

Ball-by-Ball Statistical Analysis of an IPL Match

Amaan Shaikh | Danish Ali Shaikh

2025-12-27

1. Introduction

This report presents a descriptive statistical analysis of a single IPL match (Sunrisers Hyderabad vs Royal Challengers Bengaluru, Match ID: 1426268) using ball-by-ball data.

The objective is to apply univariate and bivariate statistical techniques taught in the course and to explore relationships between variables without implying causality.

2. Data Loading and Preprocessing

```
rm(list = ls(all.names = TRUE))

ipl_data <- read.csv("data/IPL.csv")

ipl_data$current_run_rate <- round(ipl_data$team_runs / ipl_data$ball_no, 2)

included_columns <- c(
  "batter",
  "bowler",
  "batter_balls",
  "batter_runs",
  "batting_team",
  "current_run_rate",
  "over",
  "runs_total",
  "team_runs"
)

match <- subset(
  x = ipl_data,
  subset = match_id == 1426268,
  select = included_columns
)
```

Each observation represents one delivery in the match.

3. Univariate Analysis

3.1 Runs per Ball

```
table(match$runs_total)
```

```
##
##    0    1    2    4    6
##  51 117   16   43   38
```

```
round(prop.table(table(match$runs_total)), 2) * 100
```

```
##
##    0    1    2    4    6
## 19 44    6 16 14
```

This variable provides the most meaningful univariate insight, showing the distribution of scoring outcomes per delivery.

3.2 Boundary Events

```
match$is_boundary <- ifelse(match$runs_total %in% c(4, 6), 1, 0)
```

```
table(match$is_boundary)
```

```
##
##    0    1
## 184   81
```

```
prop.table(table(match$is_boundary))
```

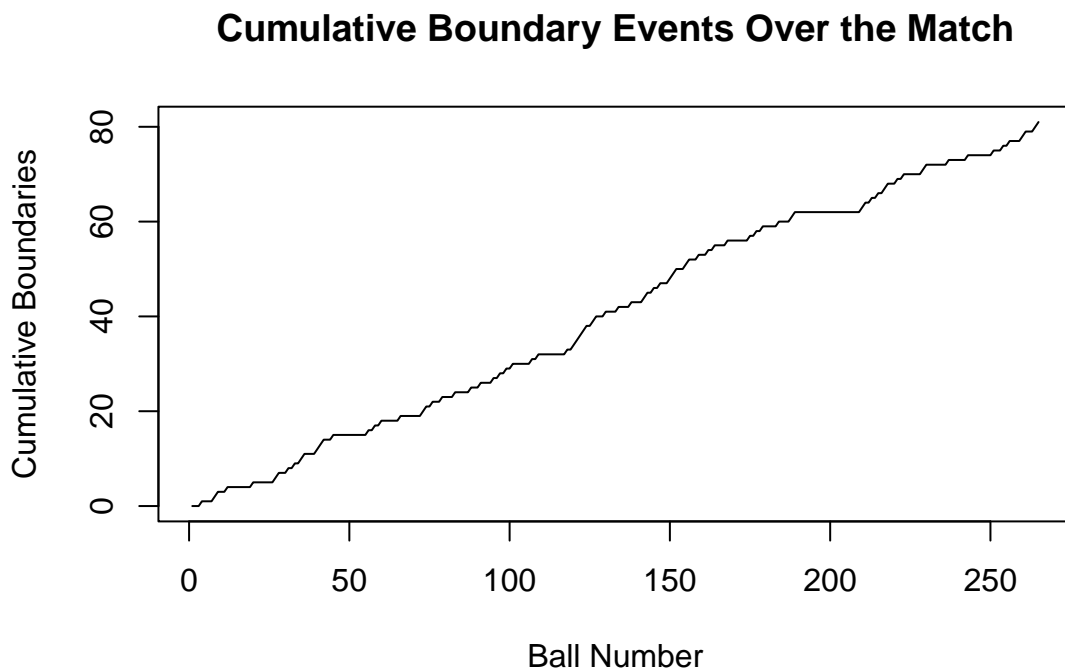
```
##
##           0           1
## 0.6943396 0.3056604
```

Boundaries form a minority of deliveries but contribute heavily to scoring.

