Statistical Techniques to Explore Relationships among Variables

By

Dr. Mojgan Afshari

Agenda

- Pearson Product-Moment Correlation
- Simple Liner Regression
- Chi-Square Test for Independence

Pearson Product-Moment Correlation

- Purpose – **determine relationship between two metric variables**
- Requirement:

  DV -Interval/Ratio

  IV -Interval/Ratio

Assumptions

1. Level of measurement

(IV and DV should be interval/ ratio)

2. Independence of observations

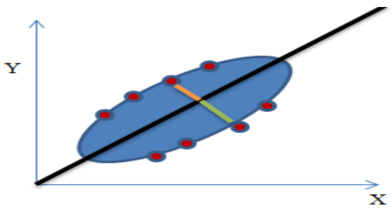
3. *Normality*

4. Linearity

- The relationship between the two variables should be linear. This means that when you look at a scatterplot of scores you should see a straight line (roughly), not a curve.

5. Homoscedasticity

- The variability in scores for variable X should be similar at all values of variable Y. Check the scatterplot. It should show a fairly even cigar shape along its length



Components of Pearson *r* analysis

Pearson corr. coefficient, *r* → Direction; Strength/ Magnitude

↑ *Descriptive*

↓ *Inferential*

Hypothesis Test:

H_O: $\rho = 0$

H_A: $\rho \neq 0$

$\rho > 0$

$\rho < 0$

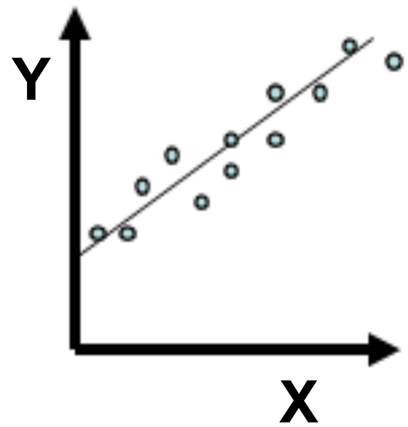
} Choose it

Note

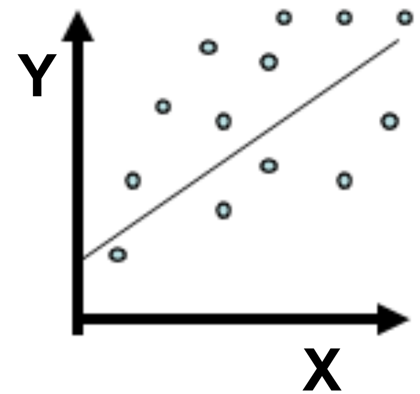
- Before performing a correlation analysis, it is a good idea to generate a **scatterplot**. This enables you to check for violation of the assumptions of linearity and homoscedasticity.

- Inspection of the scatterplots also gives you a better idea of the nature of the relationship between your variables.

