Introduction

Major League Baseball is a professional baseball organization that runs from late March/early April to late September/early October, followed by the postseason, which can run to early November.[1] Each regular season consists of 30 teams and 162 games. For this project, we will use the dataset that consists of some important team statistics of 2008 season. We will try to analyze them by figuring out which factors influence number of wins and help a team to advance into playoffs. This paper will go into details of the dataset, statistical model used to analyze the data, and a discussion about test results. We will conclude by stating how game statistics can be used to influence coaches in preparation for next season.

Dataset Description

There are a total of 30 teams/observations and each observation has 6 variables (Runs, Hits, Walks, Errors, Saves, Wins) in **MLB2008** dataset. Below is the description of variables in the dataset:

Variable	Description
Runs	Total number of runs scored by each team in 2008 regular season
Hits	Total number of hits by all batters in a team (A hit, also called a base hit, is credited to a batter when the batter safely reaches first base after hitting the ball into fair territory, without the benefit of an error or a fielder's choice.[2])
Walks	Total number of walks for each team, given by pitchers or awarded to batters[3] (A walk occurs when a pitcher throws four pitches out of the strike zone, none of which are swung at by the hitter. After refraining from swinging at four pitches out of the zone, the batter is awarded first base.[4])
Errors	Total number of errors for each team (In baseball statistics, an error is an act, in the judgment of the official scorer, of a fielder misplaying a ball in a manner that allows a batter or baserunner to advance one or more bases or allows an at bat to continue after the batter should have been put out.[5])
Saves	Total number of saves for each team (A save is awarded to the relief pitcher who finishes a game for the winning team, under certain circumstances. A pitcher cannot receive a save and a win in the same game.[6])
Wins	Total number of wins for each team in 2008 regular season

Statistical Model for Experiment

One of the models we will use is *Multiple Linear Regression*. SAS Enterprise Guide will be used to derive an equation, which will give some insights on how the variables/predictors are related to number of wins (Y). These are the important information we will focus on:

Information	Definition
Squared Semi-partial Corr Type II	Describes the relationship between the predictor and Y; highest value means that predictor is the most related to Y and lowest value means it is least related

95% Confidence Limits	A 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population.[7] If this interval contains 0, it means there is no evidence that it will be a good predictor in the presence of other predictors.
Variance Inflation	It is related to multicollinearity, which is a problem created by the existence of substantial correlations among the set of predictors (Xs). If the value is greater than 10, we'll have a problem. For values less than 10, there is no multicollinearity problem.
Parameter Estimate	Slope for each variable
Adj. R-Sq	Adjusted R-Square value that specifies how good a model is based on its predictors; the higher the value is, the better that model is.
Root MSE	This value depicts how much the predictive model (value) differs from the actual model

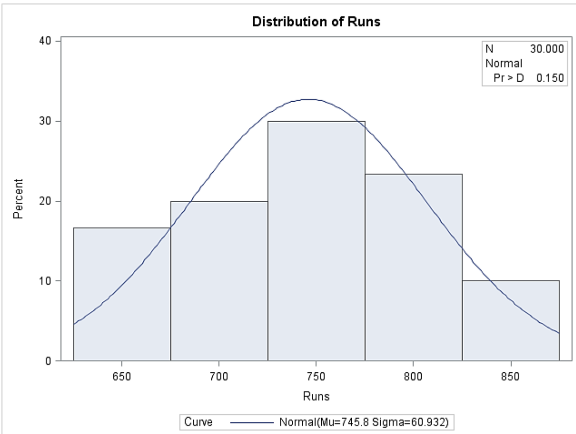
Other models we will use are T-Test and Hypothesis test. Both use hypothesis statements to determine the problem outcome. Then based on the p-value, they accept or reject a hypothesis by comparing it with α.

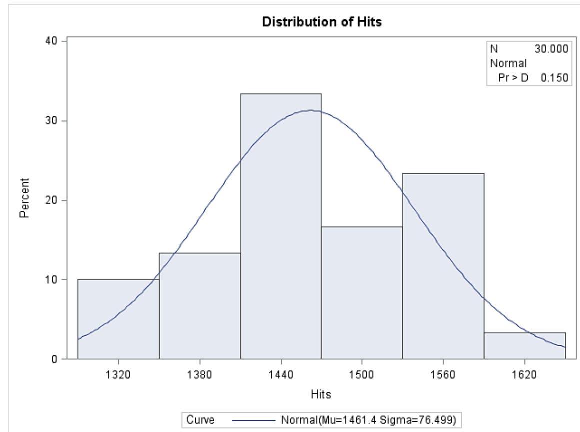
Variable	**Definition**
p-value	The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. P is also described in terms of rejecting H_0 when it is actually true, however, it is not a direct probability of this state.[8]
t-value	The t-value measures the size of the difference relative to the variation in sample data
Hypothesis statements	Statements that researchers are trying to answer through experiments to see which one is satisfied. There are two hypothesis, Null hypothesis and Alternative hypothesis.
α value (or α level)	Also known as significance level, it is the probability of rejecting the null hypothesis when the null hypothesis is true, i.e. it is the probability of making wrong decision[9]

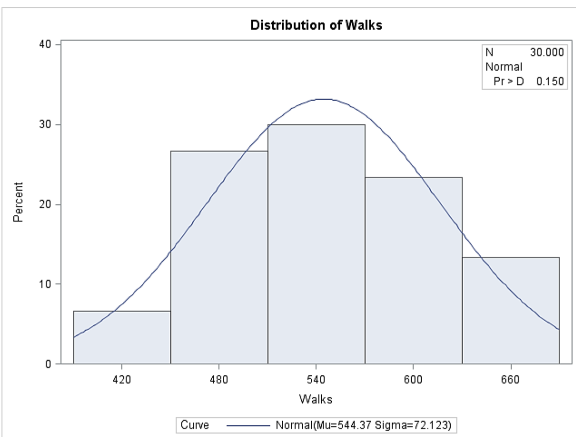
Analysis

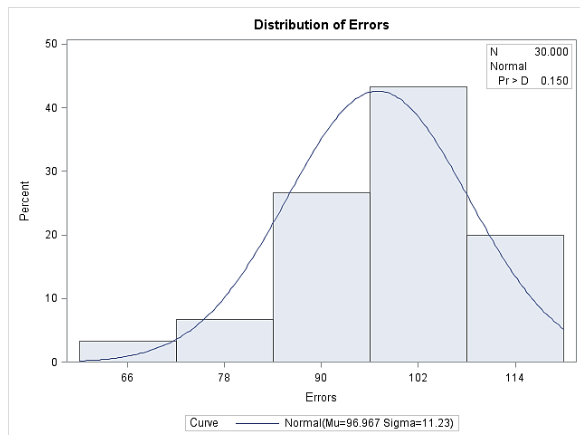
We will start our experiment by running tests on our dataset using SAS. To get a better understanding on how our data spreads out, we ran "PROC UNIVARIATE" procedure to get basic measures. We got an average of 745.8, 1461.433, 544.3667, 96.96667, 34.9 for Runs, Hits, Walks, Errors, and Saves, respectively.

Then we can test normality for each column. To test normality, we will compare Kolmogorov-Smirnov p-value with α value. If K-S > α, data can be assumed to normal enough. Else, data is not normal. Since we don’t have any specification for α, we will use the default value of 0.05.



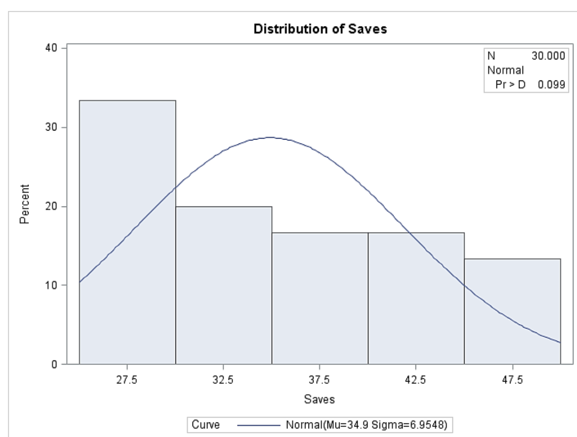






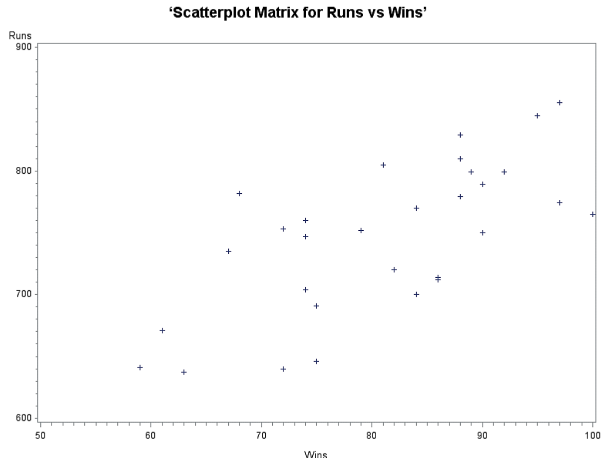
As we can see from the figures, the values are 0.15, 0.15, 0.15, 0.15, 0.099. All of them are greater than 0.05. So, we can say that they are normal.

