Statistical Computing

Compiled Revision Notes

Table of Contents

Week 1: Fundamental Concepts & Discrete Distributions 4

- Definitions 4
- Measures of Centrality and Variation 4
- Probability Theory Basics 5
- Conditional Probability and Independence 5
- Advanced Probability Theorems 6
- Discrete Random Variables 6
- Cumulative Distribution Function (CDF) 7
- Discrete Probability Distributions 7

Week 2: Continuous, Sampling & Hypothesis Testing 8

- Continuous Random Variables 8
- The Normal Distribution 8
- Sampling Distributions 9
- Hypothesis Testing Basics 9
- Confidence Intervals 10
- Hypothesis Tests for Proportions 10
- One Sample t-Test 11
- Comparing Two Means: Independent Samples 12
- Comparing Two Means: Paired Samples 12

Week 3: Enumerative Data Analysis and MLE 13

- Enumerative Data Analysis (Chi-Squared) 13
- Chi-Squared Goodness-of-Fit Test 13
- M&Ms Example 14
- Chi-Squared Test of Independence 14
- Maximum Likelihood Estimation 16
- MLE Examples 17
- R Implementation 18

Week 4: Advanced MLE & Numerical Methods 19

- Complex MLE & The Need for Optimization 19
- Numerical Optimization Methods 19
- MLE Optimization in R 20
- Why We Love MLE (Theoretical Properties) 20
- The Likelihood Ratio Test (LRT) 21

Week 5: Generalised Linear Models 21

- Linear Regression Review and Diagnostics 21
- Limitations of Linear Regression 22
- Introduction to Generalised Linear Models (GLMs) 22
- Logistic Regression 23
- Nested Models and Interactions 23

Fitting GLMs 24
Complete Exam Quick Reference Table 24

Week 1: Fundamental Concepts & Discrete Distributions

Definitions

- **Population:** The complete collection of all individuals or items under consideration in a statistical study.
- **Sample:** A subset of the population from which information is actually collected.

Parameters vs Statistics

- **Population Parameters** (constants, usually unknown):
  - $\mu \to$ population mean
  - $\sigma \to$ population standard deviation
- **Sample Statistics** (random variables):
  - $\bar{x} \to$ sample mean
  - $s \to$ sample standard deviation

Measures of Centrality and Variation

Measures of Centrality (Centre)

1. **Mean ($\bar{x}$):** Arithmetic average

$$\bar{x} = \frac{\sum x_i}{n}$$

2. **Median:** Middle value when data is ordered.
3. **Mode:** Most frequently occurring value.

Measures of Variation (Spread)

- **Range:** Max − Min. Very sensitive to outliers.
- **Sample Variance (s^2):**

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- **Sample Standard Deviation (s):**

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Probability Theory Basics

Core Rules

- **Sample Space (Ω):** Set of all possible outcomes. $P(\Omega) = 1$
- **Empty Set ($\emptyset$):** Impossible event. $P(\emptyset) = 0$
- **Probability Bounds:** $0 \leq P(A) \leq 1$
- **Complement Rule:** $P(A^c) = 1 - P(A)$. Notation: A^c, $\bar{A}$, or A'.

Combining Events

- **Intersection (AND):** $A \cap B$
- **Union (OR):** $A \cup B$

Addition Rules

- **General Addition Rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Disjoint (Mutually Exclusive) Events:**
  - Cannot occur together
  - $A \cap B = \emptyset$
  - $P(A \cup B) = P(A) + P(B)$

Conditional Probability and Independence

Conditional Probability

Probability that B occurs given that A has occurred:

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

Multiplication Law

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

Independence

Two events are independent if one does not affect the other:

- $P(B \mid A) = P(B)$, or
- $P(A \cap B) = P(A) \times P(B)$

Advanced Probability Theorems

Bayes' Theorem

Used to reverse conditional probabilities:

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

Law of Total Probability

If $A_1, A_2, \dots, A_n$ partition the sample space:

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \times P(A_i)$$

Discrete Random Variables

Definition

A **Random Variable (X)** is a numerical model for a measurement.

