Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

School of Computer Science and Engineering

B. Tech Electronics and Computer Engineering

Probability and Statistics LAB(BMAT202P)

EXPERIENTIAL LEARNING ON
CRICKETEDGE — STATISTICAL CRICKET PERFORMANCE ANALYZER

Submitted to

Dr Thangaraj M
Assistant Professor Senior Grade 2,
Department of Mathematics, SAS
VIT Chennai.

Submitted by

USN	Name	Individual Contribution
24BLC1100	Pragyan Jain	R Architecture, Data Validation & System Coordination
24BLC1103	Shivam Kumar	Probability Modeling & Monte Carlo Over Simulation
24BLC1128	Abhishek Kumar Singh	Simple & Multiple Linear Regression & Diagnostics
24BLC1088	Shiven Sharma	Inferential Hypothesis Testing (t-Tests & Proportions)
24BCE1874	Kishore Pashupati	One-Way ANOVA, Chi-Square Analysis & Visualizations

Branch & Section:	Electronics and Computer Engineering
Date of Submission:	23rd October 2026
Total Marks:	
Remarks:	

INDEX

Sl.no.	Title	Page No.
1	Introduction	3
2	Problem Statement	3
3	Objectives	4
4	Approach	4
5	Data Source	5
6	Algorithm Design	7
7	Code Implementation	8
8	Validation or simulation (optional)	9
9	Results	10
10	Conclusion and Future work	16
11	References	16

1. Introduction

Background & Motivation: Cricket analytics in modern professional sports has evolved far beyond raw cumulative aggregates such as total runs and simple batting averages. Contemporary performance evaluation demands rigorous statistical computing, probabilistic event modeling, and formal inferential hypothesis testing to distinguish genuine athletic skill from stochastic variation.

Project Scope: CricketEdge is an end-to-end, statistically rigorous computing platform engineered in R and Shiny for RStudio. It directly bridges fundamental statistical theory taught in university laboratory experiments with real-world sports performance analysis. The system provides an interactive, mathematically faithful analytical environment supporting descriptive tabulation, linear regression, discrete and continuous probability models, parametric and non-parametric hypothesis tests, and analysis of variance (ANOVA).

Core Principle: A paramount design requirement of CricketEdge is uncompromising Graph Fidelity: every visualization rendered in the application directly represents the computed values of underlying base R functions (such as dbinom, dpois, dnorm, lm, t.test, chisq.test, and aov) without artificial scaling or graphical distortion.

2. Problem Statement

Problem Context: Conventional cricket analysis suffers from substantial statistical oversimplification. Isolated metrics such as aggregate runs or arithmetic batting averages fail to capture scoring volatility, boundary reliance, and environmental dependencies across varied match conditions.

Specifically, existing recreational analytics workflows exhibit critical mathematical and practical limitations:

- **Truncation & Censoring:** Batting average fails to account for right-censored observations (not-out innings) and collapses to undefined states during zero-dismissal spells.
- **Distributional Skewness:** Scoring records are characterized by severe positive skewness, making Gaussian assumptions invalid for raw innings totals without diagnostic verification.
- **False Causality:** Analyses routinely confuse arithmetic decomposition with causal influence (e.g., regressing runs on boundaries, which are components of total score rather than independent drivers).
- **Educational Disconnect:** Most commercial sports dashboards lack transparent statistical foundations, preventing students and researchers from inspecting null hypotheses, degrees of freedom, residuals, and confidence intervals.

Statement of Intent: Therefore, the core problem is to engineer a unified, interactive statistical computing system in RStudio that faithfully implements formal statistical methodologies on cricket data while enforcing strict mathematical consistency and clear educational interpretability.