Now we can check how the columns are related to number of Wins. To test this, we can run “PROC GPLOT” procedure on SAS and compare each variable with number of wins. Let’s test the relationship between Runs scored and numbers of Wins for 2008 season.



We can see from the scatter plot that number of wins increase with runs. There appears to be a linear relation between the two and this relation can be referred to as positive association. We can state similar conclusions(i) for based on the scatter plots.

Further analysis can be done by dividing dataset into two groups based on winning percentages, high winning teams (winners) and low winning teams (losers). By running a quick "PROC UNIVARIATE" on data, we found that the median for number of Wins is 83. Using this information, one group can be formed with teams that have wins less than or equal to 83. We will create a new dataset called BaseballWins with a new variable, WinningTeam. Teams with less or equal to 83 wins will have "N" for WinningTeam and others will have "Y".



We will setup our hypothesis for Errors variable first.

Null Hypothesis: There is no difference between the average errors occurred by winners and losers

Alternative Hypothesis: There is a difference between the average errors occurred by winners and losers

We will use the default value for α (0.05).

Method	**Variances**	**DF**	**t Value**	**Pr > |t|**
Pooled	Equal	28	2.16	0.0398
Satterthwaite	Unequal	23.031	2.16	0.0417
Cochran	Unequal	14	2.16	0.0489

After running the T-Test(v) on BaseballWins, we got a p-value (Pooled Method with Equal Variances) of 0.0398, which is greater than 0.05. As a result, we can reject the null hypothesis. This means that there exists difference between the average errors of winners and losers.

We can run the same test based on averages of Saves(vi), Runs(vii), Hits(viii), and Walks(ix). If we setup our null and alternative hypothesis in a similar way, we will be able to reject null hypothesis for all variables based on p-value and comparing it with alpha level.

Now, we will use SAS Enterprise Guide to run Multiple Linear Regression on the dataset. Wins is our Dependent variable (Y) and Runs, Hits, Walks, Errors, and Saves are our Explanatory variables (X). Let's run SAS EG to the regression equation from **<u>Parameter Estimates</u>**: $Y = 64.535 + 0.056 * Runs - 0.027 * Hits - 0.035 * Walks + 0.038 * Errors + 0.862 * Saves$

To elaborate this equation, we can say the following:

- Each team will have at least 64.535 wins if they don't have any runs, hits, walks, errors, and saves

- If everything else stays constant except Runs, number of wins will improve by 0.056 for each additional run. Similarly, wins will improve by 0.038, 0.862 for additional Error and Save, consequently.
- If everything else stays constant except Hits, number of wins will decrease by 0.027 for each additional hit. Similarly, winning rate will decrease by 0.035 for additional Walk

From the second table, we can see that this model has an adjusted R-Square value of ~0.9 (close to 1). From a general observation, we can say that this is a good model (even though it can still be improved). Root MSE value of 3.545 tells us that this model will miss the actual prediction by that much. These values suggest that this a good model and we can do further analysis on the dataset.

Conclusion

This dataset provided us with some useful information that helped us determine which teams have higher chance to advance into playoff. Our T-Test analysis showed that Runs, Errors, and Saves improve winning rates. But for teams with high winning rates, Errors wasn't an influencing factor. Saves was the most contributing factor (0.862 increase in wins for each save) based on the regression equation. Based on the analysis, we can suggest which predictors will influence winning rates and help them to perform better.

Citations

1. https://en.wikipedia.org/wiki/Major_League_Baseball_schedule
2. https://en.wikipedia.org/wiki/Hit_(baseball)
3. http://m.mlb.com/glossary/standard-stats/walk
4. https://en.wikipedia.org/wiki/Base_on_balls
5. https://en.wikipedia.org/wiki/Error_(baseball)
6. http://m.mlb.com/glossary/standard-stats/save
7. https://www.graphpad.com/guides/prism/7/statistics/stat_more_about_confidence_interval.htm?toc=0&printWindow
8. https://www.statsdirect.com/help/basics/p_values.htm
9. http://blog.minitab.com/blog/michelle-paret/alphas-p-values-confidence-intervals-oh-my

Relevant Bibliography

Shmueli, Patel, and Bruce, 2010. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, 2010.

Freund, Mohr, Wilson, 2010. Statistical Methods, 2010.

Appendix

This part includes all the codes and extra materials used for this assignment.

Dataset:

2008 Team	Runs	Hits	Walks	Errors	Saves	Wins
Arizona	720	1403	451	113	33	82
Atlanta	753	1439	586	107	28	72
Baltimore	782	1538	687	100	29	68
Boston	845	1369	548	85	47	95
Chicago Cubs	855	1329	548	99	44	97
Chicago White Sox	810	1469	457	108	33	88
Cincinnati	704	1542	557	114	31	74
Cleveland	805	1530	444	94	31	81
Colorado	747	1547	562	96	28	74
Detroit	760	1541	644	113	27	74
Florida	770	1421	586	117	36	84
Houston	712	1453	492	67	38	86
Kansas City	691	1473	515	96	29	75
California Angels	765	1455	457	91	47	100
Los Angeles Dodgers	700	1381	480	101	35	84
Milwaukee	750	1415	528	101	45	90
Minnesota	829	1563	403	108	37	88
New York Mets	799	1415	590	83	43	89
New York Yankees	789	1478	489	83	39	90
Oakland	646	1364	576	98	28	75
Philadelphia	799	1444	533	90	47	92
Pittsburgh	735	1631	657	107	27	67
San Diego	637	1466	561	89	28	63
Seattle	671	1544	626	99	26	61
San Francisco	640	1416	652	96	30	72
St. Louis	779	1517	496	85	41	88
Tampa Bay	774	1349	526	90	40	97
Texas	752	1525	625	99	33	79
Toronto	714	1330	467	84	40	86
Washington	641	1496	588	96	27	59

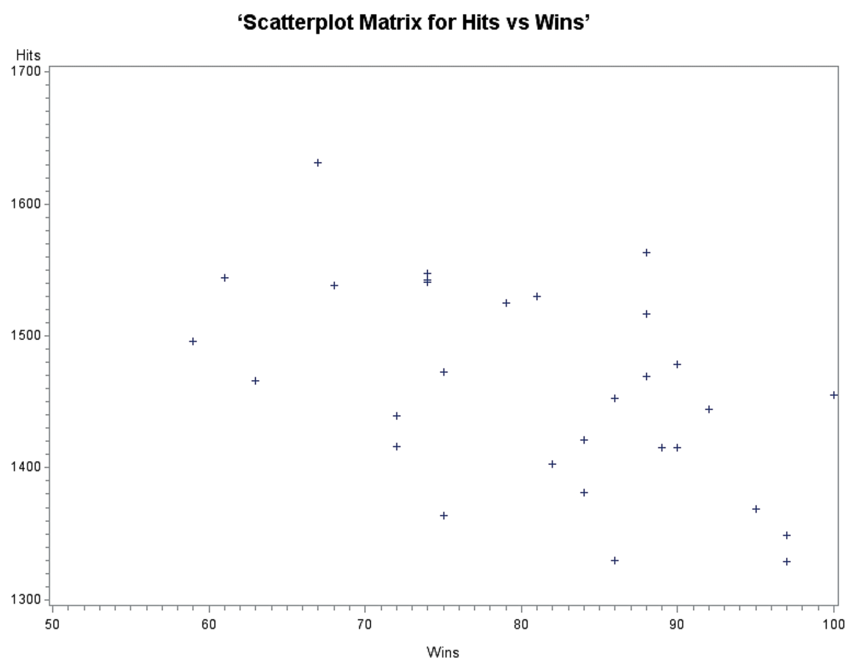
Parameter estimates from SAS EG:

Parameter Estimates						
Variable	DF	Parameter Estimate	Squared Semi-partial Corr Type II	Variance Inflation	95% Confidence Limits	
Intercept	1	64.53528	.	0	24.30142	104.76913
Runs	1	0.05574	0.04421	2.10780	0.02336	0.08812
Hits	1	-0.02739	0.02193	1.61672	-0.04998	-0.00480
Walks	1	-0.03468	0.04036	1.25219	-0.05576	-0.01360
Errors	1	0.03818	0.00105	1.41326	-0.10568	0.18203
Saves	1	0.86235	0.08455	3.43636	0.50014	1.22455

Root MSE	3.54564	R-Square	0.9160
Dependent Mean	81.00000	Adj R-Sq	0.8984
Coeff Var	4.37734		

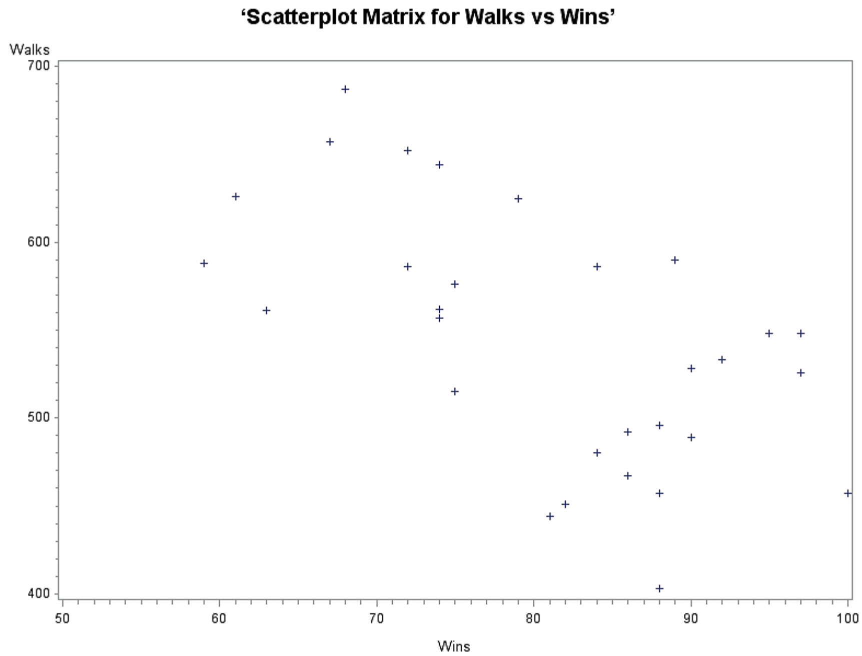
Paired T-Test for whole dataset:

i) Relationship between Hits and Wins



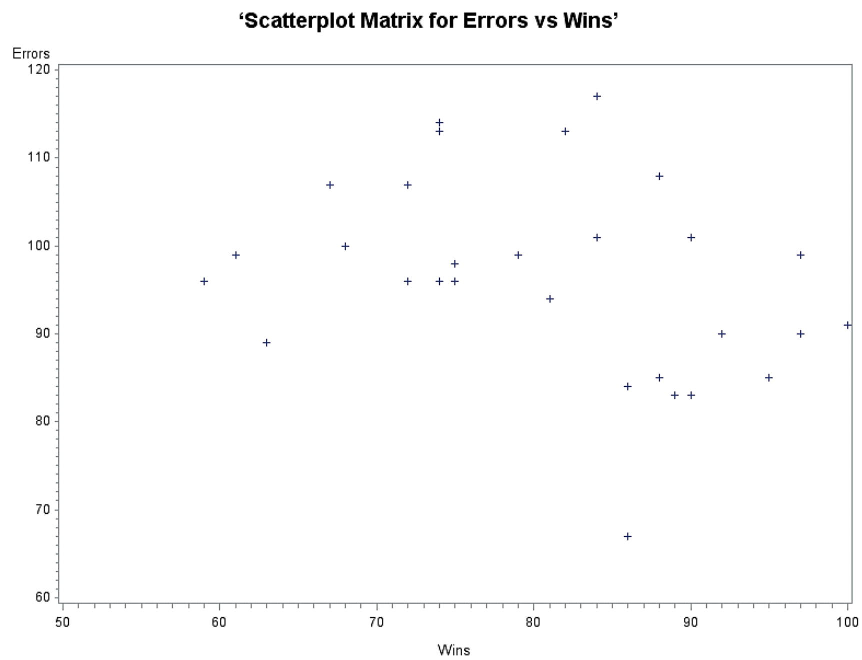
This looks to have a negative association between the variables even though it is not perfectly linear.

ii) Relationship between Walks and Wins



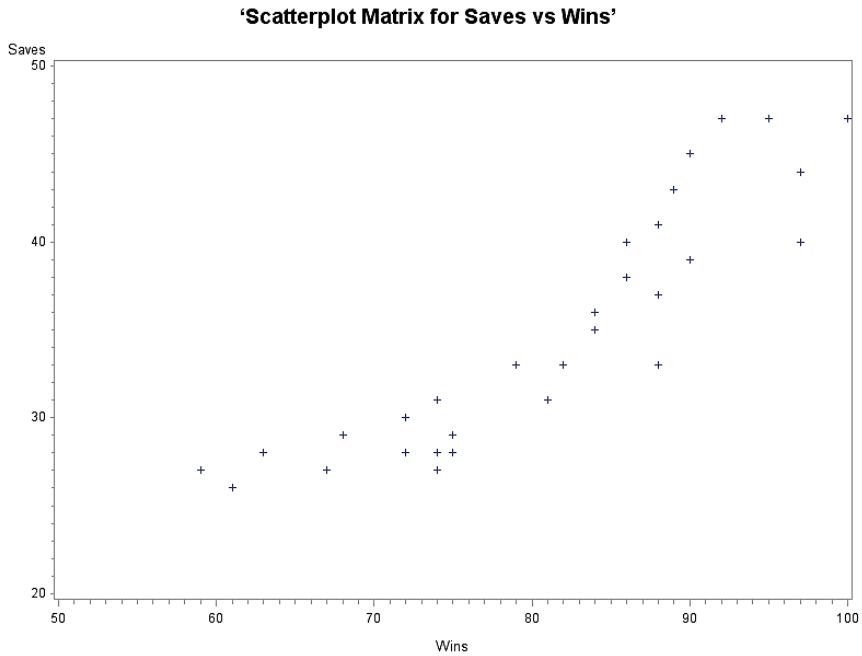
There looks to have a negative association between Walks and Wins.

iii) Relationship between Errors and Wins



There isn't any positive/negative association nor any linear relation between the variables.

iv) Relationship between Saves and Wins



There seems to have somewhat strong positive linear association between them.