- **Discrete RV:** Takes a finite or countably infinite number of values.
- **Bernoulli RV:** Simplest discrete RV.
  Takes value:
  - 1 for success
  - 0 for failure

Probability Mass Function (pmf)

$$f(x) = P(X = x)$$

Expected Value (Mean)

The long-run average or centre of gravity:

$$E(X) = \mu = \sum x \cdot P(X = x)$$

Example (Fair die):

$$E(X) = 3.5$$

Cumulative Distribution Function (CDF)

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

- For discrete RVs, the CDF has a **step shape**.
- **At least rule:** $P(X \geq k) = 1 - P(X < k)$

Discrete Probability Distributions

Binomial Distribution

Used for the number of successes in n trials.

Assumptions (Always state in exams)

1. Fixed number of trials (n)
2. Constant probability of success (p)
3. Trials are independent

Model

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

Parameters

- **Mean:** $\mu = np$
- **Standard Deviation:**

$$\sigma = \sqrt{np(1-p)}$$

Poisson Distribution

Used for counting arrivals in a fixed interval of time or space.

Assumptions

1. Probability proportional to interval size
2. Probability of two or more arrivals in a very small interval is negligible
3. Non-overlapping intervals are independent

Model

$$P(X = x) = \frac{e^{-\alpha t} (\alpha t)^x}{x!}$$

- α = average rate per unit
- t = length of interval

Key Property (Very Exam Important)

$$Rate = \lambda = E(X) = Var(X) = \alpha t$$

Week 2: Continuous, Sampling & Hypothesis Testing

Continuous Random Variables

Definition

A continuous random variable can take values anywhere in a continuum, such as height, temperature, or sales.

- **Density Function ($f(x)$):**
  A curve where the area under the curve between two points represents probability.
- **Total Area:**
  The total area under $f(x)$ is always 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Uniform Distribution

The simplest continuous distribution where probability is constant between a and b.

- **PDF:**

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

The Normal Distribution

Properties

- Defined by **Mean (μ)** and **Variance (σ^2)**.
- Notation: $X \sim N(\mu, \sigma^2)$

Empirical Rule (68, 95, 99.7)

- 68% of data lies within $\mu \pm 1\sigma$
- 95% of data lies within $\mu \pm 2\sigma$
- 99.7% of data lies within $\mu \pm 3\sigma$

Sampling Distributions

Central Limit Theorem (CLT)

Regardless of the population distribution, if sample size n is large, the distribution of the sample mean $\bar{X}$ is approximately normal.

- **Mean of $\bar{X}$:** $E(\bar{X}) = \mu$
- **Variance of $\bar{X}$:**

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

- **Standard Error:** $\frac{\sigma}{\sqrt{n}}$
- **Z Statistic:**

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Hypothesis Testing Basics

Core Concepts

- **Null Hypothesis (H_0):**
  Assumed true. Always contains equality ($=$, $\leq$, $\geq$).
- **Alternative Hypothesis (H_1):**
  The claim we seek evidence for. Always contains inequality ($\neq$, $<$, $>$).

Errors

- **Type I Error (α):**
  Rejecting H_0 when it is actually true.
- **Type II Error (β):**
  Failing to reject H_0 when it is actually false.

The p-value

The probability of observing a result at least as extreme as the one obtained, assuming H_0 is true.

- **Decision Rule:**
  Reject H_0 if

$$\text{p-value} < \alpha$$

Confidence Intervals

Definition

An interval constructed around $\bar{x}$ where we are reasonably confident the true population mean μ lies.

- **Interpretation:**
  In repeated sampling, 95% of such intervals would contain μ.

Formula (Known σ or Large n)

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

Example: Cola Cans

- $\bar{x} = 299.64$
- $n = 100$
- $\sigma = 1.2$

Resulting interval:

$$[299.40, 299.88]$$

Since 300 is not in the interval, reject H_0.

Hypothesis Tests for Proportions

Used for categorical data.

Two Approaches

Test	Method	Best For
`prop.test`	Normal (χ^2) approximation	Large samples ($n \geq 30$)
`binom.test`	Exact binomial probabilities	Small samples or when exact results are needed

Example: Thanos Snap

- $H_0 : p = 0.5$
- $H_1 : p \neq 0.5$
- Observed: 64 vanished out of 100

R Code (Approximate):

```
prop.test(64, 100, p = 0.5)
```

- p-value = 0.0069
  Reject H_0.

R Code (Exact — preferred for small samples):

```
binom.test(64, 100, p = 0.5)
```

- p-value = 0.0105
  Reject H_0.

