Implementation And Elucidation Of Estimation and Prediction Based Statistical Tests Discussion

Engr.Muhammad Mujtaba Asad [1], Dr.Razali Bin Hassan [2], Engr.Fahad Sherwani [3]

Universiti Tun Hussein Onn Malaysia,Parit Raja, Batu Pahat Johor, Malaysia.

Mujtabaasad11@gmail.com, razalih@uthm.edu.my,fahadsherwanis@gmail.com

Abstract: In this paper researchers explain and explore about the statistical test they are mostly used in social sciences and educational researches in all over the world. But generally students and new researchers cannot get the accurate and detailed concept of these implemented tests. In this paper researchers talk about two main tests CHI square and regression test, with few case studies and explanation with SPSS representation and manual calculations, These Statistical test are commonly used for estimation and predictions in quantitative researches. Mostly management and education students use to work with these statistical tests but these days engineering field or researches is also using these test for estimation purposes. In this research paper researchers took two real time mini researches and cases as a real data for the clarification of actual concept of CHI square and Regression statistical tests. This research is designed for the enhanced outcome in the performance of statistic students they are working on prediction based researches and projects. From this article student can easily understand the perception of estimation and prediction test for statistical analysis.

[Engr.Muhammad Mujtaba Asad, Dr.Razali Bin Hassan, Engr. Fahad Sherwani. **Implementation And Elucidation Of Estimation and Prediction Based Statistical Tests Discussion.** *Rep Opinion* 2014;6(3):1-9]. (ISSN: 1553-9873). http://www.sciencepub.net/report. 1

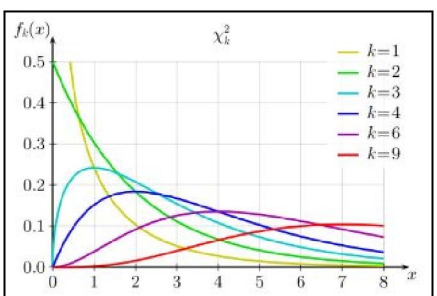
Key words: Estimation, Prediction, Statistical Tests, Chi Square, Regression.

1. Introduction CHI Square Tests

The Chi-square test is often used to answer this type of question, "How well does this model fit the facts?". Chi-square is denoted by X^2, with X pronounced 'ki'.

Karl Pearson came up with a distribution that is used to find the probability for X^2 tests involving a new curve called X^2-curve. These are the properties of the X^2-distribution:

1. X^2 is not symmetric.
2. Values of X^2 can never be negative.
3. The X^2-distribution is different for each degree of freedom.



X^2-curve

From this distribution, it can be seen that each degree of freedom has its own X^2-curve with degree of freedom calculated by this formula:

df=k-1

df: degree of freedom

k: number of categories

The formula for Chi-square test is:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

O: observed frequency

E: expected frequency

There are two types of Chi-square test. They are Chi-square goodness-of-fit and Chi-square test-of-independence or association.

1.1. Goodness-of-fit test

This test is used to see if an observed distribution fit an expected distribution. This test is used on a one group design with one single observation done on one variable. The data involved are in the form of frequencies. The distribution compared here are between observed frequencies and expected frequencies.

The steps involved in this test are:

1. Determine hypothesis null
2. Determine hypothesis alternative
3. Choose alpha level
4. Calculate test statistics
5. Determine critical value
6. Interpret results
7. Conclude findings

Case I

This is the number of childbirth in the US in 2010 which is taken from the Birth Month Statistic found in Statistic Brain in thousands. The source for this statistic is the Center for Disease Control and it was last updated on 28th of April 2013. From this table, does the month where mothers give birth affect the number of childbirth in 2010?

The data gathered is the frequency of childbirth according to months.

The data type is nominal.

The data is non-parametric.

The data is not normally distributed.

Therefore, use Chi-square goodness-of-fit test.