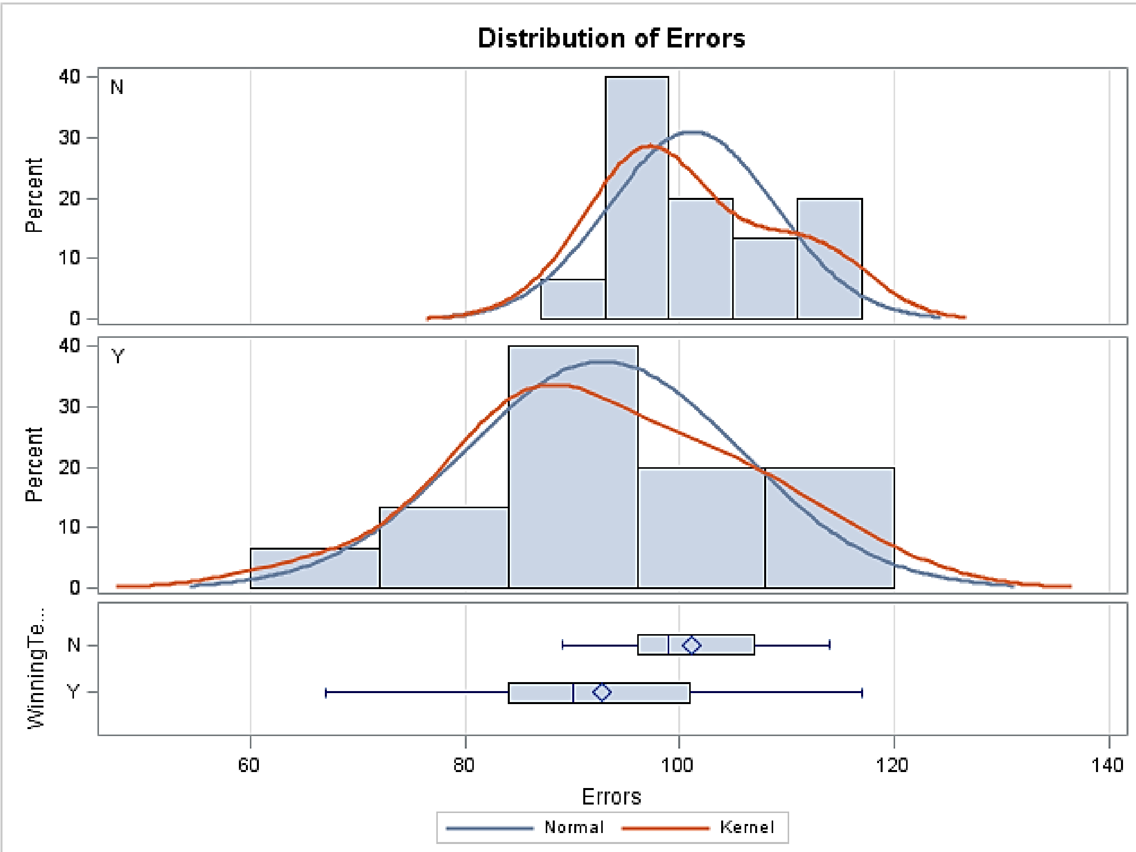
<u>T-Test for teams with low and high wins:</u>

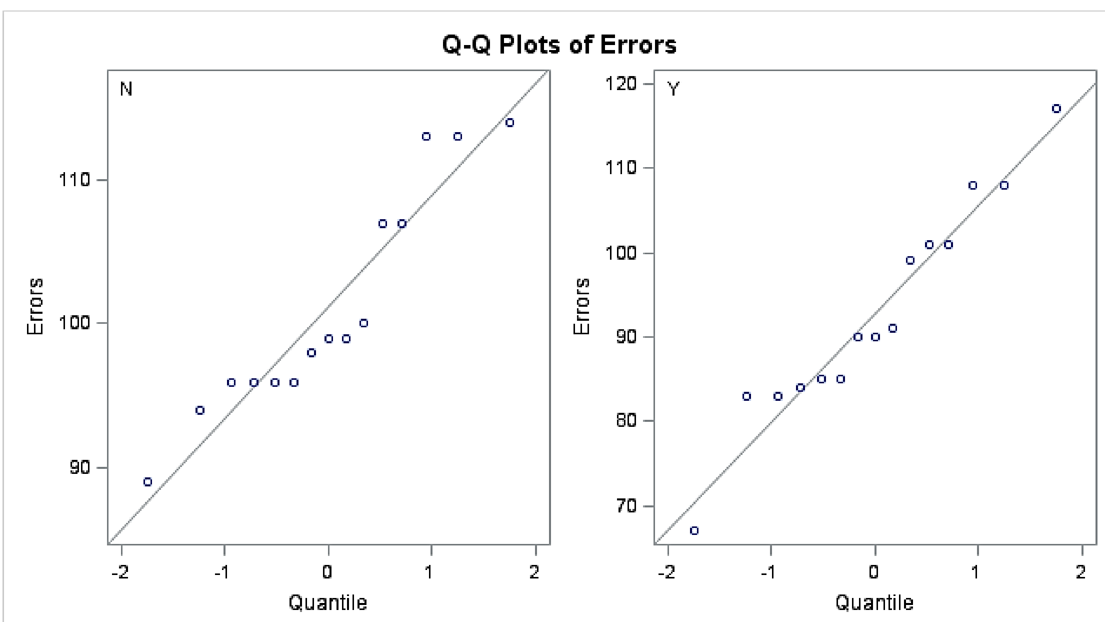
v) T-test for Errors between two groups

WinningTeam	N	Mean	Std Dev	Std Err	Minimum	Maximum
N	15	101.1	7.7447	1.9997	89.0000	114.0
Y	15	92.8000	12.8074	3.3068	67.0000	117.0
Diff (1-2)		8.3333	10.5832	3.8644		

Winning Team	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
N		101.1	96.8444	105.4	7.7447	5.6701	12.2142	5.5499	11.8822
Y		92.8000	85.7075	99.8925	12.8074	9.3766	20.1985	9.1778	19.6494
Diff (1-2)	**Pooled**	8.3333	0.4174	16.2493	10.5832	8.3986	14.3133	8.3066	14.1264
Diff (1-2)	**Satterthwaite**	8.3333	0.3397	16.3270					

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	2.73	0.0699





vi) T-test for Saves based on two groups

Our hypothesis statements for this experiment are:

Null Hypothesis: There is no difference in the average saves between winners and losers

Alternative Hypothesis: There is a difference in the average saves between winners and losers

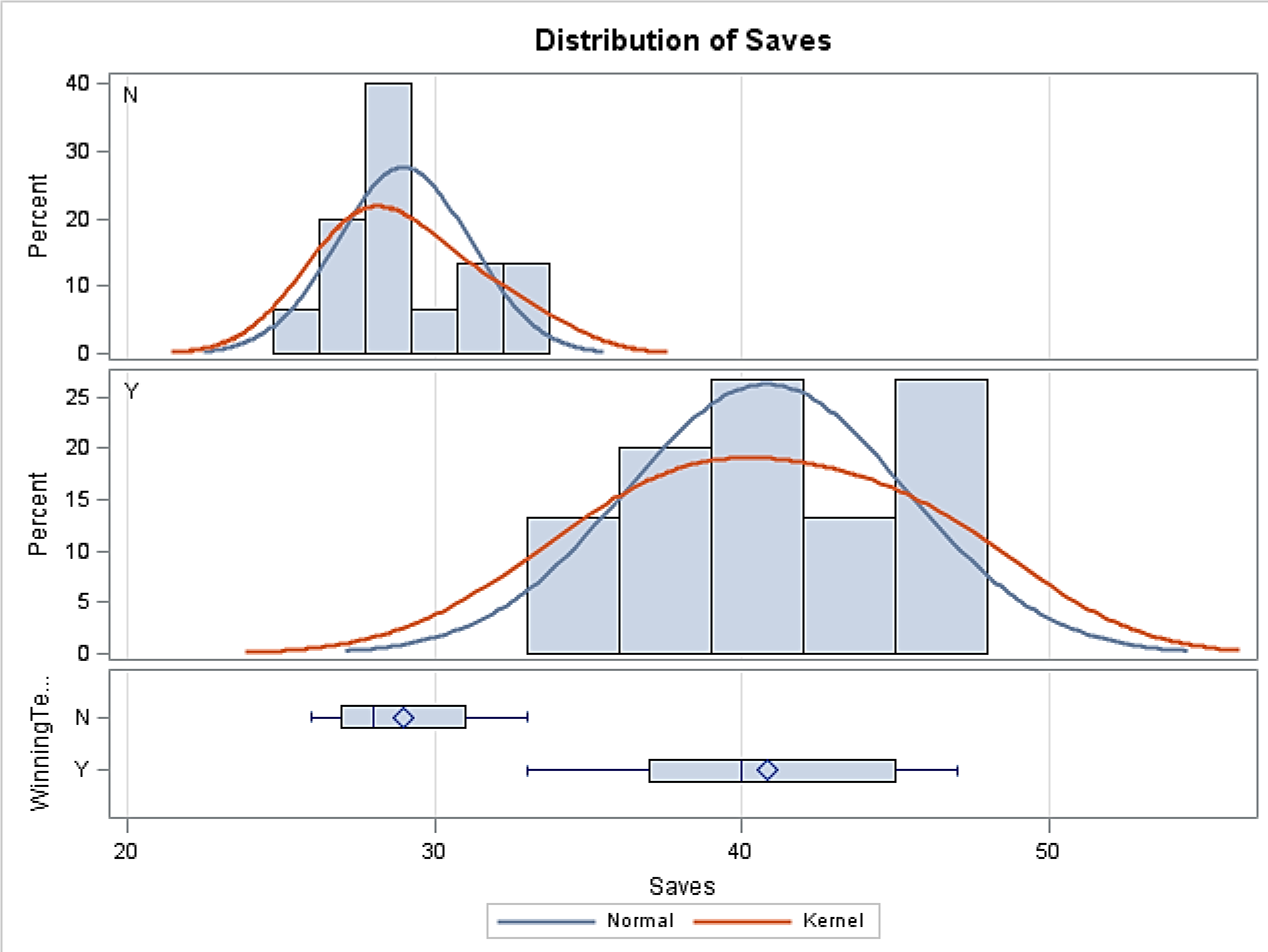
We will use the default value for α (0.05).