`binom.test` Arguments

```
binom.test(x, n, p = 0.5, alternative = "two.sided")
```

- `x` — number of successes observed
- `n` — number of trials
- `p` — hypothesised probability of success under H_0
- `alternative` — `"two.sided"`, `"less"`, or `"greater"`

When to Use Each

- Use **`binom.test`** when n is small (roughly $n < 30$), or when you need exact p-values.
- Use **`prop.test`** for large samples; it also supports comparing two proportions (`prop.test(c(x1, x2), c(n1, n2))`).

One Sample t-Test

Used when population variance σ^2 is unknown.

- Uses Student's t distribution
- Degrees of freedom: $df = n - 1$

Assumptions

1. Data is numeric and continuous.
2. Data is normally distributed.

Normality Test: Shapiro-Wilk

- If p-value > 0.05, assume normality.

Example: Corrib River Radiation

- $H_0 : \mu \geq 5$
- $H_1 : \mu < 5$

R Code:

```
t.test(corrib, mu = 5, alternative = "less")
```

- p-value = 0.002
  Reject H_0. Water is safe.

Comparing Two Means: Independent Samples

Used to compare two separate groups.

Steps

1. **Check Normality:**
   Shapiro-Wilk test on both groups.
2. **Check Variances:**
   - Levene's Test (robust)
   - Bartlett's Test (requires normality)

   If p-value > 0.05, assume equal variances.
3. **Run t-Test:**
   Choose the appropriate variant based on the variance test result.

Which t-Test to Use

Situation	Test	R Code
Unequal variances (or unsure)	Welch Two Sample t-test — **does NOT assume equal variances**	`t.test(x, y, alternative = ...)`
Equal variances confirmed	Pooled (Student's) t-test — assumes equal variances	`t.test(x, y, var.equal = TRUE, alternative = ...)`

> **Default in R:** `t.test()` uses Welch's test (`var.equal = FALSE`) — safe to use in all cases.

R Code (Welch — unequal/unknown variances):

```
t.test(x, y, alternative = "less")
```

R Code (Pooled — equal variances confirmed):

```
t.test(x, y, var.equal = TRUE, alternative = "less")
```

Comparing Two Means: Paired Samples

Used when observations are dependent or matched.

Logic

Performs a one sample t-test on the differences between paired observations.

Example: Diet Study

- $H_1 : \mu_{\text{diff}} > 0$

R Code:

```
t.test(before, after, paired = TRUE, alternative = "greater")
```

- p-value = 0.02
  Reject H_0. Diet worked.

Warning

Using an independent t-test on paired data is incorrect and can increase the chance of a Type II error.

Week 3: Enumerative Data Analysis and MLE

Enumerative Data Analysis (Chi-Squared)

Qualitative vs Quantitative

Previously we analysed **quantitative data** (height, weight, marks).

Now we analyse **qualitative (categorical) data**:

- Data consists of **counts / frequencies**
- Examples: Eye colour, Yes/No, Defective/Not defective

We compare **Observed vs Expected frequencies**.

The Chi-Squared Distribution (χ^2)

- Not symmetric
- Right skewed
- Range: $0 \to \infty$
- Depends on **degrees of freedom (df)**
- As df increases, it becomes more Normal shaped
- Right tail area = significance level α

Chi-Squared Goodness-of-Fit Test

Purpose

Tests whether observed categorical data matches a claimed distribution.

Test Statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- $O =$ Observed frequency
- $E =$ Expected frequency
- $df = k - 1$

Large χ^2 means observed differs strongly from expected.

M&Ms Example

Claim (H_0):

30% Brown, 20% Yellow, 20% Red, 10% Orange, 10% Green, 10% Blue

Hypotheses:

- H_0: Distribution matches claim
- H_1: Distribution differs

Reject H_0 if $\chi^2_{calc} > \chi^2_{critical}$.

R Code

```r
chocolate <- c(67, 36, 43, 24, 23, 7)
probs <- c(0.3, 0.2, 0.2, 0.1, 0.1, 0.1)
chisq.test(chocolate, p = probs)
```

Chi-Squared Test of Independence

Purpose

Tests whether two categorical variables are related.

Hypotheses

- H_0: Variables are independent
- H_1: Variables are dependent

Expected Counts Formula

For contingency table:

$$E_{ij} = \frac{\text{(Row Total)(Column Total)}}{\text{Grand Total}}$$

Degrees of freedom:

$$df = (r-1)(c-1)$$

Assumptions

1. Categorical variables
2. Independent observations
3. Rule of 5:
   - At least 80% of expected counts ≥ 5
   - No expected count < 1

If violated, combine categories or use Fisher's test.

