Statistics Report

**Project: Statistical Analysis of T20WC (2007-2021)**

Contributors:

1) Joshasree Maturi (19bcs069)

2) S. Pranay Sai Teja (19bcs102)

3) Mohammed Abdul Sohail (19bcs122)

4) Munagapati Anupama (19bcs123)

5) Rayudu Mounika (19bcs124)

Abstract:

All the Statistical Tests taught to us such as Chi-Square, T-Test, Wilcoxon Test, Estimation of Sample Size, etc. have been performed on the Dataset “T20WC (2007-2021)” using “IBM SPSS Statistics 26”. Each Test has been segregated in their respective UNIT Folder and each Test Folder contains:

A) Dataset

B) SPSS Output

C) Description/Report

The Report for each Test can be viewed in its Individual Folder as well as in the present Overall Report.

Index

UNIT-2:

1. Z-Test

Z-test for One sample

2. T-Test

T Test for One Sample

T Test for Two Independent Samples

T Test for Two Dependent Samples (Paired Samples)

UNIT-3:

1. Run Test

2. Sign Test - One sample

3. Sign Test - Paired sample

4. The Wilcoxon Rank Sum Test (or) Mann-Whitney U test

5. The Wilcoxon Signed-rank Test

UNIT-4:

Chi Square Test

1. Test for Goodness

2. Test for Independence

3. Test for homogeneity

4. Test for 2X2 Contingency Table

UNIT-5:

1. Type-II error

2. estimation of sample size

I) One sample (one mean and one proportions)

II) Two samples (two mean and two proportions)

Descriptions:

UNIT-2:

1. Z-Test

***Z-test for One Sample***

Procedure:

1. State the claim mathematically and verbally. Identify the null and alternative hypotheses.

2. Specify the level of significance.

3. Sketch the sampling distribution.

4. Determine the critical value(s).

5. Determine the rejection regions(s).

6. Find the standardized test statistic.

7. Make a decision to reject or fail to reject the null hypothesis.

8. Interpret the decision in the context of the original claim.

Question:

Among the Top Batsmen in T20WC (2007-2021), in a Sample of 30 Batsmen, the average matches played by an Individual is equal to the Average matches played by all 49 Batsmen in the Population. Is there enough evidence to support this claim at α = 0.05?

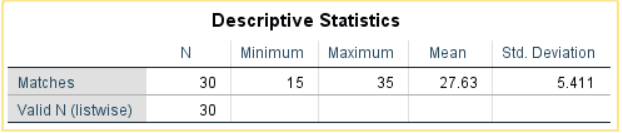
Answer:

Ho (Null Hypothesis): Sample mean population (Matches played by top 30 batsman) is equal to the Population mean population (Matches played by all 49 Batsman)

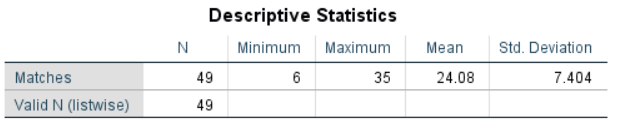
H1: Matches played is different

Output:

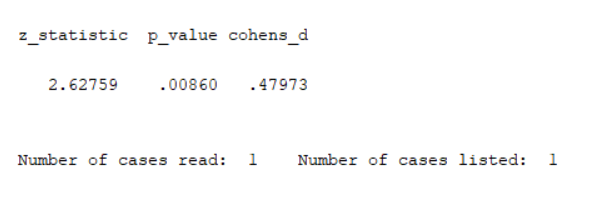
Sample Data (Matches):



Population Data (Matches):



Z-test (1 Sample)



Conclusion:

At α = 0.05, the Critical Value in One-Sample Z-Test is 1.645. (Critical)

From the above Z-Test conducted using SPSS, the obtained/calculated value is 2.627. (Calculated)

As Calculated value > Critical Value, at 5% Level of Significance, there’s enough evidence to reject the claim that in a Sample of 30 Batsmen, the average matches played by an Individual is equal to the Average matches played by all 49 Batsmen in the Population.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2. T-Test

***T-test for One Sample***

Procedure:

1. State the claim mathematically and verbally. Identify the null and alternative hypotheses.

2. Specify the level of significance.

3. Identify the degrees of freedom and sketch the sampling distribution.

4. Determine any critical values.

5. Determine any rejection region(s).

Question:

Among the Top Batsmen in T20WC (2007-2021), the average matches played by an individual is 30. In a sample of 10 Batsmen, the sample mean was 30.9 and Standard Deviation was 3.872. Is there enough evidence to support this claim at α = 0.05?

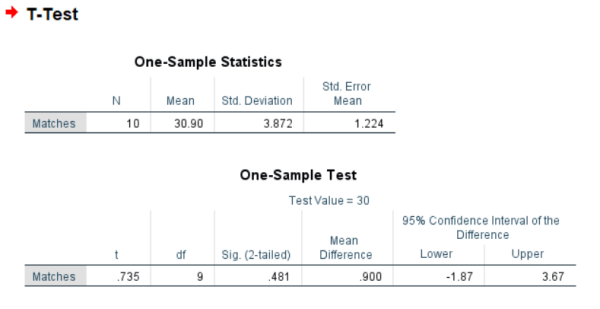
Answer:

Ho (Null Hypothesis): µ = 30 (Claim)

Ha (Alternate Hypothesis): µ ≠ 30

Output:

T-Test (1 Sample)



Conclusion:

At α = 0.05, the Critical Value in One-Sample T-Test with Degrees of Freedom 9 is 1.833. (Critical)

From the above T-Test conducted using SPSS, the obtained/calculated value is 0.481. (Calculated)

As Calculated value < Critical Value, at 5% Level of Significance, there’s not enough evidence to reject the claim that Average matches played by an individual is 30.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***T-test for Two Independent Samples***

Procedure:

1) State the research question.

2) State the statistical hypotheses.

Ho: µ1 ≥ µ2

Ha: µ1 < µ2

3) Set the decision rule.

α = 0.5

Df [Degrees of Freedom] = (n1 - 1) + (n2 - 1)

T(critical)

4) Calculate the test statistic.

5) Decide if the result is significant.

6) Interpret the obtained results.

Question:

Among the Top Batsmen in T20WC (2007-2021), in a Sample of 10 Batsmen, the Strike Rate of Players who have played greater than or equal to 32 matches is greater than or equal as the one’s played lesser than 32 matches. At α = 0.05, is there enough evidence to support this claim?

Answer:

µ1: The Strike Rate of Players who have played greater than or equal to 32 matches.

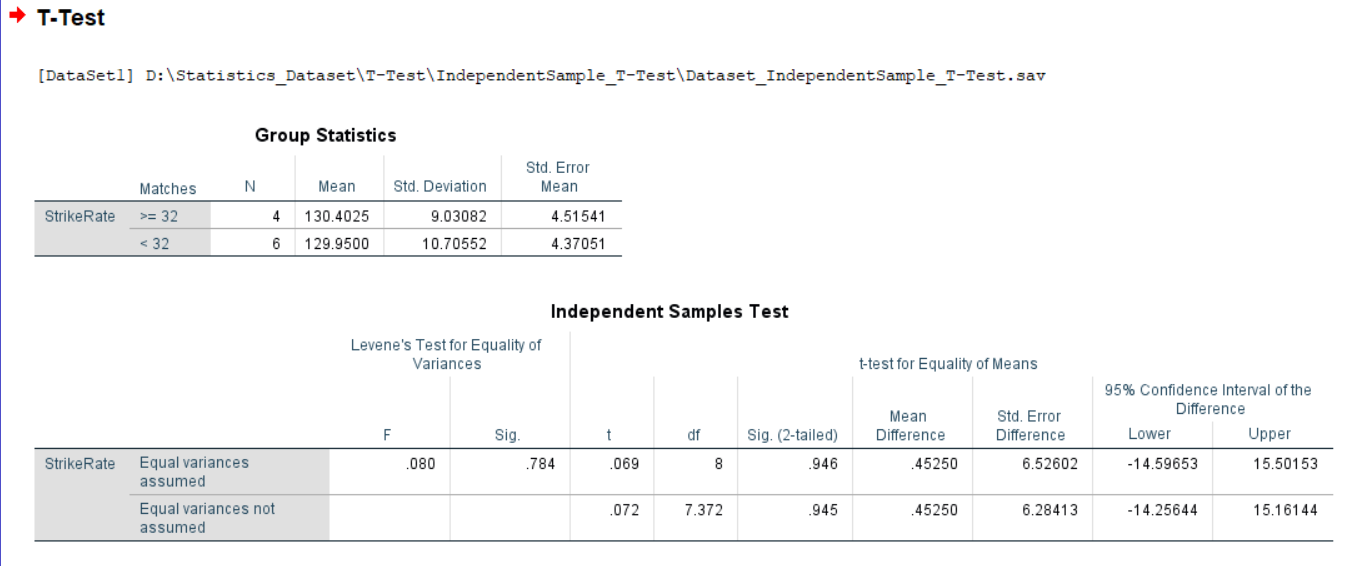
µ2: The Strike Rate of Players who have played less than 32 matches.

