

Data Concepts and Graphical Summaries

Lesson objectives

- I can show how data are used and misused in statistical studies
- I can identify different types of data
- I understand that there is variability in data
- I can see that you can analyse single sources of data or related sources

1.1

Lesson objectives

Teachers' notes

Lesson notes

MHR Page 203 #s 1 - 6

Definitions

Numerical (Quantitative) Data

- Data in the form of any **number**

Categorical (Qualitative) Data

- Data that can be **sorted** into distinct groups or **categories**

Ordinal Data

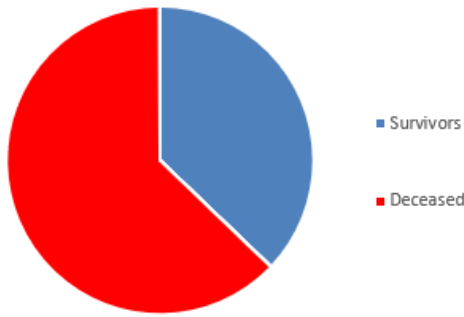
- Qualitative data that **can be ranked**
- Examples: poor, fair, good, very good

Nominal Data

- Qualitative data that **cannot be ranked**
- Examples: blue eyes, green eyes, brown eyes

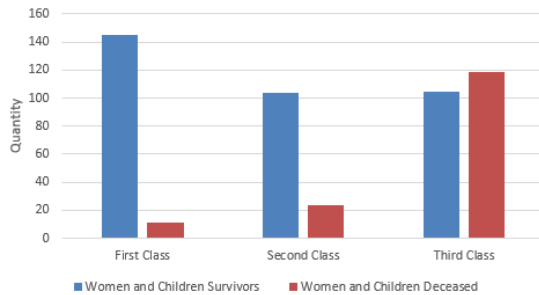
Investigate on Page 196

1. Fate of the Passengers on the Titanic



I chose a pie chart to illustrate how many passengers died compared to those who survived. Nearly two-thirds of the passengers died!

2. The Fate of Women and Children



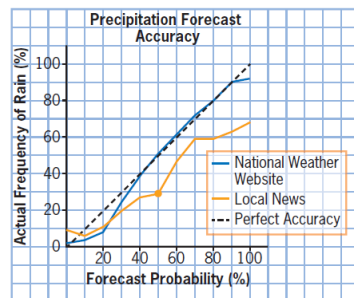
Women and children first policy pays dividends with huge survival.

3. There is often more than one way to spin data. We are going to look at ways to analyse data and cut through the headlines to find out what it is really telling us.

Example 1

Variability in Data

Sometimes people can use accurate information to tell different stories. Probability of precipitation (PoP) is a common measurement used in weather forecasts. The graph shows the accuracy of a national weather website and local news channels compared to perfect accuracy.



Source: Data from *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*

- What does the yellow dot indicate?
- Which outlet is more accurate, the national weather website or the local news?
- If both outlets base their forecasts on information collected by Environment Canada, what reasons can you suggest for their differences?

a) The yellow dot represents when the local news station predicts a 50% chance of PoP, it only rained 30% of the time.

Your Turn

Researchers often have conflicting opinions even though they use the same data. Research one of the following topics to find conflicting opinions:

- climate change
- vaccinations
- fluoridated water

Climate change: 97% of scientists agree that these trends are caused by human actions. Others argue that they are due to natural causes rather than humans.

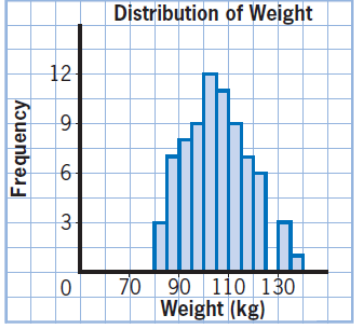
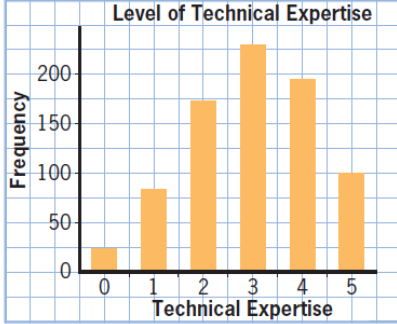
b) The national weather website is more accurate than the local news based on this data. When they predict the PoP from 40% - 90% they are invariably correct. The local news generally over-predicts the likelihood of rain, in some cases by as much as 30% points. Local news were more accurate when the PoP was 20% or less.

c) Despite using the same data, the calculations used may be different which would result in different predictions.

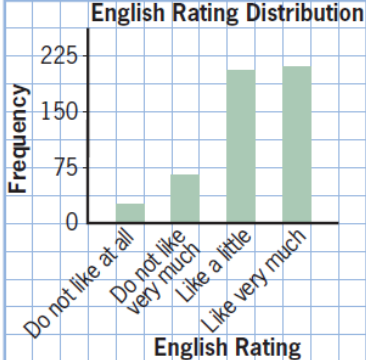
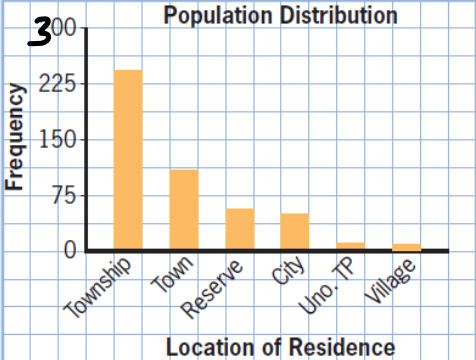
There are two main types of data: **numerical (quantitative)** and **categorical (qualitative)**. Numerical data are either continuous or discrete.

