Data Analysis

ROSEDALE
ACADEMY

Drafting a Plan - Example

The table on the right lists data related to vehicle collisions in Ontario in one year.

Outline an action plan for investigating the relationship between driver age and number of accidents.

Vehicle Collisions in Ontario			
Age	Licensed Drivers	Number in Collisions	% of Drivers in Age Group in Collision
16	85 050	1 725	2.0
17	105 076	7 641	7.3
18	114 056	9 359	8.2
19	122 461	9 524	7.8
20	123 677	9 320	7.5
21–24	519 131	36 024	6.9
25–34	1 576 673	90 101	5.7
35–44	1 895 323	90 813	4.8
45–54	1 475 588	60 576	4.1
55–64	907 235	31 660	3.5
65–74	639 463	17 598	2.8
75 & older	354 581	9 732	2.7
Total	**7 918 314**	**374 073**	**4.7** (average)

Hypothesis

The graduated licence system in Ontario has resulted in a dramatic decrease in the number of accidents involving teenage drivers.

Data Collection

The table (Vehicles Collisions in Ontario) is a starting point for data collection. However, the data given in the table is only for one year.

- You will need accident data for other years, including the years before and after the introduction of graduated licences.
- You will need the data separated by age.
- You may also want the data separated by gender, or by region.
- You may want to consider other variables, such as driving for pleasure/work, accidents by time of day, accidents involving impaired drivers, accidents by type of vehicle, and so on.

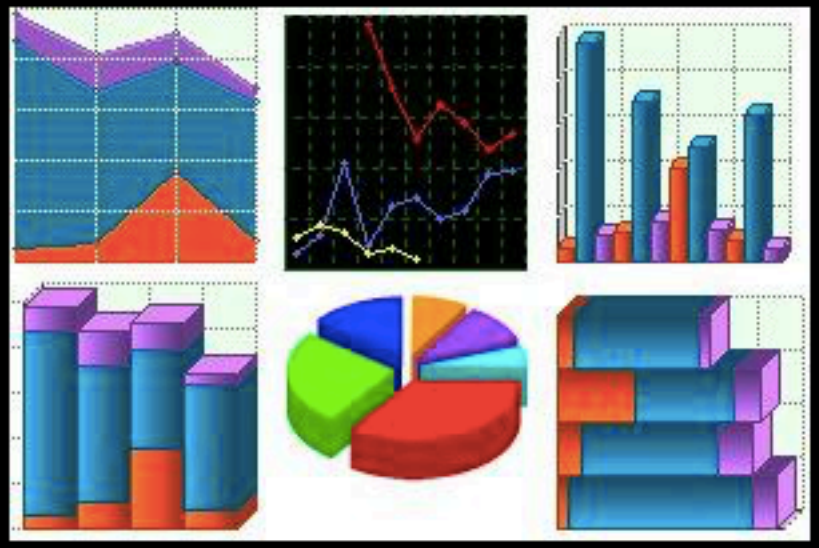
Data Organization

The data should be organized to allow you to test your hypothesis. This means you will need to isolate the effects of the graduated licence system from other factors such as population that may have changed over the years you are examining.

Data Presentation

When presenting the data, you need to keep in mind that you will be presenting these data both in your written report and in your PowerPoint presentation.

You may want to use bar graphs, circle graphs, line graphs, or tables to present your data.



Data Presentation

You also want to choose ways of presenting your data that help address your hypothesis. Note that this does not mean distorting the data by using inappropriate scales or ignoring outliers and data that does not support your hypothesis.

You are testing the validity of your hypothesis, not trying to convince your audience of the correctness of your claim.

Data Analysis

Use the tools you have learned in this course to analyse your data. Summary statistics, measures of dispersion, analysis of outliers, regression, and other techniques may be appropriate.

It may be possible to perform a formal hypothesis test on your data. You should be able to relate your results to probability theory and probability distributions, where appropriate.