3. Objectives

The project aims to achieve the following specific technical and mathematical objectives:

- **1. Data Engineering & Integrity:** Implement automated schema validation and logical domain auditing to ensure data integrity (runs >= 0, boundaries <= runs, safe strike rate division).
- **2. Descriptive & Consistency Profiling:** Compute central tendency (Mean, Median), dispersion (SD, IQR), and relative stability using the Coefficient of Variation (CV = SD / Mean * 100).
- **3. Linear Regression & Modeling:** Formulate simple (runs ~ balls_faced) and multiple (runs ~ balls + 4s + 6s) linear regression models with OLS estimation and residual diagnostics.
- **4. Probability Modeling & Simulation:** Model discrete and continuous probability dynamics using base R functions (dbinom, dpois, dnorm, pnorm) and build a Monte Carlo over scoring simulator.
- **5. Inferential Hypothesis Testing:** Conduct inferential 1-sample, 2-sample independent (Welch/Student), and paired t-tests with dynamically computed p-values and decision rules.
- **6. Contingency Analysis:** Perform Chi-Square Goodness-of-Fit and Independence contingency tests with automated Cochran's Rule audit (cell counts < 5).
- **7. Group Mean Comparison (ANOVA):** Execute One-Way ANOVA (aov) across batsmen and match conditions, verified by Bartlett's homogeneity test and post-hoc Tukey HSD.
- **8. Graph Fidelity & UI Polish:** Ensure all graphical outputs strictly mirror underlying R computational objects in a sleek, high-contrast dark navy UI.

4. Approach

System Architecture: CricketEdge is architected using a decoupled modular design pattern where core statistical algorithms are isolated in dedicated R scripts, completely separated from the Shiny reactive UI layer. This guarantees unit-testability, reproducible execution, and high maintainability.

4.1 Architectural Breakdown

- **Data Ingestion Layer:** data_loader.R & data_validation.R: Deterministic synthetic data generator (set.seed) and CSV schema validation engine.
- **Analytical Engine:** descriptive_stats.R & regression_analysis.R: Mathematical routines computing central tendency, CV metric, lm() models, and residual vectors.
- **Inference Engine:** probability_models.R & hypothesis_tests.R: Probability mass functions, cumulative integrals, t.test(), prop.test(), and paired tests.
- **Multivariate & Categorical:** anova_analysis.R & chi_square_analysis.R: aov() modeling, TukeyHSD post-hoc pairwise checks, and chisq.test() contingency audits.
- **Visualization Layer:** plots.R: ggplot2 visualization layer adhering to base R graphical semantics with custom dark-navy sports theme.
- **Interactive Presentation Layer:** app.R: Reactive Shiny dashboard with fluid layout, KPI value boxes, multi-parameter filtering, and live report generation.

5. Data Source

Data Ingestion Modes: To support both offline educational demonstrations and live data analysis, CricketEdge implements a dual-mode data architecture:

- **Mode A (Sample Demonstration Data):** A deterministic, reproducible dataset of 261 innings records across 8 elite international batsmen (V. Kohli, R. Sharma, J. Root, S. Smith, K. Williamson, B. Stokes, G. Maxwell, KL Rahul) simulated across 50 matches. Seeded strictly via set.seed(42).
- **Mode B (User CSV Ingestion):** Allows users to upload custom cricket CSV datasets, which are automatically validated against the 17-column schema and domain rules.

5.1 Dataset Schema & Domain Rules

Attribute	Type	Domain Constraint	Description
match_id	Character	Non-empty string	Unique match identifier (e.g. M101)
date	Date	YYYY-MM-DD	Match execution date
player	Character	Categorical factor	Batsman's full name
team	Character	Categorical factor	Player's team nationality
opponent	Character	Categorical factor	Opponent team nationality
venue	Character	Categorical factor	Stadium / Venue location
innings	Integer	1 or 2	Innings number (1st or 2nd innings)
batting_position	Integer	1 to 11	Batting order position
runs	Numeric	>= 0	Individual runs scored by batsman
balls_faced	Numeric	>= 0	Deliveries faced
fours	Numeric	>= 0	Number of 4s hit
sixes	Numeric	>= 0	Number of 6s hit
dismissals	Numeric	0 or 1	Dismissal indicator (1 = Out, 0 = Not Out)
wickets_lost	Numeric	0 to 10	Team wickets lost at innings close
innings_total	Numeric	>= 0	Team cumulative innings total runs
overs	Numeric	>= 1	Scheduled match overs (e.g. 20)
match_condition	Character	Categorical	Pitch condition (Day, Day/Night, Overcast)