Ho: µ1 ≥ µ2

Ha: µ1 < µ2

Output:

T-Test (Independent Two Sample)



Conclusion:

At α = 0.05, the Critical Value in One-Sample T-Test with Degrees of Freedom 8 is 2.306. (Critical)

From the above T-Test conducted using SPSS, the obtained/calculated value is 0.69. (Calculated)

As Calculated value < Critical Value, at 5% Level of Significance, there’s not enough evidence to reject the claim that the Strike Rate of Players who have played greater than or equal to 32 matches is greater than or equal as the one’s played lesser than 32 matches.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***T-test for Two Paired Samples***

Procedure:

1) State the research question.

2) State the statistical hypotheses.

Ho: µD = 0

Ha: µD ≠ 0

3) Set the decision rule.

α = 0.5

Df [Degrees of Freedom] = number of difference scores - 1

T(critical)

4) Calculate the test statistic.

5) Decide if the result is significant.

6) Interpret the obtained results.

Question:

Among the Top Batsmen in T20WC (2007-2021), in a Sample of 10 Batsmen, the Runs and Matches of players have been paired and are considered to be similar. At α = 0.05, is there enough evidence to support this claim?

Answer:

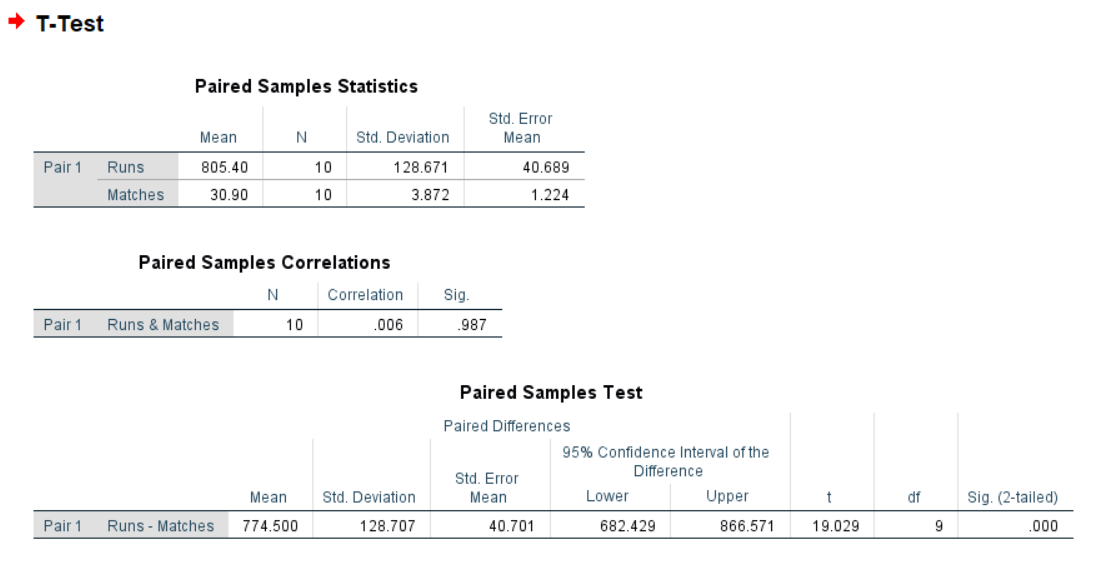
µD: Paired values of Runs and Matches.

Ho: µD = 0

Ha: µD ≠ 0

Output:

T-Test (Paired Two Sample)



Conclusion:

At α = 0.05, the Critical Value in One-Sample T-Test with Degrees of Freedom 9 is 2.262. (Critical)

From the above T-Test conducted using SPSS, the obtained/calculated value is 19.029. (Calculated)

As Calculated value > Critical Value, at 5% Level of Significance, there’s enough evidence to reject the claim that the Runs and Matches of players which have been paired are considered to be similar.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

UNIT-3:

***Runs Test***

A runs test is a statistical approach for determining if a set of data is generated at random from a given distribution. The runs test investigates the occurrence of identical occurrences separated by distinct events.

Procedure:

Step 1: State the hypotheses and identify the claim.

Step 2: Find the number of runs.

(Note: When the data are numerical, find the median. Then compare each data value with the median and classify it as above or below the median. Other methods such as odd-even can also be used.)

Step 3: Find the critical value. Use Runs Table.

Step 4: Make the decision. Compare the actual number of runs

with the critical value.

Step 5: Summarize the results.

Question: -

A data set is provided to the president of the International Cricket Council (ICC) to check the randomness of the matches played by the players. Test the claim, at α = 0.05

Solution: -

Ho: The Matches played by the players are in a random way (The claim)

H1: The Matches played by the players are not in a random way.

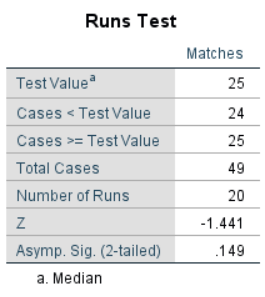


Table 1.0: Runs test of the given data with median

From the table 1.0 the median, n1 and n2 values are as follows

Median = 25  
 n1= 24(no of values less than median)  
 n2= 25(no of values greater than median)

And Number of Runs are 20

Since the n1 and n2 values are more than 20, which is not covered in the critical value table, so we compare the Z value with ***Z***1-*α*/2 (given α = 0.05, so according to the Z-table the ***Z***1-*α*/2 (Value is 1.960)

From the table 1.0, |Z| =1.441

Here “|Z|< ***Z***1-*α*/2“

So, the decision is not to rejects the null hypothesis(H0)

Therefore, the conclusion is that the matches played by the players are in random way.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Sign Test***

The sign test for single samples is the simplest nonparametric test for determining the value of a median for a specific sample.

Procedure:

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value(s). For the single-sample test, compare each value with the conjectured median. If the value is larger than the conjectured median, replace it with a positive sign. If it is smaller than the conjectured median, replace it with a negative sign. For the paired-sample sign test, subtract the after values from the before values, and indicate the difference with a positive or negative sign or 0, according to the value.

Check the data to see whether they support the null hypothesis. If they do, do not reject the null hypothesis. If not, continue with step 3.

Step 3: Compute the test value. Count the numbers of positive and negative signs found in step 2, and use the smaller value as the test value.  
Step 4: Make the decision. Compare the test value with the critical value in Table. If the test value is less than or equal to the critical value, reject the null hypothesis.

Step 5: Summarize the results.

Question:

The Cricket players say that the median for the number of innings they have played is 31, from the given data show whether there is enough evidence to reject the claim, at α= 0.05.