	Number of Childbirth
Month	
January	319.297
February	299.235
March	335.786
April	308.809
May	334.437
June	336.251
July	347.934
August	362.798
September	350.711
October	347.354
November	330.832
December	335.111
TOTAL	4,008.555

Data analysis/hypothesis testing

a) Manual

STEP 1

Determine hypothesis null

There is no statistically significant difference between the observed number of childbirth in the US each month in 2010 and the expected number of childbirth in the US each month in 2010.

STEP 2

Determine hypothesis alternative

There is a statistically significant difference between the observed number of childbirth in the US each month in 2010 and the expected number of childbirth in the US each month in 2010.

STEP 3

Choose alpha level

Since the alpha level is not stated in case description, so we choose alpha level: 0.05

STEP 4

Calculate test statistics

	Number of Childbirth	
Month	Observed	Expected
January	319.297	334.046
February	299.235	334.046
March	335.786	334.046
April	308.809	334.046
May	334.437	334.046
June	336.251	334.046
July	347.934	334.046
August	362.798	334.046
September	350.711	334.046
October	347.354	334.046
November	330.832	334.046
December	335.111	334.046
TOTAL	4,008.56	4,008.55

$$x^2 = \sum \frac{(O - E)^2}{E}$$

O = observed number of childbirth

E = expected number of childbirth

For the sake of calculation, we calculate the values with decimal points.

E = 4,008.55/12 = 334.046

$X^2 = (319.297-334.046)^2/334.046 + (299.235-334.046)^2/334.046 + (335.786-334.046)^2/334.046 + (308.809-334.046)^2/334.046 + (334.437-334.046)^2/334.046 + (336.251/334.046)^2/334.046 + (347.934-334.046)^2/334.046 + (362.798-334.046)^2/334.046 + (350.711-334.046)^2/334.046 + (347.354-334.046)^2/334.046 + (330.832-334.046)^2/334.046 + (335.111-334.046)^2/334.046$

$X^2 = 0.651 + 3.628 + 0.009 + 1.907 + 0.000 + 0.015 + 0.577 + 2.475 + 0.831 + 0.530 + 0.031 + 0.003$

$X^2 = 10.66$

STEP 5

Determine critical value

df=n-1

n=12

Therefore, df=11.

From the X^2 distribution table, the critical value is 19.68.

STEP 6

Interpret results

$X^2_{calculated} > X^2_{critical}$. Therefore, reject H_0.

STEP 7

Conclude findings

There is a statistically significant difference between the observed number of childbirth in the US each month in 2010 and the expected number of childbirth in the US each month in 2010.

The number of childbirth in the US on 2010 is affected by the month.**b) SPSS**

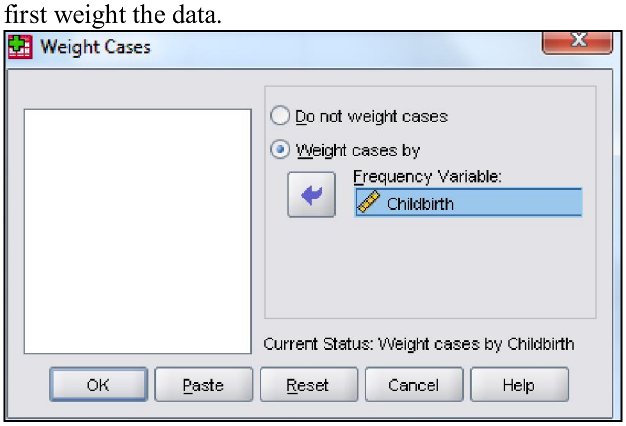
STEP 1

Plot data

Month	Childbirth
January	319.297
February	299.235
March	335.786
April	308.809
May	334.437
June	336.251
July	347.934
August	362.798
September	350.711
October	347.354
November	330.832
December	335.111

STEP 2

Since the data is already summarized, we must first weight the data.



