MAILAM ENGINEERING COLLEGE

Mailam – 604 304

(Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai
& TATA Consultancy Services Accredited Institution)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

II YEAR / III SEM

AD3301 DATA EXPLORATION AND VISUALIZATION

SYLLABUS

UNIT IV

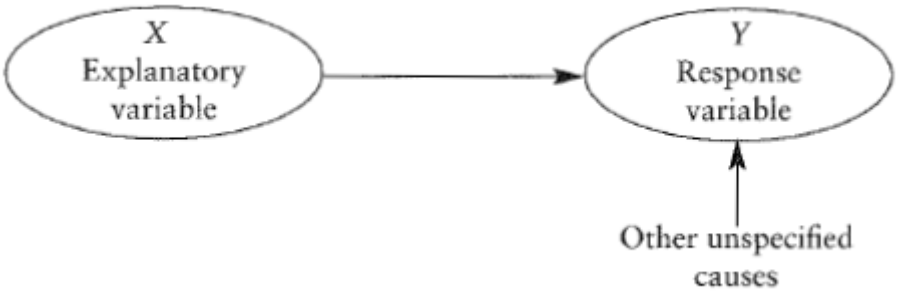
BIVARIATE ANALYSIS

SYLLABUS:

Relationships between Two Variables - Percentage Tables - Analyzing Contingency Tables - Handling Several Batches - Scatterplots and Resistant Lines – Transformations.

PART A

1. **What is bivariate analysis?**
   - Bivariate analysis is one of the statistical analysis where two variables are observed.
   - One variable here is dependent while the other is independent.
   - These variables are usually denoted by X and Y.
2. **What is causal path model?**
   - Causal reasoning is done by the construction of a schematic model of the hypothesized causes and effects: a causal path model.



   - The variables are represented inside boxes or circles and labeled;
   - Arrows run from the variables is the causes to those to be effects;

3. What is Proportions, Percentages and Probabilities?

- Proportion – the number in each category is divided by the total number of cases N.
- Percentages – Proportions multiplied by 100.
- Probabilities – Represent the relative size of different subgroups in a population.

4. Define a Contingency Table.

- A contingency table is a tabular representation of categorical data.
- Contingency table is similar to the three-dimensional bar chart.
- Contingent defines as 'true only under existing or specified conditions'.
- A contingency table displays frequencies for combinations of two categorical variables.
- Analysts also refer to contingency tables as cross tabulation and two-way tables.
- A contingency table shows the distribution of each variable conditional upon each category of the other.

5. What is a percentage table?

- The commonest way to make contingency tables readable is to cast them in **percentage form**.
- There are three different ways in which this can be done.
  - **Total percentages** - The table was constructed by dividing each cell frequency by the grand total.
  - **Row percentages** - The table was constructed by dividing each cell frequency by its appropriate row total. Tables that are constructed by percentaging the rows are usually read down the columns. This is called an 'outflow' table.
  - **Column percentages** - The table was constructed by dividing each cell frequency by its appropriate column total. This is called an 'inflow' table.

6. What are the guidelines for a well designed table?

- Labeling
- Sources
- Sample Data
- Missing Data
- Opinion Data
- Layout

7. What is a Chi-square test?

- Contingency tables can be used to perform a **Chi-square test** to determine whether there is a significant association between the two variables.
- Calculate the Chi-square statistic using the formula:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where:

- O_{ij} is the observed frequency for the combination of categories i and j
- E_{ij} is the expected frequency for the combination of categories i and j
- n is the number of rows • m is the number of columns

8. Define degree of freedom.

- Degrees of freedom are the number of independent values that a statistical analysis can estimate.

9. What is a T- test?

- A T-test is the final statistical measure for determining differences between two means that may or may not be related.
- It is a statistical method in which samples are chosen randomly, and there is no perfect normal distribution.

T-test formula

The formula for a two-sample t-test where the samples are independent (as in the example of boys' and girls' mathematics test scores) is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where X_1 and X_2 are the means of the two samples and S_{X1X2} is known as the pooled standard deviation and is calculated as follows:

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

here S_{X1} is the standard deviation of one sample and S_{X2} is the standard deviation of the other sample. In these formulae n_1 is the sample size of the first sample and n_2 is the sample size of the second sample. In simple terms therefore the size of the t-statistic depends on the size of the difference between the two means adjusted for the amount of spread and the sample sizes of the two samples.

10. What is a resistant line?

Resistant Line

- To explore paired data and to suspect a relationship between X and Y, the focus is on how to fit a line to data in a "resistant" fashion, so the fit is relatively insensitive to extreme points.

Fitting a Resistant Line - Line fitting involves joining two typical points.

- The X-axis is roughly divided into three parts
  - X values are ordered along with its corresponding Y values.
    - Divide X axis to three approximately equal length
    - The left and right should be balanced with equal number of data points
- Conditional Summary points for X and Y are found.
- A Line is drawn connecting a Left and Right Summary points

11. What is meant by log transformation?

- One method for transforming data or re-expressing the scale of measurement is to take the logarithm of each data point.
- This keeps all the data points in the same order but stretches or shrinks the scale by varying amounts at different points.

12. What are the goals of transformation?

- Data batches can be made more symmetrical.
- The shape of data batches can be made more Gaussian.
- Outliers that arise simply from the skewness of the distribution can be removed, and previously hidden outliers may be forced into view'.
- Multiple batches can be made to have more similar spreads.
- Linear, additive models may be fitted to the data.