Solution:

H0: The median of the innings played by them is equal to 31(The claim)

H1: The median of the innings played by them is not equal to 31

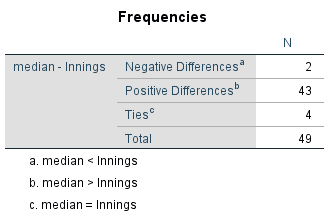
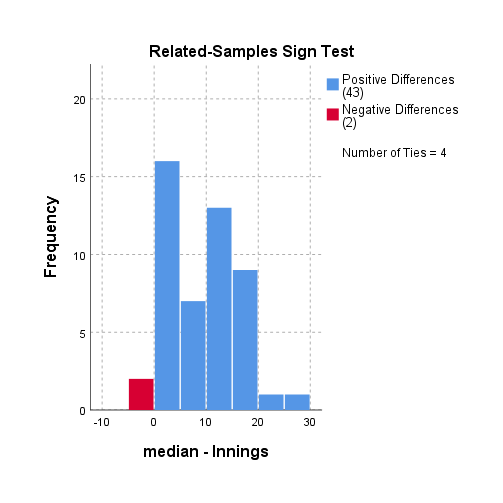


Table 1.0: Negative and positive values for give data



Bar graph for sign data

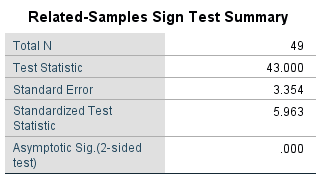


Table 1.1: sign test summary

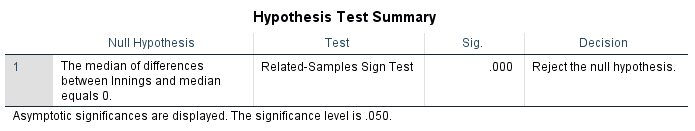


Table1.2: Hypothesis test summary

From Table 1.0   
No of negative signs=2  
No of positive signs =43

The negative value is not equal to the positive value.

From table 1.1

The asymptotic sig value is less than the significant level I.e 0.00<0.05

From Table 1.2, The decision is shown that to reject the null hypothesis

We can conclude that there is enough evidence to reject the claim that the median of innings played by the Cricket players is 31.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Sign Test for Paired Samples***

The two-sample paired sign test is used to test the null hypothesis that the probability of a random value from the population of paired differences being above the specified value is equal to the probability of a random value being below the specified value.

Procedure:

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value(s).

For the paired-sample sign test, subtract the after values from the before values, and indicate the difference with a positive or negative sign or 0, according to the value. Use Sign-test Table and n=total number of positive and negative signs.

Check the data to see whether they support the null hypothesis. If they do, do not reject the null hypothesis. If not, continue with step 3.

Step 3: Compute the test value. Count the numbers of positive and negative signs found in step 2, and use the smaller value as the test value.

Step 4: Make the decision. Compare the test value with the critical value in Sign-test Table. If the test value is less than or equal to the critical value, reject the null hypothesis.

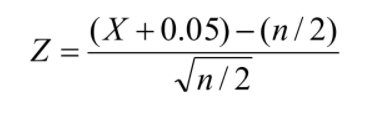
Step 5: Summarize the results.

Note: If the sample size n is 26 or more, use Z-Table and the following formula for the test value:

Where,

X =smaller number of + or - signs

n =sample size



Question:

There was a statement being spread all over the ICC that the cricket players have scored more strike rate after 2012 compared to the matches played before 2012. The director of ICC has decided to give a conclusion to this statement and collected the data of strike rate before 2012 and after 2012 of top 20 players. The data is given, at α=0.05, can the director conclude that the strike rate is more after 2012?

Solution:

Ho: The strike rate is more after 2012. (The claim)

H1: The strike rate is less after 2012.

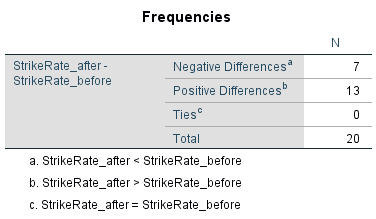
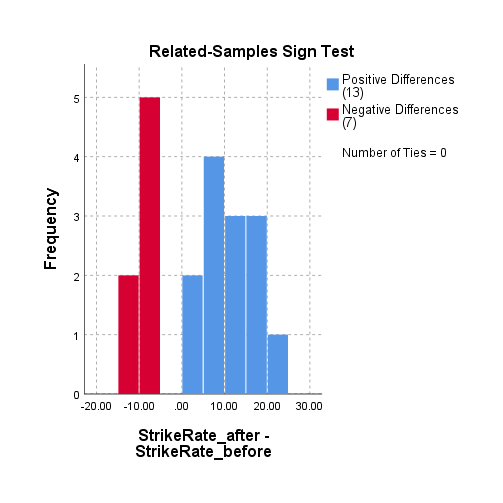


Table 1.0 Negative and Positive differences



Graph: Bar graph for Negative and Positive differences

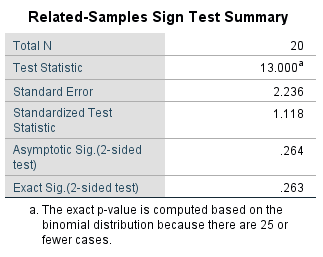


Table 2.0 Related samples for Sign Test

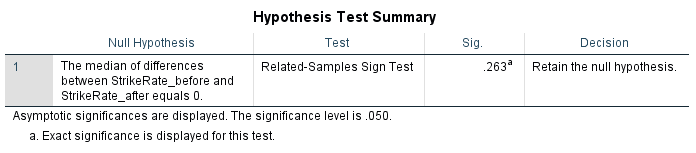


Table 3.0 Hypothesis test summary

From table 1.0

The negative differences are 7   
And the positive differences are 13

From table 2.0   
The exact significant value is 0.263

The significant level α=0.05

The exact significant value > the significant level

From table 3.0

The decision is taken as to “retain the null hypothesis”

We can conclude that there is enough evidence to Accept the claim, the statement is absolutely correct that the strike rate is more after 2012.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Mann-Whitney U Test***

**Definition:**

A popular nonparametric test to compare outcomes between two independent groups is the Mann-Whitney U test. The Mann-Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations. In the Wilcoxon tests, the values of the data for both samples are combined and then ranked. If the null hypothesis is true—meaning that there is no difference in the population distributions—then the values in each sample should be ranked approximately the same.

**Procedure:**

Step 1: State the hypotheses and identify the claim.

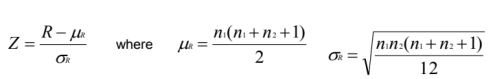
Step 2: Find the critical value(s). Use Z-Table.

Step 3: Compute the test value.

1. Combine the data from the two samples, arrange the combined data in order, and rank each value.
2. Sum the ranks of the group with the smaller sample size.

(Note: If both groups have the same sample size, either one can be used.)

1. Use these formulas to find the test value.



Where , R: sum of ranks for smaller sample size (n1 or n2)

n1: First sample size

n2: Second sample size (n1 ≥ 10 and n2 ≥ 10)

(Note: If both samples are of same size, either size can be used as n1.)

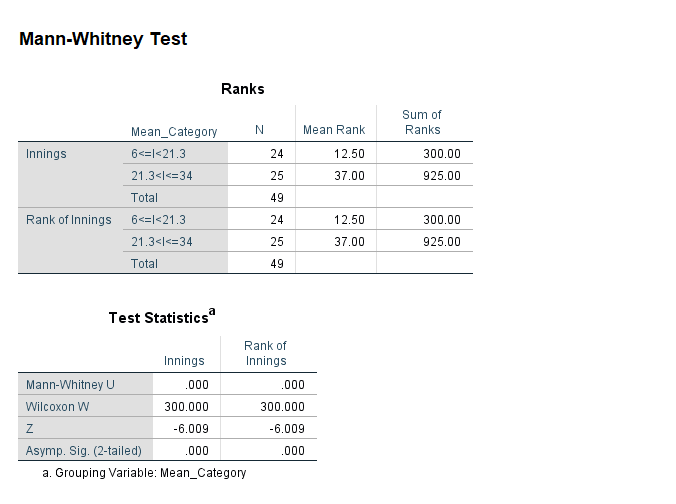
Step 4: Make the decision.

Step 5: Summarize the results.

**Problem:**

The players are divided into two independent groups i.e., Group 1 and Group 2 based on the number of Innings played in their WC T20 (2007-21) Career. A study was conducted to see whether there is a difference in the distributions (number of innings played) between the two groups. At 95% confidence interval, can it be concluded that there is a difference?

**Solution:**



**Observation:**

Zcritical = -1.960

Zcalculated = -6.009

Zcalculated **<** Zcritical

Therefore, reject our null hypothesis.