Fisher's Exact Test

Used for small samples.

```r
# Independent
wilcox.test(group_A, group_B, alternative = "two.sided")

#Paired
wilcox.test(group_A, group_B, alternative = "two.sided", paired = TRUE)
```

Mann-Whitney U Test / Wilcoxon Test

Used for median

`wilcox.test`

Effect Size: Statistical vs Practical Significance

The Problem with Large Samples

- **Statistical significance** tells you if a difference exists.

- **Practical importance** tells you if the difference matters.
- With very large n, even tiny differences can produce small p values.
- Example: A 2 second improvement may be statistically significant but practically useless.

The Solution: Effect Size Effect size measures the **magnitude** of a difference.

Chi-Squared Tests: Phi Coefficient For 2×2 tables:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Guidelines:

- 0.1 small
- 0.3 medium
- 0.5 large

---

t Tests: Cohen's d Used when comparing two means.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

For independent samples, use the pooled standard deviation.

Guidelines:

- 0.2 small
- 0.5 medium
- 0.8+ large

Maximum Likelihood Estimation

The Core Idea

How do we find the "best" parameters such as μ or λ?

MLE finds the parameter that makes your data most likely.

- **Fisher's Principle:** Choose parameter θ that makes the observed data most probable.

- Goal: Find θ that maximizes $P(\text{data} \mid \theta)$ i.e. the Likelihood Function $L(\theta)$.

MLE Step by Step

Likelihood Function Write the probability of the entire dataset.

If observations are independent:

$$L(\theta) = \prod f(x_i \mid \theta)$$

Log-Likelihood Take the natural log:

$$\ell(\theta) = \sum \ln\left(f(x_i \mid \theta)\right)$$

Why?

- Differentiating a product is messy
- Differentiating a sum is easier
- Logs turn products into sums

Differentiate Find derivative with respect to θ:

$$\frac{d\ell}{d\theta}$$

Solve Set derivative equal to 0 and solve for θ.

This gives the MLE estimate.

MLE Examples

Poisson Distribution (Horse Kicks)

- Data: Counts of deaths by horse kicks (von Bortkiewicz data)
- Model:

$$X \sim \text{Poisson}(\lambda)$$

MLE Result

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum x_i = \bar{x}$$

Takeaway: For Poisson, the MLE for λ is the **sample mean**.

Normal Distribution

We estimate two parameters: μ and σ^2.

Estimating the Mean

$$\hat{\mu} = \bar{x}$$

Takeaway: MLE mean equals the sample mean.

Estimating the Variance

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(x_i - \bar{x})^2$$

Bias Issue

- MLE divides by n → biased (underestimates variance)

- Sample variance:

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

Uses Bessel's correction and is unbiased.

Conclusion: For large n, difference is negligible.

R Implementation

For complex models, solve numerically.

Note: R minimizes functions, so use the **negative log-likelihood**.

```r
library(stats4)

# 1. Define Negative Log-Likelihood
nloglik <- function(lambda) {
    return(-sum(dpois(data, lambda, log = TRUE)))
}
```

```
# 2. Run Optimizer
fit <- mle(nloglik, start = list(lambda = 1))
summary(fit)
```

Week 4: Advanced MLE & Numerical Methods

Complex MLE & The Need for Optimization

When Math Fails (The Gamma Distribution)

- The Gamma distribution models right-skewed data, for example insurance claims.
- It uses two parameters: α (shape) and β (scale).
- **The Problem:** When you take the derivative of the Gamma log-likelihood and set it equal to 0, there is no simple closed-form solution. You cannot solve it by hand.
- **The Solution:** Numerical optimization. We use a computer to find where the derivative is approximately zero, which corresponds to the peak of the likelihood.

Numerical Optimization Methods

When we cannot find the maximum likelihood mathematically, we use algorithms to walk uphill to the peak.

Gradient Ascent / Descent

- **How it works:** Finds the direction of the steepest slope, the gradient, and takes a step in that direction.
- **Pros:** Simple to implement; only needs first derivatives.
- **Cons:** Slow, linear convergence; choosing the right step size is tricky.