13. Define and list the ladder of powers.

The ladder of powers

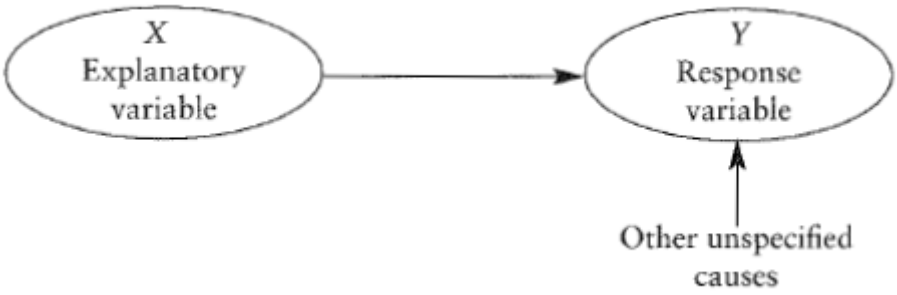
- It is family of power transformations that can help promote symmetry and sometimes Gaussian shape in many different data batches.

Power	Expression	Name
•		
•		
3	X^3	cube
2	X^2	square
1	X^1	raw data
0.5	$\sqrt{X}$	square root
0	X^0	? (log)
–0.5	$1/\sqrt{X}$	reciprocal root
–1	$1/X$	reciprocal
–2	$1/X^2$	reciprocal square
•		
•		

- Going up the ladder of powers corrects downward straggle, whereas going down corrects upward straggle.

PART B

1. **Explain in detail about Relationships between Two Variables (bivariate relationships)**
   - Bivariate Analysis is used to explore the relationship between 2 different variables.
   - Relationships between two variables (bivariate relationships) are one variable can be considered a cause and the other an effect.
   - The variable that is presumed to be the cause the explanatory variable (and denote it X) and the one that is presumed to be the effect the response variable (denoted Y); they are termed independent and dependent variables respectively.
   - Causal reasoning is often assisted by the construction of a schematic model of the hypothesized causes and effects: a causal path model.



   - The variables are represented inside boxes or circles and labeled;
   - Arrows run from the variables is the causes to those to be effects;
   - Positive effects are drawn as unbroken lines and negative effects are drawn as dashed lines.
   - A number is placed on the arrow to denote how strong the effect of the explanatory variable is.
   - An extra arrow is included as an effect on the response variable, often unlabeled,

2. **Explain in detail about Percentage Tables and Contingency Table.**
   CONTINGENCY TABLE
   - A contingency table is a tabular representation of categorical data.
   - Contingency table is similar to the three-dimensional bar chart.
   - Contingent defines as 'true only under existing or specified conditions'.
   - A contingency table displays frequencies for combinations of two categorical variables.
   - Analysts also refer to contingency tables as cross tabulation and two-way tables.

- A contingency table shows the distribution of each variable conditional upon each category of the other.
- Contingency tables classify outcomes for one variable in rows and the other in columns.
- The values at the row and column intersections are frequencies for each unique combination of the two variables.
- Each individual case is then tallied in the appropriate cells depending on its value on both variables. and the number of cases in each cell is called the cell frequency.
- Each row and column can have a total presented at the right-hand end and at the bottom respectively; these are called the marginal, and the univariate distributions can be obtained from the marginal distributions.
- Figure 4.1 shows a schematic contingency table with four rows and four columns (a four-by-four table).

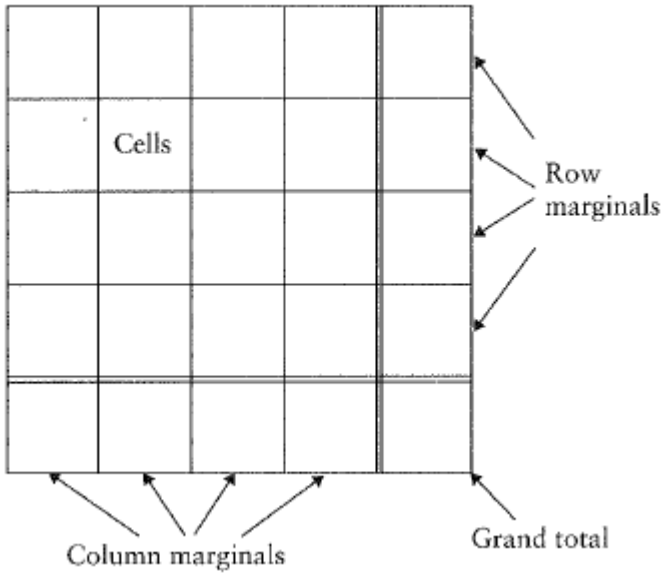


Figure 4.1 - Anatomy of a contingency table.

Finding Relationships in a Contingency Table

- In the contingency table below, the two categorical variables are gender and ice cream flavor preference.
- Below table 4.1 is a two-way table (2 X 3) where each cell represents the number of times males and females prefer a particular ice cream flavor.

Table 4.1 - A two-way table (2 X 3)

Gender	Chocolate	Strawberry	Vanilla	Total
Female	37	17	12	66
Male	21	18	32	71
Total	58	35	44	137

- The contingency table 4.2 below uses the same raw data as the previous table and displays both row and column percentages.

Table 4.2 – Contingency Table

Gender	Chocolate	Strawberry	Vanilla	Row Total
Female	Raw: 37 Row%: 56% Col%: 63.8%	Raw: 17 Row%: 25.8% Col%: 48.6%	Raw: 12 Row%: 18.2% Col%: 28.8%	Raw: 66 Row%: 100%
Male	Raw: 21 Row%:29.6% Col%: 36.2%	Raw: 18 Row%: 25.4% Col%: 51.4%	Raw: 32 Row%: 45.0% Col%: 71.2%	Raw: 71 Row%: 100%
Total	Raw: 58 Col%: 100%	Raw: 35 Col%: 100%	Raw: 44 Col%: 100%	137

Graph a Contingency Table

- Use bar charts to display a contingency table.
- The following figure 4.2 shows the row percentages for the previous two-way table.