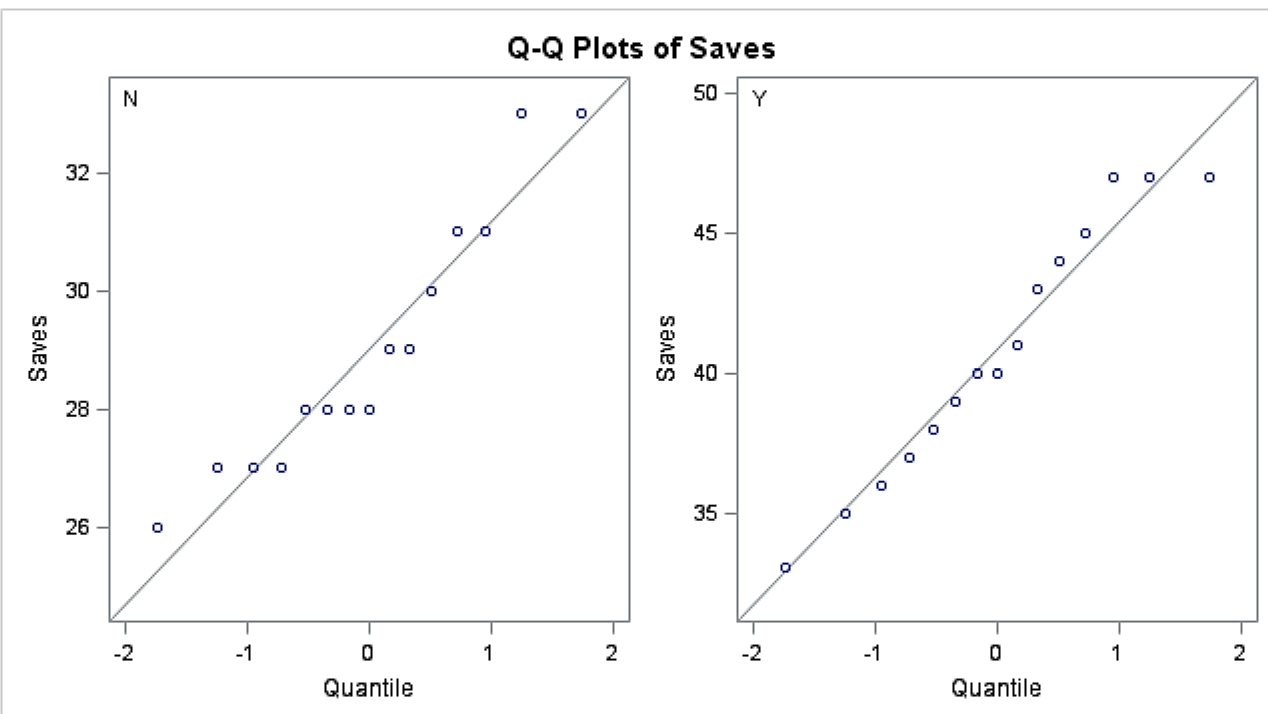
WinningTeam	N	Mean	Std Dev	Std Err	Minimum	Maximum
N	15	29.0000	2.1712	0.5606	26.0000	33.0000
Y	15	40.8000	4.5701	1.1800	33.0000	47.0000
Diff (1-2)		-11.8000	3.5777	1.3064		

WinningTeam	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
N		29.0000	27.7976	30.2024	2.1712	1.5896	3.4243	1.5559	3.3312
Y		40.8000	38.2692	43.3308	4.5701	3.3459	7.2075	3.2749	7.0115
Diff (1-2)	**Pooled**	-11.8000	-14.4760	-9.1240	3.5777	2.8392	4.8387	2.8081	4.7755
Diff (1-2)	**Satterthwaite**	-11.8000	-14.5250	-9.0750					

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-9.03	<.0001
Satterthwaite	Unequal	20.014	-9.03	<.0001
Cochran	Unequal	14	-9.03	<.0001

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	4.43	0.0087


Distribution of Saves
N
Y
Percent
WinningTe...
Saves
Normal
Kernel


Q-Q Plots of Saves
N
Y
Saves
Quantile

After running the T-Test on BaseballWins, we got a p-value (Pooled Method Equal Variances) of <0.0001, which is less than 0.05. As a result, we can reject the null hypothesis. This means that the relationship between winners and losers based on average saves is statistically significant, i.e. the average save by low winning teams are different than the average saves of high winning teams.

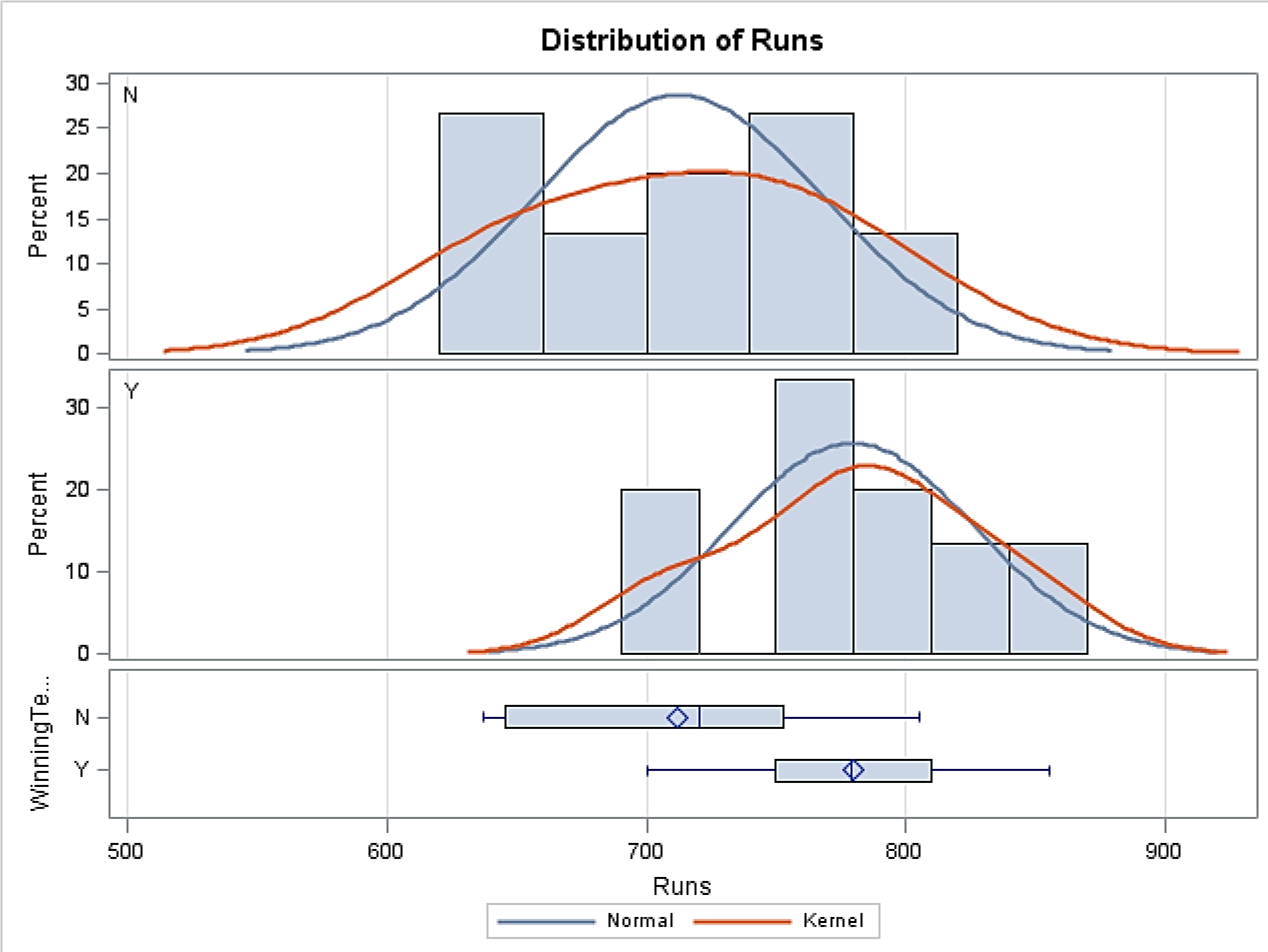
vii) T-test for average Runs between two groups