**Conclusion:**

At 95% confidence interval, there is no evidence that there is a difference between number of innings played by the players in Group 1 and Group 2

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Wilcoxon Signed-Rank Test***

**Definition:**

The Wilcoxon signed rank test (also called the Wilcoxon signed rank sum test) is a non-parametric test to compare data. The Wilcoxon signed rank test should be used if the differences between pairs of data are non-normally distributed.

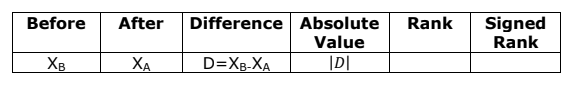
**Procedure:**

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value from Wilcoxon Signed-rank-Table.

Step 3: Compute the test value.

1. Make a table, as shown.

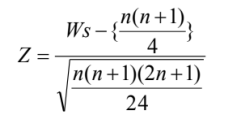


1. Find the differences (before-after), and place the values in the Difference column.
2. Find the absolute value of each difference, and place the results in the Absolute value column.
3. Rank each absolute value from lowest to highest, and place the rankings in the Rank column.
4. Give each rank a positive or negative sign, according to the sign in the Difference column.
5. Find the sum of the positive ranks and the sum of the negative ranks separately.
6. Select the smaller of the absolute values of the sums, and use this absolute value as the test value Ws.

Step 4: Make the decision. Reject the null hypothesis if the test value is less than or equal to the critical value.

Step 5: Summarize the results.

Note: When n≥30, use Z-Table and the test value



Where

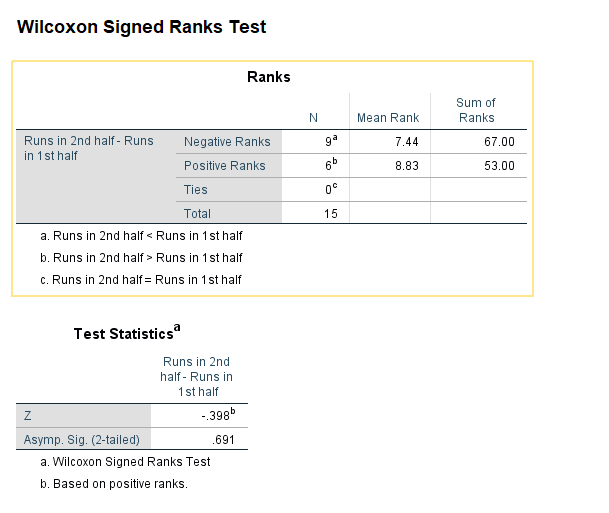
n = number of pairs where difference is not 0

Ws = smaller sum in absolute value of signed ranks

**Problem:**

The top 15 players of WC T20 (2007-2021) were taken into observation. The runs scored by the top 15 players in the first half of their WC T20 career and in the second half of their WC T20 career were recorded into two different groups and compared. At α = 0.05, can it be concluded that there is a difference between the runs scored by the players in the 1st half and 2nd half of their careers.

**Solution:**



**Observation:**

α = 0.05 Wstab = 25

Here, total number of observations (n) = 15

Smaller of the absolute values of the sums (Ws) = 53

Here at α = 0.05 and Wstab = 25, Ws  > Wstab, we can’t reject our null hypothesis.

**Conclusion:**

At α = 0.05, it is evident that there is a significance difference in between the runs scored by top 15 players in the 1st half and 2nd half of their WC T20(2007-21) careers.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

UNIT-4:

***Chi-Square Test for Goodness of Fit***

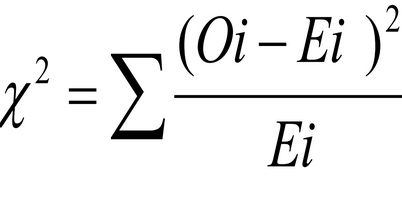
Procedure:

The Chi-Square Goodness-of-Fit Test

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value. The test is always right-tailed.

Step 3: Compute the test value. Find the sum of the values.



Step 4: Make the decision.

Step 5: Summarize the results.

Problem:

We are interested to see if the four categories {[0,10], (10,20], (20,30], (30, ∞)} of "Total No of Matches a player played". The chance of playing is equally distributed among the players. A sample of 49 players is selected, and the dataset is used. At α=0.05 Is there enough evidence to reject the claim that the frequencies in all categories are equal?

Solution:

Step-1: State the hypotheses and identify the claim.

Ho: there is equal frequency in all categories(claim)

H1: Equal frequency doesn’t occur.

Step-2: Find the critical value.

The degrees of freedom are 3, and α=0.05.

Therefore, Critical value from Chi-square-Table value = 7.815

Step-3: Compute the test value.

From the output of the test in SPSS viewer, we have chi-square=16.388

Step-4: Make the decision.

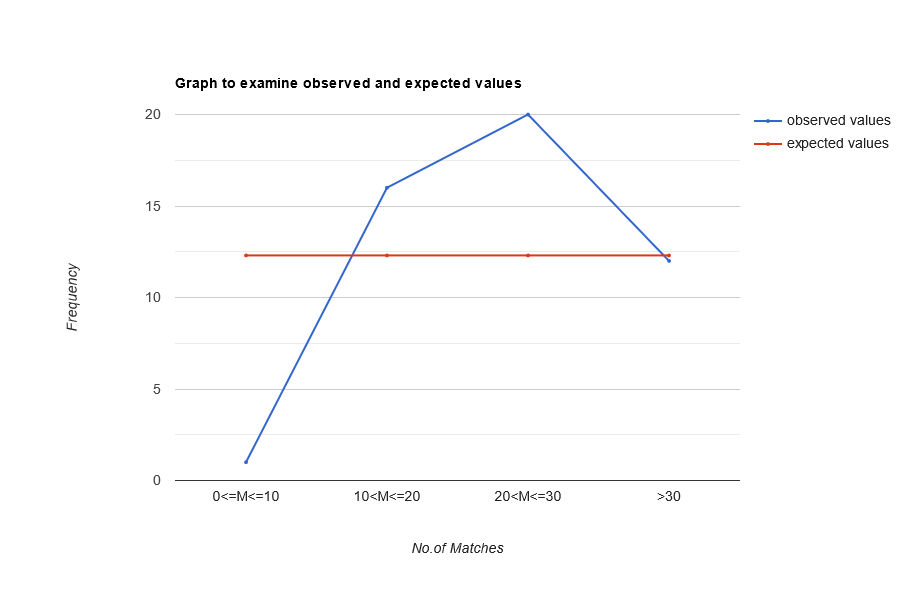
p-value = 0.001 and α=0.05

The decision is to reject the null hypothesis, since

7.815<16.388

Step-5: Summarize the results.

There is enough evidence to reject the claim that there is equal frequency in all categories. The fit of equal frequencies is “not good enough.”



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Chi-Square Test for Goodness of Fit***

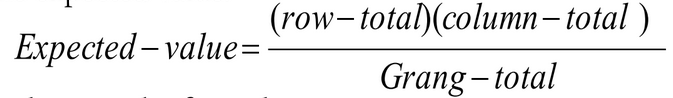
Procedure:

Step 1: State the hypotheses and identify the claim.

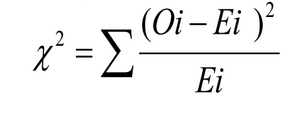
Step 2: Find the critical value in the right tail. Use a Chi-square Table.

Step 3: Compute the test value. To compute the test value, first find the expected values.

For each cell of the contingency table, use the formula to get the expected value.



To find the test value, use the formula



Step 4: Make the decision.

Step 5: Summarize the results.

Problem:

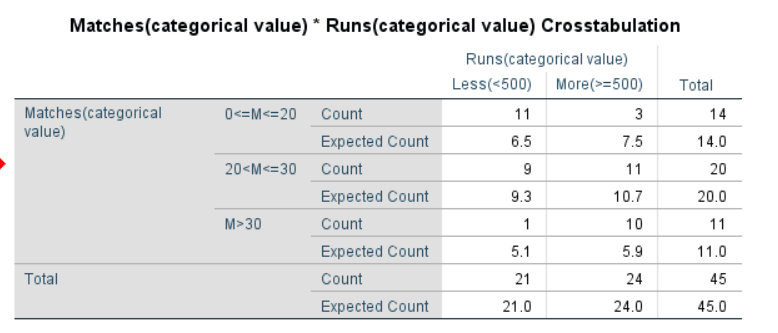
From the population of players in T20-WC batting dataset, a sample of 45 players have been selected and their corresponding matches and runs are noted down. At a level of significance 0.05, check whether there is any relationship between the Matches played by the player and runs scored by them.