Newton's Method

- **How it works:** Uses curvature, the Hessian matrix of second derivatives, to fit a quadratic curve and jump straight to its maximum.
- **Pros:** Very fast, quadratic convergence; fewer, smarter steps.
- **Cons:** Fails if the Hessian matrix is not invertible or near saddle points; computationally expensive because it requires second derivatives.

BFGS (Quasi-Newton)

- **How it works:** Achieves Newton-like speed without computing second derivatives. It approximates the Hessian matrix using previous gradient information.
- **Pros:** Fast, robust, and requires no second derivatives. This is the default in R's `optim()` and `mle()`.

Nelder-Mead (Simplex)

- **How it works:** Uses no derivatives. It constructs a simplex, a geometric shape of points, that reflects and shrinks over the surface to find the peak.
- **Pros:** Extremely robust; works on non-smooth functions and poor starting values.
- **Cons:** Slow; struggles in high-dimensional problems with many parameters.

MLE Optimization in R

The Negative Log-Likelihood Trick

- R's optimization functions such as `optim()` and `nlm()` are designed to minimize, not maximize.
- To compute the Maximum Likelihood Estimate, we minimize the Negative Log-Likelihood.
- If $\ell(\theta)$ is the log-likelihood, we minimize $-\ell(\theta)$.

Using `log=TRUE`

- When computing likelihoods in R, always use `log=TRUE` inside density functions, for example `dgamma(x, shape, scale, log=TRUE)`.
- This computes the log-probability directly, which is more numerically stable than computing a very small probability and then taking its logarithm.

Optimization Pitfalls

- **Local Maxima:** The algorithm may converge to a smaller local peak instead of the global maximum.
- **Solution:** Try multiple starting values. If all runs converge to the same point, you likely found the global maximum. If not, the likelihood may be multimodal.
- **Check Convergence:** In R, `optim()$convergence == 0` indicates successful convergence. Any non-zero value indicates failure.

Why We Love MLE (Theoretical Properties)

Even when computed numerically, MLE has excellent theoretical properties.

1. **Consistency:** As sample size $n \to \infty$, $\hat{\theta} \to \theta$.
2. **Equivariance:** If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
3. **Asymptotic Normality:** For large samples, $\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{I(\theta)}\right)$, where $I(\theta)$ is the Fisher Information.
4. **Asymptotic Efficiency:** For large samples, the MLE achieves the minimum possible variance among regular estimators.

The Likelihood Ratio Test (LRT)

Concept

Used to compare two nested models to determine whether additional parameters significantly improve model fit.

- H_0 (Restricted Model): Parameters are fixed, for example a fair coin with $p = 0.5$.
- H_1 (Unrestricted Model): Parameters are estimated using MLE, for example $p = \hat{p}$.

The Test Statistic

$$\Lambda = -2\left[\ell(\hat{\theta}_0) - \ell(\hat{\theta})\right]$$

- $\ell(\hat{\theta})$: Log-likelihood of the unrestricted model.
- $\ell(\hat{\theta}_0)$: Log-likelihood of the restricted model.

The Distribution

- Under H_0, $\Lambda \sim \chi^2_{df}$
- Degrees of freedom df equal the number of restrictions imposed under H_0.

Profile Likelihood & Confidence Intervals

- Since the LRT statistic follows a χ^2 distribution asymptotically, we can invert the test to construct confidence intervals without assuming normality.
- In R, `confint(fit)` computes profile likelihood confidence intervals.

Week 5: Generalised Linear Models

Linear Regression Review and Diagnostics

- Unlike machine learning which focuses on prediction, this module focuses on inference to understand relationships, quantify uncertainty, and compare competing models.
- **Linear Model:**

  $$y = \beta_0 + \beta_1 x + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$.
- Interpreting the summary output:
  - The p-value tests the null hypothesis $H_0 : \beta = 0$.

  - While often informally described as the probability that the coefficient occurred by chance, it precisely measures the probability of observing an estimate this far from zero if the true coefficient were actually zero.
  - The F-statistic tests whether all slope coefficients are simultaneously zero ($H_0 : \beta_1 = \cdots = \beta_p = 0$).

- Diagnostic plots are critical for checking assumptions:
  - **Residuals vs Fitted:** Looks for non-linear patterns; should display random scatter.
  - **Normal Q-Q:** Checks if errors are normally distributed; points should follow the diagonal line.
  - **Scale-Location:** Checks for constant variance (homoscedasticity); a funnel shape indicates a violation.
  - **Residuals vs Leverage:** Identifies highly influential points pulling the regression line.

