# STAT340 Comprehensive Study Guide

## Probability, Statistics, and Monte Carlo Methods

---

## Table of Contents



---

# 1. Probability and Random Variables

## 1.1 Probability Spaces and Events

### 1.1.1 Sample Space

**Definition 1.1** (Sample Space): The **sample space** $\Omega$ is the set of all possible outcomes of a random experiment.

**Examples:** - Coin flip: $\Omega = \{\mathrm{H}, \mathrm{T}\}$ - Die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$ - Continuous measurement: $\Omega = \mathbb{R}$ or $\Omega = [a, b]$

### 1.1.2 Events

**Definition 1.2** (Event): An **event** $A$ is a subset of the sample space $\Omega$, i.e., $A \subseteq \Omega$.

**Definition 1.3** (Event Operations): - **Union**: $A \cup B$ represents "A or B occurs" - **Intersection**: $A \cap B$ represents "A and B both occur" - **Complement**: $A^c$ represents "A does not occur" - **Mutually Exclusive**: Events $A$ and $B$ are mutually exclusive if $A \cap B = \emptyset$

### 1.1.3 Axioms of Probability

**Axiom 1** (Non-negativity): For any event $A$, $P(A) \geq 0$.

**Axiom 2** (Normalization): $P(\Omega) = 1$.

**Axiom 3** (Countable Additivity): For mutually exclusive events $A_1, A_2, \ldots$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

### 1.1.4 Fundamental Probability Rules

**Theorem 1.1** (Addition Rule): For any events $A$ and $B$,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Proof Intuition**: The sum $P(A) + P(B)$ counts $P(A \cap B)$ twice, so we subtract it once.

**Theorem 1.2** (Complement Rule): For any event $A$,

$$P(A^c) = 1 - P(A)$$

**Theorem 1.3** (Conditional Probability): For events $A$ and $B$ with $P(B) > 0$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

**Interpretation**: $P(A \mid B)$ is the probability of $A$ occurring given that $B$ has occurred. We "zoom in" on the portion of the sample space where $B$ is true and renormalize.

**Theorem 1.4** (Multiplication Rule): For events $A$ and $B$,

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

## 1.2 Random Variables

### 1.2.1 Definition and Types

**Definition 1.4** (Random Variable): A **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to each outcome in the sample space.

**Definition 1.5** (Types of Random Variables): - **Discrete**: $X$ takes countably many values (e.g., integers) - **Continuous**: $X$ takes uncountably many values (e.g., real numbers in an interval)

### 1.2.2 Probability Mass Function (PMF)

**Definition 1.6** (PMF): For a discrete random variable $X$, the **probability mass function** is

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

**Properties**: 1. $p_X(x) \geq 0$ for all $x$ 2. $\sum_{\text{all } x} p_X(x) = 1$

**Example 1.1**: Fair six-sided die with $X =$ outcome.

$$p_X(k) = \frac{1}{6} \text{ for } k \in \{1, 2, 3, 4, 5, 6\}$$

### 1.2.3 Probability Density Function (PDF)

**Definition 1.7** (PDF): For a continuous random variable $X$, the **probability density function** $f_X(x)$ satisfies

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

**Properties**: 1. $f_X(x) \geq 0$ for all $x$ 2. $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$ 3. $P(X = x) = 0$ for any specific value $x$ (continuous case)

**Note**: For continuous random variables, probabilities are computed over intervals, not at individual points.

### 1.2.4 Cumulative Distribution Function (CDF)

**Definition 1.8** (CDF): For any random variable $X$, the **cumulative distribution function** is

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

**Properties**: 1. $F_X(x)$ is non-decreasing: if $x_1 < x_2$, then $F_X(x_1) \leq F_X(x_2)$ 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

**Relationships**: - **Discrete**: $F_X(x) = \sum_{k \leq x} p_X(k)$ - **Continuous**: $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$ and $f_X(x) = \frac{d}{dx} F_X(x)$

**Theorem 1.5** (Probability from CDF): For $a < b$,

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

## 1.3 Expected Value and Variance

### 1.3.1 Expected Value

**Definition 1.9** (Expected Value): The **expected value** (or mean) of a random variable $X$ is

- **Discrete**: $E[X] = \sum_{\text{all } x} x \cdot p_X(x)$
- **Continuous**: $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$

**Interpretation**: $E[X]$ represents the long-run average value of $X$ over many independent repetitions.

**Example 1.2**: Fair die with $X =$ outcome.

$$E[X] = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

**Theorem 1.6** (Law of the Unconscious Statistician): For a function $g$ and random variable $X$,

- **Discrete**: $E[g(X)] = \sum_{\text{all } x} g(x) \cdot p_X(x)$
- **Continuous**: $E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$

**Theorem 1.7** (Linearity of Expectation): For random variables $X$ and $Y$ and constants $a, b$,

$$E[aX + bY] = a \cdot E[X] + b \cdot E[Y]$$

**Important**: This holds **regardless** of whether $X$ and $Y$ are independent.

### 1.3.2 Variance

**Definition 1.10** (Variance): The **variance** of a random variable $X$ is

$$\text{Var}(X) = E[(X - \mu)^2] \text{ where } \mu = E[X]$$

**Theorem 1.8** (Computational Formula for Variance):

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

**Proof**:

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 = E[X^2] - (E[X])^2 \end{aligned}$$

**Definition 1.11** (Standard Deviation): The **standard deviation** of $X$ is

$$\text{SD}(X) = \sigma_X = \sqrt{\text{Var}(X)}$$

**Interpretation**: Standard deviation measures the typical deviation from the mean, in the same units as $X$.

**Theorem 1.9** (Properties of Variance): 1. $\text{Var}(aX + b) = a^2 \text{Var}(X)$ for constants $a, b$ 2. $\text{Var}(X) \geq 0$ with equality iff $X$ is constant 3. If $X$ and $Y$ are **independent**, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$

## 1.4 Common Discrete Distributions

### 1.4.1 Bernoulli Distribution

**Definition 1.12**: $X \sim \text{Bernoulli}(p)$ if $X \in \{0, 1\}$ with

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

where $0 \leq p \leq 1$.

**Properties**: - $E[X] = p$ - $\text{Var}(X) = p(1-p)$ - PMF: $p_X(k) = p^k(1-p)^{1-k}$ for $k \in \{0, 1\}$

**Applications**: Single trial with two outcomes (success/failure, yes/no, heads/tails).

### 1.4.2 Binomial Distribution

**Definition 1.13**: $X \sim \text{Binomial}(n, p)$ represents the number of successes in $n$ independent Bernoulli$(p)$ trials.

**PMF**:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

**Properties**: - $E[X] = np$ - $\text{Var}(X) = np(1-p)$

**Derivation of Expected Value**:

$$E[X] = E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n] = np$$

where each $X_i \sim \text{Bernoulli}(p)$.

**Applications**: Number of heads in $n$ coin flips, number of defective items in a sample, number of successes in fixed trials.

### 1.4.3 Geometric Distribution

**Definition 1.14**: $X \sim \text{Geometric}(p)$ represents the number of **failures** before the first success in independent Bernoulli$(p)$ trials.