STEP 3

Analyze data

Month	Observed	Expected
January	299	334.0
February	309	334.0
March	319	334.0
April	331	334.0
May	334	334.0
June	335	334.0
July	336	334.0
August	336	334.0
September	347	334.0
October	348	334.0
November	351	334.0
December	363	334.0
TOTAL	4008.0	

Chi-square = 10.743

df = 11

From the X^2 distribution table, the critical value is 19.68.

STEP 4

Interpret results

$X^2_{calculated} > X^2_{critical}$. Therefore, reject H_0.

STEP 5

Conclude findings

There is a statistically significant difference between the observed number of childbirth in the US each month in 2010 and the expected number of childbirth in the US each month in 2010.

1.2 Test-of-association

Another name for this test is Chi-square test-of-independence. This test is used to test the association between two categorical variables. In order for the data to be analyzed using this test, the data has to pass two assumptions which are:

1. The two variables must be either nominal data type or ordinal data type (categorical data).
2. The two variables must have two or more independent categories.

The steps involved in this test are:

1. Determine hypothesis null
2. Determine hypothesis alternative
3. Choose alpha level
4. Calculate test statistics
5. Determine critical value
6. Interpret results
7. Conclude findings

Case

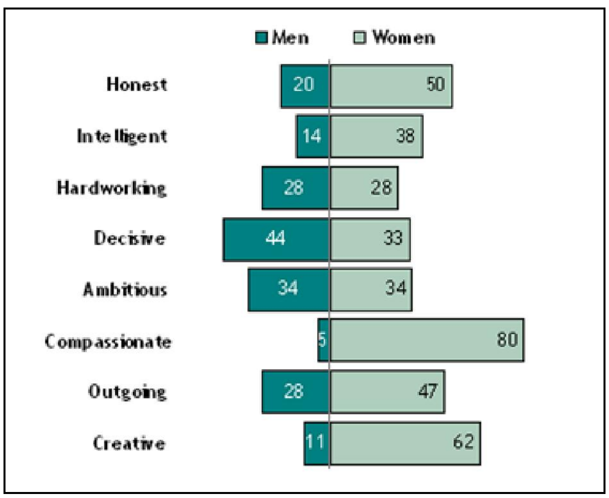
A survey was done on 2,250 adults on the subject of gender and leadership traits. The result is as shown in the diagram below. This result is shown in percentage. For example, under the 'Honest' trait, 20% of the respondents said men are more honest than women while 50% of the respondents said women are more honest than men. Responses such as "Equally True" and "Don't know" are not shown in this result. From the result of the survey, is there an association between gender and leadership traits?

There are two variables involved (gender and leadership).

There are two independent categories under the gender variable (men and women).

There are eight independent categories under the leadership traits variable (honest, intelligent, hardworking, decisive, ambitious, compassionate, outgoing and creative).

1- The data type is nominal.
2- The data is non-parametric.
3- The data is not normally distributed.
4- Therefore, use Chi-square test-of-association.



Data analysis/hypothesis testing

a) Manual

STEP 1

Determine hypothesis null

There is no statistically significant difference between the observed survey results of leadership traits between men and women and the expected survey results of leadership traits between men and women.

STEP 2

Determine hypothesis alternative

There is a statistically significant difference between the observed survey results of leadership traits between men and women and the expected survey results of leadership traits between men and women.

STEP 3

Choose alpha level

Since the alpha level is not stated in case description, so we choose alpha level: 0.05

STEP 4

Calculate test statistics

$$f_{ij} = \frac{n_i . n_j}{N}$$

Where i=row, j = column,

ni=total of row I; nj=total of row j

N = total of all frequencies.

