

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A1b: Preliminary preparation and analysis of data- Descriptive statistics**

**VARUN YOGANANDA**

**V01107265**

**Date of Submission: 18-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **1** |
| **3.** | Interpretations | **1** |

**INTRODUCTION**

This analysis focuses on analyzing IPL cricket data to gain significant insights into player performance and financial benefits. Using R/Python, sophisticated statistical programming languages, IPL organisers' dataset will be cleaned and organised round-wise to contain detailed statistics such as batsman, ball, runs, and wickets per player per match. The research seeks to determine the top three run-getters and wicket-takers from each IPL round. By fitting the most appropriate statistical distributions for the runs scored and wickets taken by these top performers during the last three IPL tournaments, we will acquire a better understanding of performance trends. Furthermore, the initiative will look into the relationship between players' on-field performance and their salary, examining how remuneration connects with cricket contributions.

**OBJECTIVES**

1. Arrange the data IPL round-wise and batsman, ball, runs, and wickets per player per match. Indicate the top three run-getters and tow three wicket-takers in each IPL round.
2. Fit the most appropriate distribution for runs scored and wickets taken by the top three batsmen and bowlers in the lost three IPL tournaments.
3. Find the relationship between a player’s performance and the salary he gets in your data.

**RESULTS & INTERPRETATION**

**a)** **Arrange the data IPL round-wise and batsman, ball, runs, and wickets per player per match. Indicate the top three run-getters and top three wicket-takers in each IPL round. (From R)**

**Code:**

> # Summarise player runs and wickets

> player\_runs <- grouped\_data %>%

+ group\_by(Season, Striker) %>%

+ summarise(runs\_scored = sum(runs\_scored, na.rm = TRUE)) %>%

+ ungroup()

> player\_wickets <- grouped\_data %>%

+ group\_by(Season, Bowler) %>%

+ summarise(wicket\_confirmation = sum(wicket\_confirmation, na.rm = TRUE)) %>%

+ ungroup()

> # Sort player runs for season 2023

> player\_runs\_2023 <- player\_runs %>%

+ filter(Season == '2023') %>%

+ arrange(desc(runs\_scored))

>

> # Get top 3 run-getters and bottom 3 wicket-takers per season

> top\_run\_getters <- player\_runs %>%

+ group\_by(Season) %>%

+ top\_n(3, runs\_scored) %>%

+ ungroup()

**Result:**

> print(top\_run\_getters)

# A tibble: 51 × 3

Season Striker runs\_scored

*<chr>* *<chr>* *<dbl>*

1 2007/08 G Gambhir 534

2 2007/08 SE Marsh 616

3 2007/08 ST Jayasuriya 514

4 2009 AB de Villiers 465

5 2009 AC Gilchrist 495

6 2009 ML Hayden 572

7 2009/10 JH Kallis 572

8 2009/10 SK Raina 528

9 2009/10 SR Tendulkar 618

10 2011 CH Gayle 608

> print(bottom\_wicket\_takers)

# A tibble: 58 × 3

Season Bowler wicket\_confirmation

*<chr>* *<chr>* *<dbl>*

1 2007/08 IK Pathan 20

2 2007/08 JA Morkel 20

3 2007/08 SK Warne 20

4 2007/08 SR Watson 20

5 2007/08 Sohail Tanvir 24

6 2009 A Kumble 22

7 2009 A Nehra 22

8 2009 RP Singh 26

9 2009/10 A Mishra 20

10 2009/10 Harbhajan Singh 20

**Interpretation:**

The data provides valuable insights into the best run scorers and wicket-takers across IPL seasons. For example, in the 2007/08 season, SE Marsh scored the most runs (616), closely followed by G Gambhir and ST Jayasuriya. In later seasons, stars like as ML Hayden (572 runs in 2009) and SR Tendulkar (618 runs in 2009/10) dominated the run rankings. The trend indicates that different players thrived in different seasons, emphasising the IPL's competitive character.   
  
Sohail Tanvir topped the bowling attack in 2007/08 with 24 wickets, while IK Pathan, JA Morkel, SK Warne, and SR Watson all got 20 wickets. In the 2009 season, RP Singh led the rankings with 26 wickets, while A Kumble and A Nehra taking 22 wickets each.

These statistics underscore the dynamic and unpredictable nature of the IPL, where both seasoned players and emerging talents consistently vie for the top spots in both batting and bowling categories.