3.3 Cumulative Boundary Progression

```
match$cum_boundaries <- cumsum(match$is_boundary)

plot(match$cum_boundaries,
      type = "l",
      xlab = "Ball Number",
      ylab = "Cumulative Boundaries",
      main = "Cumulative Boundary Events Over the Match")
```



This plot is descriptive only and does not imply momentum or strategy.

4. Bivariate Analysis

4.1 Runs per Ball by Batting Team

(Qualitative \times Numeric)

```
tapply(match$runs_total, match$batting_team, mean)
```

## Royal Challengers Bengaluru	Sunrisers Hyderabad
## 1.969925	2.174242

```
tapply(match$runs_total, match$batting_team, median)
```

```
## Royal Challengers Bengaluru      Sunrisers Hyderabad
##                               1                1
```

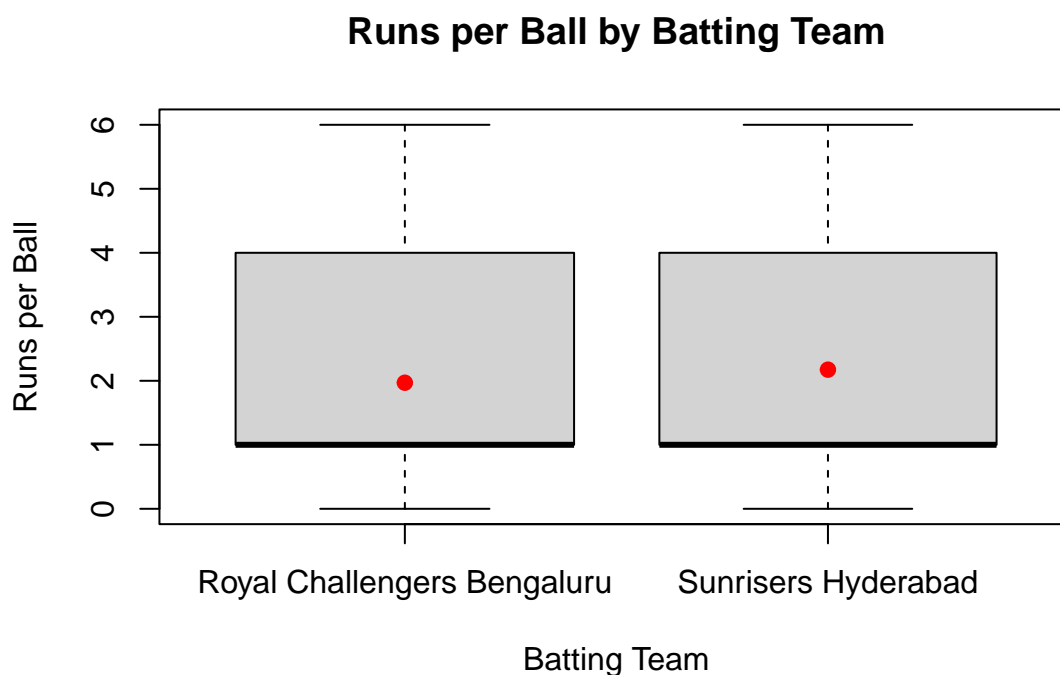
```
tapply(match$runs_total, match$batting_team, sd)
```

```
## Royal Challengers Bengaluru      Sunrisers Hyderabad
##                1.995980                2.087755
```

```
bp <- boxplot(runs_total ~ batting_team,
              data = match,
              main = "Runs per Ball by Batting Team",
              xlab = "Batting Team",
              ylab = "Runs per Ball",
              varwidth = TRUE,
              col = "lightgray")

group_means <- tapply(match$runs_total, match$batting_team, mean)

points(seq_along(bp$names),
       group_means[bp$names],
       pch = 19,
       col = "red")
```