*Expected frequency for respondents saying men are more honest than women.

i = 1 j = 1

$f_{1.1}$ = (70*184)/556

= 23.165

*Expected frequency for respondents saying women are more honest than men.

i = 1 j = 2

$f_{1.2}$ = (70*372)/556

= 46.83

	Observed			**Expected**		
	Gender			Gender		
Leadership traits	**Men**	**Women**	**Total**	**Men**	**Women**	**Total**
Honest	20	50	70	*23.165	**46.835	70
Intelligent	14	38	52	17.209	34.791	52
Hardworking	28	28	56	18.532	37.468	56
Decisive	44	33	77	25.482	51.518	77
Ambitious	34	34	68	22.504	45.496	68
Compassionate	5	80	85	28.129	56.871	85
Outgoing	28	47	75	24.820	50.180	75
Creative	11	62	73	24.158	48.842	73
Total	184	372	556	184	372	556

X^2 =
$(20-23.165)^2/23.165 + (50-46.835)^2/46.835$
$+ (14-17.209)^2/17.209 + (38-34.791)^2/34.791$
$+ (28-18.532)^2/18.532 + (28-37.468)^2/37.468$
$+ (44-25.482)^2/25.482 + (33-51.518)^2/51.518$
$+ (34-22.504)^2/22.504 + (34-45.496)^2/45.496$
$+ (5-28.129)^2/28.129 + (80-56.871)^2/56.871$
$+ (28-24.820)^2/24.820 + (47-50.180)^2/50.180$
$+ (11-24.158)^2/24.158 + (62-48.842)^2/48.842$
$X^2 = 77.407$

STEP 5

Determine critical value

df=(r-1)(c-1)

r=2

c=8

df=(2-1)(8-1)=(1)(7)=7

Therefore, df=7.

From the X^2 distribution table, the critical value is 14.07.

STEP 6

Interpret results

$X^2_{\text{calculated}} > X^2_{\text{critical}}$. Therefore, reject H_0.

STEP 7

Conclude findings

There is a statistically significant difference between the observed survey results of leadership traits between men and women and the expected survey results of leadership traits between men and women.

There is an association between gender and leadership traits

b) SPSS Analysis

STEP 1

Plot data

Gender	Traits	Frequency
Men	Honest	20
Women	Honest	50
Men	Intelligent	14
Women	Intelligent	38
Men	Hardworking	28
Women	Hardworking	28
Men	Decisive	44
Women	Decisive	33
Men	Ambitious	34
Women	Ambitious	34
Men	Compassionate	5
Women	Compassionate	80
Men	Outgoing	28
Women	Outgoing	47
Men	Creative	11
Women	Creative	62

STEP 2

STEP 3

Analyze Data

Chi-square = 77.407

df = 7

From the X^2 distribution table, the critical value is 14.07.

STEP 5

Interpret results

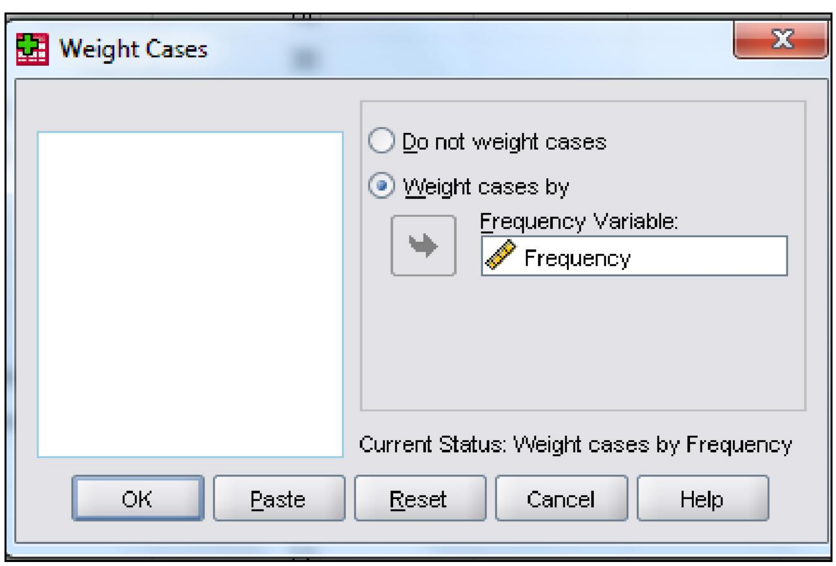
$X^2_{\text{calculated}} > X^2_{\text{critical}}$. Therefore, reject H_0.