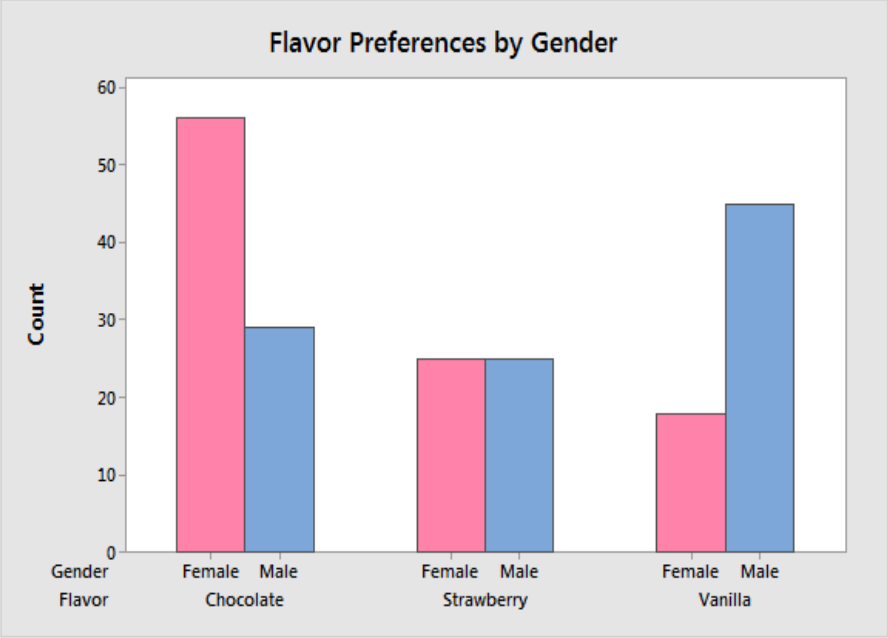


Figure 4.2 - Clustered bar chart

PERCENTAGE TABLES

- The commonest way to make contingency tables readable is to cast them in percentage form.
- There are three different ways in which this can be done.
- Total percentages - The table was constructed by dividing each cell frequency by the grand total.
- Row percentages - The table was constructed by dividing each cell frequency by its appropriate row total. Tables that are constructed by percentaging the rows are usually read down the columns. This is called an 'outflow' table.
- Column percentages - The table was constructed by dividing each cell frequency by its appropriate column total. This is called an 'inflow' table.
- The inflow and outflow tables focus attention on the data in different ways, and the researcher have to be clear about what questions were being addressed in the analysis to inform which way the percentages were calculated.

GOOD TABLE MANNERS

- **Labeling**
  - A clear title should summarize the contents.
  - It should be as short as possible, while at the same time making clear when the data were collected, the geographical unit covered, and the unit of analysis.
  - Other parts of a table also need clear, informative labels.
  - The variables included in the rows and columns must be clearly identified.
- **Sources**
  - The reader needs to be told the source of the data.
  - It is not good enough to say that it was from *Social Trends.* The volume and year, and either the table or page, and sometimes even the column in a complex table must be included.
  - When the data are first collected from a published source, all these things should be recorded, or a return trip to the library will be needed.
- **Sample data**
  - If data are based on a sample drawn from a wider population, it always needs special referencing.
  - The reader must be given enough information to assess the adequacy of the sample.
  - The following details should be available somewhere:
    - The method of sampling
    - the achieved sample size,
    - the response rate or refusal rate,
    - the geographical area which the sample covers
    - the frame from which it was drawn.

- **Missing data**
  - Don't exclude cases from analysis, miss out particular categories of a variable or ignore particular attitudinal items in a set without good reason and without telling the reader what you are doing and why.
- **Definitions**
  - There can be no hard and fast rule about how much definitional information to include in the tables.
  - They could become unreadable if too much were included.
- **Opinion data**
  - When presenting opinion data, always give the exact wording of the question put to respondents, including the response categories if these were read out.
- **Ensuring frequencies can be reconstructed**
  - It should always be possible to convert a percentage table back into the raw cell frequencies.
  - To retain the clarity of a percentage table, present the minimum number of base *Ns* needed for the entire frequency table to be reconstructed.
- **Showing which way the percentages run**
  - Proportions add up to 1 and percentages add up to 100.
- **Layout**
  - The effective use of space and grid lines can make the difference between a table that is easy to read and one which is not.
  - Avoid underlining words or numbers.
  - Clarity is often increased by reordering either the rows or the columns.
    1. Closer figures are easier to compare;
    2. Comparisons are more easily made down a column;
    3. A variable with more than three categories is best put in the rows so that there is plenty of room for category labels

3. Explain in detail about the Analysis of Contingency Tables with example.

Contingency Tables Analysis:

- Contingency tables analysis is a central branch of categorical data analysis and is focused on the analysis of data represented as contingency tables.
- Contingency tables analysis is widely used in marketing research, in biomedical research, including drug trials, and in social sciences.
- A contingency table is a tool used to summarize and analyze the relationship between two categorical variables.
- It is a type of cross-tabulation that displays the frequencies or counts of the combinations of categories for the two variables.
- To create a contingency table, the following steps are typically followed:
  - Identify the two categorical variables to be analyzed.