**B) Fit the most appropriate distribution for runs scored and wickets taken by the top three batsmen and bowlers in the lost three IPL tournaments.**

**(Code from R)**

> # Define a function to get the best distribution

> get\_best\_distribution <- function(data) {

+ dist\_names <- c('norm', 'lnorm', 'gamma', 'weibull', 'exponential', 'logis', 'cauchy')

+ dist\_results <- list()

+ params <- list()

+ for (dist\_name in dist\_names) {

+ fit <- fitdist(data, dist\_name)

+ ks\_test <- ks.test(data, dist\_name, fit$estimate)

+ p\_value <- ks\_test$p.value

+ cat("p value for", dist\_name, "=", p\_value, "\n")

+ dist\_results[[dist\_name]] <- p\_value

+ params[[dist\_name]] <- fit$estimate

+ }

+ best\_dist <- names(which.max(unlist(dist\_results)))

+ best\_p <- max(unlist(dist\_results))

+ cat("\nBest fitting distribution:", best\_dist, "\n")

+ cat("Best p value:", best\_p, "\n")

+ cat("Parameters for the best fit:", params[[best\_dist]], "\n")

+ return(list(best\_dist, best\_p, params[[best\_dist]]))

+ }

> # Function to fit the best distribution

> get\_best\_distribution <- function(data) {

+ # Fit different distributions

+ fit\_norm <- fitdist(data, "norm")

+ fit\_pois <- fitdist(data, "pois")

+ fit\_exp <- fitdist(data, "exp")

+

+ # Compare the distributions

+ gof\_stat <- gofstat(list(fit\_norm, fit\_pois, fit\_exp), fitnames = c("Normal", "Poisson", "Exponential"))

+

+ # Print the goodness-of-fit statistics

+ print(gof\_stat)

+

+ # Return the best fit distribution

+ best\_fit <- names(which.min(gof\_stat$aic))

+ return(best\_fit)

+ }

>

> # Fit the distribution to Q de Kock's runs scored and get the best distribution

> best\_distribution <- get\_best\_distribution(Q\_de\_Kock\_runs)

**Result:**

Goodness-of-fit statistics

Normal Poisson Exponential

Kolmogorov-Smirnov statistic 0.1280142 0.4254026 0.0805889

Cramer-von Mises statistic 0.4175224 6.0350887 0.1594708

Anderson-Darling statistic 2.6398461 Inf Inf

Goodness-of-fit criteria

Normal Poisson Exponential

Akaike's Information Criterion 989.2156 2914.264 925.9846

Bayesian Information Criterion 994.5235 2916.918 928.6386

**Interpretation:**

The goodness-of-fit statistics and criteria for top batsmen's runs indicate that the Exponential distribution is the best fit of the examined distributions. This is supported by the lowest values for the Kolmogorov-Smirnov (0.0805889), Cramer-von Mises (0.1594708), and Anderson-Darling statistics. Furthermore, the Exponential distribution has the lowest Akaike's Information Criterion (AIC: 925.9846) and Bayesian Information Criterion (BIC: 928.6386), highlighting its advantage over the Normal and Poisson distributions. The high values of these statistics for the Poisson distribution show that it is the least appropriate match, whilst the Normal distribution is intermediate but not as appropriate as the Exponential distribution.

**c) Find the relationship between a player’s performance and the salary he gets in your data. (Code from Python)**

# Create a new column in df\_salary with matched names from df\_runs

df\_salary['Matched\_Player'] = df\_salary['Player'].apply(lambda x: match\_names(x, df\_runs['Striker'].tolist()))

# Merge the DataFrames on the matched names

df\_merged = pd.merge(df\_salary, df\_runs, left\_on='Matched\_Player', right\_on='Striker')

df\_merged.info()

# Calculate the correlation

correlation = df\_merged['Rs'].corr(df\_merged['runs\_scored'])

print("Correlation between Salary and Runs:", correlation)

**Result:**

Correlation between Salary and Runs: 0.30612483765821674

**Interpretation:**

The correlation coefficient between wage and Runs Scored (0.3061) indicates a relatively good association between a player's wage and the number of runs scored in IPL matches. This suggests that, on average, athletes who score more runs get greater compensation. However, the association is weak, showing that other criteria than runs scored influence compensation, such as match-winning performances, consistency, and overall team contribution. As a result, while runs add significantly to a player's pay, they are not the only factor, emphasising the multifaceted nature of salary determination in professional cricket.