Limitations of Linear Regression

- Linear regression assumes the response Y is continuous and unbounded.
- It assumes a direct linear relationship between predictors and the mean response.
- It assumes errors are normally distributed with constant variance.
- Forcing binary data (pass/fail), count data, or positive right-skewed data into a linear model leads to invalid predictions and violated assumptions.

Introduction to Generalised Linear Models (GLMs)

- Introduced by Nelder and Wedderburn (1972), GLMs unify various regression models under one framework.

- A GLM consists of three core components:
  - **Random Component:** Specifies that the response Y_i follows a distribution from the exponential family (e.g., Normal, Binomial, Poisson, Gamma).
  - **Systematic Component:** The linear predictor combining the predictors.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  - **Link Function:** Connects the mean μ_i to the linear predictor η_i.

$$g(\mu_i) = \eta_i$$

Common GLM Families

- **Normal:** Uses the Identity link ($g(\mu) = \mu$) for continuous, roughly symmetric data.
- **Binomial:** Uses the Logit link for binary responses or proportions.
- **Poisson:** Uses the Log link ($g(\mu) = \log \mu$) for count data.
- **Gamma:** Uses the Log link for positive, right-skewed continuous data.

Logistic Regression

- Used when the response Y is binary, modelling the probability of success $p = P(Y = 1)$.
- It applies the logit (log-odds) transformation to ensure predictions remain within the valid $[0, 1]$ bounds.
  - **Logit Link:**

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Note on Probability and Odds Conversions: Converting between probability and odds is essential for interpretation. (Note: The formula for p relies on addition in the denominator, correcting the subtraction noted in class).

- **Odds Formula:**

$$\text{odds} = \frac{p}{1-p}$$

- **Probability Formula:**

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

Interpreting Coefficients

- Logistic regression coefficients are interpreted as Odds Ratios.
  - **Odds Ratio:**

$$e^{\beta_j}$$

- If $e^{\beta_j} > 1$: Increasing x_j by 1 unit increases the odds of success.
- If $e^{\beta_j} < 1$: Increasing x_j by 1 unit decreases the odds of success.
- If $e^{\beta_j} = 1$ ($\beta_j = 0$): The predictor has no effect.
- To compute confidence intervals directly on the odds ratio scale in R, use `exp(confint(model))`.

Nested Models and Interactions

- **Likelihood Ratio Test (LRT):**
  - Used to formally compare a reduced model (H_0) against a full model (H_1) to determine if adding predictors significantly improves the fit.
  - In R, this is executed using the `anova()` function with a Chi-Squared test. `anova(model1, model2, test="Chisq")`
- **Interactions:**
  - An interaction implies that the effect of one predictor depends on the level of another predictor.

  - Specified in R using `*` (which includes both main effects and the interaction) or `:` (for the interaction term only).

Fitting GLMs

- GLMs are fitted using Maximum Likelihood Estimation (MLE).
- The objective is to find the parameter values $\hat{\beta}$ that minimize the negative log-likelihood.
- **Iteratively Reweighted Least Squares (IRLS):**
  - This is the specific numerical algorithm used to optimize the log-likelihood for GLMs.
  - R's `glm()` function uses IRLS, which converges more accurately and efficiently for these specific models than general-purpose optimizers like `mle()`.

Complete Exam Quick Reference Table

Concept / Test	Formula or R Function	Use Case	Key Exam Notes
Sample Mean	$\bar{x} = \frac{\sum x_i}{n}$	Estimate μ	Centre of data
Sample Variance	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$	Spread	Uses $n-1$
Sample Std Dev	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$	Spread in units	Root of variance
Addition Rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Combine events	Avoid double counting
Complement Rule	$P(A^c) = 1 - P(A)$	At least one questions	Often simplifies
Conditional Prob	$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$	Given info	Order matters
Independence	$P(A \cap B) = P(A)P(B)$	Check independence	Only if unrelated
Bayes Theorem	$P(A \mid B) = \frac{P(B\mid A)P(A)}{P(B)}$	Reverse conditional	Common trap
Binomial Mean	$\mu = np$	Expected successes	Fixed n, p
Binomial SD	$\sigma = \sqrt{np(1-p)}$	Spread	Memorise
Poisson Mean	$\mu = \lambda$	Arrivals	Mean = variance
Uniform PDF	$f(x) = \frac{1}{b-a}$	Constant density	Area = probability
Z Statistic	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	Mean tests	Known σ
Confidence Interval	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	Estimate mean	Check if μ_0 inside
Shapiro Wilk Test	`shapiro.test(x)`	Normality	H_0: Normal
Levene Test	`leveneTest()`	Compare variances	Robust
Bartlett Test	`bartlett.test()`	Compare variances	Needs normality