**PMF**:

$$P(X = k) = (1-p)^k p \text{ for } k = 0, 1, 2, \ldots$$

**Properties**: - $E[X] = \frac{1-p}{p}$ - $\text{Var}(X) = \frac{1-p}{p^2}$

**Memoryless Property**: $P(X > m + n \mid X > m) = P(X > n)$ for all $m, n \geq 0$.

**Applications**: Number of failures before first success, waiting times for discrete events.

**Note**: Some texts define Geometric as the number of trials until first success, giving $E[X] = \frac{1}{p}$.

### 1.4.4 Poisson Distribution

**Definition 1.15**: $X \sim \text{Poisson}(\lambda)$ where $\lambda > 0$ is the rate parameter.

**PMF**:

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

**Properties**: - $E[X] = \lambda$ - $\text{Var}(X) = \lambda$ - **Unique property**: Mean equals variance

**Theorem 1.10** (Poisson Limit Theorem): If $X_n \sim \text{Binomial}(n, p_n)$ where $n \to \infty$, $p_n \to 0$, and $np_n \to \lambda$, then

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

**Applications**: - Number of events in a fixed time interval (emails per hour, customers per day) - Rare events with large sample size - Approximation to Binomial when $n$ large, $p$ small, $np$ moderate

### 1.4.5 Discrete Uniform Distribution

**Definition 1.16**: $X \sim \text{Uniform}(\{a, a+1, \ldots, b\})$ if each value has equal probability.

**PMF**:

$$P(X = k) = \frac{1}{b - a + 1} \text{ for } k \in \{a, a+1, \ldots, b\}$$

**Properties**: - $E[X] = \frac{a+b}{2}$ - $\text{Var}(X) = \frac{(b-a+1)^2-1}{12}$

**Applications**: Fair die rolls, random selection from finite set.

## 1.5 Common Continuous Distributions

### 1.5.1 Uniform Distribution

**Definition 1.17**: $X \sim \text{Uniform}(a, b)$ if $X$ is equally likely over $[a, b]$.

**PDF**:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

**CDF**:

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases}$$

**Properties**: - $E[X] = \frac{a+b}{2}$ - $\text{Var}(X) = \frac{(b-a)^2}{12}$

**Applications**: Random numbers, modeling complete uncertainty over an interval.

### 1.5.2 Exponential Distribution

**Definition 1.18**: $X \sim \text{Exponential}(\lambda)$ where $\lambda > 0$ is the rate parameter.

**PDF**:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

**CDF**:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

**Properties**: - $E[X] = \frac{1}{\lambda}$ - $\text{Var}(X) = \frac{1}{\lambda^2}$

**Theorem 1.11** (Memoryless Property): For $s, t \geq 0$,

$$P(X > s + t \mid X > s) = P(X > t)$$

**Relationship to Poisson**: If events occur according to a Poisson process with rate $\lambda$, the time until the next event follows Exponential$(\lambda)$.

**Applications**: Waiting times, lifetimes, time between events.

### 1.5.3 Normal (Gaussian) Distribution

**Definition 1.19**: $X \sim N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance.

**PDF**:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } x \in \mathbb{R}$$

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

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

**Properties**: - $E[X] = \mu$ - $\text{Var}(X) = \sigma^2$ - Symmetric about $\mu$ - Approximately 68% of probability within $[\mu - \sigma, \mu + \sigma]$ - Approximately 95% of probability within $[\mu - 2\sigma, \mu + 2\sigma]$ - Approximately 99.7% of probability within $[\mu - 3\sigma, \mu + 3\sigma]$

**Theorem 1.12** (Standardization): If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

**Theorem 1.13** (Linear Transformation): If $X \sim N(\mu, \sigma^2)$, then

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

**Theorem 1.14** (Sum of Normals): If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

**Applications**: Heights, weights, test scores, measurement errors, approximation to many distributions (CLT).

---

# 2. Independence and Conditional Probability

## 2.1 Independence

### 2.1.1 Independence of Events

**Definition 2.1** (Independent Events): Events $A$ and $B$ are **independent** if

$$P(A \cap B) = P(A) \cdot P(B)$$

**Equivalent condition**: $P(A \mid B) = P(A)$ (when $P(B) > 0$).

**Interpretation**: Knowing that $B$ occurred does not change the probability of $A$.

**Theorem 2.1**: If $A$ and $B$ are independent, then: 1. $A$ and $B^c$ are independent 2. $A^c$ and $B$ are independent 3. $A^c$ and $B^c$ are independent

**Definition 2.2** (Mutual Independence): Events $A_1, A_2, \ldots, A_n$ are **mutually independent** if for any subset $I \subseteq \{1, 2, \ldots, n\}$,

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

**Warning**: Pairwise independence does not imply mutual independence.

### 2.1.2 Independence of Random Variables

**Definition 2.3** (Independent Random Variables): Random variables $X$ and $Y$ are **independent** if for all sets $A, B \subseteq \mathbb{R}$,

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

**Equivalent conditions**: - **Discrete**: $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all $x, y$ - **Continuous**: $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all $x, y$ - **CDF**: $F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)$ for all $x, y$

**Theorem 2.2**: If $X$ and $Y$ are independent, then: 1. $E[XY] = E[X] \cdot E[Y]$ 2. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ 3. $E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$ for any functions $g, h$

**Important**: $E[XY] = E[X]E[Y]$ does **not** imply independence (converse is false).

## 2.2 Covariance and Correlation

### 2.2.1 Covariance

**Definition 2.4** (Covariance): The **covariance** between random variables $X$ and $Y$ is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

**Theorem 2.3** (Computational Formula):

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

**Proof**:

$$\begin{aligned}
\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\
&= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\
&= E[XY] - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
&= E[XY] - \mu_X \mu_Y = E[XY] - E[X]E[Y]
\end{aligned}$$

**Properties**: 1. $\text{Cov}(X, X) = \text{Var}(X)$ 2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (symmetric) 3. $\text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y)$ 4. If $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$

**Warning**: $\text{Cov}(X, Y) = 0$ does **not** imply independence (e.g., $Y = X^2$ with $X$ symmetric around 0).

**Theorem 2.4** (Variance of Sum):

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

**Corollary**: For independent $X$ and $Y$,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

### 2.2.2 Correlation

**Definition 2.5** (Correlation Coefficient): The **correlation** between $X$ and $Y$ is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

**Theorem 2.5** (Properties of Correlation): 1. $-1 \leq \rho(X, Y) \leq 1$ 2. $|\rho(X, Y)| = 1$ iff $Y = aX + b$ for some constants $a \neq 0, b$ 3. $\rho(X, Y) = 0$ iff $\text{Cov}(X, Y) = 0$ (uncorrelated) 4. Correlation is unitless

**Interpretation**: - $\rho > 0$: Positive linear relationship - $\rho < 0$: Negative linear relationship - $\rho = 0$: No linear relationship - $|\rho| \approx 1$: Strong linear relationship - $|\rho| \approx 0$: Weak linear relationship

**Important**: Correlation measures **linear** relationships only. Variables can be strongly related nonlinearly yet have $\rho = 0$.

