2, ..., X_n$ are n independent standard normal random variables, then the sum of their squares follows a chi-squared distribution with n degrees of freedom.

Notation: $V \sim \chi^2(n)$

Mathematical Form:

$$V = X_1^2 + X_2^2 + ... + X_n^2 \sim \chi^2(n), \text{ where } X_i \sim N(0,1)$$

Properties:

Property	Value
Mean	n (degrees of freedom)
Variance	2n
Support	$[0, \infty)$
Shape	Right-skewed, becomes more symmetric as df increases

R Implementation:

```r
```

```
# PDF
dchisq(x, df=n)

# CDF: P(X ≤ x)
pchisq(x, df=n)

# Quantile: find x given probability
qchisq(p, df=n)
```

Example:

```
r

# 95th percentile with 7 degrees of freedom
qchisq(0.95, df=7)  # 14.06714
```

Applications:

- Goodness-of-fit tests
- Tests of independence
- Variance estimation

3.3.3 Student's t Distribution

Definition: A continuous probability distribution that arises when estimating the mean of a normally distributed population when sample size is small and population standard deviation is unknown.

Properties:

Property	Value
Symmetry	Symmetric around 0
Shape	Bell-shaped but heavier tails than normal
Convergence	Approaches N(0,1) as df → ∞
Mean	0 (for df > 1)
Variance	df/(df-2) (for df > 2)

Key Characteristics:

- More prone to extreme values than normal distribution
- "Heavier tails" provide more conservative estimates
- With more degrees of freedom, becomes closer to standard normal

R Implementation:

```r
# PDF
dt(x, df=n)

# CDF: P(X ≤ x)
pt(x, df=n)

# Quantile: find x given probability
qt(p, df=n)
```

Example:

```r
# 2.5th and 97.5th percentiles with 5 df
qt(c(0.025, 0.975), df=5)  # -2.570582, 2.570582
```

Applications:

- Student's t-test
- Confidence intervals for means
- Linear regression analysis

3.3.4 Snedecor's F Distribution

Definition: A continuous probability distribution that arises as the ratio of two chi-squared distributions divided by their respective degrees of freedom.

Named after: Ronald Fisher and George W. Snedecor

Notation: $F \sim F(df_1, df_2)$

Properties:

Property	Value
Support	$[0, \infty)$
Shape	Right-skewed
Parameters	Two degrees of freedom (df_1, df_2)
Mean	$df_2/(df_2-2)$ for $df_2 > 2$

R Implementation:

```r
# PDF
df(x, df1=n1, df2=n2)

# CDF: P(X ≤ x)
pf(x, df1=n1, df2=n2)

# Quantile: find x given probability
qf(p, df1=n1, df2=n2)
```

Example:

```r
# 95th percentile with (12, 5) degrees of freedom
qf(0.95, df1=12, df2=5)  # 4.677704
```

Applications:

- F-test for comparing variances
- Analysis of Variance (ANOVA)
- Regression analysis (overall model significance)

---

4. KNOWLEDGE BASE REQUIREMENTS

4.1 Mathematical Prerequisites

Topic	Required Knowledge	Complexity
Calculus	Integration, differentiation	Medium
Algebra	Factorial, combinatorics, exponentials	Medium
Probability Theory	Basic probability, conditional probability	High
Set Theory	Sample spaces, events	Low

4.2 Statistical Foundations

Concept	Description	Prerequisite For
Expected Value	Weighted average of outcomes	All distributions
Variance	Measure of spread	All distributions
Standard Deviation	Square root of variance	Normal distribution
Degrees of Freedom	Parameters in distribution	χ^2, t, F distributions

4.3 Programming Skills (R)

Function Type	Prefix	Description
Density/Mass	d	Probability at a point
Cumulative	p	Probability up to a point
Quantile	q	Value at a given probability
Random	r	Generate random samples

5. TECHNIQUES AND METHODS

5.1 Probability Calculation Techniques

Scenario	Technique	Formula
$P(X = x)$ discrete	PMF	Use dbinom, dpois
$P(X \leq x)$	CDF (lower tail)	Use pbinom, ppois, pnorm
$P(X > x)$	Upper tail	$1 - P(X \leq x)$ or lower.tail=FALSE
$P(a \leq X \leq b)$	Interval	$P(X \leq b) - P(X \leq a)$
Find x for $P(X \leq x) = p$	Quantile	Use qnorm, qt, qchisq, qf

5.2 Distribution Selection Guide

Data Characteristic	Recommended Distribution
Binary outcomes	Bernoulli
Count of successes in n trials	Binomial
Count of events in fixed interval	Poisson
Continuous symmetric data	Normal
Sum of squared standard normals	Chi-squared
Small sample mean estimation	Student's t
Ratio of variances	F

5.3 Standardization Technique

Z-Score Transformation:

$$Z = (X - \mu) / \sigma$$

This converts any normal distribution $N(\mu, \sigma^2)$ to standard normal $N(0, 1)$.