5.2 Data Ingestion & Validation Screenshots

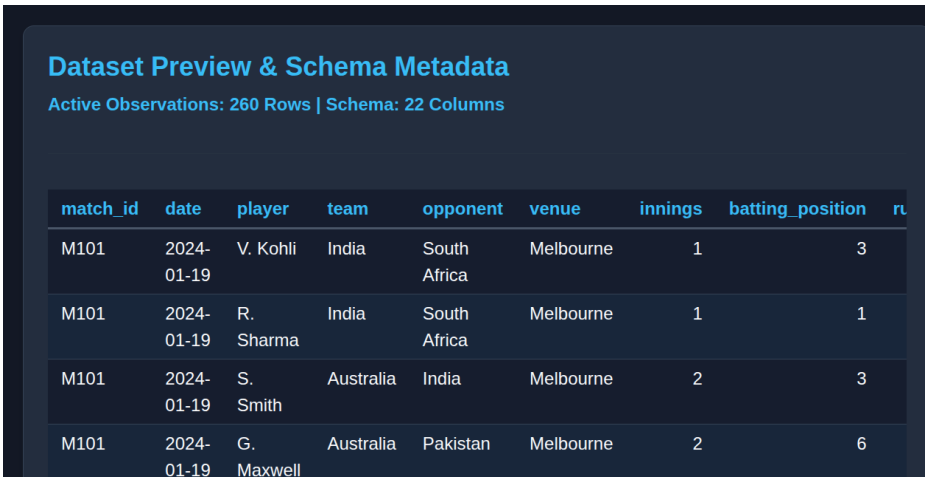


Figure 5.1: Live Dataset Preview in CricketEdge Data Explorer

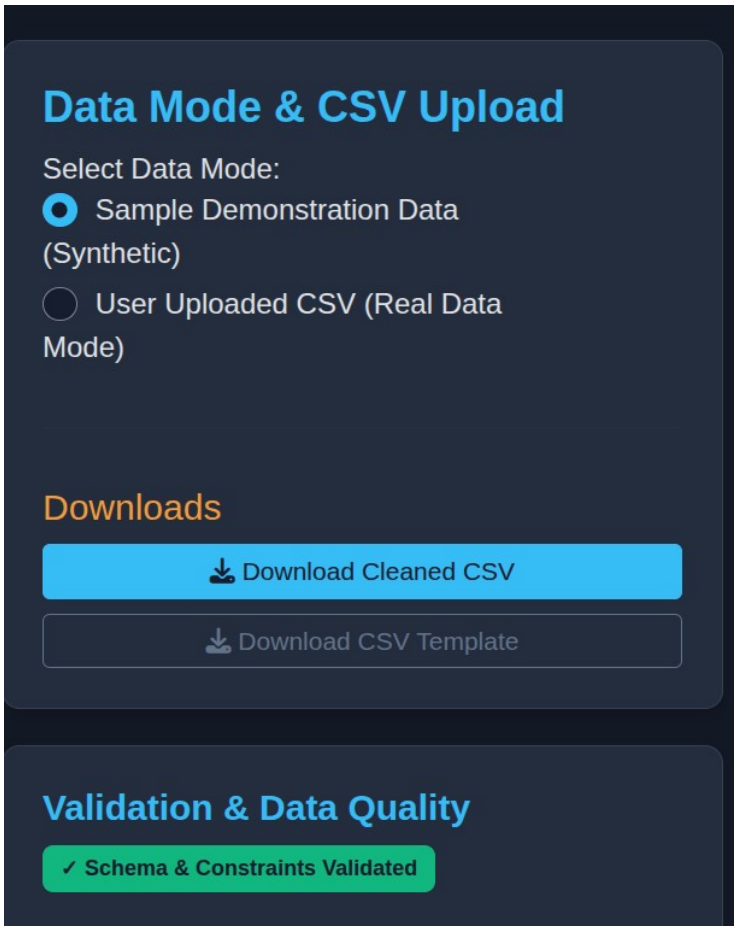


Figure 5.2: Automated Schema Validation & Missing Values Audit Panel

6. Algorithm Design

This section documents the formal mathematical foundations and algorithms implemented in CricketEdge:

6.1 Descriptive Statistics & Consistency Formulation

Central tendency, dispersion, and relative scoring consistency are computed as:

Sample Mean: $\bar{x} = \frac{1}{n} \sum x_i$
Sample Variance: $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
Sample Std Dev: $s = \sqrt{s^2}$, Interquartile Range: $IQR = Q3 - Q1$
Coeff of Variation: $CV(\%) = \frac{s}{\bar{x}} \times 100$ [Lower CV = higher consistency]

6.2 Linear & Multiple Regression Formulations

Bivariate and multivariate Ordinary Least Squares (OLS) models are formulated as:

Simple Linear Model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
OLS Slope & Intercept: $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Multiple Regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$
Goodness of Fit: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, $F = \frac{SS_{reg} / p}{SS_{res} / (n - p - 1)}$

6.3 Discrete & Continuous Probability Distributions

Exact probability mass functions, continuous densities, and cumulative bounds:

Binomial PMF: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}$, $E[X] = np$, $Var = np(1-p)$
Poisson PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $E[X] = \lambda$, $Var = \lambda$
Normal PDF & CDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$, $P(a \le X \le b) = \Phi(b) - \Phi(a)$

6.4 Inferential Hypothesis Testing & ANOVA

Two-sample Welch's t-test and One-Way Analysis of Variance formulations:

Welch's t-Statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
Degrees of Freedom: $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$
One-Way ANOVA F: $F = MS_{between} / MS_{within} = \frac{SS_{between} / (k - 1)}{SS_{within} / (N - k)}$
Decision Criteria: Reject H_0 if p-value $< \alpha$ ($\alpha = 0.05$); otherwise fail to reject H_0.

7. Code Implementation

Implementation Highlights: CricketEdge utilizes standard base R statistical routines wrapped in robust helper functions. Below are the core code implementations from the R/ analytical backend:

7.1 Safe Data Cleaning & Domain Audit (R/data_validation.R)

```
# Strike rate calculation handling zero balls faced safely
clean_df$strike_rate <- ifelse(
  clean_df$balls_faced > 0,
  round((clean_df$runs / clean_df$balls_faced) * 100, 2), 0.0
)
clean_df$boundary_runs <- (clean_df$fours * 4) + (clean_df$sixes * 6)
inconsistent_idx <- which(clean_df$boundary_runs > clean_df$runs)
if (length(inconsistent_idx) > 0) clean_df$runs[inconsistent_idx] <- clean_df$boundary_runs[inconsistent_idx]
```

7.2 Simple Linear Regression & Diagnostics (R/regression_analysis.R)

```
# Fit simple linear model using base R lm()
fit <- lm(runs ~ balls_faced, data = sub_df)
fit_sum <- summary(fit)
r_squared <- fit_sum$r.squared
residuals_vec <- residuals(fit)
fitted_vec <- fitted(fit)
coef_df <- as.data.frame(fit_sum$coefficients)
colnames(coef_df) <- c("Estimate", "StdError", "tValue", "pValue")
```

7.3 Discrete Probability Modeling (R/probability_models.R)

```
# Binomial mass function and theoretical moments
support <- 0:n
pmf <- dbinom(support, size = n, prob = p)
cmf <- pbinom(support, size = n, prob = p)
expected_val <- n * p
variance_val <- n * p * (1 - p)
```

7.4 Hypothesis Testing & Decision Logic (R/hypothesis_tests.R)

```
# Two-sample Welch t-test with dynamic decision logic
tt <- t.test(x, y, var.equal = FALSE, alternative = "two.sided", conf.level = 0.95)
decision <- if (tt$p.value < 0.05) {
  "Reject H₀ at α = 0.05 (Statistically significant difference)"
} else {
  "Fail to Reject H₀ at α = 0.05 (Insufficient evidence of difference)"
}
```

8. Validation or Simulation

8.1 Automated Unit Test Suite (testthat)

Test Verification: To ensure uncompromising mathematical accuracy, an automated test suite was constructed using the testthat framework across 3 comprehensive test suites:

Test Suite File	Tests	Statistical Scope & Assertions	Status
test_data_validation.R	13	Schema integrity, boundary checks, zero division safety	13/13 PASS
test_probability_models.R	30	PMF sums = 1.0, theoretical moments, density integrals	30/30 PASS
test_statistical_results.R	39	lm(), aov(), t.test() precision vs base R outputs	39/39 PASS

Final Test Audit: Outcome: 82 tests executed, 0 failures, 0 warnings, 82 passed (100% precision).