- Collect and summarize the data. Count the number of observations in each combination of categories for the two variables.
- Organize the data in a table with the categories of the first variable listed along the rows and the categories of the second variable listed along the columns.
- Enter the counts or frequencies in the cells of the table.

Chi-square test of independence hypotheses

- Contingency tables can be used to perform a **Chi-square test** to determine whether there is a significant association between the two variables.
- To do this, the following steps are typically followed:
  - Calculate the expected frequencies for each combination of categories using the formula:

$$E_{ij} = \frac{R_i C_j}{n}$$

- Where:
  - Eij is the expected frequency for the combination of categories i and j
  - Ri is the row total for category i
  - Cj is the column total for category j
  - n is the total sample size
- Calculate the Chi-square statistic using the formula:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Where:
  - O_{ij} is the observed frequency for the combination of categories i and j
  - E_{ij} is the expected frequency for the combination of categories i and j
  - n is the number of rows • m is the number of columns
- Determine the critical value of the Chi-square statistic based on the significance level (alpha) of the test and the degrees of freedom.
  - The degrees of freedom (df): For a chi-square test of independence, the df is (number of variable 1 groups − 1) * (number of variable 2 groups − 1).
  - Significance level (α): By convention, the significance level is usually .05.
- Compare the calculated Chi-square statistic to the critical value to determine whether to reject or fail to reject the null hypothesis.

- Null hypothesis (H0): Variable 1 and variable 2 are not related in the population; The proportions of variable 1 are the same for different values of variable 2.
- Alternative hypothesis (Ha): Variable 1 and variable 2 are related in the population; The proportions of variable 1 are not the same for different values of variable 2.

- If the X2 value is greater than the critical value, then the difference between the observed and expected distributions is statistically significant ($p < \alpha$).

The data allows to reject the null hypothesis that the variables are unrelated and provides support for the alternative hypothesis that the variables are related.

- If the X2 value is less than the critical value, then the difference between the observed and expected distributions is not statistically significant ($p > \alpha$).

The data doesn't allow to reject the null hypothesis that the variables are unrelated and doesn't provide support for the alternative hypothesis that the variables are related.

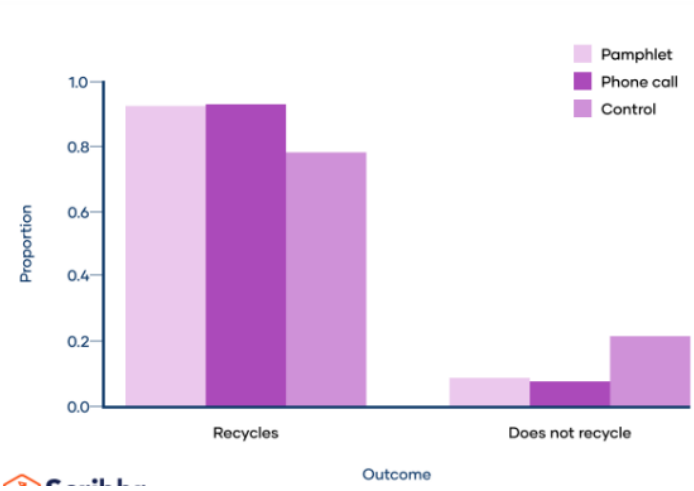
- Contingency tables are also used to analyze the relationship between a categorical variable and a continuous variable.
- In this case, the continuous variable is typically grouped into intervals or categories, and a contingency table is created to summarize the frequencies or counts for each combination of categories.

Example

Six months after the intervention, the city looks at the outcomes for the 300 households (only four households are shown here):

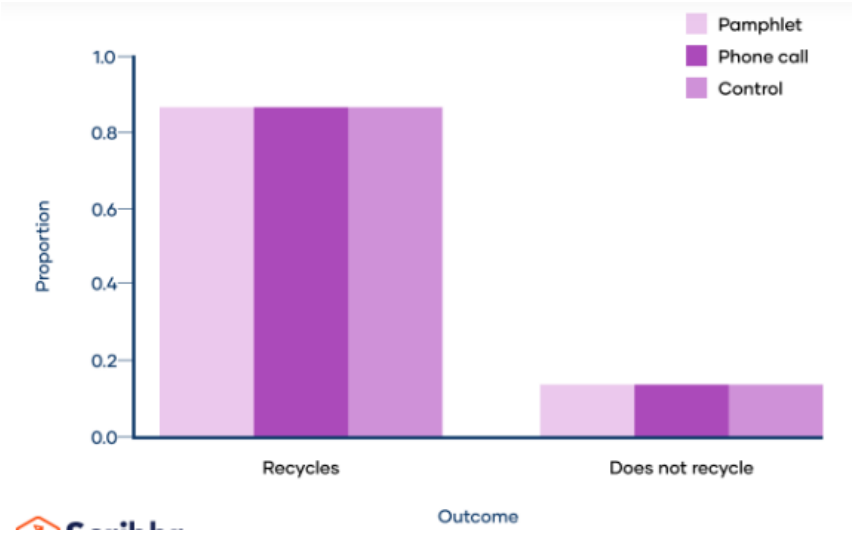
Observed Frequency

Intervention	Recycles	Does not recycle	Row totals
Flyer (pamphlet)	89	9	98
Phone call	84	8	92
Control	86	24	110
Column totals	259	41	*N* = 300



Expected Frequency

Intervention	Recycles	Does not recycle	Row totals
Flyer (pamphlet)	89	9	98
	$\frac{(98 \times 259)}{300} = 84.61$	$\frac{(98 \times 41)}{300} = 13.39$	
Phone call	84	8	92
	$\frac{(92 \times 259)}{300} = 79.43$	$\frac{(92 \times 41)}{300} = 12.57$	
Control	86	24	110
	$\frac{(110 \times 259)}{300} = 94.97$	$\frac{(110 \times 41)}{300} = 15.03$	
Column totals	259	41	$N = 300$



Follow these five steps to calculate the test statistic:

Step 1: Create a table

Create a table with the observed and expected frequencies in two columns.