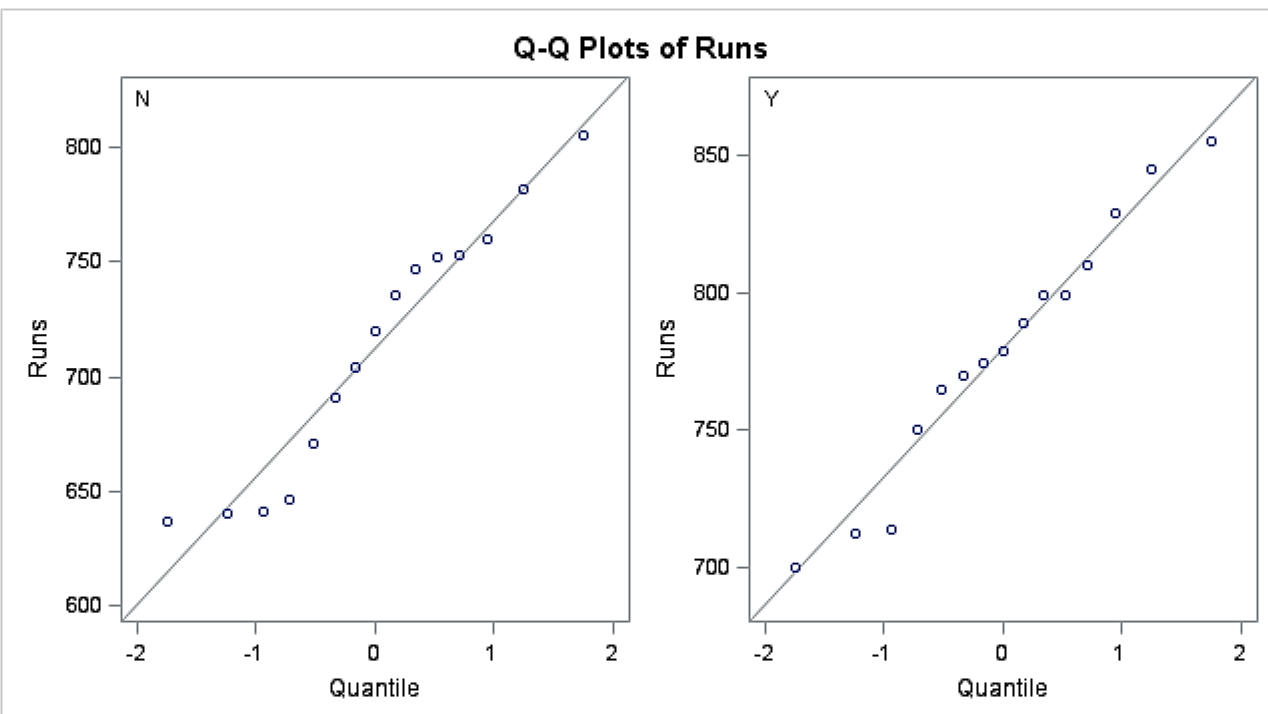
WinningTeam	N	Mean	Std Dev	Std Err	Minimum	Maximum
N	15	712.3	55.6410	14.3665	637.0	805.0
Y	15	779.3	46.7435	12.0691	700.0	855.0
Diff (1-2)		-67.0667	51.3852	18.7632		

WinningTeam	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
N		712.3	681.5	743.1	55.6410	40.7363	87.7514	39.8724	85.3657
Y		779.3	753.4	805.2	46.7435	34.2221	73.7191	33.4964	71.7149
Diff (1-2)	**Pooled**	-67.0667	-105.5	-28.6320	51.3852	40.7782	69.4960	40.3314	68.5887
Diff (1-2)	**Satterthwaite**	-67.0667	-105.6	-28.5804					

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-3.57	0.0013
Satterthwaite	Unequal	27.191	-3.57	0.0013
Cochran	Unequal	14	-3.57	0.0030

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.42	0.5229


Distribution of Runs
N
Y
Percent
WinningTe...
Runs
Normal
Kernel


Q-Q Plots of Runs
N
Y
Runs
Quantile

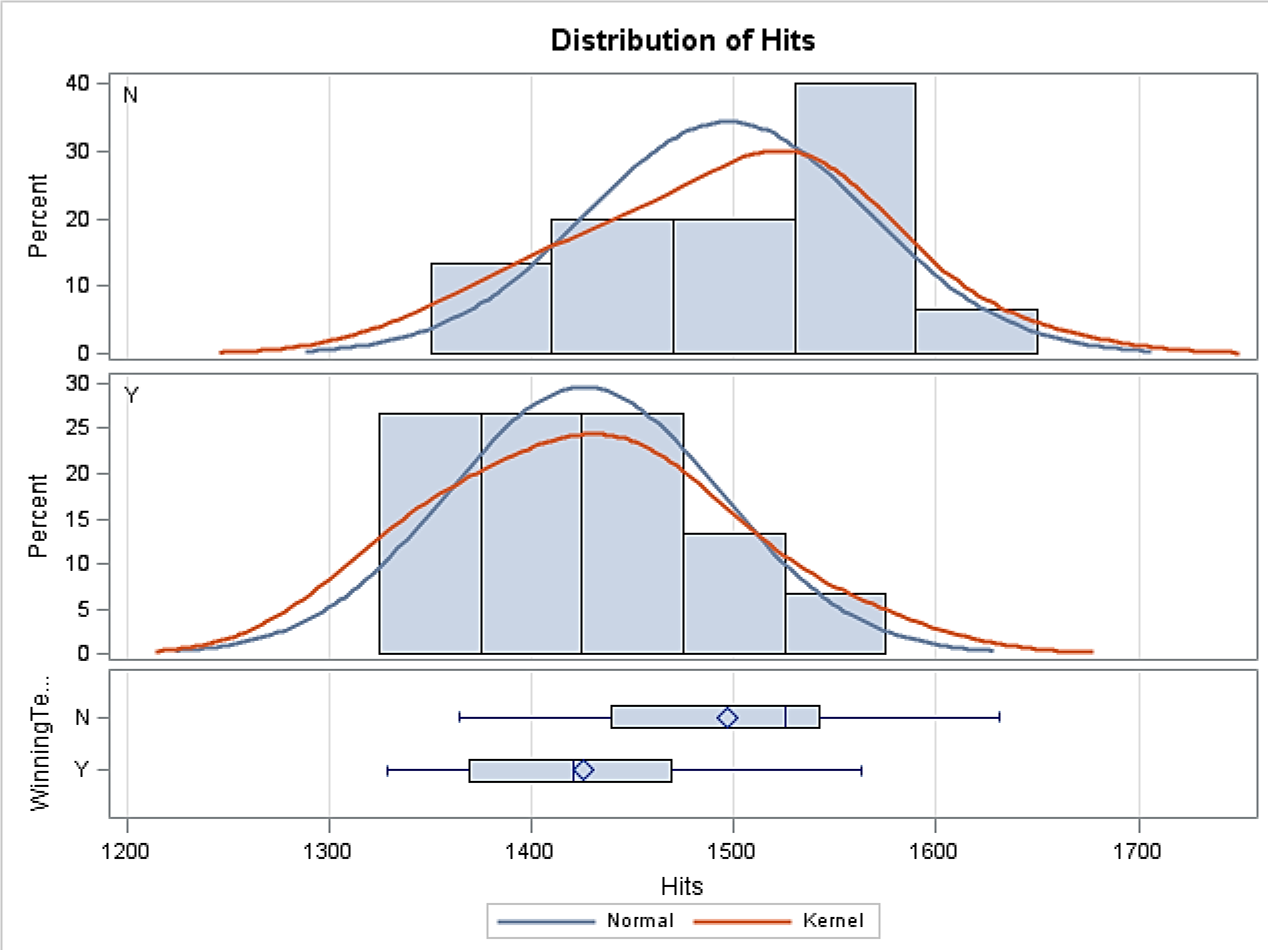
viii) T-test based on average Hits between two groups