Note: Matches are categorized into 3 categories and Runs scored are divided into 2 categories

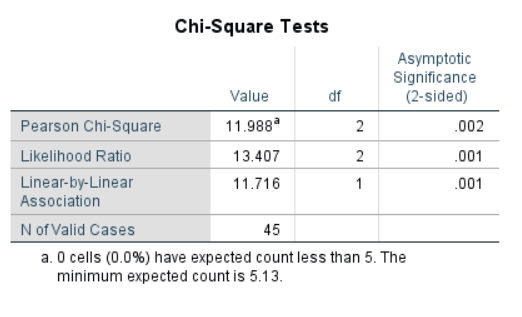
Solution:

H0: There is no significant relationship between the Matches played by the player and runs scored by them.

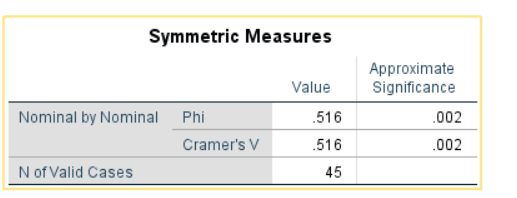
H1: There is a significant relationship between the Matches played by the player and runs scored by them.



*Table: 1.0 - Table containing both observed and expected values (from SPSS)*



*Table: 1.1*



*Table: 1.2*

From table 1.1, we got chi-square value as 11.988

And α=0.05, degrees of freedom =2

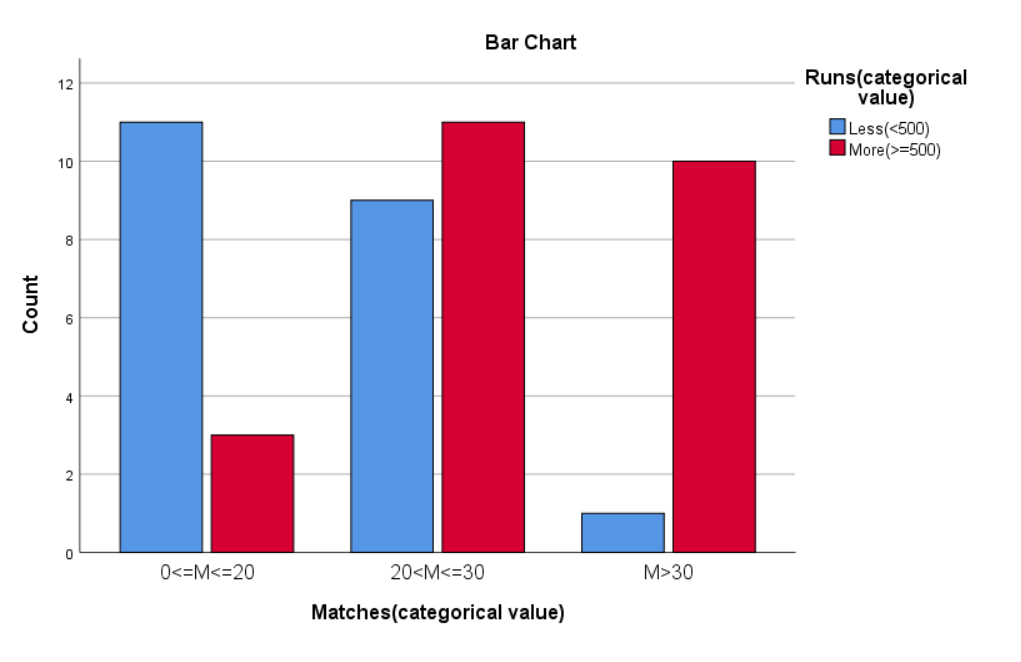
From the chi-square value table, Critical value = 5.991

Here, critical value< calculated value.

Hence, the decision is to reject the null hypothesis.

Therefore, the conclusion is that there is a significant relationship between the Matches played by the player and runs did by them.

***Bar Graph to examine the relationship between matches and runs:***



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

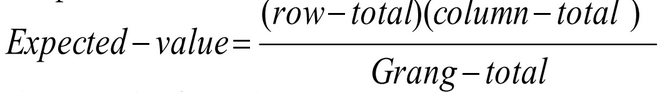
***Chi-Square Test for Homogeneity***

Procedure:

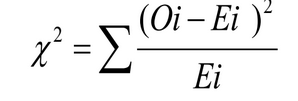
Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value in the right tail. Use Chi-square Table.

Step 3: Compute the test value. To compute the test value, first find the expected values. For each cell of the contingency table, use the formula to get the expected value.



To find the test value, use the formula



Step 4: Make the decision.

Step 5: Summarize the results.

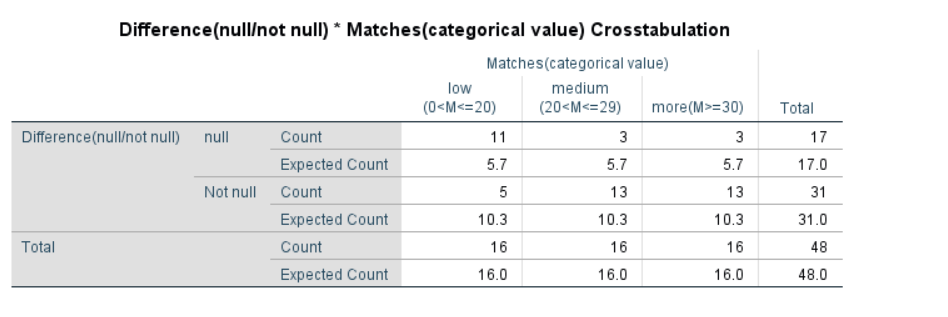
Problem:

We selected three samples of 16 players in categories of matches played by them (i.e., low, medium, more based on the number) and noted the difference between their matches and innings (Difference) and categorized the value into null or nor null. At α=0.05, test the claim that the proportion of (matches)-category is same for all (difference of matches and innings)-category.

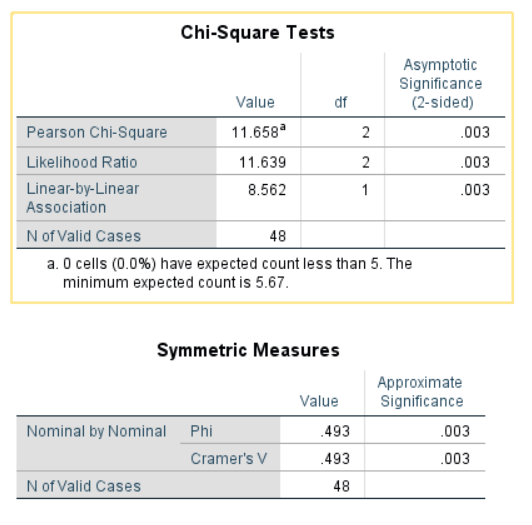
Solution:

H0: All proportions of (matches)-category are same for all (difference of matches and innings)-category.

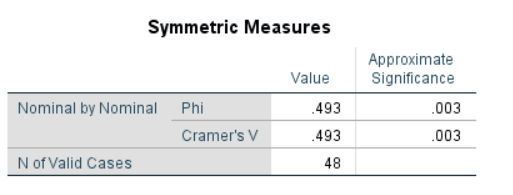
H1: All proportions are not equal.