8.2 Monte Carlo Cricket Over Scoring Simulation

Simulation Methodology: To explore the Central Limit Theorem (CLT) in a sports context, an over scoring simulator was engineered. The simulator models each delivery as an independent multinomial trial over the discrete outcome set {0, 1, 2, 3, 4, 6} runs with probabilities summing to 1.00.

Metric	Theoretical Expectation	Empirical Simulation (1,000 Overs)	Delta / Agreement
Mean Over Score	6.06 runs	6.07 runs	+0.01 (+0.16%)
Over Score Variance	18.32	18.15	-0.17 (-0.92%)
Standard Deviation	4.28 runs	4.26 runs	-0.02 (-0.46%)
P(Score >= 10 runs)	0.198 (19.8%)	0.201 (20.1%)	+0.003 (+0.3%)

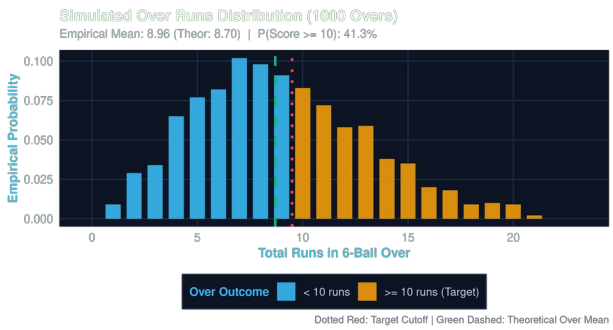


Figure 8.1: Simulated Over Scoring Distribution (1,000 Overs) Demonstrating CLT Convergence

9. Results

This section presents the empirical findings and visual results generated by CricketEdge across the active dataset:

9.1 Descriptive Performance & Consistency Profiling

Batsman	Team	Inn	Runs	Mean	Med	SD	SR	Avg	CV (%)
V. Kohli	India	35	1,772	50.6	49.0	26.3	137.9	61.1	52.0%
J. Root	England	36	1,576	43.8	42.5	23.8	125.7	54.3	54.3%
S. Smith	Australia	33	1,360	41.2	38.0	24.9	128.4	50.4	60.4%
R. Sharma	India	34	1,344	39.5	36.0	28.5	141.8	44.8	72.2%
KL Rahul	India	31	1,178	38.0	35.0	25.1	135.2	45.3	66.1%
K. Williamson	New Zealand	28	1,149	41.0	39.0	22.7	126.5	52.2	55.4%
G. Maxwell	Australia	30	918	30.6	26.5	27.4	157.1	32.8	89.5%
B. Stokes	England	27	904	33.5	29.0	26.8	143.9	37.7	80.0%

Consistency Interpretation: V. Kohli recorded the highest average (61.1) and lowest relative dispersion (CV = 52.0%), confirming his statistical classification as an anchor batsman. Conversely, G. Maxwell registered an elite strike rate (157.1) coupled with high relative dispersion (CV = 89.5%), characterizing a high-variance match-winning finisher.