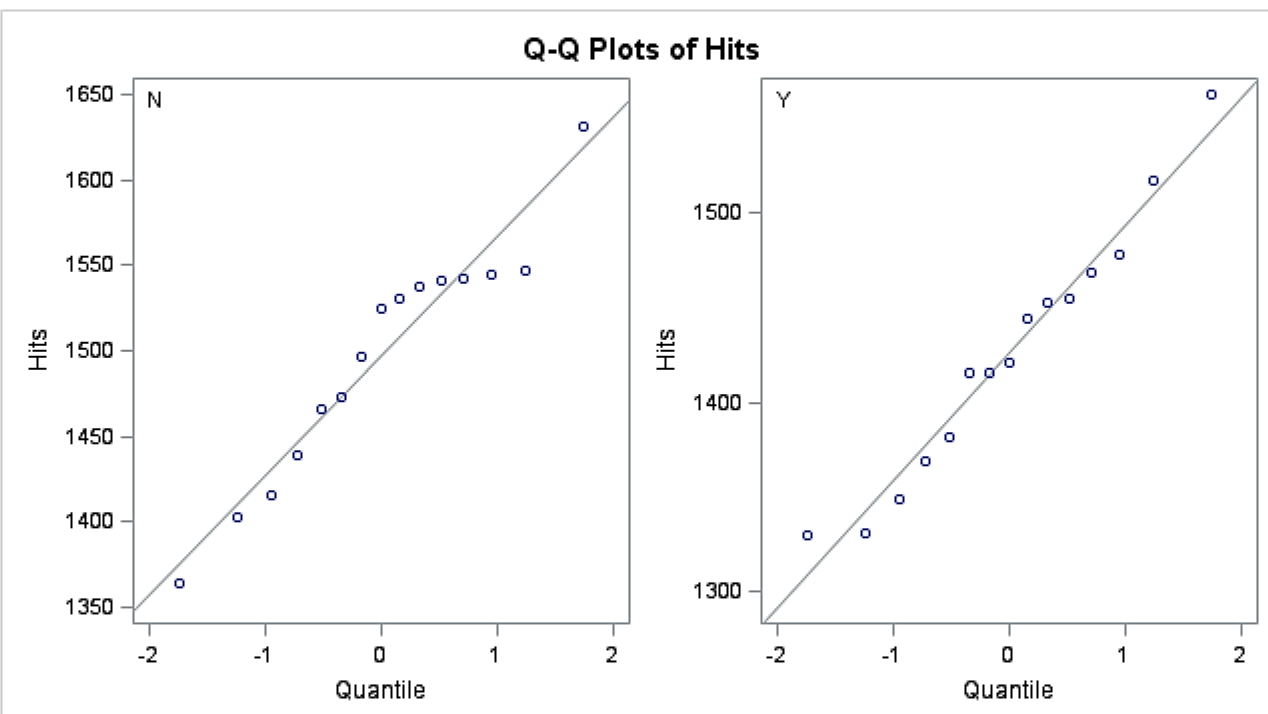
WinningTeam	N	Mean	Std Dev	Std Err	Minimum	Maximum
N	15	1497.0	69.7792	18.0169	1364.0	1631.0
Y	15	1425.9	67.3963	17.4017	1329.0	1563.0
Diff (1-2)		71.1333	68.5981	25.0485		

WinningTeam	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
N		1497.0	1458.4	1535.6	69.7792	51.0872	110.0	50.0038	107.1
Y		1425.9	1388.5	1463.2	67.3963	49.3427	106.3	48.2962	103.4
Diff (1-2)	**Pooled**	71.1333	19.8238	122.4	68.5981	54.4381	92.7757	53.8416	91.5645
Diff (1-2)	**Satterthwaite**	71.1333	19.8210	122.4					

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	2.84	0.0083
Satterthwaite	Unequal	27.966	2.84	0.0083
Cochran	Unequal	14	2.84	0.0131

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.07	0.8984


Distribution of Hits
Percent
N
Y
WinningTe...
Hits
1200
1300
1400
1500
1600
1700
Normal
Kernel


Q-Q Plots of Hits
N
Y
Hits
Quantile

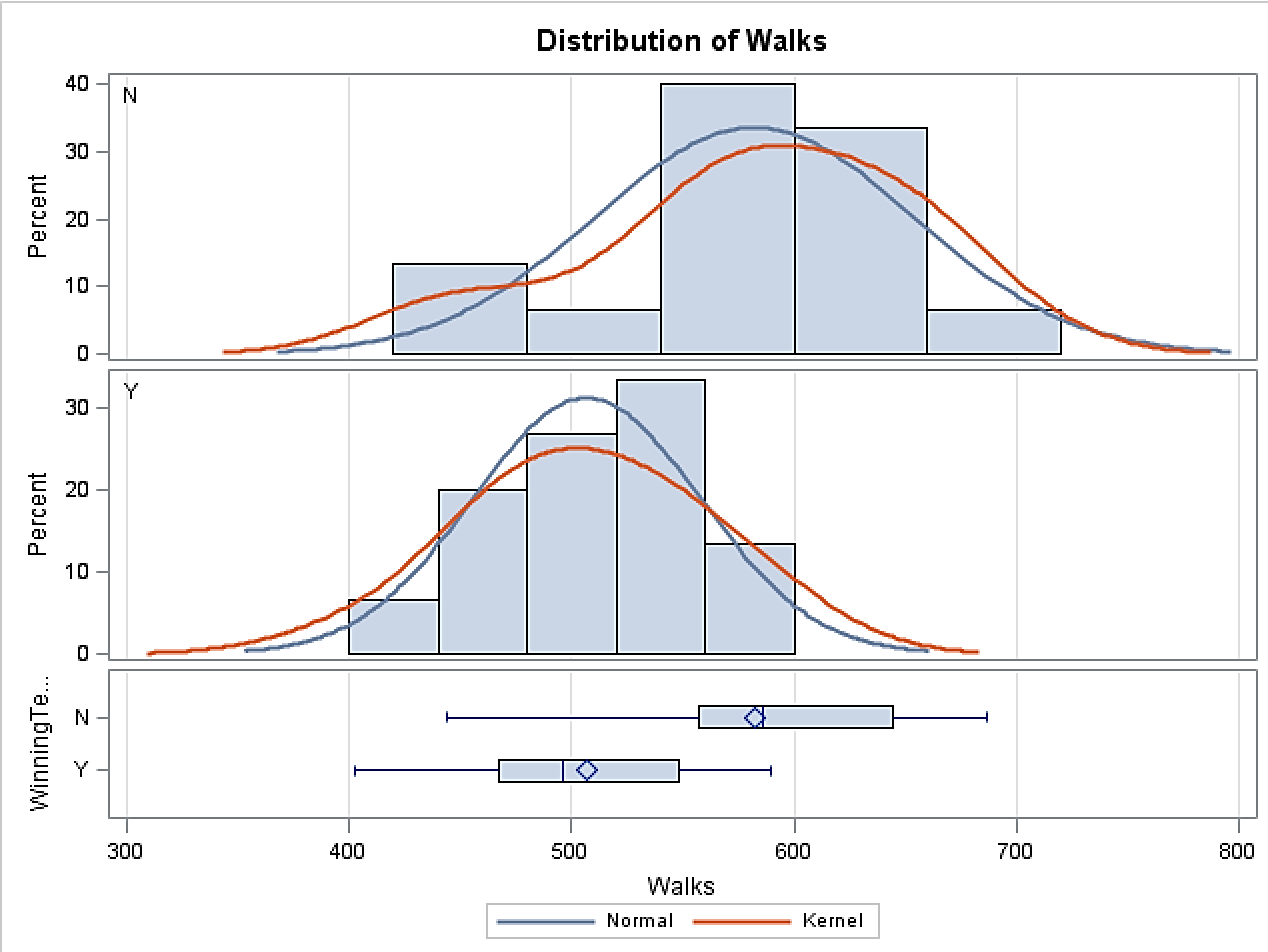
ix) T-test based on average Walks between two groups