*Table: 1.0 - Table containing both observed and expected values (from SPSS)*



*Table: 1.1*



*Table: 1.2*

From table 1.1, we got chi-square value as 11.658

And α=0.05, degrees of freedom =2

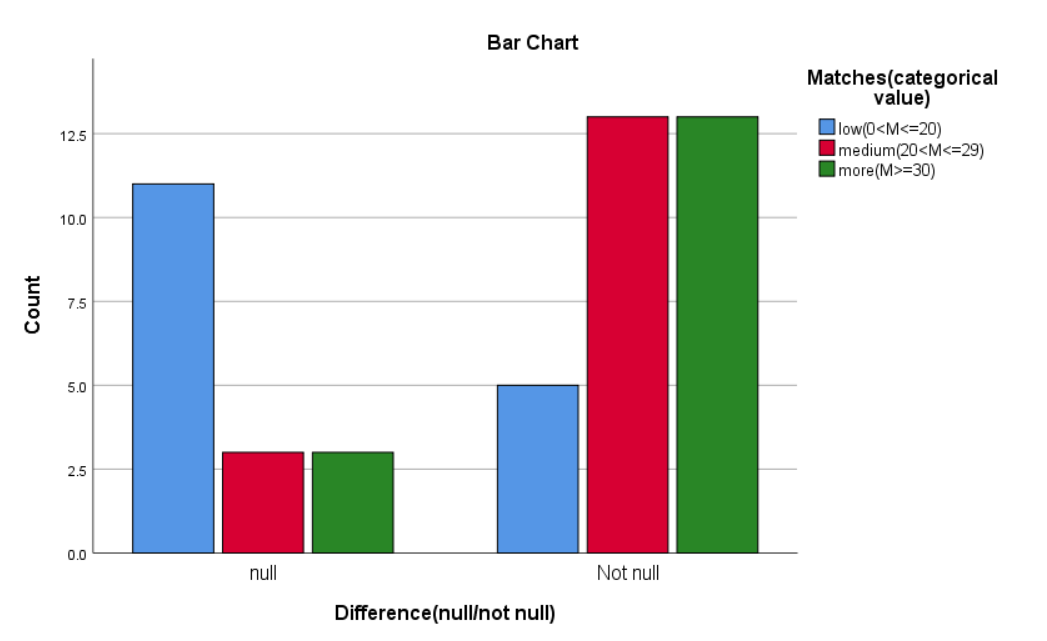
From the chi-square value table, Critical value = 5.991

Here, critical value< calculated value.

Hence, the decision is to reject the null hypothesis.

Therefore, the conclusion is that all proportions (matches)-category are not same for all (difference of matches and innings)-category.

***Bar Graph to examine the relationship between matches and difference(between matches and innings):***



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Chi-Square Test for 2X2 Contingency Table***

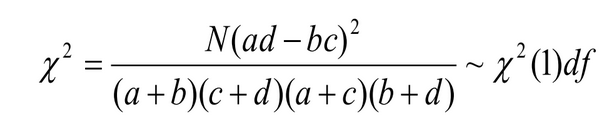
Procedure:

Step 1: State the hypotheses and identify the claim.

Step 2: Find the critical value in the right tail. Use Chi-square Table.

Step 3: Compute the test value. To compute the test value of the

contingency table, use the formula to get the test value.



Step 4: Make the decision.

Step 5: Summarize the results.

Problem:

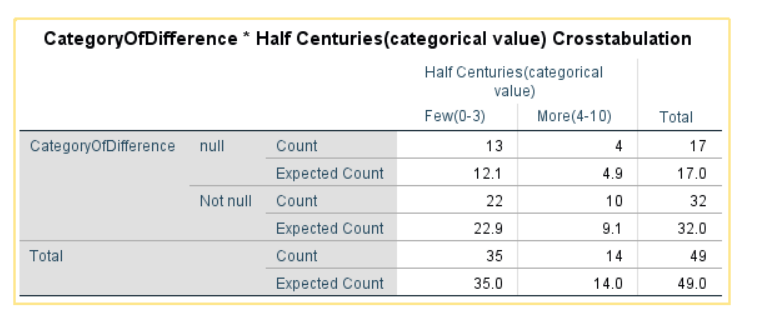
From the dataset given we can see the information of each player. And we are interested to find if there is any evidence of a relationship between difference-category value and the half centuries scored-category value. At α=0.05 level, is there evidence of a relationship?

Solution:

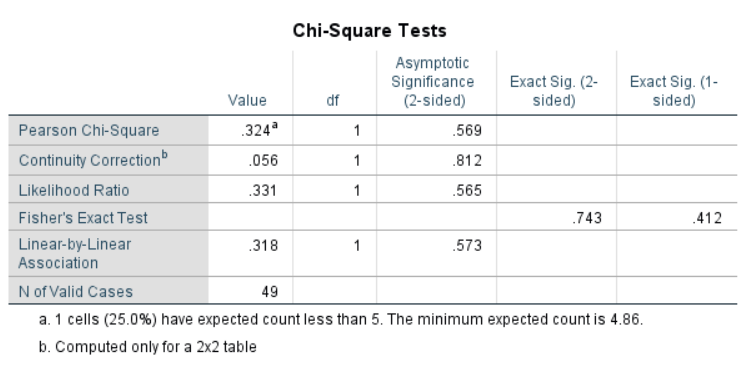
Ho: There is no relationship between difference of matches and innings-category value and the half centuries scored-category value.

H1: There is a relationship between difference of matches and innings-category value and the half centuries scored-category value.

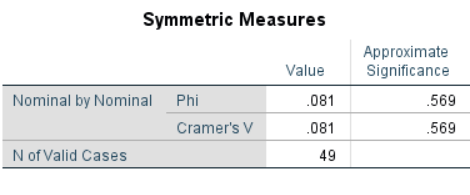
* The critical value at α=0.05 level with 1 degree of freedom is 3.841.



*Table: 1.0-Observed and expected values (from SPSS)*



*Table: 1.1-Tests Statistics values (from SPSS)*



*Table: 1.2*

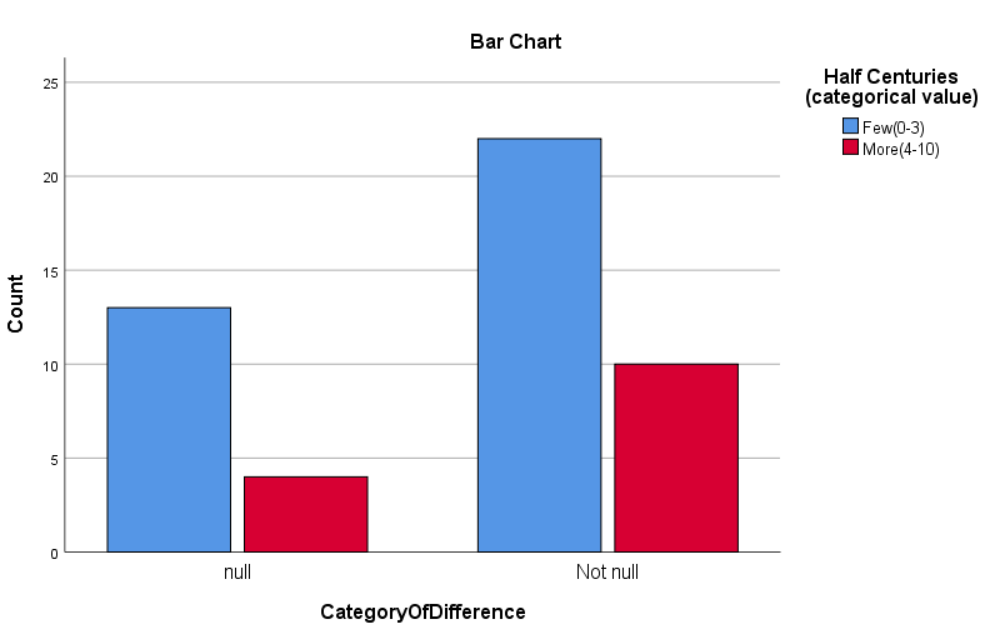
From Table 1.1, we got Chi-Square value as 0.324

Here, critical value> calculated value.

Hence, the decision is not to reject the null hypothesis.

Therefore, the conclusion is that there is no relationship between difference of matches and innings-category value and the half centuries scored-category value.