---

6. CHALLENGES AND COMMON PITFALLS

6.1 Conceptual Challenges

Challenge	Description	Resolution
Discrete vs Continuous	Confusing PMF and PDF	Remember: PMF gives $P(X=x)$, PDF requires integration
CDF Interpretation	Understanding cumulative probabilities	$F(x) = P(X \leq x)$, always $\leq$
Parameter Confusion	Mixing up parameters	Binomial: n, p; Poisson: λ; Normal: μ, σ^2
Tail Probabilities	Upper vs lower tail	Lower = $P(X \leq x)$; Upper = $P(X > x)$

6.2 Calculation Errors

Error Type	Example	Correction
Off-by-one	$P(X \geq 17)$ calculated as $P(X > 17)$	$P(X \geq 17) = P(X > 16)$ for discrete
Wrong tail	Using lower.tail for upper probability	Set lower.tail=FALSE
Missing transformation	Using N(0,1) formulas for $N(\mu,\sigma^2)$	Standardize first or specify parameters
df confusion	Wrong degrees of freedom in t or F	$\chi^2(n)$, $t(n-1)$, $F(n_1-1, n_2-1)$

6.3 R Programming Pitfalls

Pitfall	Example	Correct Usage
sd vs variance	pnorm uses sd, not variance	pnorm(x, mean=μ, sd=σ)
Default parameters	Assuming default mean/sd	Always specify explicitly
Vectorization errors	Expecting single value	R functions are vectorized

7. COMPLEXITY RANKINGS

7.1 Overall Topic Complexity (1-10 Scale)

Topic	Conceptual	Computational	Application	Overall
Random Variables	5	2	4	3.7
CDF/PMF/PDF	6	4	5	5.0
Bernoulli	3	2	4	3.0
Binomial	5	5	6	5.3
Poisson	5	4	6	5.0
Normal	6	5	8	6.3
Chi-squared	7	4	7	6.0
Student's t	7	4	8	6.3
F Distribution	8	5	8	7.0

7.2 Learning Sequence Recommendation

Level 1 (Foundation):
├── Random Variables
├── Probability Distribution Concepts
└── CDF, PMF, PDF

Level 2 (Discrete):
├── Bernoulli Distribution
├── Binomial Distribution
└── Poisson Distribution

Level 3 (Continuous Basic):
├── Normal Distribution
└── Standardization

Level 4 (Continuous Advanced):
├── Chi-squared Distribution
├── Student's t Distribution
└── Snedecor's F Distribution

Level 5 (Integration):
└── Distribution Relationships

---

8. R PROGRAMMING REFERENCE

8.1 Discrete Distribution Functions

Distribution	PMF	CDF	Notes
Binomial	dbinom(x, size, prob)	pbinom(x, size, prob)	size=n trials, prob=p success
Poisson	dpois(x, lambda)	ppois(x, lambda)	lambda=λ rate

8.2 Continuous Distribution Functions

Distribution	PDF	CDF	Quantile	Parameters
Normal	dnorm(x, mean, sd)	pnorm(x, mean, sd)	qnorm(p, mean, sd)	mean=μ, sd=σ
Chi-squared	dchisq(x, df)	pchisq(x, df)	qchisq(p, df)	df=degrees of freedom
Student's t	dt(x, df)	pt(x, df)	qt(p, df)	df=degrees of freedom
F	df(x, df1, df2)	pf(x, df1, df2)	qf(p, df1, df2)	df1, df2=degrees of freedom

8.3 Common R Patterns

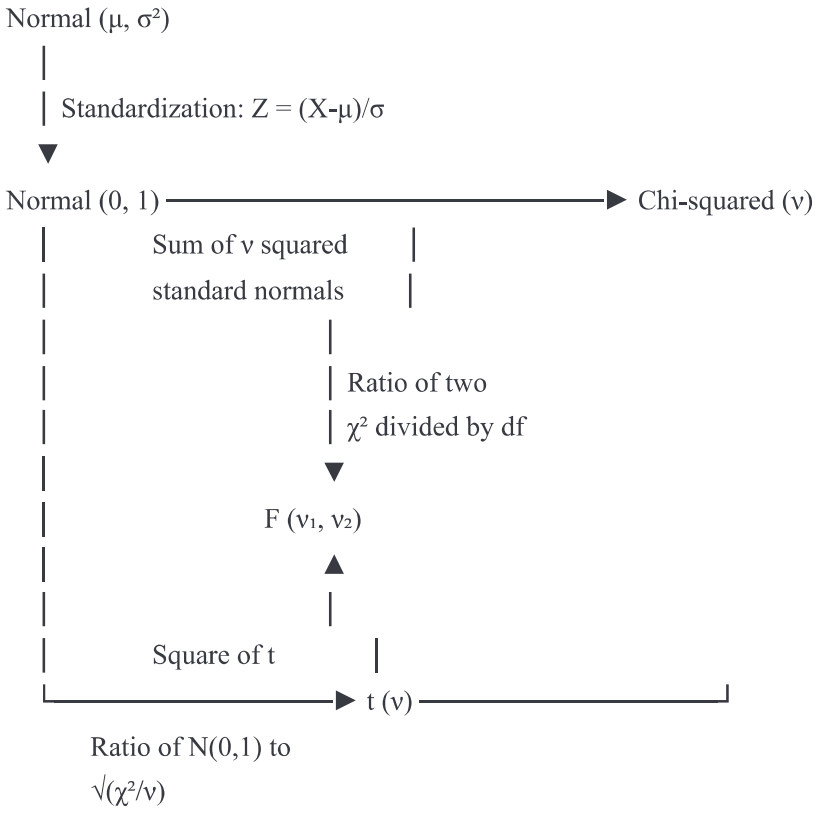
```
r
```

```r
# Probability calculations
pnorm(x, mean=μ, sd=σ, lower.tail=TRUE)   # P(X ≤ x)
pnorm(x, mean=μ, sd=σ, lower.tail=FALSE)  # P(X > x)

# Interval probability
pnorm(b, mean=μ, sd=σ) - pnorm(a, mean=μ, sd=σ)  # P(a < X ≤ b)

# Two-tailed probability
2 * pnorm(z, mean=0, sd=1, lower.tail=FALSE)  # P(|Z| > z)

# Critical values
qnorm(0.975, mean=0, sd=1)  # z for 95% confidence (two-tailed)
qt(0.975, df=n-1)           # t for 95% confidence
qchisq(0.95, df=n-1)        # χ² for 95% confidence
qf(0.95, df1=k-1, df2=n-k)  # F for 95% confidence
```