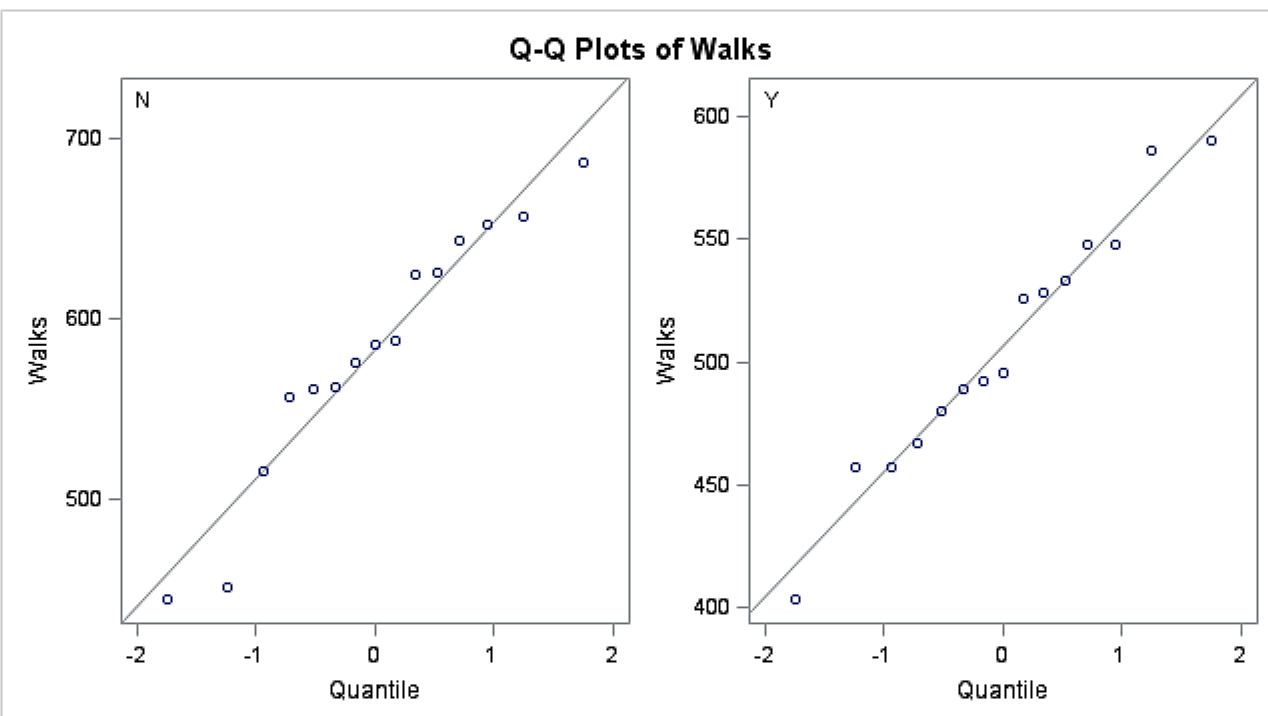
WinningTeam	N	Mean	Std Dev	Std Err	Minimum	Maximum
N	15	582.1	71.4278	18.4426	444.0	687.0
Y	15	506.7	51.2580	13.2348	403.0	590.0
Diff (1-2)		75.4000	62.1663	22.6999		

WinningTeam	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev		95% UMPU CL Std Dev	
N		582.1	542.5	621.6	71.4278	52.2942	112.6	51.1851	109.6
Y		506.7	478.3	535.1	51.2580	37.5273	80.8389	36.7315	78.6412
Diff (1-2)	**Pooled**	75.4000	28.9013	121.9	62.1663	49.3339	84.0770	48.7934	82.9793
Diff (1-2)	**Satterthwaite**	75.4000	28.6856	122.1					

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	3.32	0.0025
Satterthwaite	Unequal	25.397	3.32	0.0027
Cochran	Unequal	14	3.32	0.0050

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.94	0.2267


Distribution of Walks
N
Y
Percent
Percent
WinningTe...
Walks
Normal
Kernel


Q-Q Plots of Walks
N
Y
Walks
Walks
Quantile
Quantile

Codes used:

```
PROC IMPORT OUT= Baseball
            DATAFILE= "C:\Users\akhan12\Desktop\BB2008.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

PROC PRINT DATA=Baseball;
RUN;


*******************************************************************
* analyze each column;
*ODS select BasicMeasures;
PROC UNIVARIATE DATA=Baseball;
VAR Runs;
RUN;

*ODS select BasicMeasures;
PROC UNIVARIATE DATA=Baseball;
VAR Hits;
RUN;

*ODS select BasicMeasures;
PROC UNIVARIATE DATA=Baseball;
VAR Walks;
RUN;

*ODS select BasicMeasures;
PROC UNIVARIATE DATA=Baseball;
VAR Errors;
RUN;

*ODS select BasicMeasures;
PROC UNIVARIATE DATA=Baseball;
VAR Saves;
RUN;


******************************************************************
* check and see if they fit the normal distribution;
TITLE "How Normal is the Runs Histogram?";
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Baseball;
HISTOGRAM Runs/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;


TITLE "How Normal is the Hits Histogram?";
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Baseball;
HISTOGRAM Hits/ normal(percents=20 40 60 80 midpercents);
```

```
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;


TITLE "How Normal is the Walks Histogram?";
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Baseball;
HISTOGRAM Walks/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;



TITLE "How Normal is the Errors Histogram?";
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Baseball;
HISTOGRAM Errors/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;


TITLE "How Normal is the Saves Histogram?";
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Baseball;
HISTOGRAM Saves/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;


*********************************************************
* is there any relation between wins and other columns;
TITLE 'Scatterplot Matrix for Runs vs Wins';
PROC GPLOT DATA = Baseball;
      PLOT Runs * Wins;
RUN;


TITLE 'Scatterplot Matrix for Hits vs Wins';
PROC GPLOT DATA = Baseball;
      PLOT Hits * Wins;
RUN;


TITLE 'Scatterplot Matrix for Walks vs Wins';
PROC GPLOT DATA = Baseball;
      PLOT Walks * Wins;
RUN;


TITLE 'Scatterplot Matrix for Errors vs Wins';
PROC GPLOT DATA = Baseball;
      PLOT Errors * Wins;
RUN;


TITLE 'Scatterplot Matrix for Saves vs Wins';
PROC GPLOT DATA = Baseball;
```

```
      PLOT Saves * Wins;
RUN;


*********************************************************************
* divide the teams in two groups, higher wins and lower wins.
then check if two groups have the same error percentage;

****USED TO CHECK THE MIDDLE POINT****;
PROC UNIVARIATE DATA=Baseball;
VAR Wins;
RUN;
**************************************;

/*
WE WILL CREATE TWO DATASETS BASED ON NUMBER OF WINS. ONE GROUP, BASEBALLLOWER
WILL
HAVE NUMBER OF WINS <= 83 (83 IS THE MEDIAN), AND OTHER GROUP, BASEBALLHIGHER
WILL
HAVE MORE THAN 83 WINS.
*/
DATA BaseballWins;
SET Baseball;
IF (Wins <= 83)
      THEN WinningTeam = 'N';  *TEAMS CATAGORIZED AS LOSERS;
ELSE WinningTeam = 'Y';  *TEAMS CATAGORIZED AS WINNERS;
RUN;


PROC SORT DATA=BaseballWins;
BY WinningTeam;
RUN;


PROC PRINT DATA=BaseballWins;
RUN;

*COMPARISON FOR ERRORS;
TITLE "T-test comparison between Errors for different groups";
ods graphics on;
PROC TTEST Data=BaseballWins cochran ci=equal umpu;
class WinningTeam;
var Errors;
RUN;


*COMPARISON FOR SAVES;
TITLE "T-test comparison between Errors for different groups";
ods graphics on;
PROC TTEST Data=BaseballWins cochran ci=equal umpu;
class WinningTeam;
var Saves;
RUN;


*COMPARISON FOR RUNS;
TITLE "T-test comparison between Errors for different groups";
```

```
ods graphics on;
PROC TTEST Data=BaseballWins cochran ci=equal umpu;
class WinningTeam;
var Runs;
RUN;


*COMPARISON FOR HITS;
TITLE "T-test comparison between Errors for different groups";
ods graphics on;
PROC TTEST Data=BaseballWins cochran ci=equal umpu;
class WinningTeam;
var Hits;
RUN;


*COMPARISON FOR WALKS;
TITLE "T-test comparison between Errors for different groups";
ods graphics on;
PROC TTEST Data=BaseballWins cochran ci=equal umpu;
class WinningTeam;
var Walks;
RUN;
```