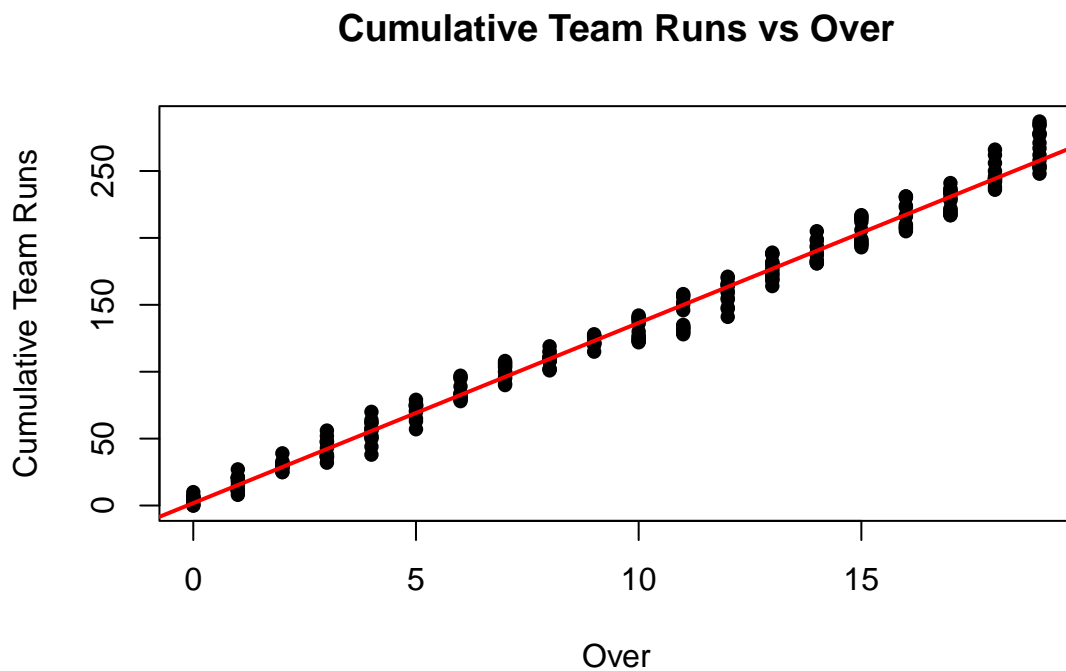


4.2 Cumulative Team Runs vs Over

(Numeric \times Numeric)

```
plot(match$over,
      match$team_runs,
      pch = 16,
      xlab = "Over",
      ylab = "Cumulative Team Runs",
      main = "Cumulative Team Runs vs Over")

lm_runs_over <- lm(team_runs ~ over, data = match)
abline(lm_runs_over, col = "red", lwd = 2)
```



```
cor(match$over, match$team_runs)
```

```
## [1] 0.9935467
```

```
summary(lm_runs_over)
```

```
##
## Call:
## lm(formula = team_runs ~ over, data = match)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.4500  -5.7379   0.1864   5.4288  29.2621
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.81355    1.07415   1.688  0.0925 .
## over        13.46970    0.09482 142.057  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.801 on 263 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9871
## F-statistic: 2.018e+04 on 1 and 263 DF,  p-value: < 2.2e-16
```

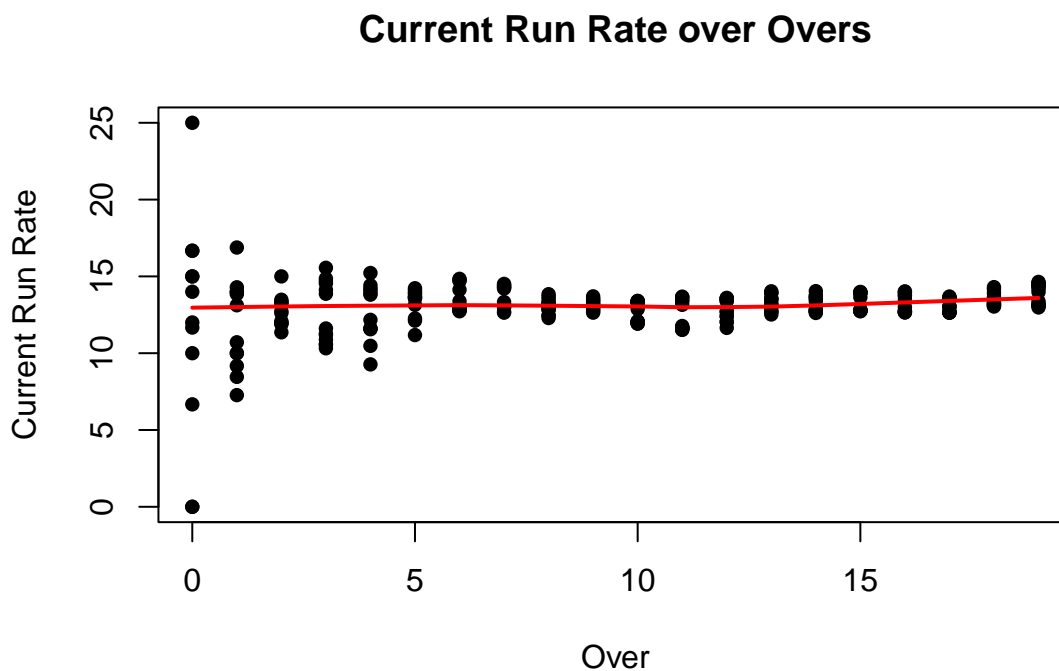
High correlation is expected because team runs are cumulative.