# 2.3 Conditional Probability and Bayes' Rule

## 2.3.1 Conditional Probability

**Definition 2.6** (Conditional PMF/PDF): - **Discrete**: $p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$ - **Continuous**: $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

**Theorem 2.6** (Conditional Expectation):

$$E[X \mid Y = y] = \begin{cases} \sum_x x \cdot p_{X|Y}(x \mid y) & \text{(discrete)} \\ \int x \cdot f_{X|Y}(x \mid y)\, dx & \text{(continuous)} \end{cases}$$

## 2.3.2 Law of Total Probability

**Theorem 2.7** (Law of Total Probability): If $B_1, B_2, \ldots, B_n$ partition the sample space (mutually exclusive and exhaustive), then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

**Application**: Useful for computing probabilities when conditioning on different scenarios.

## 2.3.3 Bayes' Rule

**Theorem 2.8** (Bayes' Rule): For events $A$ and $B$ with $P(B) > 0$,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

**Extended Form**: If $A_1, A_2, \ldots, A_n$ partition the sample space,

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) P(A_j)}$$

**Terminology**: - $P(A)$: **Prior probability** of $A$ - $P(B \mid A)$: **Likelihood** of $B$ given $A$ - $P(A \mid B)$: **Posterior probability** of $A$ given $B$ - $P(B)$: **Marginal probability** of $B$ (normalizing constant)

**Example 2.1** (Medical Testing): - Disease prevalence: $P(D) = 0.01$ - Test sensitivity: $P(+ \mid D) = 0.95$ - Test false positive rate: $P(+ \mid D^c) = 0.05$

Find $P(D \mid +)$ (probability of disease given positive test):

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = \frac{0.0095}{0.0590} \approx 0.161$$

**Interpretation**: Despite a positive test, only about 16% chance of having the disease due to low base rate.

# 3. Monte Carlo Simulations

## 3.1 Foundations of Monte Carlo Methods

### 3.1.1 Law of Large Numbers

**Theorem 3.1** (Weak Law of Large Numbers): Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0$$

**Interpretation**: The sample average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the population mean $\mu$ as $n \to \infty$.

**Practical Implication**: With sufficiently many simulations, the average of simulated values approximates the true expected value.

### 3.1.2 Standard Error

**Definition 3.1** (Standard Error): The standard error of the sample mean $\bar{X}_n$ is

$$\mathrm{SE}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

where $\sigma = \sqrt{\mathrm{Var}(X_i)}$.

**Interpretation**: - $\bar{X}_n$ is typically within $\approx 2 \cdot \mathrm{SE}(\bar{X}_n)$ of $\mu$ (95% confidence) - To halve the standard error, need 4 times as many simulations - Standard error decreases as $1/\sqrt{n}$

**Example 3.1**: With $\sigma = 10$ and $n = 100$ simulations,

$$\mathrm{SE}(\bar{X}_{100}) = \frac{10}{\sqrt{100}} = 1$$

Our estimate is accurate to within $\approx 2$ units (95% confidence).

### 3.1.3 Pseudorandom Number Generation

**Definition 3.2**: A **pseudorandom number generator** (PRNG) produces a deterministic sequence of numbers that appear random.

**Key Concepts**: - **Seed**: Initial value that determines the sequence - **Reproducibility**: Same seed produces same sequence - In R: `set.seed(seed_value)` sets the seed

**Example 3.2** (R code):

```r
set.seed(123)
runif(5)  # Always produces same 5 random numbers
# [1] 0.2875775 0.7883051 0.4089769 0.8830174 0.9404673
```

### 3.1.4 Inverse Transform Method

**Theorem 3.2** (Probability Integral Transform): If $X$ has CDF $F_X$, then $U = F_X(X) \sim \text{Uniform}(0, 1)$.

**Corollary**: If $U \sim \text{Uniform}(0, 1)$ and $F_X$ is a CDF, then $X = F_X^{-1}(U)$ has CDF $F_X$.

**Algorithm 3.1** (Inverse Transform Method): 1. Generate $U \sim \text{Uniform}(0, 1)$ 2. Compute $X = F_X^{-1}(U)$ 3. Then $X$ follows the distribution with CDF $F_X$

**Example 3.3** (Exponential Distribution): - CDF: $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$ - Inverse: $F_X^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$ - Generate: $U \sim \text{Uniform}(0, 1)$, then $X = -\frac{1}{\lambda}\ln(1-U) \sim \text{Exponential}(\lambda)$

## 3.2 Monte Carlo Estimation

### 3.2.1 Estimating Expected Values

**Problem**: Estimate $\mu = E[g(X)]$ where $X$ has known distribution.

**Monte Carlo Algorithm**: 1. Generate i.i.d. samples $X_1, X_2, \ldots, X_n$ from the distribution of $X$ 2. Compute $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n g(X_i)$ 3. By LLN, $\hat{\mu}_n \to \mu$ as $n \to \infty$

**Standard Error**: $\text{SE}(\hat{\mu}_n) = \frac{\sigma_g}{\sqrt{n}}$ where $\sigma_g = \sqrt{\text{Var}(g(X))}$

**Example 3.4**: Estimate $E[X^2]$ where $X \sim \text{Exponential}(2)$.

```r
set.seed(123)
n <- 10000
x <- rexp(n, rate = 2)
estimate <- mean(x^2)
se <- sd(x^2) / sqrt(n)
cat("Estimate:", estimate, "SE:", se)
# True value: Var(X) + (E[X])^2 = 1/4 + 1/4 = 0.5
```

### 3.2.2 Estimating Probabilities

**Problem**: Estimate $p = P(\text{event})$ for some event.

**Monte Carlo Algorithm**: 1. Generate $n$ random outcomes 2. Count how many satisfy the event condition: $k = \sum_{i=1}^n \mathbb{1}(\text{event}_i)$ 3. Estimate: $\hat{p}_n = \frac{k}{n}$

**This is equivalent to estimating $E[\mathbb{1}(\text{event})]$ where $\mathbb{1}$ is the indicator function.**

**Standard Error**: For Bernoulli($p$) random variables,

$$\text{SE}(\hat{p}_n) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

**Example 3.5** (Birthday Problem): Estimate probability that in a group of 23 people, at least 2 share a birthday.

```
set.seed(123)
n_sim <- 10000
shared <- replicate(n_sim, {
  birthdays <- sample(1:365, size = 23, replace = TRUE)
  length(unique(birthdays)) < 23
})
prob_estimate <- mean(shared)
se <- sqrt(prob_estimate * (1 - prob_estimate) / n_sim)
# Estimate ≈ 0.507, SE ≈ 0.005
```

### 3.2.3 Monte Carlo Integration

**Problem**: Estimate $I = \int_a^b g(x)\, dx$.

**Monte Carlo Algorithm**: 1. Note that $I = (b-a) \cdot E[g(U)]$ where $U \sim \text{Uniform}(a, b)$ 2. Generate $U_1, \ldots, U_n \sim \text{Uniform}(a, b)$ 3. Estimate: $\hat{I}_n = (b-a) \cdot \frac{1}{n} \sum_{i=1}^n g(U_i)$