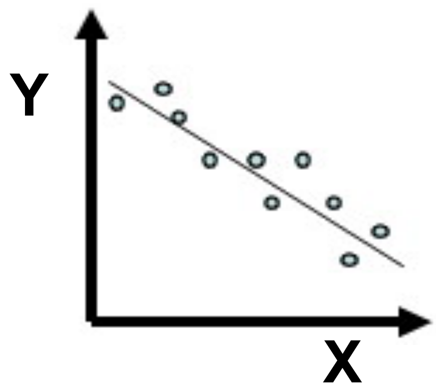
Scatter plot/ Scatter gram



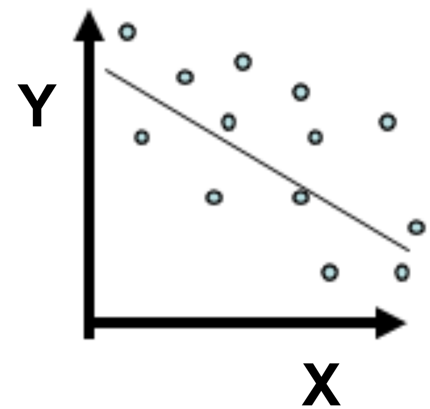
Strong positive correlation



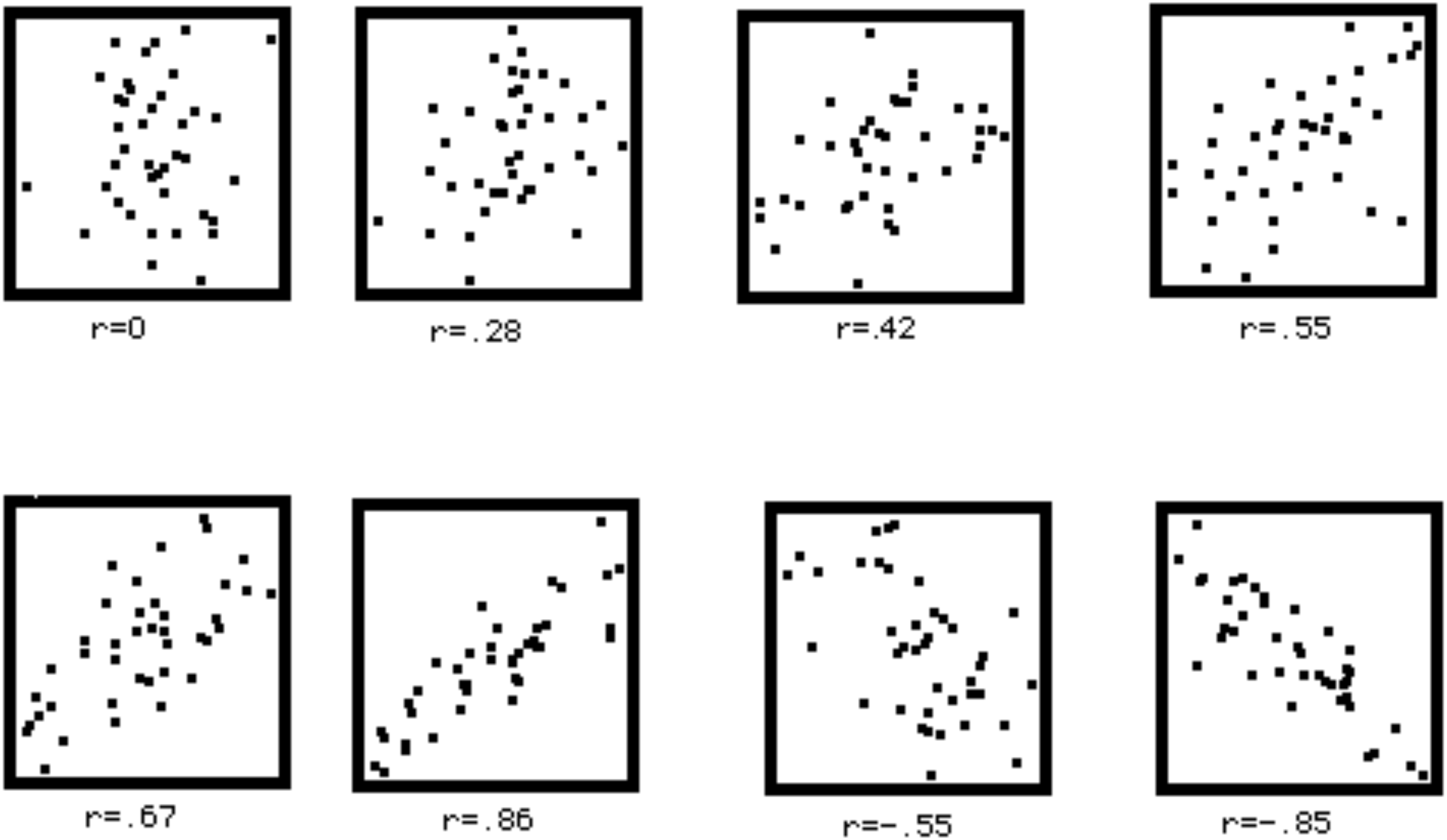
Weak positive correlation



Strong negative correlation



Weak negative correlation


r=0
r=.28
r=.42
r=.55
r=.67
r=.86
r=-.55
r=-.85

- **Pearson correlation coefficients** (r) can only take on values from -1 to + l.

$$-1 \leq r \leq 1$$

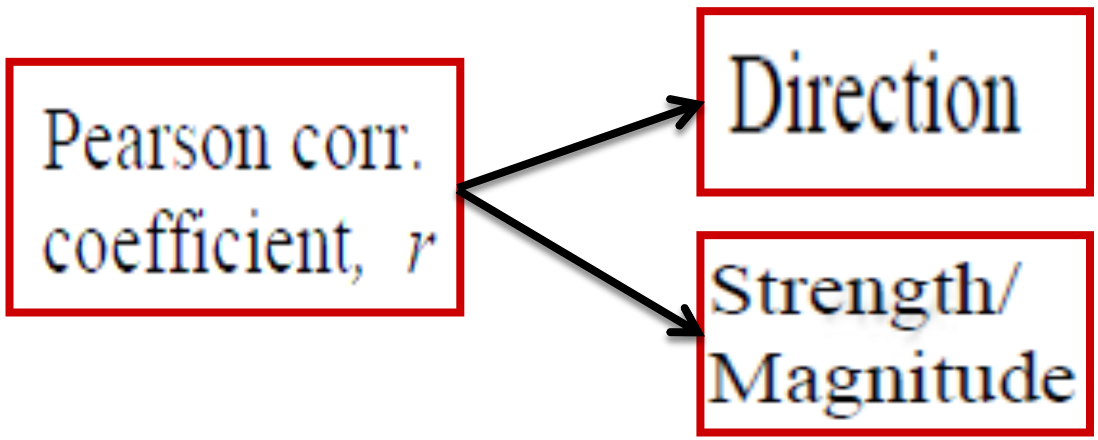
- The sign out the front indicates whether there is a positive correlation (an increase in X will also increase in the value of Y) or a negative correlation (as one variable increases, the other decreases).

- The **size of the absolute value** provides an indication of the **strength of the relationship**.

- A perfect correlation of 1 or -1 indicates that the value of one variable can be determined exactly by knowing the value on the other variable. A scatterplot of this relationship would show a straight line.

- A correlation of 0 indicates no relationship between the two variables.

1. Descriptive



Guildford Rule of Thumb

r	*Strength of Relationship*
< .2	Negligible Relationship
.2 - .4	Low relationship
.4 - .7	Moderate relationship
.7 - .9	High relationship
> .9	Very high relationship