Concept / Test	Formula or R Function	Use Case	Key Exam Notes
One Sample t Test	`t.test(x, mu=...)`	Mean vs constant	Unknown σ
Independent t Test	`t.test(x, y)`	Two groups, unequal/unknown variances	Welch default; no equal variance assumption
Independent t Test (Equal Var)	`t.test(x, y, var.equal=TRUE)`	Two groups, equal variances confirmed	Pooled; only after Levene/Bartlett p > 0.05
Paired t Test	`t.test(x, y, paired=TRUE)`	Before vs after	Uses differences
Proportion Test	`prop.test(x, n)`	Test proportion	Large samples ($n \geq 30$), normal approx
Exact Binomial Test	`binom.test(x, n, p=...)`	Test proportion exactly	Small samples; exact binomial p-value
Chi Square Statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Categorical tests	Large = big difference
Goodness of Fit	`chisq.test(x, p=probs)`	Match distribution	$df = k - 1$
Independence Test	`chisq.test(matrix)`	Relationship test	$df = (r-1)(c-1)$
Fisher Exact Test	`fisher.test(matrix)`	Small samples	Use if counts < 5
Effect Size (Phi)	$\phi = \sqrt{\frac{\chi^2}{n}}$	Strength of association	0.1 small, 0.3 med, 0.5 large
Cohen's d	$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$	Effect size for mean differences	0.2 small, 0.5 medium, 0.8 large.
Likelihood	$L(\theta) = \prod f(x_i \mid \theta)$	Parameter estimation	Maximise
Log Likelihood	$\ell(\theta) = \log L(\theta)$	Simplify math	Turns product into sum
MLE Normal Mean	$\hat{\mu} = \bar{x}$	Estimate mean	Same as sample mean
MLE Normal Variance	$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$	Estimate variance	Biased
BFGS	`optim(method="BFGS")`	General-purpose MLE optimization	Fast, robust, no second derivatives required
Nelder-Mead	`optim(method="Nelder-Mead")`	Non-smooth likelihoods	Very robust but slower, weak in high dimensions

Concept / Test	Formula or R Function	Use Case	Key Exam Notes
Negative Log-Likelihood	$-\sum \log f(x_i \mid \theta)$	Convert maximization to minimization	R minimizes by default
Convergence Check	`fit$convergence == 0`	Verify optimizer success	0 indicates successful convergence
Equivariance (MLE)	If $\hat{\theta}$ is MLE, then $g(\hat{\theta})$ is MLE of $g(\theta)$	Transformations of parameters	Core theoretical property
LRT Statistic	$\Lambda = -2[\ell(\hat{\theta}_0) - \ell(\hat{\theta})]$	Compare nested models	Based on log-likelihood difference
LRT Distribution	$\Lambda \sim \chi^2_{df}$	Compute p-values	df equals number of restrictions
Profile Confidence Intervals	`confint(fit)`	Construct CIs via LRT	Does not rely on normal approximation
Linear F-Statistic	$H_0 : \beta_1 = \cdots = \beta_p = 0$	Overall model significance	Tests if at least one predictor matters
GLM Setup	$g(\mu_i) = \eta_i$	Linking mean to predictors	Connects distribution to linear equation
Logit Link	$\log\left(\frac{p}{1-p}\right) = \eta$	Logistic regression link	Bounds predictions to $[0, 1]$
Odds	$\text{odds} = \frac{p}{1-p}$	Probability to Odds	Ratio of success to failure
Probability from Odds	$p = \frac{\text{odds}}{1+\text{odds}}$	Odds to Probability	Reverses the odds calculation
Odds Ratio	e^{β_j}	Interpreting logistic coefficients	> 1 increases odds, < 1 decreases odds
Model Comparison (LRT)	`anova(mod1, mod2, test="Chisq")`	Full vs Reduced model	Tests if added variables improve fit
GLM Fitting Algorithm	Iteratively Reweighted Least Squares (IRLS)	Optimization in `glm()`	More efficient than standard `mle()`