**Derivation**:

$$I = \int_a^b g(x)\, dx = \int_a^b g(x) \cdot \frac{1}{b-a} \cdot (b-a)\, dx = (b-a) \cdot E[g(U)]$$

**Example 3.6**: Estimate $\int_0^1 x^2\, dx$ (true value = 1/3).

```
set.seed(123)
n <- 10000
u <- runif(n, 0, 1)
estimate <- 1 * mean(u^2)  # (b-a) = 1
# Estimate ≈ 0.333
```

**Example 3.7** (Estimating $\pi$): Generate random points $(x, y)$ in $[-1, 1] \times [-1, 1]$ and count how many fall inside the unit circle.

$$\pi = 4 \cdot P(x^2 + y^2 \leq 1) \text{ for } (x, y) \sim \text{Uniform}([-1, 1]^2)$$

```
set.seed(123)
n <- 100000
x <- runif(n, -1, 1)
y <- runif(n, -1, 1)
inside <- (x^2 + y^2 <= 1)
```

```
pi_estimate <- 4 * mean(inside)
# Estimate ≈ 3.14
```

## 3.3 Practical Considerations

### 3.3.1 Choosing the Number of Simulations

**Guidelines**: - **Quick estimate**: $n = 1{,}000$ - **Reasonable accuracy**: $n = 10{,}000$ - **High precision**: $n = 100{,}000$ or more

**Rule of thumb**: For estimating probability $p$, need $n \geq \frac{100}{p}$ to get stable estimate.

### 3.3.2 When Monte Carlo Fails

**Condition**: Monte Carlo requires $E[g(X)]$ to exist.

**Counterexample**: Cauchy distribution has no defined mean. Sampling from Cauchy and computing sample means will not converge.

**Example 3.8**: The Cauchy distribution has PDF

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

The mean $E[X]$ does not exist. Monte Carlo estimation of the mean will not work.

---

# 4. Hypothesis Testing

## 4.1 Fundamental Concepts

### 4.1.1 Hypotheses

**Definition 4.1**: A **hypothesis** is a statement about a population parameter or distribution.

**Definition 4.2** (Null and Alternative Hypotheses): - **Null hypothesis** ($H_0$): The “status quo” or “no effect” hypothesis - **Alternative hypothesis** ($H_a$ or $H_1$): The “research” or “effect exists” hypothesis

**Examples**: 1. $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ (two-sided) 2. $H_0 : \mu \leq \mu_0$ vs. $H_a : \mu > \mu_0$ (one-sided, upper tail) 3. $H_0 : p = 0.5$ vs. $H_a : p \neq 0.5$

**Convention**: We test $H_0$ and either reject it or fail to reject it. We never “accept” $H_0$.

### 4.1.2 Test Statistics

**Definition 4.3**: A **test statistic** $T = T(X_1, \ldots, X_n)$ is a function of the data that measures evidence against $H_0$.

**Properties of good test statistics**: 1. Large values indicate evidence against $H_0$ 2. Has a known distribution under $H_0$ 3. Is sensitive to violations of $H_0$

**Common test statistics**: - Difference in means: $T = \bar{X}_1 - \bar{X}_2$ - Sample proportion: $T = \hat{p}$ - Standardized statistic: $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

### 4.1.3 P-Values

**Definition 4.4**: The **p-value** is

$$p\text{-value} = P(\text{observe test statistic at least as extreme as observed} \mid H_0 \text{ true})$$

**Formal definition**: If $T_{\text{obs}}$ is the observed test statistic, - **Two-sided**: $p\text{-value} = P(|T| \geq |T_{\text{obs}}| \mid H_0)$ - **Upper tail**: $p\text{-value} = P(T \geq T_{\text{obs}} \mid H_0)$ - **Lower tail**: $p\text{-value} = P(T \leq T_{\text{obs}} \mid H_0)$

**Interpretation**: - Small p-value (typically $< 0.05$): Strong evidence against $H_0$ - Large p-value: Insufficient evidence to reject $H_0$

**Common misconceptions**: - ❌ P-value is NOT $P(H_0 \text{ true} \mid \text{data})$ - ❌ P-value is NOT the probability of making an error - ✓ P-value IS the probability of seeing data this extreme if $H_0$ is true

### 4.1.4 Significance Level

**Definition 4.5**: The **significance level** $\alpha$ is the threshold for rejecting $H_0$.

**Decision rule**: Reject $H_0$ if $p\text{-value} < \alpha$.

**Common choices**: $\alpha = 0.05$ (standard), $\alpha = 0.01$ (conservative), $\alpha = 0.10$ (exploratory)

**Interpretation**: $\alpha$ is the probability of rejecting $H_0$ when it is true (Type I error rate).

## 4.2 Permutation Tests

### 4.2.1 Two-Sample Permutation Test

**Setup**: Two independent samples from populations with distributions $F_1$ and $F_2$. - Sample 1: $X_1, \ldots, X_{n_1}$ - Sample 2: $Y_1, \ldots, Y_{n_2}$

**Hypotheses**: - $H_0$: $F_1 = F_2$ (distributions are identical) - $H_a$: $F_1 \neq F_2$ (distributions differ)

**Key insight under** $H_0$: Group labels are meaningless, so any permutation of labels is equally likely.

**Algorithm 4.1** (Permutation Test): 1. Compute observed test statistic $T_{\text{obs}}$ (e.g., $\bar{X} - \bar{Y}$) 2. Pool all data: $Z = (X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$ 3. For $b = 1, \ldots, B$ (e.g., $B = 10{,}000$): - Randomly permute $Z$ to get

$Z^{(b)}$ - Assign first $n_1$ to group 1, rest to group 2 - Compute permuted test statistic $T^{(b)}$ 4. Compute p-value: - Two-sided: $p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(|T^{(b)}| \geq |T_{\text{obs}}|)$ - One-sided (upper): $p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(T^{(b)} \geq T_{\text{obs}})$

**Example 4.1** (R code):

```
set.seed(123)
# Data
group1 <- c(12, 15, 18, 14, 16)
group2 <- c(20, 22, 19, 25, 21)

# Observed test statistic
T_obs <- mean(group2) - mean(group1)

# Permutation test
all_data <- c(group1, group2)
n1 <- length(group1)
n_perm <- 10000

T_perm <- replicate(n_perm, {
  perm_indices <- sample(1:length(all_data), n1)
  new_group1 <- all_data[perm_indices]
  new_group2 <- all_data[-perm_indices]
  mean(new_group2) - mean(new_group1)
})

# P-value (two-sided)
p_value <- mean(abs(T_perm) >= abs(T_obs))
```

## 4.2.2 Randomization Tests

**Context**: Treatment vs. control in randomized experiment.

**Difference from permutation test**: Randomization test accounts for the specific randomization scheme used in the experiment.