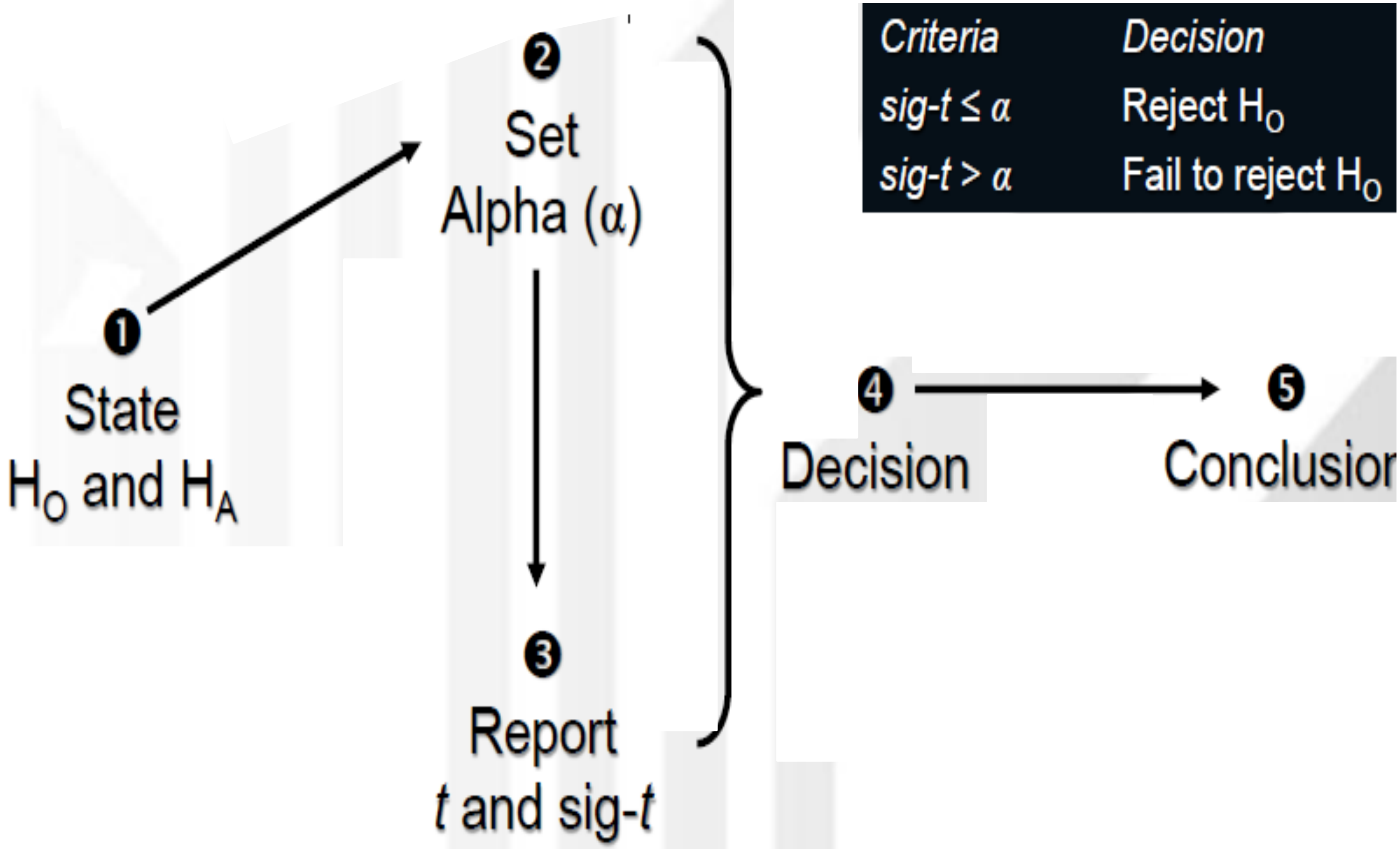
2. Inferential

Hypothesis Test

Excel

What to Expect?

Hypothesis Test



Criteria	*Decision*
sig-t ≤ *α*	Reject H_O
sig-t > *α*	Fail to reject H_O

Steps in Hypothesis Testing

1. State the null and alternative hypotheses

$$H_O: \rho = 0$$
$$H_A: \rho \neq 0$$
$$\rho < 0$$
$$\rho > 0$$

2. Set confidence interval

Generally, confident level is set at .05

Step 3: Report *t* and sig-*t*

Simply report:

1. *t*
2. sig-*t*

Step 4: Decision

- Only two (2) possible decisions.
- Reject or Fail to Reject H_O

Reject H_O: sig-$t \leq \alpha$

Fail to reject H_O: sig-$t > \alpha$

Criteria	*Decision*
sig-$t \leq \alpha$	Reject H_O
sig-$t > \alpha$	Fail to reject H_O

Step 5: Conclusion

Reject H_0

It can be concluded that there is not significant relationship between IV and DV at 0.05 level of significance

Fail to reject H_0

It can be concluded that there is significant relationship between IV and DV at 0.05 level of significance

Correlation using Excel:
How to run a correlation analysis using Excel
and write up the findings for a report

Exercise

1. Data were collected from a randomly selected sample to determine relationship between average assignment and test scores in statistics. Distribution for the data is presented in the table below. Assuming the data are normally distributed,

Data set:

Assign	Test
8.5	88
6	66
9	94
10	98
8	87
7	72
5	45
6	63
7.5	85
5	77

1) **Plot a scatter diagram** to represent the following pair of scores of the two variables.
2) Calculate an appropriate **correlation coefficient**,
3) Describe the nature of relationship between the two variables, and
4) Test the hypothesis on the relationship at **.01 level of significance.**

2) Explain the following concept. You may use graphs to illustrate each concept

a) Perfect positive linear correlation
b) Perfect negative linear correlation
c) Strong positive linear correlation
d) Strong negative linear correlation
e) Weak positive linear correlation
f) Weak negative linear correlation
g) No linear correlation

3) For a sample data set, the linear correlation coefficient r has a positive value.

Which of the following is true about the slope b of the regression line estimated

for the same sample data?

a) The value of b will be positive

b) The value of b will be negative

c) The value of b can be positive or negative

- 3) The data on ages (in years) and prices (in hundred of dollars for eight cars of a specific model) are shown below:

Age:	**8**	**3**	**6**	**9**	**2**	**5**	**6**	**3**
Prices:	**18**	**94**	**50**	**21**	**145**	**42**	**36**	**99**

a) Do you expect the ages and prices of cars to be positively or negatively related? Explain.

b) Calculate the linear correlation coefficient.

c) Test at the 5% significance level whether ρ is negative

4) The following table lists the Advertising and Marketing scores for 7 students in a statistics class.

- **Advertising score: 79 95 81 66 87 94 59**
- **Marketing score: 85 97 78 76 94 84 67**

a) Do you expect the Advertising and Marketing scores to be positively or negatively correlated?

b) Plot a scatter diagram. By looking at the scatter diagram, do you expect the correlation coefficient between these 2 variables to be close to zero, 1, or -1.

c) Find the correlation coefficient. Is the value of r consistent with what you expected in parts a and b?

d) Using the 1% significance level, test whether the linear correlation coefficient is Positive

5) X= Price

Y= Sale

Correlations

		X	Y
X	Pearson Correlation	1	-.283**
	Sig.(2-tailed)	.	.004
	N	100	100
Y	Pearson Correlation	-.283**	1
	Sig.(2-tailed)	.004	.
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

a) State a research question appropriate for this analysis.

b) What are the variables involved in this analysis? State their respective scale of measurement.

c) Report and describe the coefficient from the analysis.

d) Test the relationship between the variables at .01 level of significance.

i. State the null and alternative hypotheses

ii. What would be your decision and justify your answer

iii. What can you conclude?