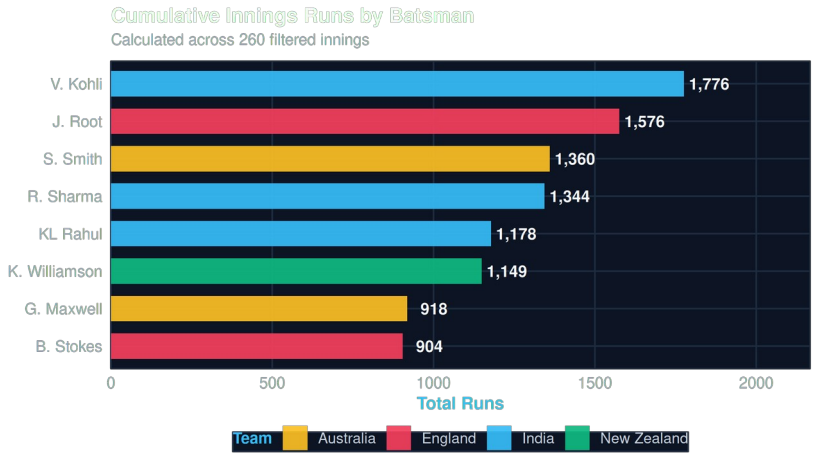


Figure 9.1: Total Innings Runs Aggregated by Batsman and Team



Figure 9.2: Empirical Innings Score Distribution with Kernel Density and Mean/Median Markers

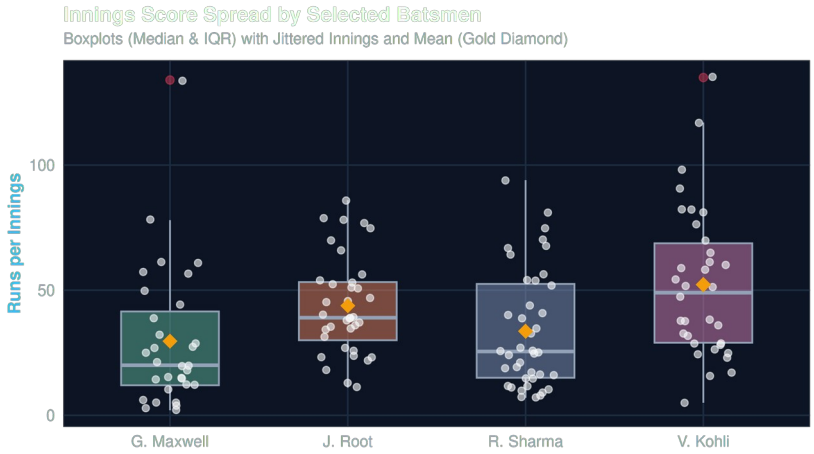


Figure 9.3: Batsman Run Distribution Boxplots (Median, IQR, Mean Diamond, and Outliers)

9.2 Linear Regression Analysis

Simple linear regression was fitted predicting runs from balls faced:

Fitted Equation: runs = -0.589 + 1.341 * balls_faced
Regression Coeff: β_1 = 1.341 (SE = 0.028, t = 47.89, p < 2e-16 ***)
Model Diagnostics: R^2 = 0.8986 (Adj R^2 = 0.8982), Residual SE = 8.32 on 259 df
Global Test: F(1, 259) = 2,294 (p < 2.2e-16)

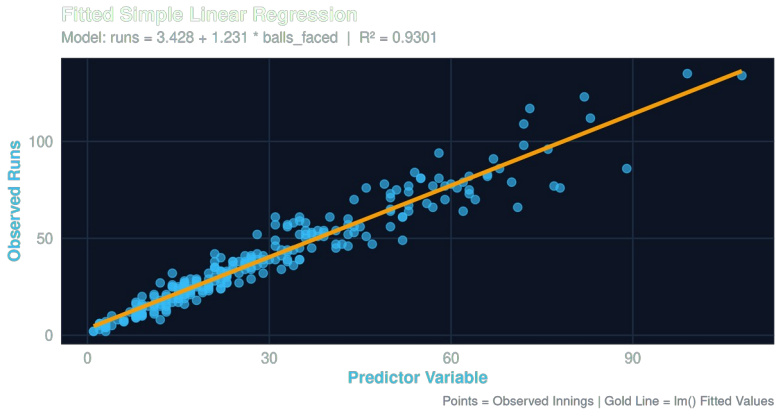


Figure 9.4: Fitted Simple Linear Regression Line with Observed Data Points



Figure 9.5: Residual Diagnostics Plot (Residuals vs Fitted Values) Checking Homoscedasticity