Example: Step 1

Intervention	Outcome	Observed	Expected
Flyer	Recycles	89	84.61
	Does not recycle	9	13.39
Phone call	Recycles	84	79.43
	Does not recycle	8	12.57
Control	Recycles	86	94.97
	Does not recycle	24	15.03

Step 2: Calculate $O - E$

In a new column called "$O - E$", subtract the expected frequencies from the observed frequencies.

Example: Step 2

Intervention	Outcome	Observed	Expected	$O - E$
Flyer	Recycles	89	84.61	4.39
	Does not recycle	9	13.39	-4.39
Phone call	Recycles	84	79.43	4.57
	Does not recycle	8	12.57	-4.57
Control	Recycles	86	94.97	-8.97
	Does not recycle	24	15.03	8.97

Step 3: Calculate $(O - E)^2$

In a new column called "$(O - E)^2$", square the values in the previous column.

Example: Step 3

Intervention	Outcome	Observed	Expected	$O - E$	$(O - E)^2$
Flyer	Recycles	89	84.61	4.39	19.27
	Does not recycle	9	13.39	-4.39	19.27
Phone call	Recycles	84	79.43	4.57	20.88
	Does not recycle	8	12.57	-4.57	20.88
Control	Recycles	86	94.97	-8.97	80.46
	Does not recycle	24	15.03	8.97	80.46

Step 4: Calculate $(O - E)^2 / E$

In a final column called "$(O - E)^2 / E$", divide the previous column by the expected frequencies.

Example: Step 4

Intervention	Outcome	Observed	Expected	$O - E$	$(O - E)^2$	$(O - E)^2 / E$
flyer	Recycles	89	84.61	4.39	19.27	0.23
	Does not recycle	9	13.39	-4.39	19.27	1.44
Phone call	Recycles	84	79.43	4.57	20.88	0.26
	Does not recycle	8	12.57	-4.57	20.88	1.66
Control	Recycles	86	94.97	-8.97	80.46	0.85
	Does not recycle	24	15.03	8.97	80.46	5.35

Step 5: Calculate X^2

Finally, add up the values of the previous column to calculate the chi-square test statistic (X2).

Example: Step 5

$X^2 = 0.23 + 1.44 + 0.26 + 1.66 + 0.85 + 5.35$

$X^2 = 9.79$

Example: Finding the critical chi-square value

- Since there are three intervention groups (flyer, phone call, and control) and two outcome groups (recycle and does not recycle) there are (3 − 1) * (2 − 1) = 2 degrees of freedom.
- For a test of significance at α = .05 and df = 2, the X2 critical value is 5.99.

Compare the chi-square value to the critical value

- Example: Comparing the chi-square value to the critical value
- $X^2 = 9.79$ Critical value = 5.99
- The X^2 value is greater than the critical value.

Decide whether to reject the null hypothesis

- The X^2 value is greater than the critical value.
- Therefore, the city **rejects** the null hypothesis that whether a household recycles and the type of intervention they receive are **unrelated**.
- The city concludes that their interventions have an effect on whether households choose to recycle.

4. Explain the methods used in Handling Several Batches or communicating to a group of people.

Boxplot

- It is important to display data well when communicating it to others.
- The boxplot is a device for conveying the information in the five number summaries economically and effectively.
- Refer Figure 4.3 for the Anatomy of the data.

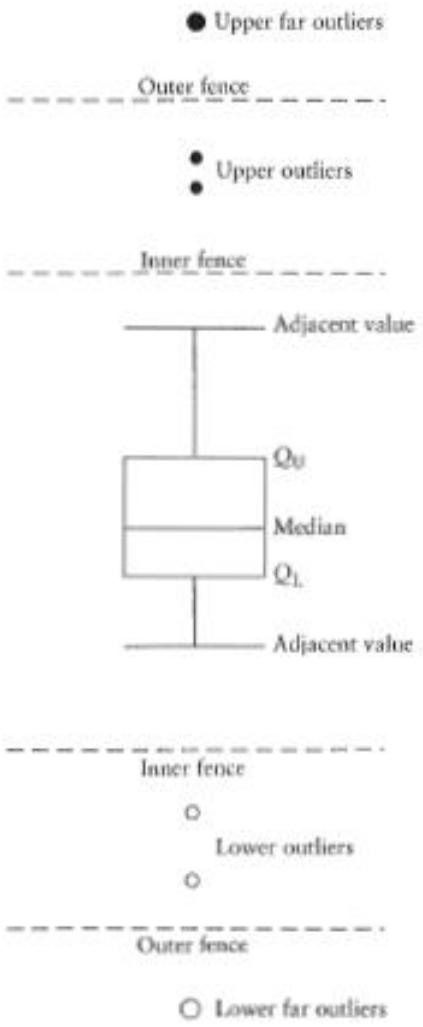


Figure 4.3 – Anatomy of Box plot

- The middle 50 per cent of the distribution is represented by a box.
- The median is shown as a line dividing that box.
- Whiskers are drawn connecting the box to the end of the main body of the data.
- They extend to the adjacent values, the data points which come nearest to the inner fence while still being inside or on them.
- Outliers are points that are unusually distant from the rest of the data.
- Then the points beyond which the outliers fall (the inner fences) and the points beyond which the far outliers fall (the outer fences) are identified; inner fences lie one step beyond the quartiles and outer fences lie two steps beyond the quartiles.