Simple Linear Regression

Introduction

Purpose

- **To determine relationship between IV and DV**
- **To predict value of the dependent variable (Y) based on value of independent variable (X)**
- **To assess how well the dependent variable can be explained by knowing the value of the independent variable**

Requirement

- Scales of measurement for variables:

DV -interval or ratio

IV -interval or ratio

Concepts of

Simple Linear Regression

Refer to Coefficient table → Prediction Equation

$$\hat{Y} = b_0 + b_1 X$$

R and R^2

↑ *Descriptive*

↓ *Inferential*

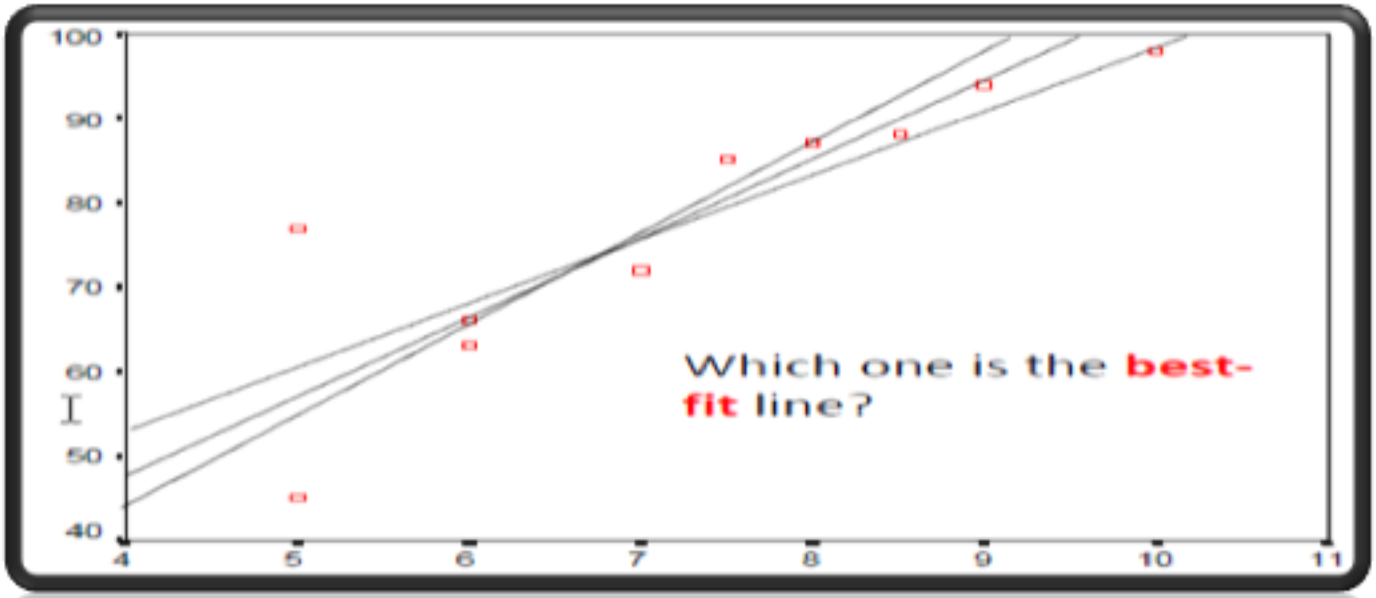
Hypothesis Test:

- Regression Model
- Slope

Descriptive

Bases for Best Fit Line

- Use Scatter Plot to identify the line of best fit
- The line is also called the **least squares regression line**
- The purpose of this line is to show the overall trend or pattern in the data and to allow the reader to make predictions about future trends in the data.



Prediction equation

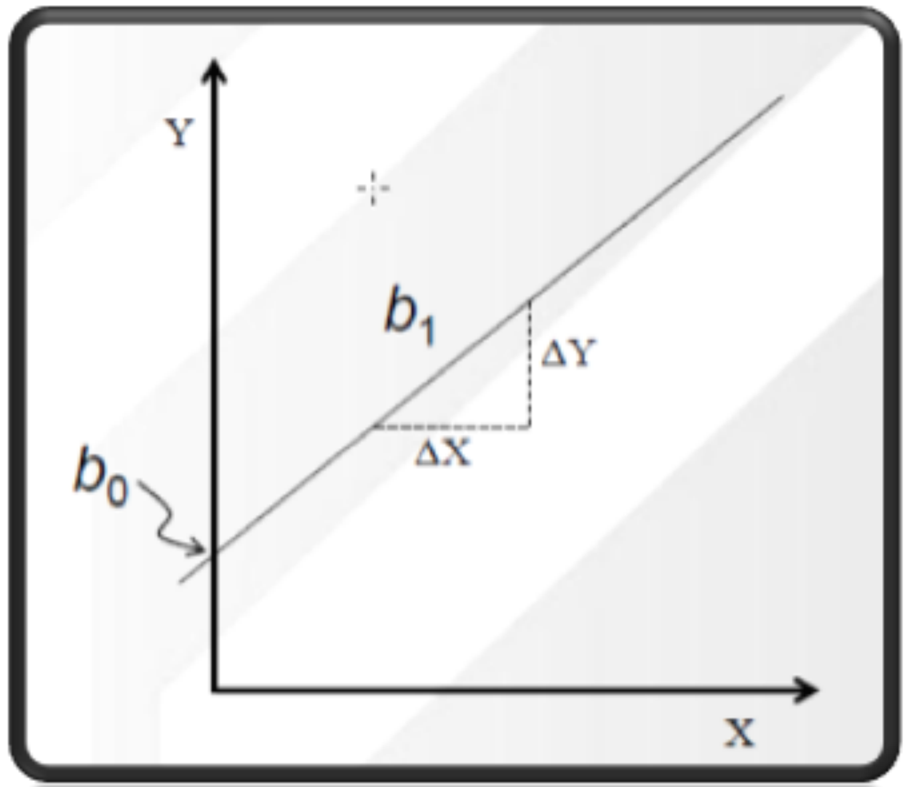
Regression Model involves two statistics and parameters, namely: Intercept and slope of the regression line.