**Algorithm**: Same as permutation test, but: - Treatments were randomly assigned - Under $H_0$, treatment assignments are random - Permute treatment labels to simulate null distribution

**Example 4.2**: Drug trial with $n_1$ receiving treatment, $n_2$ receiving placebo. - $H_0$: Treatment has no effect - Under $H_0$, any assignment of treatments is equally likely - Permute treatment labels to get null distribution of test statistic

## 4.2.3 One-Sample Tests

**Setup**: Single sample $X_1, \ldots, X_n$ from population with parameter $\theta$.

**Hypotheses**: - $H_0$: $\theta = \theta_0$ - $H_a$: $\theta \neq \theta_0$

**Monte Carlo (Parametric) Test**: 1. Assume distributional form under $H_0$ (e.g., Normal, Poisson) 2. Estimate any nuisance parameters from data 3. Simulate $B$ datasets from the null model 4. Compute test statistic for each simulated dataset 5. Compare observed test statistic to simulated distribution

**Example 4.3** (Testing Poisson mean): - $H_0$: $\lambda = 25$ - Data: 15 observations with sample mean $\bar{x} = 28.5$ - Test statistic: $T = \bar{X}$

```
set.seed(123)
n <- 15
lambda0 <- 25
x_bar_obs <- 28.5

# Simulate under H0
x_bar_sim <- replicate(10000, mean(rpois(n, lambda = lambda0)))

# P-value (two-sided)
p_value <- 2 * min(mean(x_bar_sim >= x_bar_obs),
                   mean(x_bar_sim <= x_bar_obs))
```

## 4.3 Type I and Type II Errors

### 4.3.1 Error Types

**Truth Table**:

| | $H_0$ **True** | $H_0$ **False** |
|---|---|---|
| **Reject** $H_0$ | Type I Error (false positive) | Correct (true positive) |
| **Fail to Reject** $H_0$ | Correct (true negative) | Type II Error (false negative) |

**Definition 4.6** (Type I Error): **Type I error** occurs when we reject $H_0$ when it is true.

$$\alpha = P(\text{Type I Error}) = P(\text{Reject } H_0 \mid H_0 \text{ true})$$

**Definition 4.7** (Type II Error): **Type II error** occurs when we fail to reject $H_0$ when it is false.

$$\beta = P(\text{Type II Error}) = P(\text{Fail to reject } H_0 \mid H_a \text{ true})$$

**Relationship**: - $\alpha$ is the significance level we choose - $\beta$ depends on the true parameter value, sample size, and test design - Decreasing $\alpha$ generally increases $\beta$ (tradeoff)

### 4.3.2 Power

**Definition 4.8** (Power): The **power** of a test is

$$\text{Power} = 1 - \beta = P(\text{Reject } H_0 \mid H_a \text{ true})$$

**Interpretation**: Power is the probability of correctly detecting an effect when it exists.

**Factors affecting power**: 1. **Effect size**: Larger effects → higher power 2. **Sample size**: Larger $n$ → higher power 3. **Significance level**: Larger $\alpha$ → higher power (but more Type I errors) 4. **Variability**: Smaller $\sigma$ → higher power

**Theorem 4.1**: For testing $H_0 : \mu = \mu_0$ vs. $H_a : \mu = \mu_a$ with known $\sigma$,

$$\text{Power} = P\left(Z > z_{\alpha/2} - \frac{|\mu_a - \mu_0|}{\sigma/\sqrt{n}}\right)$$

where $Z \sim N(0,1)$ and $z_{\alpha/2}$ is the critical value.

**Example 4.4**: Test $H_0 : \mu = 100$ vs. $H_a : \mu \neq 100$ with $\alpha = 0.05$, $\sigma = 15$, $n = 25$. If true mean is $\mu_a = 110$, power is:

$$\text{Power} = P\left(|Z| > 1.96 - \frac{|110 - 100|}{15/\sqrt{25}}\right) = P(|Z| > 1.96 - 3.33) \approx 0.80$$

### 4.3.3 Sample Size Determination

**Problem**: How large should $n$ be to achieve desired power $(1 - \beta)$ at significance level $\alpha$?

**Formula for two-sample t-test** (testing $\mu_1 - \mu_2 = 0$):

$$n = 2\left(\frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta}\right)^2$$

where $\delta = |\mu_1 - \mu_2|$ is the effect size.

**Example 4.5**: Detect difference $\delta = 5$ with $\sigma = 10$, $\alpha = 0.05$, power = 0.80.

$$n = 2\left(\frac{(1.96 + 0.84) \times 10}{5}\right)^2 = 2 \times 5.6^2 \approx 63 \text{ per group}$$

## 4.4 One-Tailed vs. Two-Tailed Tests

**Definition 4.9** (Tails): - **Two-tailed**: $H_a : \theta \neq \theta_0$ (detect any difference) - **One-tailed (upper)**: $H_a : \theta > \theta_0$ (detect increase only) - **One-tailed (lower)**: $H_a : \theta < \theta_0$ (detect decrease only)

**P-value calculations**: - **Two-tailed**: $p = 2 \times P(T \geq |T_{\text{obs}}| \mid H_0)$ - **One-tailed (upper)**: $p = P(T \geq T_{\text{obs}} \mid H_0)$ - **One-tailed (lower)**: $p = P(T \leq T_{\text{obs}} \mid H_0)$

**When to use**: - **Two-tailed**: Default; when effect direction is unknown or both directions matter - **One-tailed**: When direction is specified a priori and only one direction is meaningful

**Warning**: Choosing one-tailed after seeing data is invalid (inflates Type I error).

---

# 5. Point Estimation

## 5.1 Estimators and Estimates

### 5.1.1 Basic Definitions

**Definition 5.1** (Parameter): A **parameter** $\theta$ is a fixed unknown quantity describing a population.

**Definition 5.2** (Statistic): A **statistic** is any function of the data: $T = T(X_1, \ldots, X_n)$.

**Definition 5.3** (Estimator): An **estimator** $\hat{\theta}$ is a statistic used to estimate parameter $\theta$. - Estimator is a random variable (depends on random sample) - Has its own distribution (called the **sampling distribution**)

**Definition 5.4** (Estimate): An **estimate** is the realized value of an estimator for a specific dataset.

**Example 5.1**: - Parameter: $\mu$ (population mean) - Estimator: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample mean) - Estimate: If data is $\{3, 7, 5\}$, then estimate is $\bar{x} = 5$

### 5.1.2 Common Estimators

| Parameter | Estimator | Name |
|---|---|---|
| Population mean $\mu$ | $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ | Sample mean |
| Population variance $\sigma^2$ | $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ | Sample variance |
| Population proportion $p$ | $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ | Sample proportion |
| Population maximum $M$ | $\max(X_1, \ldots, X_n)$ | Sample maximum |

## 5.2 Properties of Estimators

### 5.2.1 Bias

**Definition 5.5** (Bias): The **bias** of estimator $\hat{\theta}$ for parameter $\theta$ is

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

**Definition 5.6** (Unbiased Estimator): $\hat{\theta}$ is **unbiased** if $E[\hat{\theta}] = \theta$ (bias = 0).

**Theorem 5.1**: The sample mean $\bar{X}$ is unbiased for $\mu$:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

**Theorem 5.2**: The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is unbiased for $\sigma^2$:

$$E[S^2] = \sigma^2$$

**Why $n-1$ instead of $n$?** Using $n$ leads to underestimation because $\bar{X}$ is computed from the same data (loses one degree of freedom).