---

9. DISTRIBUTION RELATIONSHIPS

9.1 Relationship Diagram



9.2 Key Relationships

Relationship	Mathematical Form
Normal to Standard Normal	$Z = (X - \mu) / \sigma$
Standard Normal to Chi-squared	$\Sigma_i Z_i^2 \sim \chi^2(n)$
Chi-squared to F	$(\chi_1^2/\nu_1) / (\chi_2^2/\nu_2) \sim F(\nu_1, \nu_2)$
t-squared to F	$t^2(\nu) \sim F(1, \nu)$
Normal to t	$Z/\sqrt{(\chi^2/\nu)} \sim t(\nu)$

9.3 Limiting Distributions

As df → ∞	Converges To
$t(\nu)$	$N(0, 1)$
$\chi^2(\nu)/\nu$	1
$F(\nu_1, \infty)$	$\chi^2(\nu_1)/\nu_1$

---

10. PRACTICE PROBLEMS AND SOLUTIONS

Problem 1: Binomial Distribution

Question: A quiz has 12 multiple choice questions with 5 options each. If guessing randomly, what is $P(X \leq 4)$?

Solution:

```
r

pbinom(4, size=12, prob=0.2)  # 0.9274445
```

Interpretation: There is a 92.7% probability of getting 4 or fewer correct by random guessing.

Problem 2: Poisson Distribution

Question: If 12 cars cross a bridge per minute on average, what is $P(X \geq 17)$?

Solution:

```r
ppois(16, lambda=12, lower=FALSE)  # 0.101291
```

Interpretation: There is a 10.1% probability of 17 or more cars crossing in any given minute.

Problem 3: Normal Distribution

Question: Vehicle speeds are N(102, 64) km/h. What is $P(X \leq 120)$?

Solution:

```r
pnorm(120, mean=102, sd=8, lower.tail=TRUE)  # 0.9877755
```

Interpretation: 98.8% of vehicles travel at 120 km/h or slower.

Problem 4: Finding Quantiles

Question: What speed separates the top 10% of vehicles (N(102, 64))?

Solution:

```r
qnorm(0.9, mean=102, sd=8)  # 112.2524 km/h
```

Problem 5: Chi-squared Distribution

Question: Find the 95th percentile of $\chi^2(7)$.

Solution:

```r
qchisq(0.95, df=7)  # 14.06714
```

Problem 6: Student's t Distribution

Question: Find the critical values for a 95% confidence interval with 5 df.

Solution:

```r
qt(c(0.025, 0.975), df=5)  # -2.570582, 2.570582
```

Problem 7: F Distribution

Question: Find the 95th percentile of F(12, 5).

Solution:

```r
qf(0.95, df1=12, df2=5)  # 4.677704
```

SUMMARY

This module provides the essential foundation for understanding probability distributions in data science. The progression from discrete to continuous distributions, along with their R implementations, prepares students for hypothesis testing, confidence interval construction, and advanced statistical modeling covered in subsequent modules.

Key Takeaways:

1. Random variables map outcomes to probabilities
2. Discrete distributions (Bernoulli, Binomial, Poisson) handle countable outcomes
3. Continuous distributions (Normal, χ^2, t, F) handle uncountable outcomes
4. R provides consistent functions (d, p, q prefixes) for all distributions
5. Understanding distribution relationships enables proper test selection

---

Document prepared as Module 2 analysis for Statistics for Data Science course Total Slides Analyzed: 73 (Pages 1-72)