$$\hat{Y} = b_0 + b_1 X$$

$\hat{Y}$ Predicted value of Y

b_0 Y-intercept

b_1 Slope (regression coefficient)



Inferential

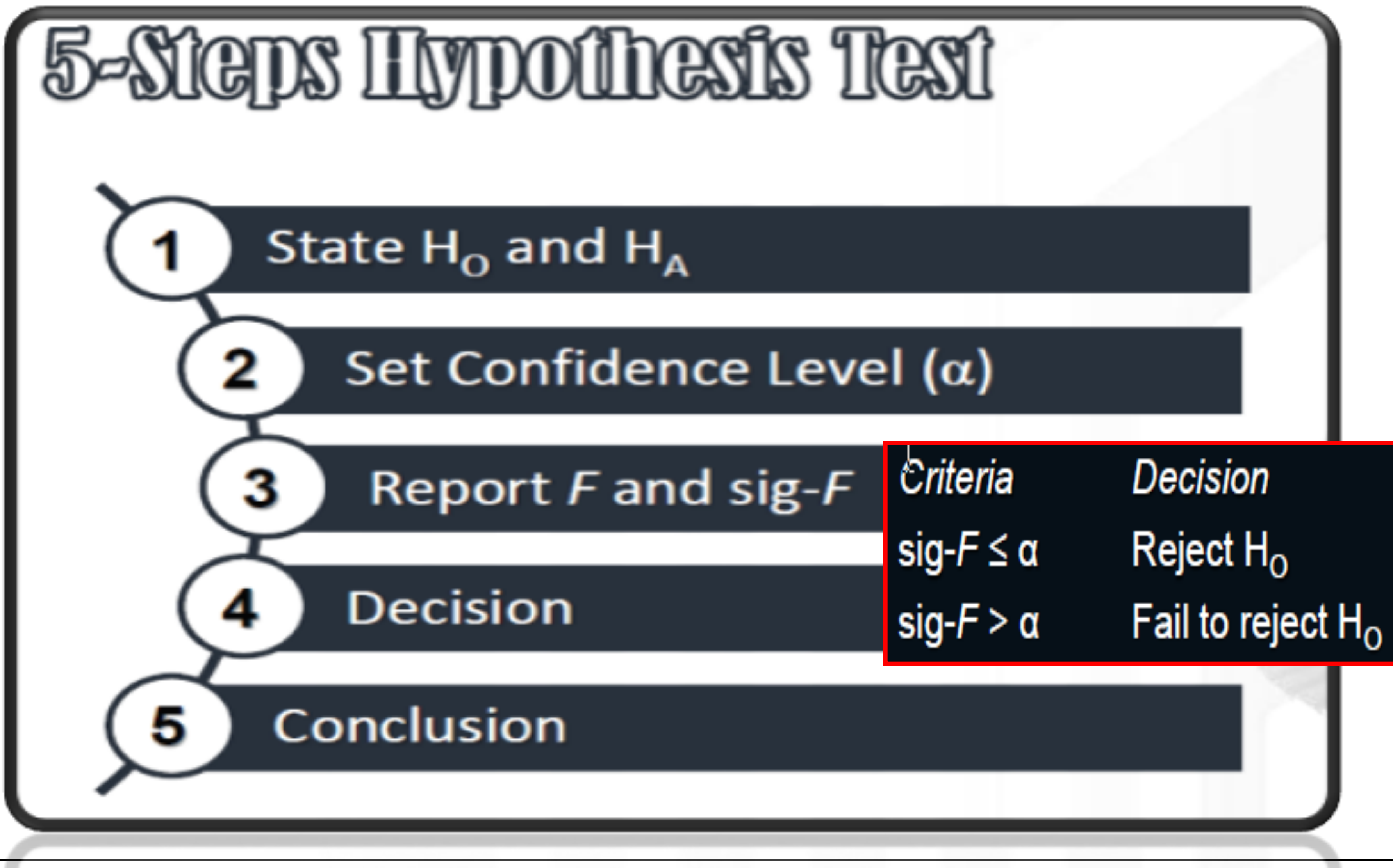
- Evaluation of **the regression model (prediction equation)** and **slope of the regression line** involve testing of hypothesis.

- The testing of hypothesis for **prediction equation** involves the **F test** while the testing of **slop** entails **t test**

Hypothesis Test:

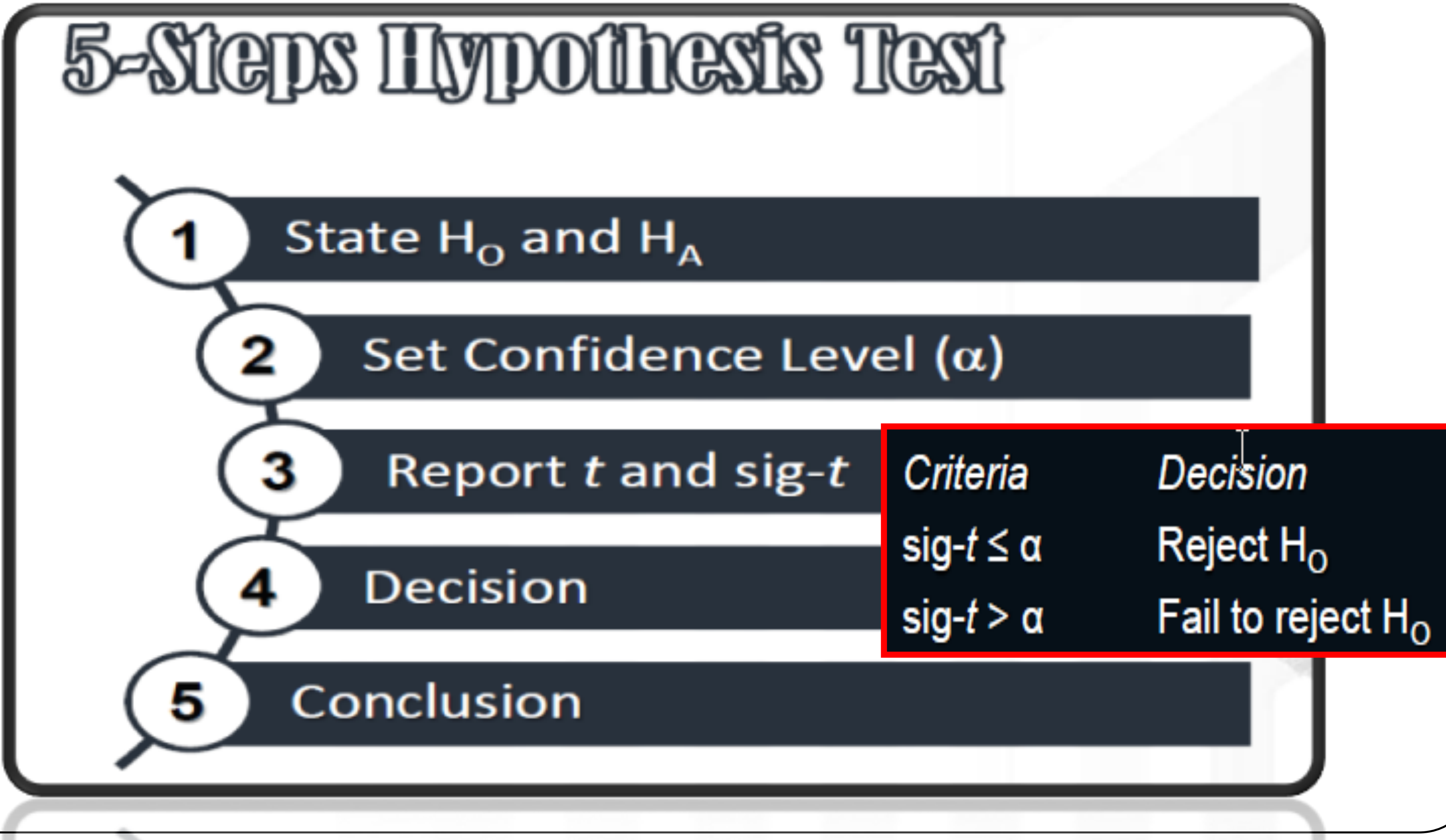
Regression Model

- The purpose of evaluating the regression model or the prediction equation is to determine whether the **model really fits the data.**



Hypothesis Test: Slope

The purpose to test hypothesis regarding the regression slope is to determine **the significance of the relationship between IV and DV.**



Example

Data were collected from a randomly selected sample to determine relationship between average assignment scores and test scores in statistics. Distribution for the data is presented in the table below.

1. Calculate b_1 and b_0 and derive the prediction equation
2. Test the hypothesis for the regression model at α= .05
3. What the values of coefficient of determination (R^2) and multiple correlation coefficient (r). Interpret the two values.
4. Test hypothesis for the slope at .05

Data set:

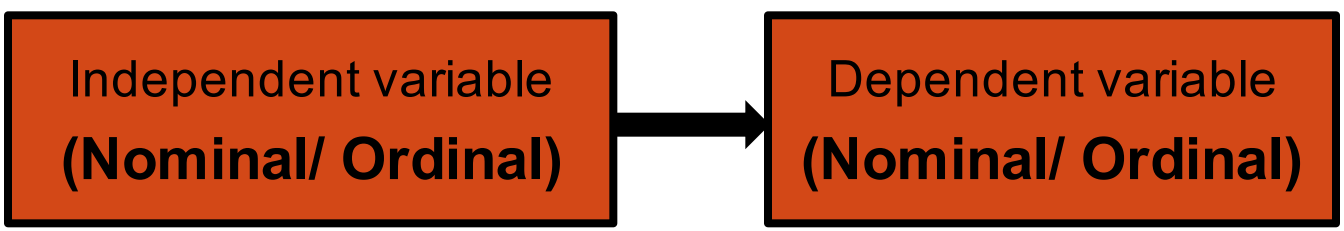
	Scores	
ID	*Assign*	*Test*
1	8.5	88
2	6	66
3	9	94
4	10	98
5	8	87
6	7	72
7	5	45
8	6	63
9	7.5	85
10	5	77

Chi-Square

Test for Independence

Chi-square test for independence

- Explore the association between ***two*** **categorical variables**. Each of these variables can have two or more categories.



Example

- **To explore the association between Gender (Male / Female) and Smoking Behaviour (Smoker/Non-Smoker).**

- **IV: Gender (Male/ Female)**
  DV: Smoking Behaviour (Smoker / Non Smoker)

- **Research questions:**

There are a variety of ways questions can be phrased:

1. Is there an association between gender and smoking behaviour?
2. Are males more likely to be smokers than females?
3. Is the proportion of males that smoke the same as the proportion of females?