4.3 Current Run Rate over Overs

(Ordinal \times Continuous)

```
plot(match$over,
      match$current_run_rate,
      pch = 16,
      xlab = "Over",
      ylab = "Current Run Rate",
      main = "Current Run Rate over Overs")

lines(lowess(match$over, match$current_run_rate),
      col = "red", lwd = 2)
```



Early overs show high volatility, followed by stabilisation.

4.4 Boundary Frequency by Batting Team

(Qualitative \times Binary)

```
table(match$batting_team, match$is_boundary)
```

```
##
##              0  1
## Royal Challengers Bengaluru 93 40
## Sunrisers Hyderabad        91 41
```

```
prop.table(table(match$batting_team, match$is_boundary), margin = 1)
```

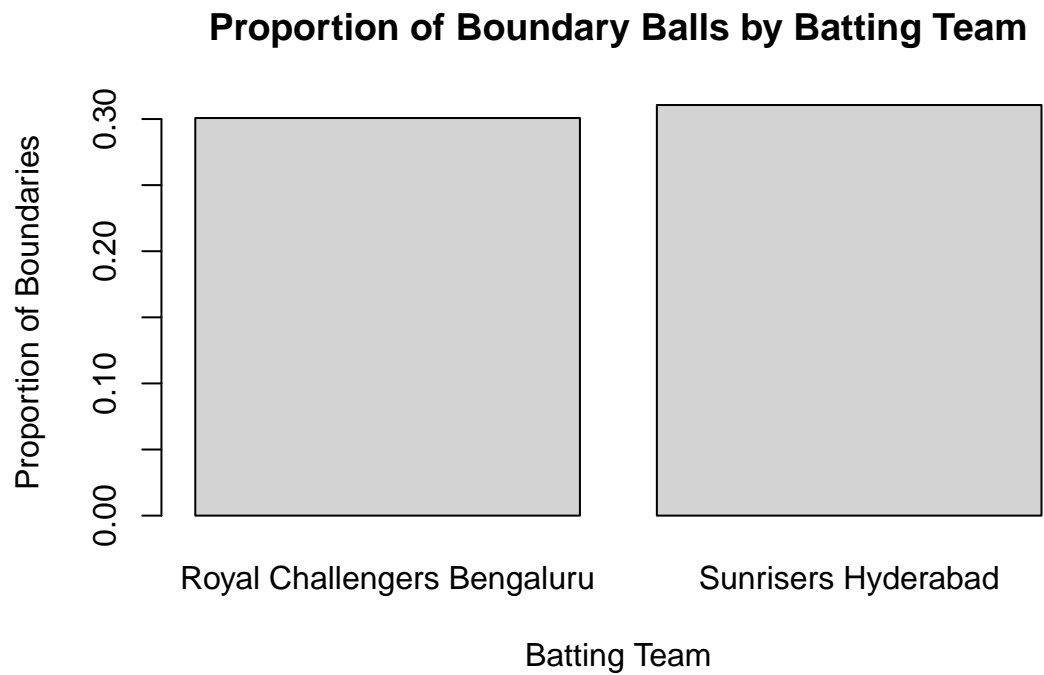
```
##
##              0          1
## Royal Challengers Bengaluru 0.6992481 0.3007519
## Sunrisers Hyderabad        0.6893939 0.3106061
```

```
barplot(
  prop.table(table(match$batting_team, match$is_boundary), margin = 1)[,2],
  main = "Proportion of Boundary Balls by Batting Team",
```

```

xlab = "Batting Team",
ylab = "Proportion of Boundaries",
col = "lightgray"
)

```



5. Conclusion

This report demonstrates the application of descriptive univariate and bivariate statistics to IPL ball-by-ball data while respecting the limitations of cumulative variables.

Key findings include:

- Runs per ball is the most informative univariate variable
- Boundary events are relatively infrequent but highly impactful
- Differences between teams are subtle and distribution-based