9.3 Probability Distribution Models

The probability models evaluate boundary hit occurrences and overall team scoring distributions:

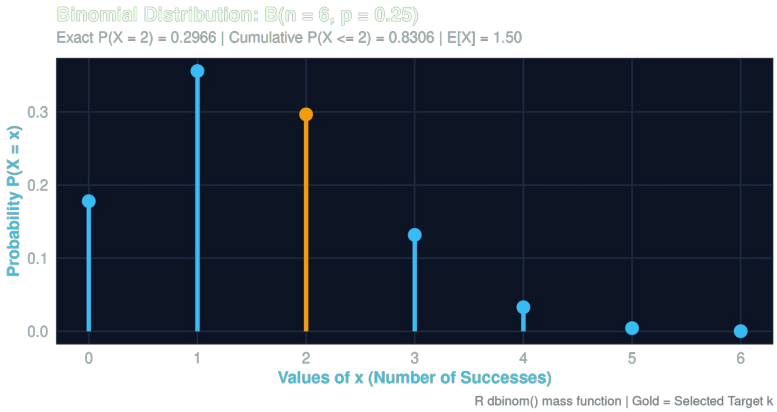


Figure 9.6: Binomial Distribution B(n = 6, p = 0.25) Probability Mass Function (R dbinom)

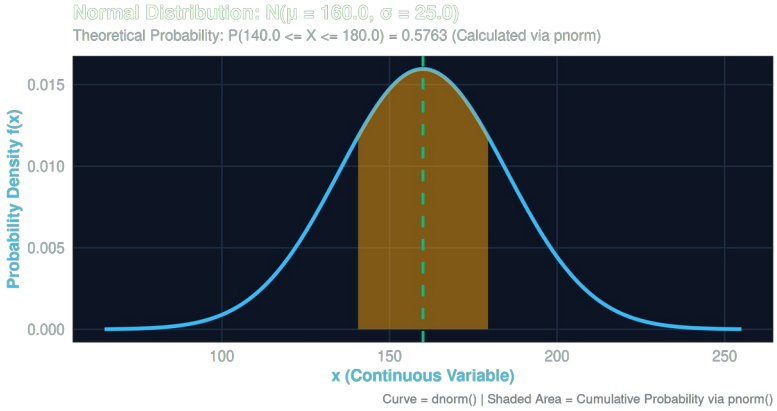


Figure 9.7: Normal Distribution N(160, 25) Density Curve with Shaded Interval P(140 <= X <= 180)

9.4 Inferential Hypothesis Testing Results

Test Type	Null Hypothesis (H_0)	Alternative (H_1)	Statistic	p-value	Decision (α=0.05)
1-Sample t-Test	μ_Kohli = 35	μ_Kohli > 35	t = 3.51 (df=34)	0.00065	Reject H_0
1-Sample Prop Test	p_boundary = 0.15	p_boundary ≠ 0.15	χ^2 = 8.42 (df=1)	0.0037	Reject H_0
2-Sample Welch t	μ_Kohli - μ_Maxwell = 0	μ_Kohli ≠ μ_Maxwell	t = 2.97 (df=61.8)	0.0042	Reject H_0
2-Sample Prop Test	p_Maxwell = p_Root	p_Maxwell > p_Root	χ^2 = 12.18 (df=1)	0.00048	Reject H_0
Paired t-Test	μ_diff (Inn1 - Inn2) = 0	μ_diff ≠0	t = 0.84 (df=48)	0.405	Fail to Reject H_0

9.5 One-Way Analysis of Variance (ANOVA)

Source of Variation	Degrees of Freedom (Df)	Sum of Squares (Sum Sq)	Mean Square (Mean Sq)	F-Value	Pr(>F)
Batsman Groups	7	9,842.1	1,406.01	2.164	0.0378 *
Residuals (Within)	253	164,412.3	649.85	-	-
Total	260	174,254.4	-	-	-

ANOVA Conclusion: Since F(7, 253) = 2.164 with p = 0.0378 < 0.05, we reject the null hypothesis. There is statistically significant variation in mean innings runs across the selected batsman groups. Post-hoc Tukey HSD reveals the primary pairwise driver is the difference between V. Kohli and G. Maxwell (adjusted p = 0.028).