STEP 6

Conclude findings

There is a statistically significant difference between the observed survey results of leadership traits between men and women and the expected survey results of leadership traits between men and women.

There is an association between gender and leadership traits.



Traits		Men	Women	Total
Ambitious	Count	34	34	68
	Expected Count	22.5	45.5	68.0
Compassionate	Count	5	80	85
	Expected Count	28.1	56.9	85.0
Creative	Count	11	62	73
	Expected Count	24.2	48.8	73.0
Decisive	Count	44	33	77
	Expected Count	25.5	51.5	77.0
Hardworking	Count	28	28	56
	Expected Count	18.5	37.5	56.0
Honest	Count	20	50	70
	Expected Count	23.2	46.8	70.0
Intelligent	Count	14	38	52
	Expected Count	17.2	34.8	52.0
Outgoing	Count	28	47	75
	Expected Count	24.8	50.2	75.0
Total	Count	184	372	556
	Expected Count	184.0	372.0	556.0

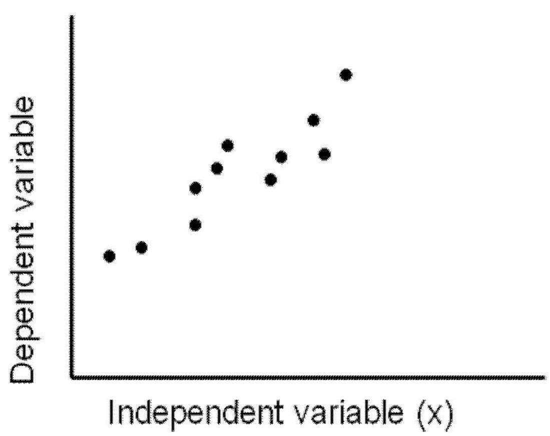
2. Introduction To Regression Tests

2.1 Definition of Regression

A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

2.2 Explanation about Concept of Regression

The two basic types of regression are linear regression and multiple regression. Linear regression uses one independent variable to explain and/or predict the outcome of Y, while multiple regression uses two or more independent variables to predict the outcome.



2.3 Regression Formula

Linear Regression: $Y = a + bX + u$

Multiple Regression: $Y = a + b_1X_1 + b_2X_2 + B_3X_3 + ... + B_tX_t + u$

Where:

Y= the variable that we are trying to predict

X= the variable that we are using to predict Y

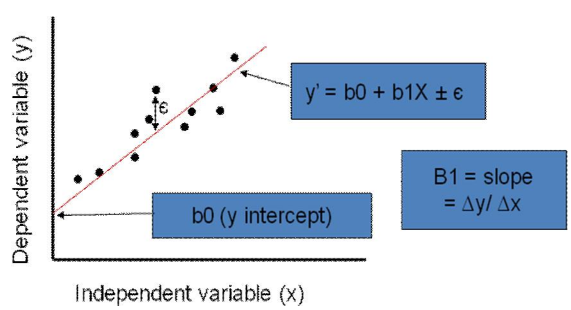
a= the intercept

b= the slope

u= the regression residual.

In multiple regression the separate variables are differentiated by using subscripted numbers.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. Regression is often used to determine how much specific factors such as the price of a commodity, interest rates, particular industries or sectors influence the price movement of an asset.



2.4 Regression Example: To find the Simple/Linear Regression of

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation.

Step 1: Count the number of values.

N = 5

Step 2: Find XY, X2

See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3: Find ΣX, ΣY, ΣXY, ΣX2.

ΣX = 311

ΣY = 18.6

ΣXY = 1159.7

ΣX2 = 19359

Step 4: Substitute in the above slope formula given.