Data Analysis: Simple Example

Let me explain this section using an example. Let's suppose I chose my topic as 'Internet Use in Canada'.

Some of the hypotheses I can developed for this topic are as follows.

Data Analysis: Hypotheses

1. How often does the average Canadian use the Internet daily?
2. For what uses do Canadians use the Internet the most often?
3. What is the profile of a Canadian who is likely to spend a lot of time using the Internet?
4. Impact of Internet on daily lives in the last 10 years.
5. The various Internet activities that most Canadians participate in.
6. What is the percentage of users in Canada who have posted pictures online?

Data Analysis: Sources

We can use several websites to collect data. The best one will be CANSIM website, which contains social and economic data from Statistics Canada. You will recall that this source will be considered secondary data. However, you can collect the data using surveys at your school to have the primary data.

But be sure there is relevant and available data to be collected. You will need to have at least 30 data points and at least 3 statistical variables (attributes) for your secondary data.

Data Analysis: One-Variable

For example in my chosen topic, one-variable analysis can be:

1. Hours of Internet use at home
2. Hours of Internet use at School

In both cases, I can represent the data on a histogram by assuming 'time spend in hours' as '**x**' variable and 'number of people' as '**y**' variable).

Data Analysis: One-Variable

Now using the provided data. We can determine the following:

- Frequency Distribution (based on number of hours)
- Mean
- Median
- Mode
- Variance and Standard Deviation
- IQR

Data Analysis: Graphs

We can have pie charts or graphs representing various other segments of the data for example *(just to name a few)*:

- Percentage of households in Canada using the Internet for various household types.
- Percentage of households in Canada using the Internet for various locations.
- Percentage of households in Canada using the Internet based on age of household head.
- Percentage of households in Canada using the Internet based on age of household head education.

Data Analysis: Two-Variable

The purpose of the two-variable analysis is to conduct the data analysis for our co-related study.

For your main variable, you are expected to create a scatter plot versus each of the other two variables that you have chosen for your study. Typically, your main variable will be plotted on the x-axis.

Data Analysis: Two-Variable

For each scatter plot, you are expected to fit a linear model to the data. On the plot, include the equation of the linear model and the correlation coefficient that measures the strength and direction of the linear relationship.

Data Analysis: Two-Variable

For example in my collected data, I can have the scatter plots for the following:

- Percentage of people in Canada connected to the Internet by Technology Use Level.
- Hours of Internet use in Canada by Average Provincial Income.
- Hours of Internet use at home by Household Income.
- Average home Internet use in Canada in a week by Age.
- Percentage of Households in Canada regularly using the Internet by Year.

Data Analysis: Probability Distribution

Let's suppose using an online survey we found that 20% of all Internet users have posted pictures online. A sample survey interviewed around 1500 Internet users. Now using Normal Distribution, we can describe the shape and spread of the distribution of the proportions in the sample. Also if we have a number of people in the sample who have posted pictures online, then we can determine the z-score and use the z-score chart to represent the percentage of users.

Data Analysis: Summary

Step 1:

You will pose a significant problem of interest that requires the organization and analysis of primary or secondary data.
You will recall that the primary data can be collected from your surveys. Whereas secondary data from a reliable source such as Statistics Canada.

Data Analysis: Summary

Step 2:

Design a plan to study the problem. For example, identify the variables and the population. Establish the procedures for gathering and analysing the data. At this step, do not forget the sample size and possible sources of bias.

Data Analysis: Summary

Step 3:

Interpret, analyse, and summarize data using Excel worksheet. For example, generate and interpret numerical and graphical statistical summaries.

Data Analysis: Summary

Step 4:

Recognize and apply a probability distribution model and calculate the expected value of a probability distribution.

If you have any questions or concerns, please do not hesitate to contact your teacher.

Click on the image to download the Checklist for Data Analysis

Click on the image to download the Summary for Data Analysis

Finish