MIDTERM PROJECT REPORT DSCI 5180

MOST WICKETS IN TEST CRICKET DATA SET

I have taken the most wickets in Test Cricket data set from Kaggle website Excel. It contains 14 columns and 60 rows.

This file contains 14 attributes in total with following descriptions for all of them as:

Rank: Ranking of the bowler as per wickets taken

Player: Name of the Player

Country: It represents the Country for which the player plays

Span: Time span during which players played

Matches: Number of matches player has played during his time period

Innings: Number of innings player has played in his career

Balls: Number of Balls player has bowled so far

Runs: Total number of runs conceded by the player in their bowling

Wickets: Total number of wickets taken by the player

Average: Runs on an average for which bowler has taken 1 wicket

Econ: On an average, how many runs each bowler conceded in an over

SR: Average number of balls bowled by bowler between two wickets

5: How often player has taken 5 wickets at-least (lower than 10)

10: How often player has taken 10 wickets at-least

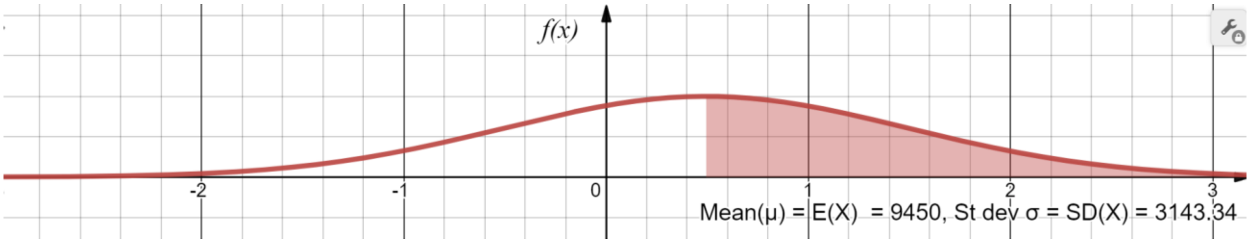
Using this data set want to answer the following questions.

Module 1:

1. The ICC test cricket batting is believed to have a normal distribution with a mean of 9450 and with a standard deviation of 3143.34. What is the probability that randomly selected players will have runs conceded more than certain value (for example 9500).

Solution: (Refer Appendix for Detailed Solution)

To calculate the probability that the randomly selected players will have runs conceded more than 9450,



The probability that randomly selected data will have runs conceded more than 9450 is 0.496.

Applying knowledge learnt from Module -2:

Module 2:

2. Construct a 95% confidence interval for the total matches of the data.

Solution: (Refer Appendix for Detailed Solution)

Mean $\bar{X} = 85$

Lower confidence interval LCL = 81.279

Upper confidence interval UCL = 88.721

Here, we are 95% confident and have sufficient evidence that the total matches of the data are between 81.279 and 88.721.

Module 3:

3. A Cricket Association mentions that the average number of runs conceded by each player is 10000. A researcher claims that the association is wrong, so they collect a sample of 51 players and calculates a mean as 9950 and a standard deviation 3159.663. Assume that the distribution is approximately normal. Test the statistician’s claim at the 0.05 level of significance.

Solution: (Refer Appendix for Detailed Solution)

The **p-value < level of significance** and hence we reject the null hypothesis. There is sufficient evidence to conclude that the mean of the average glucose levels is equal to 150.

Applying knowledge learnt from Module -5:

Module 5:

Perform regression analysis to predict the ‘Total number of wickets picked up by a player and Average number of runs conceded by each player in a match based on Number of matches’

The output in Appendix reveals $R^2 \approx 0.6987$. Thus, in the weight model, the two independent variables (Total number of wickets picked up by a player and Average number of runs conceded by each player in a match) can explain approximately 69.87% of the variation in the number of matches.

The output in Appendix D to be adjusted R^2a≈0.6882 which is 0.01% less than the value of R2. The difference between the R^2 and the adjusted R square is small. The adjusted R^2 in the regression model reveals that 68.82% of the variation.

The estimate of the coefficient on Wickets and Average is 0.2010 and 2.1747 respectively.

P Value

The output in Appendix reveals, the p-value of Wickets and Average as $1.9788E^{-16}$ and 0.00034 respectively which is less than the level of significance (∝=0.05). Therefore, we can say that both the independent variable considered in the regression model are good predictors of (dependent variable).