**Warning**: The sample standard deviation $S = \sqrt{S^2}$ is **not** unbiased for $\sigma$ (because $E[\sqrt{X}] \neq \sqrt{E[X]}$).

**Example 5.2** (Biased estimator): For Uniform$(0, M)$, the sample maximum $\hat{M} = \max(X_1, \ldots, X_n)$ is biased:

$$E[\hat{M}] = \frac{n}{n+1}M < M$$

**Adjusted estimator**: $\hat{M}_{\text{adj}} = \frac{n+1}{n}\max(X_1, \ldots, X_n)$ is unbiased.

## 5.2.2 Variance and Mean Squared Error

**Definition 5.7** (Variance of Estimator):

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

**Theorem 5.3**: For sample mean,

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

**Proof**:

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

**Definition 5.8** (Mean Squared Error): The **MSE** of estimator $\hat{\theta}$ is

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

**Theorem 5.4** (Bias-Variance Decomposition):

$$\text{MSE}(\hat{\theta}) = [\text{Bias}(\hat{\theta})]^2 + \text{Var}(\hat{\theta})$$

**Proof**:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}])^2] + 2E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] + (E[\hat{\theta}] - \theta)^2 \\
&= \text{Var}(\hat{\theta}) + 0 + [\text{Bias}(\hat{\theta})]^2
\end{aligned}$$

**Interpretation**: - MSE combines bias and variance into one measure - Small bias with large variance can have larger MSE than moderate bias with small variance - **Bias-variance tradeoff**: Sometimes accepting small bias reduces variance enough to lower MSE

### 5.2.3 Consistency

**Definition 5.9** (Consistent Estimator): $\hat{\theta}_n$ is **consistent** for $\theta$ if

$$\hat{\theta}_n \xrightarrow{P} \theta \text{ as } n \to \infty$$

That is, for any $\epsilon > 0$, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$.

**Theorem 5.5**: By the Weak Law of Large Numbers, $\bar{X}_n$ is consistent for $\mu$.

**Theorem 5.6** (Sufficient condition for consistency): If $\text{Bias}(\hat{\theta}_n) \to 0$ and $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.

**Example 5.3**: Even though $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is biased for $\sigma^2$, it is consistent because bias $= -\frac{\sigma^2}{n} \to 0$.

## 5.3 Methods of Estimation

### 5.3.1 Method of Moments

**Idea**: Match sample moments to population moments.

**Algorithm 5.1** (Method of Moments): 1. Compute first $k$ population moments in terms of parameters 2. Set them equal to corresponding sample moments 3. Solve for parameter estimates

**Example 5.4**: Exponential$(\lambda)$ has $E[X] = \frac{1}{\lambda}$. - Sample moment: $\bar{X}$ - Equation: $\frac{1}{\lambda} = \bar{X}$ - Estimate: $\hat{\lambda} = \frac{1}{\bar{X}}$

**Note**: $\hat{\lambda} = \frac{1}{\bar{X}}$ is biased (by Jensen's inequality) but consistent.

### 5.3.2 Simulating Sampling Distributions

**Purpose**: Understand behavior of estimator (bias, variance, MSE).

**Monte Carlo Algorithm**: 1. Choose true parameter value $\theta$ 2. For $b = 1, \ldots, B$ (e.g., $B = 10{,}000$): - Generate sample $X_1^{(b)}, \ldots, X_n^{(b)}$ from distribution with parameter $\theta$ - Compute $\hat{\theta}^{(b)}$ 3. Analyze $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$: - Empirical bias: $\frac{1}{B}\sum_{b=1}^B \hat{\theta}^{(b)} - \theta$ - Empirical variance: $\frac{1}{B}\sum_{b=1}^B (\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2$ - Empirical MSE: $\frac{1}{B}\sum_{b=1}^B (\hat{\theta}^{(b)} - \theta)^2$

**Example 5.5** (R code):

```
set.seed(123)
n <- 20
true_lambda <- 2
B <- 10000
```

```r
estimates <- replicate(B, {
  x <- rexp(n, rate = true_lambda)
  1 / mean(x)  # Method of moments estimator
})

bias <- mean(estimates) - true_lambda
variance <- var(estimates)
mse <- mean((estimates - true_lambda)^2)
```

---

# 6. Interval Estimation

## 6.1 Confidence Intervals

### 6.1.1 Definition and Interpretation

**Definition 6.1** (Confidence Interval): A $(1-\alpha)$ **confidence interval** for parameter $\theta$ is a random interval $[L, U]$ such that

$$P(L \leq \theta \leq U) = 1 - \alpha$$

**Common confidence levels**: 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), 99% ($\alpha = 0.01$).

**Correct interpretation**: "If we repeat this procedure many times, approximately $(1-\alpha) \times 100\%$ of intervals will contain $\theta$."

**Incorrect interpretations**: - ❌ "There is a 95% probability that $\theta$ is in $[L, U]$" (θ is fixed, not random) - ❌ "95% of the data is in $[L, U]$" (this is about population, not data)

**Key point**: Confidence refers to the **procedure**, not a specific interval. Once computed, the specific interval either contains $\theta$ or it doesn't (we just don't know which).

### 6.1.2 Monte Carlo Confidence Intervals

**Idea**: Use simulation to approximate the sampling distribution of estimator $\hat{\theta}$.

**Algorithm 6.1** (Monte Carlo CI): 1. From data, compute point estimate $\hat{\theta}$ 2. Generate $B$ datasets from the model with parameter $\hat{\theta}$ 3. For each dataset $b$, compute $\hat{\theta}^{(b)}$ 4. Compute quantiles: $L = Q_{0.025}(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$ and $U = Q_{0.975}(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)})$ 5. 95% CI: $[L, U]$

**Example 6.1** (Exponential rate):

```r
# Observed data
x <- rexp(50, rate = 2)
lambda_hat <- 1 / mean(x)
```

```r
# Monte Carlo CI
B <- 10000
n <- length(x)
lambda_sim <- replicate(B, {
  x_sim <- rexp(n, rate = lambda_hat)
  1 / mean(x_sim)
})

ci <- quantile(lambda_sim, c(0.025, 0.975))
```

**Coverage rate**: Proportion of times CI contains true parameter (should be $\approx 1 - \alpha$).

### 6.1.3 Central Limit Theorem

**Theorem 6.1** (Central Limit Theorem): Let $X_1, \ldots, X_n$ be i.i.d. with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

**Practical form**: For large $n$,

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

**Interpretation**: - Regardless of the distribution of $X_i$, $\bar{X}_n$ is approximately normal for large $n$ - "Large $n$" depends on skewness: typically $n \geq 30$ suffices, but $n > 100$ is safer for very skewed distributions

**Applications**: 1. Approximate probabilities: $P(\bar{X}_n \leq x) \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)$ 2. Construct confidence intervals 3. Perform hypothesis tests