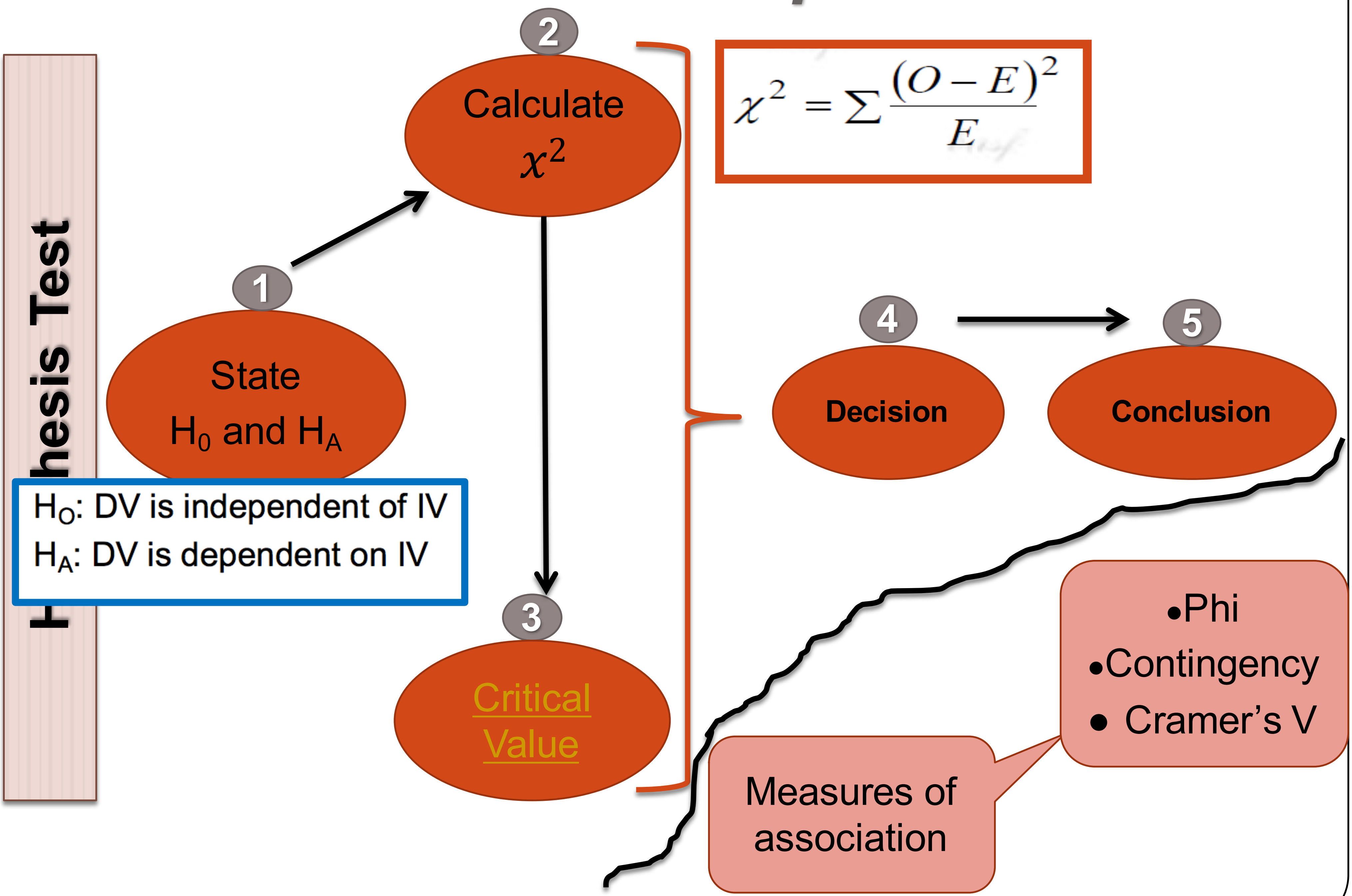
Minimum Expected Cell Frequency Assumption

The lowest expected frequency in any cell should be 5 or more. Some authors suggest less stringent criteria: at least 80 per cent of cells should have expected frequencies of 5 or more.

We could not use x^2 for the following case:

	OBSERVED	EXPECTED
AGREE	247	255
DISAGREE	6	4

What to Expect?



Step in Testing Hypothesis

1. **State the null and alternative hypotheses**

   H_O: DV is independent of IV

   H_A: DV is dependent on IV

2. **Calculate the test statistic**

   x^2 value

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

- Contingency table –two variables
- Calculation based on:

  O-Observed frequency

  E-Expected frequency

$$E = \frac{RT \times CT}{GT}$$

- Summary of the table to calculate the chi-square statistics

$$E = \frac{RT \times CT}{GT}$$

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

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

3. **Determine critical value**

- α
- $df = (R-1)(C-1)$

4. Make your decision

5. Make conclusion

Manual

Criteria	Decision
$\chi^2_{cal} > \chi^2_{critical}$	Reject H_o
$\chi^2_{cal} \le \chi^2_{critical}$	Fail to reject H_o

Excel

Criteria	Decision
Sig-$\chi^2 < \alpha$	Reject H_o
Sig-$\chi^2 \ge \alpha$	Fail to reject H_o

Effect Size

- **To determine the strength and magnitude of the association between two variables.**

Effect size statistics:

- **Phi coefficient (2 by 2 tables)**
- **Cramer's V (Tables larger than 2 by 2)**
- **Contingency coefficient (Tables larger than 2 by 2)**

Contingency $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Phi φ $$\varphi = \sqrt{\frac{\chi^2}{n}}$$

Cramer's V $$V = \sqrt{\frac{\chi^2}{n \cdot df^*}}$$

Criteria for judging the Effect Size

*df**	*small*	*medium*	*large*
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

Example 1:

- A study was conducted to test the association between Firm size and cloud computing adoption. Data collected from a randomly selected sample follow.
- 1. Test the hypothesis on the association between the two variables at .01 level of significance.
- 2. Calculate and describe an appropriate measure of

Firm Size	Cloud computing adoption		
	High	Moderate	Low
Large	93	70	12
Small	87	32	6

1. Hypotheses testing

a. State H_0 and H_A

- H_0: Cloud computing adoption is independent of Firm size
- H_A: Cloud computing adoption is dependent on Firm size

b. Test statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$

Calculate expected value for each cell: $E = \frac{RT \times CT}{GT}$

Firm Size	Cloud computing adoption			Row Totals
	High	Moderate	Low	
Large	93 (105.0)	70 (59.5)	12 (10.5)	175
Small	87 (75.0)	32 (42.5)	6 (7.5)	125
Column Totals	180	102	18	300

Group	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
LH	93	105.0	-12	144	1.371
LM	70	59.5	10.5	110.25	1.853
LL	12	10.5	1.5	2.25	.214
SH	87	75.0	12	144	1.920
SM	32	42.5	-10.5	110.25	2.594
SL	6	7.5	-1.5	2.25	.300
	300				**8.252**