Outlier

- Data Set contain points which are lot higher or lower than the main body of the data . These are called Outliers.
- Reasons for Outliers
  - They may just result from a fluke of the particular sample that was drawn.

- They may arise through measurement or transcription errors, which can occur in official statistics as well as anywhere else.
- They may occur because the whole distribution is strongly skewed.
- These particular data points do not really belong substantively to the same data batch.

Multiple boxplots

- Boxplots, laid out side by side, permit comparisons to be made with ease.
- The standard four features of each region's distribution can now be compared:
  - The level
  - The spread
  - The Shape
  - Outliers

  Example: Refer Figure 4.4

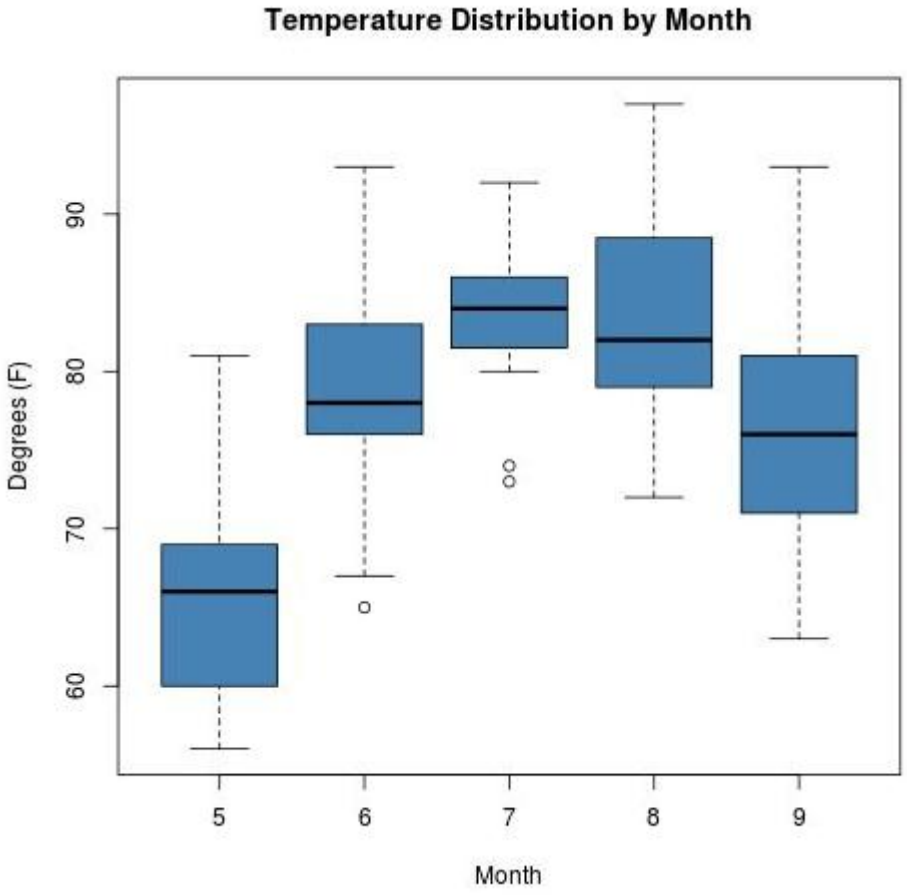


Figure 4.4 – Example of Multiple Box plot

T-Test

- A T-test is the final statistical measure for determining differences between two means that may or may not be related.
- It is a statistical method in which samples are chosen randomly, and there is no perfect normal distribution.

T-test formula

The formula for a two-sample t-test where the samples are independent (as in the example of boys' and girls' mathematics test scores) is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where X_1 and X_2 are the means of the two samples and S_{X1X2} is known as the pooled standard deviation and is calculated as follows:

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

here S_{X1} is the standard deviation of one sample and S_{X2} is the standard deviation of the other sample. In these formulae n_1 is the sample size of the first sample and n_2 is the sample size of the second sample. In simple terms therefore the size of the t-statistic depends on the size of the difference between the two means adjusted for the amount of spread and the sample sizes of the two samples.

5. Discuss in detail about Scatterplots and Linear Relationships.

Scatterplots

- To depict the information about the value of two interval level variables at once, each case is plotted on a graph known as Scatterplot.
- A Scatterplot has two axes, a vertical axes Y and a horizontal axes X.
- The cause variable (Explanatory variable) is placed on X axis and effect variable (Response Variable) is placed on Y axis.
- Scatterplot depict bivariate relationships.

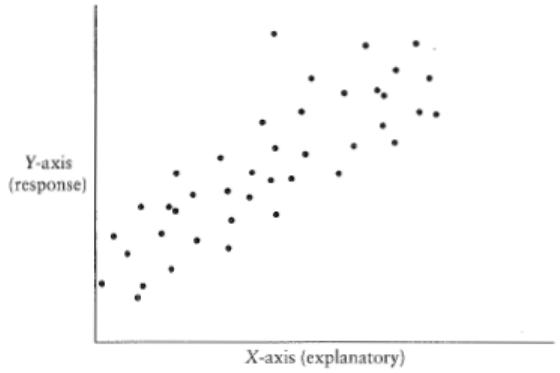


Figure 4.5 – Scatterplot – Monotonic and Positive Relationship

Example

- The data in Table 4.1 relate to the percentage of households that are headed by a lone parent and contain dependent children, and the percentage of households that have no van or car.