## 6.2 Parametric Confidence Intervals

### 6.2.1 Z-Intervals (Known Variance)

**Setup**: $X_1, \ldots, X_n$ i.i.d. with $E[X_i] = \mu$ and known $\mathrm{Var}(X_i) = \sigma^2$.

**By CLT**: $\bar{X}_n \approx N(\mu, \sigma^2/n)$ for large $n$.

**Standardization**: $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$

**Derivation**:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging:

$$P\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$(1-\alpha)$ **confidence interval for** $\mu$:

$$\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

**Critical values**: - 90% CI: $z_{0.05} = 1.645$ - 95% CI: $z_{0.025} = 1.96$ - 99% CI: $z_{0.005} = 2.576$

**Example 6.2**: Sample of $n = 36$ with $\bar{x} = 100$, known $\sigma = 15$. Find 95% CI for $\mu$.

$$100 \pm 1.96 \times \frac{15}{\sqrt{36}} = 100 \pm 4.9 = [95.1, 104.9]$$

### 6.2.2 T-Intervals (Unknown Variance)

**Setup**: $X_1, \ldots, X_n$ i.i.d. Normal$(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown.

**Problem**: Cannot use Z-interval because $\sigma$ is unknown.

**Solution**: Replace $\sigma$ with sample standard deviation $S$, but use $t$-distribution to account for extra uncertainty.

**Theorem 6.2**: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d., then

$$T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where $t_{n-1}$ is the $t$-distribution with $n-1$ degrees of freedom.

$(1-\alpha)$ **confidence interval for** $\mu$:

$$\bar{X}_n \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

**Properties of $t$-distribution**: - Symmetric, bell-shaped (like normal) - Heavier tails than normal (more uncertainty) - As $n \to \infty$, $t_{n-1} \to N(0,1)$

**Example 6.3**: Sample of $n = 16$ with $\bar{x} = 100$, $s = 15$. Find 95% CI for $\mu$. - Critical value: $t_{0.025,15} = 2.131$ - CI: $100 \pm 2.131 \times \frac{15}{\sqrt{16}} = 100 \pm 7.99 = [92.01, 107.99]$

**Note**: For large $n$ (e.g., $n > 30$), $t$ and $z$ intervals are nearly identical.

### 6.2.3 Confidence Intervals for Proportions

**Setup**: $X_1, \ldots, X_n$ i.i.d. Bernoulli$(p)$. Estimate $p$.

**Estimator**: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample proportion).

**Properties**: - $E[\hat{p}] = p$ - $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$

**By CLT**: For large $n$,

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

**Problem**: Variance depends on unknown $p$.

**Solution**: Plug in $\hat{p}$ for $p$ in the standard error (this works well for moderate $p$ and large $n$).

$(1-\alpha)$ **confidence interval for** $p$ (Wald interval):

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**Validity condition**: Requires $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$ (ensures CLT approximation is good).

**Example 6.4**: In $n = 100$ trials, observe $x = 60$ successes. Find 95% CI for $p$.

$$\hat{p} = 0.6, \quad \sqrt{\frac{0.6 \times 0.4}{100}} = 0.049$$

$$\text{CI: } 0.6 \pm 1.96 \times 0.049 = 0.6 \pm 0.096 = [0.504, 0.696]$$

**Note**: Wilson and Agresti-Coull intervals have better coverage properties for small $n$ or extreme $p$.

## 6.3 Width and Margin of Error

**Definition 6.2** (Margin of Error): The **margin of error** (ME) is half the width of the confidence interval.

For mean (known $\sigma$):

$$\text{ME} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

For proportion:

$$\text{ME} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**Factors affecting width**: 1. **Confidence level** $(1-\alpha)$: Higher confidence → wider interval 2. **Sample size** $n$: Larger $n$ → narrower interval (ME $\propto 1/\sqrt{n}$) 3. **Variability** $\sigma$: Larger $\sigma$ → wider interval

**Sample size for desired ME**: Solve $\text{ME} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ for $n$:

$$n = \left(\frac{z_{\alpha/2}\sigma}{\text{ME}}\right)^2$$

**Example 6.5**: Estimate mean with ME = 2, $\sigma = 10$, 95% confidence.

$$n = \left(\frac{1.96 \times 10}{2}\right)^2 = 96.04 \Rightarrow n = 97$$

## 6.4 Duality with Hypothesis Testing

**Theorem 6.3** (CI-Test Duality): A $(1-\alpha)$ confidence interval for $\theta$ consists of all values $\theta_0$ that would **not be rejected** in a two-sided level-$\alpha$ test of $H_0 : \theta = \theta_0$.

**Equivalently**: - If $\theta_0 \in \text{CI}$, then fail to reject $H_0 : \theta = \theta_0$ at level $\alpha$ - If $\theta_0 \notin \text{CI}$, then reject $H_0 : \theta = \theta_0$ at level $\alpha$

**Example 6.6**: 95% CI for $\mu$ is $[95.1, 104.9]$. - $H_0 : \mu = 100$ would not be rejected at $\alpha = 0.05$ (100 in CI) - $H_0 : \mu = 90$ would be rejected at $\alpha = 0.05$ (90 not in CI)

**Practical use**: Can use CI to perform many hypothesis tests simultaneously.

---

# 7. R Programming Reference

## 7.1 Probability Distributions in R

### 7.1.1 Distribution Function Naming

R uses a systematic naming convention: `prefix + distribution_name`

**Prefixes**: - `d`: Density (PDF for continuous, PMF for discrete) - `p`: Probability (CDF) - `q`: Quantile (inverse CDF) - `r`: Random generation

### 7.1.2 Common Distributions

| **Distribution** | **R name** | **Parameters** | **Example** |
|---|---|---|---|
| Bernoulli | `binom` | `size=1, prob=p` | `rbinom(10, 1, 0.3)` |
| Binomial | `binom` | `size=n, prob=p` | `dbinom(5, 10, 0.5)` |
| Geometric | `geom` | `prob=p` | `rgeom(10, 0.3)` |
| Poisson | `pois` | `lambda` | `rpois(10, 5)` |
| Uniform (discrete) | `sample` | `x, size, replace, prob` | `sample(1:6, 10, TRUE)` |
| Uniform (continuous) | `unif` | `min=a, max=b` | `runif(10, 0, 1)` |

| Distribution | R name | Parameters | Example |
|---|---|---|---|
| Exponential | `exp` | `rate=λ` | `rexp(10, 2)` |
| Normal | `norm` | `mean=μ, sd=σ` | `rnorm(10, 5, 2)` |

**Examples**:

```
# Binomial
dbinom(3, size=10, prob=0.3)  # P(X = 3) when X ~ Binom(10, 0.3)
pbinom(3, size=10, prob=0.3)  # P(X ≤ 3)
qbinom(0.5, size=10, prob=0.3)  # Median
rbinom(1000, size=10, prob=0.3)  # 1000 random values

# Normal
pnorm(1.96) - pnorm(-1.96)  # P(-1.96 < Z < 1.96) for Z ~ N(0,1)
qnorm(0.975)  # 97.5th percentile = 1.96
rnorm(100, mean=5, sd=2)  # 100 random values from N(5, 4)

# Exponential
pexp(2, rate=0.5)  # P(X ≤ 2) for X ~ Exp(0.5)
```