$\chi^2 = 8.252$

c. Critical value

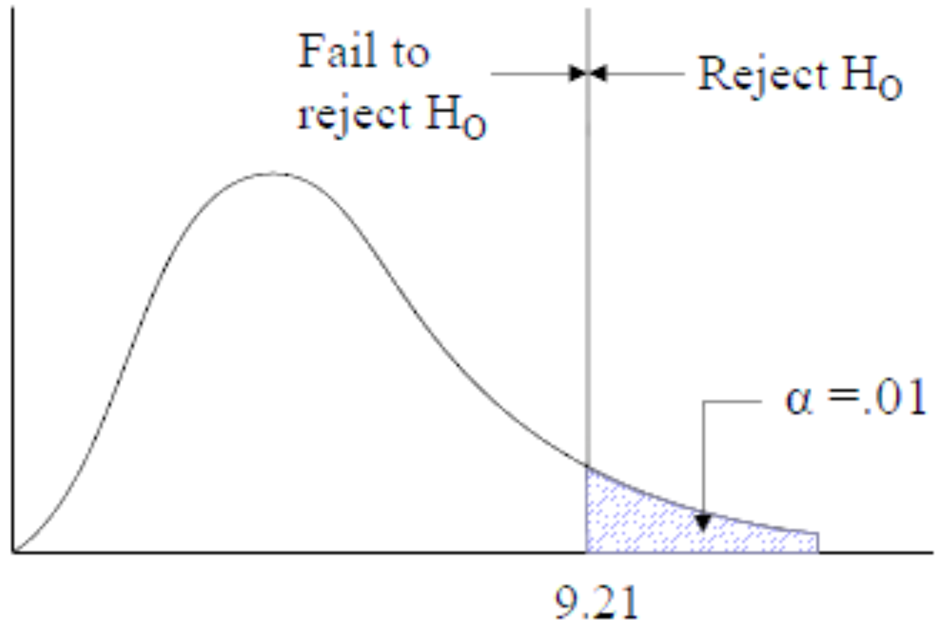
$$df = (R-1)(C-1)$$

$$= (2-1)(3-1)$$

$$= 1 \times 2$$

$$= 2$$

$$\chi^2_{2,.01} = 9.21$$



d. Decision

Since χ^2_{cal} (8.252) < $\chi^2_{critical}$ (9.21)

Fail to reject H_0

e. Conclusion

There is not enough evidence from the sample to conclude that the two variables. **Firm size and Cloud computing adoption** are dependent at .01 level of significance.

Example 2:

Dr Irwan is interested to test the relationship between gender and Online shopping Data taken from a randomly selected sample follow.

1. Test the hypothesis on the relationship at .01 level of significance.
2. Calculate and describe an appropriate measure of association between the two variables

Gender	Online shopping	
	Yes	No
Male	60	110
Female	75	55

a. State H_0 and H_A

H_O: Online shopping is independent of gender

H_A: Online shopping is dependent on gender

b. Test statistic

Calculate expected value for each cell: $E = \frac{RT \times CT}{GT}$

Gender	Online shopping		Row Totals
	Yes	No	
Male	60 (76.5)	110 (93.5)	170
Female	75 (58.5)	55 (71.5)	130
Column Totals	135	165	300

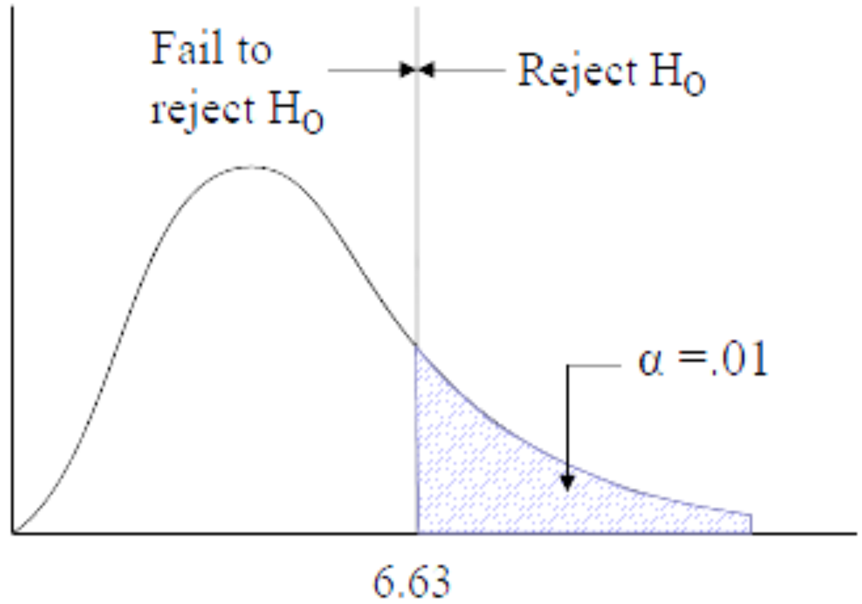
	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
MY	60	76.5	-16.5	272.25	3.558
MN	110	93.5	16.5	272.25	2.911
FY	75	58.5	16.5	272.25	4.653
FN	55	71.5	-16.5	272.25	3.807
					14.929

χ^2 = **14.929**

c. Critical value

$$df = (R-1)(C-1)$$
$$= (2-1)(2-1)$$
$$= 1 \times 1$$
$$= 1$$

$$\chi^2_{1,.01} = 6.63$$



d. **Decision**

Since χ^2_{cal} (14.929) > $\chi^2_{critical}$ (6.63)

∴ Reject H_0

e. Conclusion

There is a strong evidence from the sample to conclude that the two variables, *gender* and **Online shopping** are dependent at .01 level of significance.

2. Measure of association

For a 2 x 2 contingency table, phi coefficients is the most appropriate to be used

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

$$= \sqrt{\frac{14.933}{300}}$$

$$= .223$$

Low association between gender and Online shopping