Slope (b) = (NΣXY - (ΣX)(ΣY)) / (NΣX2 - (ΣX)2)=

((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)2)

= (5798.5 - 5784.6)/(96795 - 96721)

= 13.9/74

= 0.19

Step 5: Now, again substitute in the above intercept formula given.

Intercept (a) = (ΣY - b(ΣX)) / N

= (18.6 - 0.19(311))/5

= (18.6 - 59.09)/5

= -40.49/5

= -8.098

Step 6: Then substitute these values in regression equation formula

Regression Equation(y) = a + bx

= -8.098 + 0.19x.

Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value in the above equation.

Regression Equation(y) = a + bx

= -8.098 + 0.19(64).

= -8.098 + 12.16

= 4.06

Mini Research Case (Manually)

A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums.

Driving Experience (years)	Monthly Auto Insurance Premium
5	$64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

a. Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? Do you expect a positive or a negative relationship between these two variables?

b. Compute SS*xx*, SS*yy*, and SS*xy*.

c. Find the least squares regression line by choosing appropriate dependent and independent variables based on your answer in part a.

d. Interpret the meaning of the values of *a* and *b* calculated in part c.

e. Plot the scatter diagram and the regression line.

f. Calculate *r* and *r*2 and explain what they mean.

g. Predict the monthly auto insurance premium for a driver with 10 years of driving experience.

h. Compute the standard deviation of errors.

i. Construct a 90% confidence interval for B.

j. Test at the 5% significance level whether *B* is negative.

Solution (Manually)

a. Based on theory and intuition, we expect the insurance premium to depend on driving experience. Consequently, the insurance premium is a dependent variable and driving experience is an independent variable in the regression model. A new driver is considered a high risk by the insurance companies, and he or she has to pay a higher premium for auto insurance. On average, the insurance premium is expected to decrease with an increase in the years of driving experience. Therefore, we expect a negative relationship between these two variables. In other words, both the population correlation coefficient ρ and the population regression slope B are expected to be negative.

From the *t* distribution table, the *t* value for .05 area in the right tail of the *t* distribution and 6 *df* is 1.943. The 90% confidence interval for *B* is

Experience x	Premium y	xy	x^2	y^2
5	64	320	25	4096
2	87	174	4	7569
12	50	600	144	2500
9	71	639	81	5041
15	44	660	225	1936
6	56	336	36	3136
25	42	1050	625	1764
16	60	960	256	3600
$\Sigma x = 90$	$\Sigma y =$ 474	$\Sigma xy =$ 4739	$\Sigma x^2 =$ 1396	$\Sigma y^2 =$ 29,642

The value of a = 76.6605 gives the value of ŷ for x = 0; that is, it gives the monthly auto

Insurance premium for a driver with no driving experience. The value of b gives the change in ŷ due to a change of one unit in x. Thus, b = −1.5476 indicates that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by $1.55. Note that when b is negative, y decreases as x increases.

The scatter diagram and the regression line for the data on eight auto drivers. Note that the regression line slopes downward from left to right. This result is consistent with the negative relationship we anticipated between driving experience and insurance premium.

The value of r = −.77 indicates that the driving experience and the monthly auto insurance

Premiums are negatively related. The (linear) relationship is strong but not very strong. The value of r2 = .59 states that 59% of the total variation in insurance premiums is explained by years of driving experience and 41% is not. The low value of r2 indicates that there may be many other important variables that contribute to the determination of auto insurance premiums. For example, the premium is expected to depend on the driving record of a driver and the type and age of the car.

To construct a 90% confidence interval for B, first we calculate the standard deviation of b:

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{10.3199}{\sqrt{383.5000}} = .5270$$

$$\alpha / 2 = (1 - .90) / 2 = .05$$

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

$$b \pm ts_b \quad = -1.5476 \pm 1.943(.5270)$$

$$= -1.5476 \pm 1.0240 = \mathbf{-2.57 \text{ to } -.52}$$

We perform the following five steps to test the hypothesis about *B*.