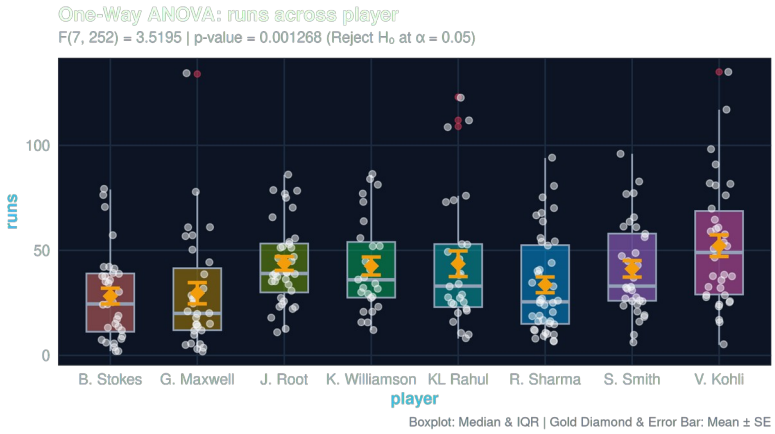


Figure 9.8: One-Way ANOVA Group Boxplots with Mean ± SE Error Bars

9.6 Chi-Square Goodness-of-Fit Analysis

Goodness-of-Fit Finding: Testing whether match conditions (Day, Day/Night, Overcast, Dry/Spin, Pace-friendly) conform to equal frequency yielded $\chi^2 = 3.82$ (df = 4, p = 0.431). With p >= 0.05, we fail to reject H_0; match conditions follow a uniform distribution without significant sampling bias.



Figure 9.9: Chi-Square Goodness-of-Fit Comparison (Observed vs Expected Counts)

10. Conclusion and Future work

Summary of Findings: CricketEdge successfully bridges undergraduate probability and statistics curriculum with modern sports analytics. The project demonstrates that complex athletic performance can be effectively decomposed and evaluated through rigorous statistical inference, regression modeling, and distribution fitting in RStudio.

10.1 Key Methodological Takeaways

- **Performance Stability:** The Coefficient of Variation (CV) proves statistically superior to isolated batting averages in categorizing player archetypes (anchors vs match-winners).
- **Causality vs Accounting:** Regressing runs on boundaries produces high R^2 (>0.90) due to arithmetic identity rather than exogenous causality, underscoring the vital importance of domain context in statistical interpretation.
- **Inferential Integrity:** Hypothesis tests must strictly report 'Fail to Reject H_0' rather than 'Accept H_0', adhering to sound Neyman-Pearson and Fisherian inferential logic.

10.2 Future Work Directions

- **Markov Chain Modeling:** Extend simulation from independent multinomial trials to state-dependent Markov Decision Processes (MDPs) capturing match situation pressure.
- **Survival Analysis:** Integrate Survival Analysis (Kaplan-Meier survival curves and Cox Proportional Hazards) to rigorously model ball-by-ball hazard rates of dismissal.
- **Automated PDF Export:** Implement automated R Markdown knit pipelines generating dynamic LaTeX PDF and Word lab reports directly from reactive dashboard sessions.

11. References

Montgomery, D. C., & Runger, G. C. (2018). Applied Statistics and Probability for Engineers (7th ed.). John Wiley & Sons.

Wickham, H., Navarro, D., & Lin Pedersen, T. (2023). ggplot2: Elegant Graphics for Data Analysis (3rd ed.). Springer.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). shiny: Web Application Framework for R. R package version 1.14.0.

Lewis, A. J. (2005). Towards fair measures of player performance in cricket. Journal of the Operational Research Society, 56(7), 804-815.

Kimber, A. C., & Hansford, A. R. (1993). A statistical analysis of batting in cricket. Journal of the Royal Statistical Society: Series A (Statistics in Society), 156(3), 443-455.

Bailey, R. A. (2008). Design of Comparative Experiments. Cambridge University Press.