***Bar Graph to examine the relationship between difference-category (matches-innings) and Half centuries-category:***



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

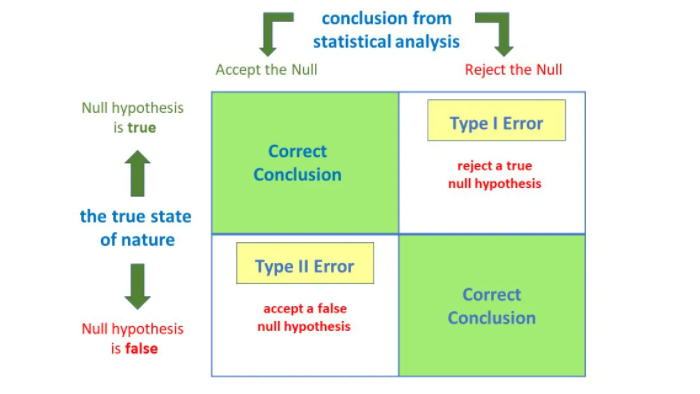
UNIT-5:

***Type -2 error***

The probability of not finding a difference that actually, exists between two groups (or between sample and population). Not Rejecting null hypothesis when it is false.

Known as the β

Power is (1- β) and is usually 80%



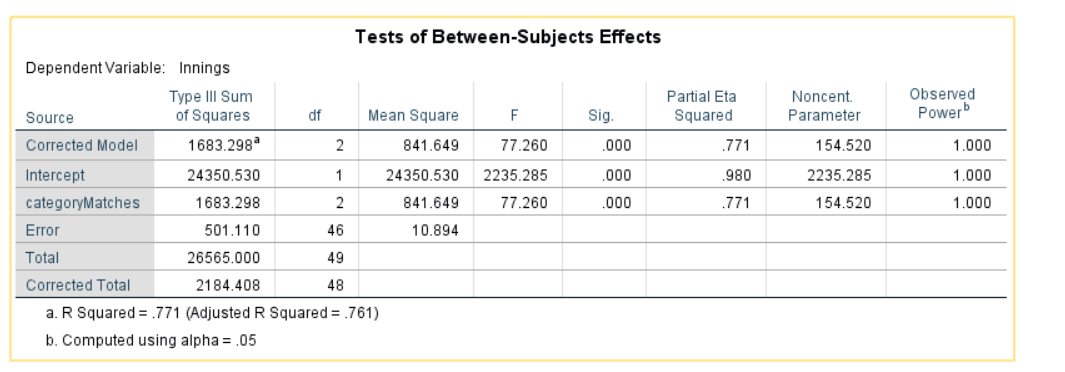
**Problem:**

From the population of players in the T20-WC batting dataset, a sample of 49 players have been selected and their corresponding matches and innings are noted down. Calculate the observed power of innings as matches (Category values) are fixed, and calculate value of β. (assuming α=0.05)

Solution:

H0 : There is no significant relationship between matches played by the players (category value) and innings scored by them.

H1 : There is a significant relationship between matches played by the players (category value) and innings scored by them.



*Table: 1 - Table containing observed power and significance*

From the table we get power of innings

Power = 1(100%)

i.e., 1-β =1 ==> β=0

This indicates the probability is 100% of rejecting the null hypothesis when it is false.

Therefore, accept H0.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Estimation of sample size***

One of the most crucial aspects of statistical analysis is determining the optimum sample size. If the sample size is too small, the results will be invalid and the population being investigated will not be appropriately represented. The likelihood that a statistical test will reject the null hypothesis when it is false is known as its power. That is, power is a measure of the likelihood of not making a type II error. The sample size and the effect size are the two most important criteria that influence a study’s power. The power of the significance test increases as the sample size grows. This is because a bigger sample size narrows the test statistic’s distribution. This reduces the standard error of the distribution and the acceptance region, increasing the level of power.

A study should only be carried out if there is a reasonable probability of obtaining relevant information. As a result, determining the optimal sample size for a study is a critical stage in its design.

**A) One sample (One Mean and One Proportions)**

For single mean:

This test is used to determine a specific sample size while considering preference proportions and other features of the dataset. We must first define an acceptable margin of error d before determining a suitable sample size. Remember that the wiggle room around the point estimate is the margin of error d. Sometimes half the width of the confidence interval may be used. The wider the C.I., The less reliable the sample statistic is, the less likely it is to provide an accurate approximation of the true value of the population parameter.

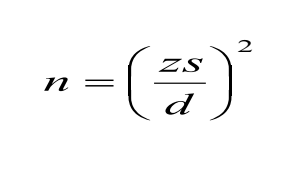
Procedure:

1. Calculate standard deviation.

2. Specify the confidence interval and normal deviation.

3. Find the error rate.

4. Compute the sample size using the formula.



where:

n = sample size

z = confidence interval in standard error units

s = standard error of the mean

d= acceptable magnitude of error

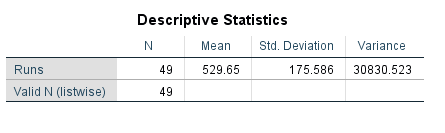
5. Conclusion.

Problem:

we want to estimate the average runs of players, and we want the error of estimation to be less than 20 runs of the true mean, with probability of 95% (e.g., error rate of 5%). [z=1.96 for 95% confidence level]

Solution:

Estimate the standard error of the mean s from the data



From the table we came to know standard deviation s= 175.586

And z=1.96; d=20 (from data)

Therefore,

n = 1.96\*1.96\*175.586\*175.586/20\*20

n =296.0955

The sample size required to estimate the mean runs with 95% confidence and error of estimation less than 20 runs is 296

For a single proportion:

This test is used to determine a specific sample size while considering preference proportions and other features of the dataset. In this scenario, we try to estimate the sample size required for a particular preference given a specific proportion in the population.

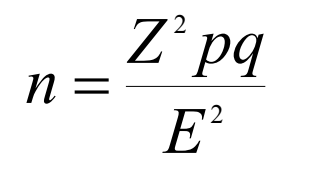
Procedure:

1. Specify the proportion.

2. Specify the confidence interval and normal deviation.

3. Find the error rate.

4. Compute the sample size using the formula:



Where:

n = number of items in samples

Z^2 = square of confidence interval in standard error units

p = estimated proportion of success

q = (1-p) or estimated the proportion of failures

E^2 = square of maximum allowance for error between true

proportion and sample proportion, or Zsp squared.

5. Conclusion.

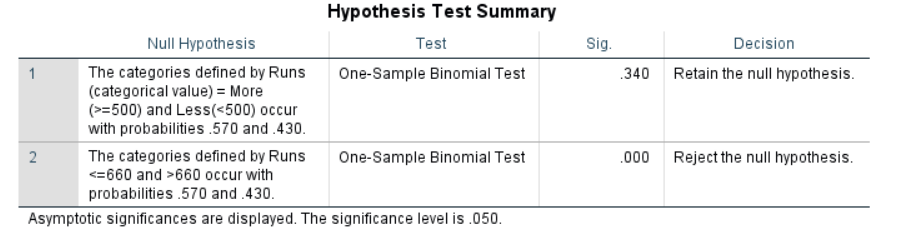
Problem:

The proportion of preference for players who scored greater than 500 runs is 57%. We want to estimate the preference p in a community within 5% with 95% confidence interval.

Solution:

Ho: The proportion of preference for players who scored greater than 500 is 57%

H1: The proportion of preference for players who scored less than 500 is 57%



From the table,

We can accept the H0.

with 95% of significance level i.e α=0.05

p=57%=0.57 => q=0.43

E=0.05 (given)

When we are having 95%significance z=1.96

So,

N = Z^2 pq/E^2

= 1.96\*1.96\*0.57\*0.43/0.05\*0.05

= 376.6304

The sample size required to estimate the preference of proportion with 95% confidence and error of estimation less than 5.0% is 376.6

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**B) Two Samples (Two Mean and Two Proportions)**

This test is used to determine a specific sample size for two different approaches on samples in the same data set. In this case, we try to estimate the sum of sample sizes while comparing two means or proportions.

For two means:

Procedure:

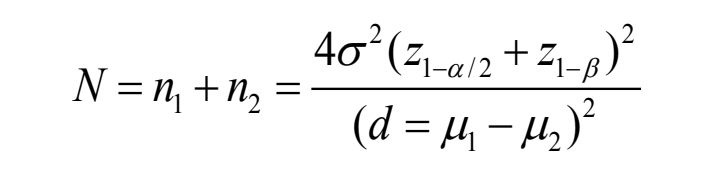
1. Specify the standard deviation.