Step 1. *State the null and alternative hypotheses.* The null and alternative hypotheses are written as follows:

$$H_0 : B = 0 \quad (B \text{ is not negative})$$

$$H_1 : B < 0 \quad (B \text{ is negative})$$

Note that the null hypothesis can also be written as $H0$: $B \geq 0$.

Step 2. *Select the distribution to use.*

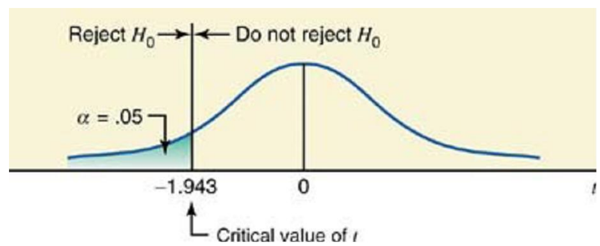
Use the t distribution to make the hypothesis test.

Step 3. *Determine the rejection and no rejection regions*. The significance level is .05. The < sign in the alternative hypothesis indicates that it is a left-tailed test.

$$\text{Area in the left tail of the } t \text{ distribution} = \alpha = .05$$

$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, the critical value of t for .05 area in the left tail of the t distribution and 6 df is −1.943.



Step 4. *Calculate the value of the test statistic.* The value of the test statistic t for b is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{-1.5476 - \overset{\text{From } H_0}{0}}{.5270} = -2.937$$

Step 5. *Make a decision*

The value of the test statistic $t = -2.937$ falls in the rejection region. Hence, we reject the null hypothesis and conclude the *B is negative.*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.768[a]	.590	.521	10.320

a. Predictors: (Constant), drivingexp

3. Conclusion

In this paper researchers talk about two main tests CHI square and regression test with few case studies and explain with SPSS and manual calculations, These Statistical test are use for estimation and predictions in quantitative researches. Mostly management and education students use to work with these statistical tests. In this research paper researchers took two real time mini researches and real data for the understanding of actual concept of CHI square and Regression. This research is designed for enhanced outcome in the performance of statistical students they are working on estimation based researches. From this article student can easily understand the perception of estimation and prediction test for statistical analysis. From the examples and cases which are mentioned above will be helpful to new researchers to understand the concept and implementation of these tests.

Correspondence to:
Engr.Muhammad Mujtaba Asad
FPTV,Universisti Tun Hussein Onn Malaysia.
Telephone: +60187600543
Emails: mujtabaasad11@gmail.com

References

1. Duke, J. D. (1978), "Tables to Help Students Grasp Size Differences in Simple Correlations," Teaching of Psychology, 5, 219-221.
2. FitzPatrick, P. J. (1960), "Leading British Statisticians of the Nineteenth Century," Journal of the American Statistical Association*, 55, 38-70.*
3. Galton, F. (1894), *Natural Inheritance* (5th ed.), New York: Macmillan and Company.
4. Karylowski, J. (1985), "Regression Toward the Mean Effect: No Statistical Background Required," *Teaching of Psychology*, 12, 229-230.
5. Paul, D. B. (1995), *Controlling Human Heredity, 1865 to the Present*, Atlantic Highlands, N.J.: Humanities Press.
6. Pearson, K. (1896), "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia," *Philosophical Transactions of the Royal Society of London*, 187, 253-318.
7. Pearson, K. (1922), *Francis Galton: A Centenary Appreciation*, Cambridge University Press.
8. Miller F. Composting of municipal solid waste and its components. Microbiology of Solid Waste. CRC Press: 1996:116-45.
9. Pearson, K. (1930), *The Life, Letters and Labors of Francis Galton*, Cambridge University Press.Hammelers H. A Theoretical Model of Composting Kinetics. Science and Engineering of Composting. The Ohio State University. 1993:37-58.
10. Williams, R. H. (1975), "A New Method for Teaching Multiple Regression to Behavioral Science Students," *Teaching of Psychology*, 2, 76-78.

3/5/2014