# 7.2 Monte Carlo Simulations

## 7.2.1 Setting the Seed

```
set.seed(123)  # For reproducibility
```

## 7.2.2 Basic Monte Carlo Template

```
set.seed(123)
n_sim <- 10000

# Estimate E[g(X)]
results <- replicate(n_sim, {
  x <- rnorm(50, mean=10, sd=2)  # Generate data
  g_x <- mean(x^2)  # Compute quantity of interest
  return(g_x)
})

# Analysis
estimate <- mean(results)
se <- sd(results) / sqrt(n_sim)
ci <- quantile(results, c(0.025, 0.975))
```

## 7.2.3 Estimating Probabilities

```r
set.seed(123)
n_sim <- 10000

# Estimate P(event)
event_occurred <- replicate(n_sim, {
  x <- rexp(1, rate=2)
  y <- rnorm(1, mean=0, sd=1)
  return(x > y)  # Returns TRUE or FALSE
})

prob_estimate <- mean(event_occurred)  # Proportion of TRUEs
se <- sqrt(prob_estimate * (1 - prob_estimate) / n_sim)
```

## 7.3 Hypothesis Testing

### 7.3.1 Permutation Test Template

```r
set.seed(123)

# Data
group1 <- c(12, 15, 18, 14, 16)
group2 <- c(20, 22, 19, 25, 21)

# Observed test statistic
T_obs <- mean(group2) - mean(group1)

# Permutation test
all_data <- c(group1, group2)
n1 <- length(group1)
n_perm <- 10000

T_perm <- replicate(n_perm, {
  shuffled <- sample(all_data)
  new_group1 <- shuffled[1:n1]
  new_group2 <- shuffled[(n1+1):length(all_data)]
  mean(new_group2) - mean(new_group1)
})

# P-value (two-sided)
p_value <- mean(abs(T_perm) >= abs(T_obs))

# Visualization
hist(T_perm, breaks=50, main="Null Distribution")
abline(v=T_obs, col="red", lwd=2)
```

### 7.3.2 Parametric Test Template

```r
set.seed(123)

# Observed data
x <- rpois(20, lambda=28)  # Example data
x_bar_obs <- mean(x)

# Null hypothesis: λ = 25
lambda0 <- 25
n <- length(x)

# Simulate under H0
x_bar_sim <- replicate(10000, mean(rpois(n, lambda=lambda0)))

# P-value (two-sided)
p_value <- 2 * min(mean(x_bar_sim >= x_bar_obs),
                   mean(x_bar_sim <= x_bar_obs))
```

## 7.4 Confidence Intervals

### 7.4.1 Monte Carlo CI Template

```r
set.seed(123)

# Observed data
data <- rexp(50, rate=2)
theta_hat <- 1 / mean(data)  # Point estimate

# Monte Carlo CI
B <- 10000
n <- length(data)

theta_sim <- replicate(B, {
  sim_data <- rexp(n, rate=theta_hat)
  1 / mean(sim_data)
})

# 95% CI
ci <- quantile(theta_sim, c(0.025, 0.975))

# Visualization
hist(theta_sim, breaks=50, main="Sampling Distribution")
abline(v=ci, col="red", lwd=2)
```

### 7.4.2 CLT-based CI

```r
# Sample statistics
x_bar <- mean(data)
```

```r
s <- sd(data)
n <- length(data)
alpha <- 0.05

# Z-interval (known σ, or large n)
z_crit <- qnorm(1 - alpha/2)  # 1.96 for 95%
ci_z <- x_bar + c(-1, 1) * z_crit * s / sqrt(n)

# T-interval (unknown σ, normal data)
t_crit <- qt(1 - alpha/2, df=n-1)
ci_t <- x_bar + c(-1, 1) * t_crit * s / sqrt(n)

# Proportion
p_hat <- mean(data)  # For Bernoulli data
ci_prop <- p_hat + c(-1, 1) * z_crit * sqrt(p_hat*(1-p_hat)/n)
```

## 7.5 Useful Functions

### 7.5.1 Summary Statistics

```r
mean(x)          # Sample mean
median(x)        # Sample median
var(x)           # Sample variance (n-1 denominator)
sd(x)            # Sample standard deviation
IQR(x)           # Interquartile range
quantile(x, c(0.25, 0.75))  # Quartiles
summary(x)       # Five-number summary + mean
```

### 7.5.2 Data Manipulation

```r
sum(x)           # Sum
length(x)        # Number of elements
unique(x)        # Unique values
table(x)         # Frequency table
sort(x)          # Sort ascending
sample(x, size=n, replace=TRUE)  # Random sample
rep(value, times=n)  # Repeat value n times
seq(from, to, by)    # Sequence
c(x, y)          # Concatenate vectors
```

### 7.5.3 Logical Operations

```r
x > 5            # Logical comparison
x >= 5 & x <= 10  # AND
x < 3 | x > 7     # OR
!condition        # NOT
```

```r
sum(condition)     # Count TRUE values
mean(condition)    # Proportion TRUE
which(condition)   # Indices where TRUE
```

### 7.5.4 Visualization

```r
hist(x, breaks=50)           # Histogram
plot(x, y)                   # Scatterplot
barplot(table(x))           # Bar plot
abline(h=value, col="red")  # Horizontal line
abline(v=value, col="blue") # Vertical line
lines(x, y)                 # Add lines to plot
```

---

# Appendix: Key Formulas and Theorems

## Probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A^c) = 1 - P(A)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A \mid B)P(B)$$

## Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

## Expected Value and Variance

$$E[aX + bY] = aE[X] + bE[Y]$$

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

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

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

## Covariance and Correlation

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

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

## Common Distributions

| Distribution | E[X] | Var(X) |
|---|---|---|
| Bernoulli(p) | $p$ | $p(1-p)$ |
| Binomial(n,p) | $np$ | $np(1-p)$ |
| Geometric(p) | $\frac{1-p}{p}$ | $\frac{1-p}{p^2}$ |
| Poisson(λ) | $\lambda$ | $\lambda$ |
| Uniform(a,b) | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponential(λ) | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Normal(μ,$\sigma^2$) | $\mu$ | $\sigma^2$ |

## Confidence Intervals

**Mean (σ known)**: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

**Mean (σ unknown)**: $\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$

**Proportion**: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

## Critical Values

95% CI: $z_{0.025} = 1.96$, 90% CI: $z_{0.05} = 1.645$, 99% CI: $z_{0.005} = 2.576$

---

# Study Tips for Exam Success

1. **Understand concepts, not just formulas**: Know why formulas work, not just what they are
2. **Practice with R**: Simulate examples to build intuition
3. **Work through examples**: The best way to learn is by doing
4. **Check your work**: Does the answer make sense? Are units correct?
5. **Master the basics**: Probability rules, expected value, variance
6. **Know when to use each method**: Understand conditions and assumptions
7. **Draw pictures**: Visualize distributions, sampling distributions, hypothesis tests

8. **Write clear interpretations**: Practice explaining results in context

---

**End of Study Guide**