2. Specify the significance level and power of the test.

3. Specify the confidence interval and critical normal values.

4. Find the true difference.

5. Compute the sample size using the formula:



n1, n2: sample sizes required for the two techniques

d: difference of means or proportions denoting accuracy of the estimate

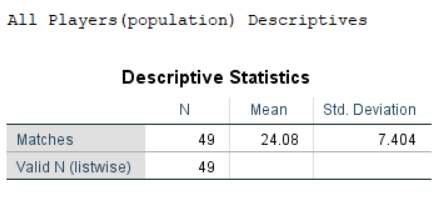
Z: normal deviate reflecting type-1 error

α: significance level

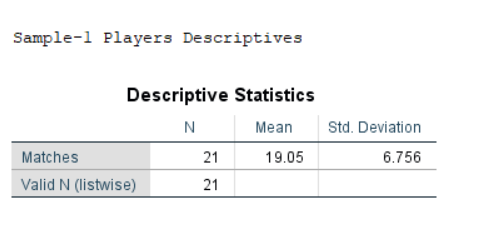
1 − β: power of the test

6. Conclusion.

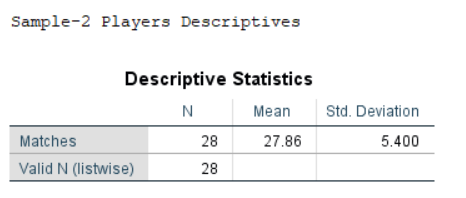
Problem:



*Table -1.0*



*Table -1.1*



*Table -1.2*

Two samples of players and their respective No. of Matches played are collected from the population data and the standard deviation for the population data is 7.404 matches (from table-1.0). we have means of the samples as 19.05 and 27.86 matches respectively (from table-1.1,1.2). How large a sample would be needed for comparing these two samples with Power = 1- β = 0.90 and α=0.05?

Solution:

When σ = 7.404, β = 0.10, α = 0.05; z 1- α /2 = 1.96

Power = 1- β; z 1- β= 1.282, d = μ1- μ2= (19.05-27.86)

N = n1+n2 = [4\* (7.404)2(1.96+1.282)2] **÷** (19.05-27.86)2

= 29.693 ≈ 30

Hence about 15 for each group (rounded value).

For two proportions:

Procedure:

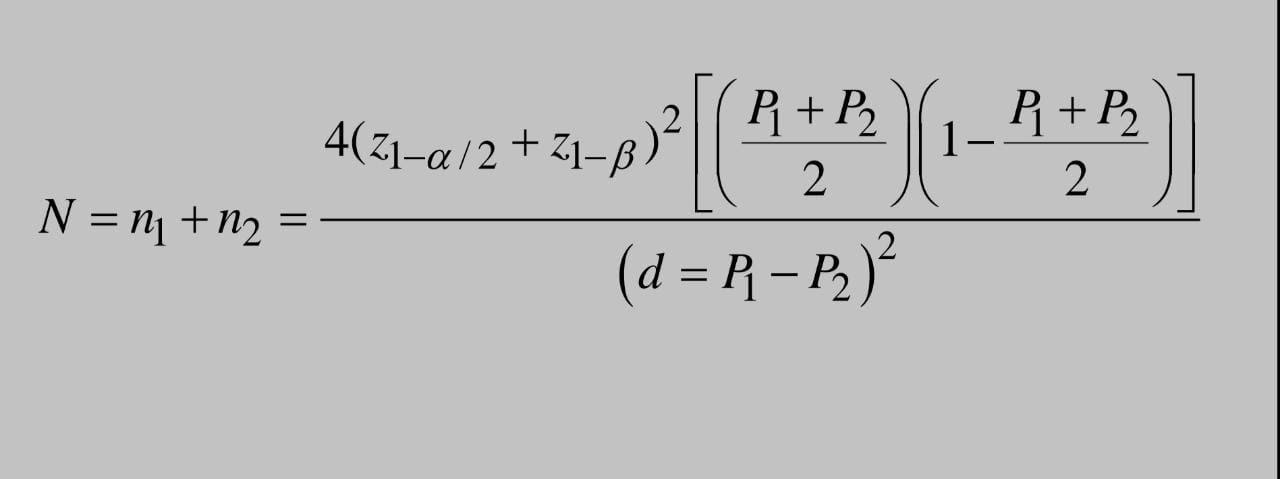
1. Specify the proportions P1, P2.

2. Specify the significance level and power of the test.

3. Specify the confidence interval and critical normal values.

4. Find the true difference.

5. Compute the sample size using the formula:



n1, n2: sample sizes required for the two techniques

P1, P2: proportions to be estimated

d: difference of means or proportions denoting accuracy of the estimate

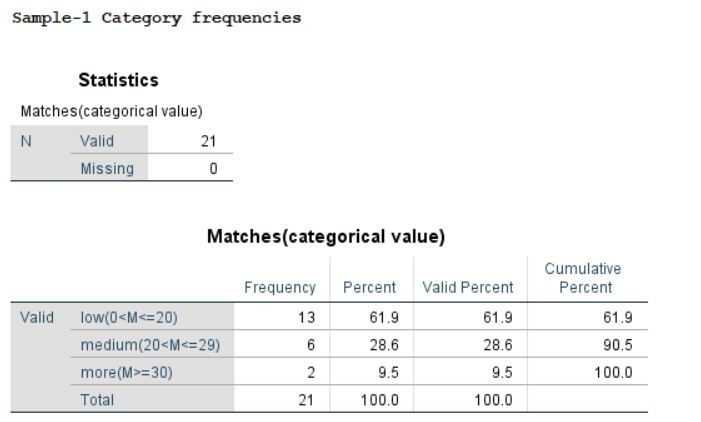
Z: normal deviate reflecting type-1 error

α: significance level

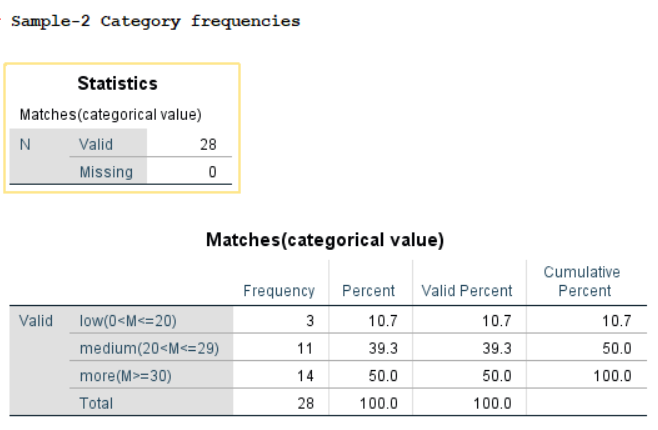
1 − β: power of the test

6. Conclusion.

Problem:



*Fig-2.0*



*Fig-2.1*

According to the descriptives from fig-2.0 and 2.1, we have 61.9% players (out of 21 players) of sample-1 have played less number of matches and 10.7% players (out of 28 players) of sample-2 have played less number of matches. How many players

should be included from each group to determine whether this difference is significant at a=0.05 if we wish to have β=0.1 chance of detecting the difference if it is real?

Solution:

Given, P1 = 0.619, P2 = 0.107, β = 0.10, α = 0.05; z 1- α /2 = 1.96

Power = 1- β; z 1- β = 1.282

🡺 (P1 + P2) / 2 = (0.619 + 0.107) / 2 = 0.363

Therefore,

N = n1+n2

= {4 \* (1.96 + 1.282)2[ (0.363) \* (1-0.363)]} ÷(0.619 - 0.107 )2

= 37.084 ≈ 38

Hence about 19 for each group (rounded value).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Acknowledgement

We would like to express our Gratitude to Prof of Statistics, Dr. Ramesh Athe, for giving us this opportunity to work with realistic Dataset and explore various Tests in the field of Statistics.

References

* [ICC Men's T20 World Cup Cricket Team Records & Stats | ESPNcricinfo.com](https://stats.espncricinfo.com/ci/engine/records/batting/most_runs_career.html?id=89;type=trophy) (For Dataset)
* PPTS provided to us by Ramesh Athe Sir in Google Classroom
* Online resources