Table 4.1 Lone parent households and households with no car or van,% by region.

Government Office Region (2001)	% Lone parent households	% Households with no car or van
North East	7.35	35.94
North West	7.67	30.21
Yorkshire/Humber	6.58	30.31
East Midlands	6.08	24.25
West Midlands	6.73	26.77
Eastern	5.29	19.80
London	7.60	37.49
South East	5.22	19.43
South West	5.42	20.21
Wales	7.28	25.95

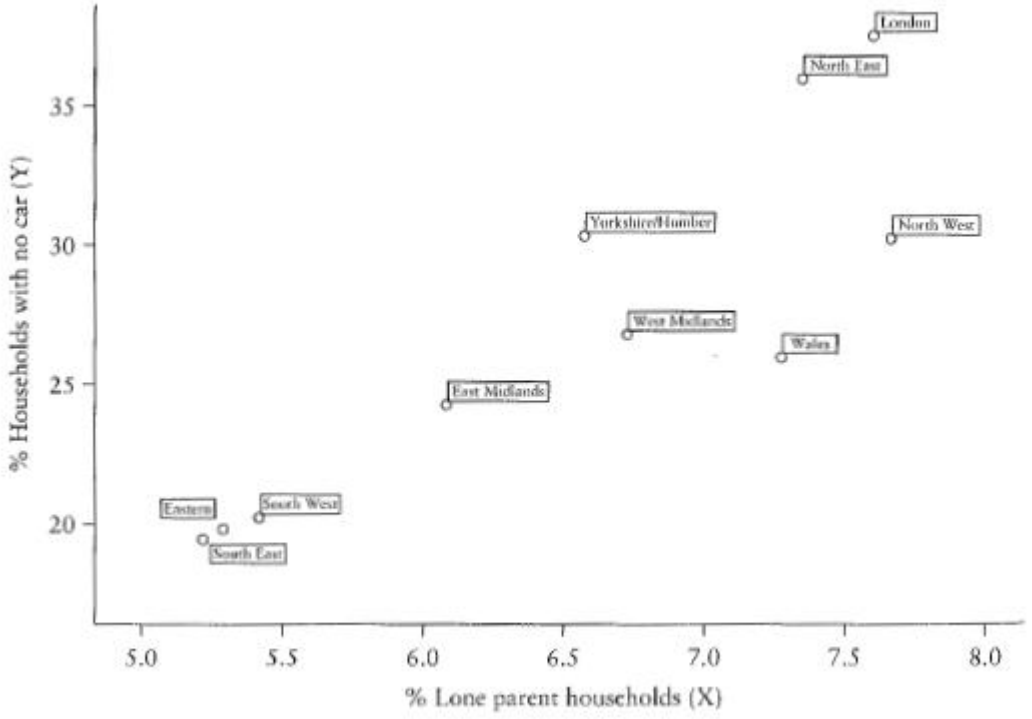


Figure 4.6 – Lone parent households by households with no car or van: scatterplot.

Linear Relationships

- Straight lines are easy to visualize and to draw on a graph and can be expressed algebraically

$$Y = a + bX$$

- X and Y are variables and a and b are coefficients that quantify any particular line.

- Figure 4.7 shows the Anatomy of Straight Line

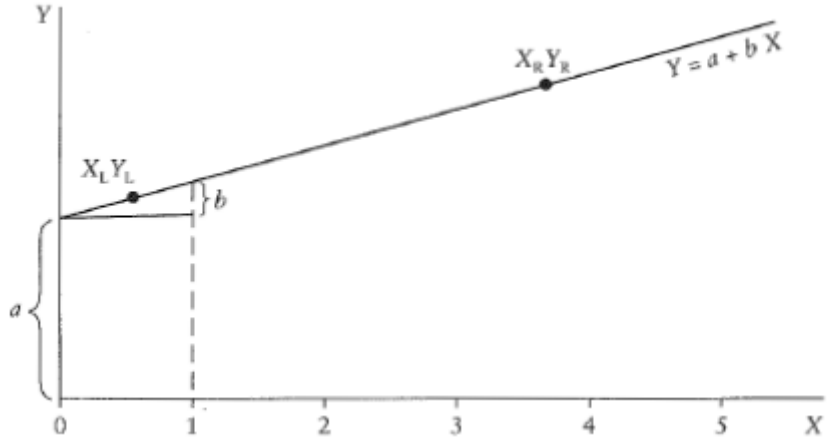


Figure 4.7 – Anatomy of Straight line

- The degree of slope or gradient of the line is given by the coefficient *b;* the steeper the slope, the bigger the value of *b.*
- The intercept *a* is the value of Y when X is zero.
- This value is also sometimes described as the constant.
- The slope of a line can be derived from any two points on it.
- If we choose two points on the line, one on the left-hand side with a low X value (called X_L' Y_L), and one on the right with a high X value (called X_R, Y_R), then the slope is

$$\frac{Y_R - Y_L}{X_R - X_L}$$

- If the line slopes from top left to bottom right, Y_R - Y_L will be negative and thus the slope will be negative.

Limitations to draw a line

- Make half the points lie above the line and half below along the full length of the line.
- Make each point as near to the line as possible
- Make each point as near to the line in Y direction as possible.
- Make the squared distance between each point and the line in Y direction as small as possible. This technique is known as Linear Regression.

6. Define Resistant Line and explain about fitting a Resistant Line.

Resistant Line

- To explore paired data and to suspect a relationship between X and Y, the focus is on how to fit a line to data in a "resistant" fashion, so the fit is relatively insensitive to extreme points.

Fitting a Resistant Line - Line fitting involves joining two typical points.

- The X-axis is roughly divided into three parts
  - X values are ordered along with its corresponding Y values.

- Divide X axis to three approximately equal length
- The left and right should be balanced with equal number of data points

- Conditional Summary points for X and Y are found.
- A Line is drawn connecting a Left and Right Summary points

Example

Table 4.2 Worksheet for calculating a resistant line.

	1 X	2 Y	3 Fit Y *	4 Residual Y
South East	5.22	19.43	18.39	1.04
Eastern	5.29	19.8	18.88	0.92
South West	5.42	20.21	19.79	0.42
East Midlands	6.08	24.25	24.40	−0.15
Yorkshire/Humber	6.58	30.31	27.89	2.42
West Midlands	6.73	26.77	28.94	−2.17
Wales	7.28	25.95	32.79	−6.84
North East	7.35	35.94	33.28	2.66
London	7.6	37.49	35.02	2.47
North West	7.67	30.21	35.51	−5.30

Obtain the Summary Points

- The summary X value is the median X in each third; in the first third of the data, the summary X value is 5.29, the value for the Eastern region.
- The median Y in each third becomes the summary Y value, here 19.8.
- The summary X and Y values for each of batches is

$$X_L = 5.29 \qquad Y_L = 19.8$$
$$X_M = 6.66 \qquad Y_M = 26.36$$
$$X_R = 7.6 \qquad Y_R = 35.94$$

- The slope

$$\frac{Y_R - Y_L}{X_R - X_L} = \frac{35.94 - 19.8}{7.6 - 5.29} = 6.99$$

- The intercept is the average of below three that is -18.1

$$a_R = Y_R - bX_R = 35.94 - (6.99 \times 7.6) = -17.1$$

$$a_M = Y_M - bX_M = 26.36 - (6.99 \times 6.66) = -20.2$$

$$a_L = Y_L - bX_L = 19.8 - (6.99 \times 5.29) = -17.1$$

- The full prediction equation is

$$\% \text{ no car} = -18.1 + (6.99 \times \% \text{ lone parent})$$

- The full set of predicted values is shown in column 3 of Table 4.2; the column is headed Y (pronounced 'Y-hat'), a common notation for fitted values.
- In the South East the percentage of households with no car is 1.04 higher than the predicted 18.39, namely 19.43.
- Residuals from the fitted values can also be calculated for each region, and these are shown in column 4 of Table 4.2.
- All the data values can be recast in the traditional DFR form:

$$\text{Data (19.43)} = \text{Fit (18.39)} + \text{Residual (1.04)}$$

- The simple technique for fitting the 'best' straight line through the Scatterplot minimizes the total sum of these residual values.

7. Define Transformation and explain in detail about transformations.

The log transformation

- One method for transforming data or re-expressing the scale of measurement is to take the logarithm of each data point.
- This keeps all the data points in the same order but stretches or shrinks the scale by varying amounts at different points.

The ladder of powers

- It is family of power transformations that can help promote symmetry and sometimes Gaussian shape in many different data batches.

Power	Expression	Name
•		
•		
3	X^3	cube
2	X^2	square
1	X^1	raw data
0.5	$\sqrt{X}$	square root
0	X^0	? (log)
–0.5	$1/\sqrt{X}$	reciprocal root
–1	$1/X$	reciprocal
–2	$1/X^2$	reciprocal square
•		
•		

Figure 4.8 - The ladder of powers

- Going up the ladder of powers corrects downward straggle, whereas going down corrects upward straggle.

The goals of transformation

- Data batches can be made more symmetrical.
- The shape of data batches can be made more Gaussian.
- Outliers that arise simply from the skewness of the distribution can be removed, and previously hidden outliers may be forced into view'.
- Multiple batches can be made to have more similar spreads.
- Linear, additive models may be fitted to the data.

Determining the best power for transformation

- When investigating a transformation to promote symmetry in a single batch, first examine the midpoint summaries - the median, the mid quartile, and the mid extreme – to check if they tend to increase or decrease.
- If they systematically increase in value, a transformation lower down the ladder should be tried.
- If the midpoint summaries the trend downwards, the transformation was too powerful, and must move back up the ladder.
- Continue experimenting with different transformations from the summary values until the best to promote symmetry and Gaussian shape.
- Curves which are monotonic and contain only one bend can be thought of as one of the four quadrants of a circle.
- To straighten out any such curves, first draw a tangent.
- Then imagine pulling the curve towards the tangent (as shown in figure 4.9).
- Notice the direction in which having to pull the curve on each axis, and move on the ladder of powers accordingly.
- To straighten the data in figure 4.9, for example, the curve has to be pulled down in the ¥-direction and up in the X-direction; linearity will therefore probably be improved by raising the Y variable to a power lower down on the ladder and/or by raising the X variable to a power higher up on the ladder.

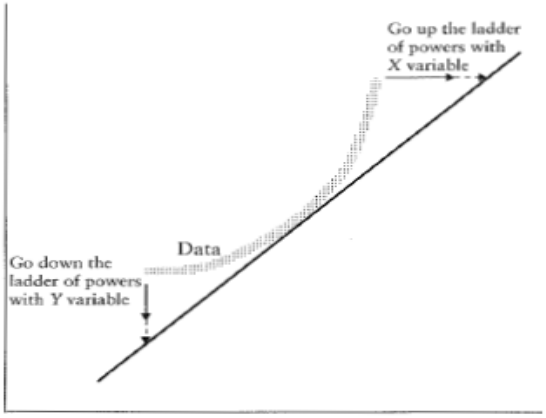


Figure 4.9 Guide to linearizing transformations for curves.