Continuous data can have any value in a range (including decimal numbers). For example, the weight of a person or the amount of time an experiment takes could have any value in a range. We often use a histogram to display continuous data. When the bars in a histogram are touching, it means that the data can be any value in a range.

Discrete data are data that only have specific values (usually whole numbers). We often represent discrete data with a bar graph. The bars do not touch, indicating that there are no possible values in between.

Continuous Numerical Data	Discrete Numerical Data
	
<p>Since the weight of a person can be any value, weight is continuous data.</p>	<p>When someone fills out a survey and uses a rating scale, the scale is measured in whole numbers, so it is discrete data.</p>

There are two main types of categorical data: **ordinal data** and **nominal data**.

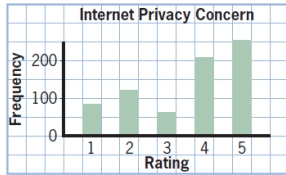
Categorical Ordinal Data	Categorical Nominal Data
	
<p>Since the person is answering using a rating scale, these are ordinal data.</p>	<p>The data are categorized by the type of place. There is no logical order, so these nominal data are placed from the highest bar to the lowest.</p>

Example 2

Comparing Types of Data

For each graph, identify the type of data, give reason(s) for your choice, and write one statement about what the data show.

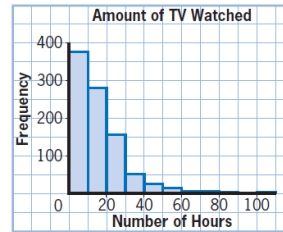
- a) A survey asks people to rate how concerned they are with Internet privacy on a scale where 1 is not concerned and 5 is very concerned.



Categorical, ordinal data because they are non-numerical and ranked

Most people are concerned with internet privacy

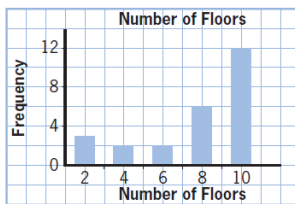
- b) A survey asks 1000 people how many hours a week they watch TV.



Numerical, continuous data because the numbers can have any value

Over a third of the respondents watch 0-10 hours of TV a week

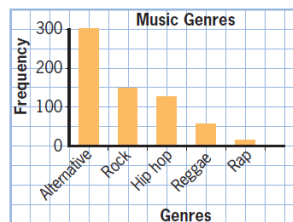
- c) A town planner records how many floors each apartment building has.



Numerical, discrete data because each value can only be a distinct whole number

Most apartment buildings in the town have 10 floors

- d) The songs in a digital music library are sorted by genre.



Categorical, nominal data because they are non-numerical and not ranked

The owner of the library likes alternative music the most

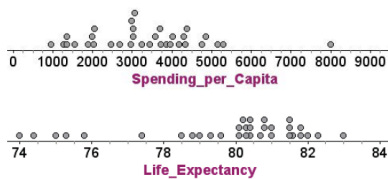
Example 3

Data With More Than One Variable

Researchers often measure more than one variable of a particular item. Then, they can analyse the data to see if the measurements are connected to each other. The table shows the life expectancy in years and the health care spending per capita in dollars from various countries.

Source: Table 2: Total expenditure on health per capita, OECDiLibrary, October 11, 2013 and Table 11: Life expectancy at birth, total population, OECDiLibrary, December 6, 2013

- a) Which country spends the most per person on health care? the least? Where is Canada on the list?
 b) The dot plots show the spending per capita and the life expectancy of each country. In each plot, each dot represents one country.



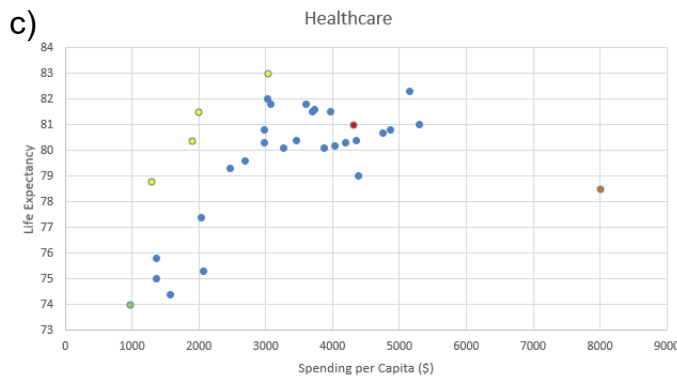
For each graph, identify the lowest and highest ranked country, as well as Canada and the United States. What story does the data seem to tell?

- c) Create a scatter plot of life expectancy versus health care spending. Which countries lie outside the general trend? How do you know?
 d) In what area of the graph would a country want to be located? Why?
 e) In Canada, everyone has access to health care services. This is not the case in the United States. How can these data be used to argue against more health care spending? How can they be used to argue in favour of more health care spending?

a) The **USA** spends the most per person on healthcare and **Mexico** spends the least. **Canada** is in the top 25% on spending.

Country	Spending per Capita (\$)	Life Expectancy (years)
Australia	3734	81.6
Austria	4345	80.4
Belgium	3874	80.1
Canada	4309	81
Chile	1283	78.8
Czech Republic	2039	77.4
Denmark	4390	79
Estonia	1371	75
Finland	3259	80.1
France	3962	81.5
Germany	4187	80.3
Greece	2977	80.3
Hungary	1567	74.4
Iceland	3597	81.8
Ireland	4037	80.2
Israel	1991	81.5
Italy	3030	82
Japan	3025	83
Korea	1895	80.4
Luxembourg	4755	80.7
Mexico	957	74
Netherlands	4870	80.8
New Zealand	2984	80.8
Norway	5300	81
Poland	1356	75.8
Portugal	2692	79.6
Slovak Republic	2063	75.3
Slovenia	2470	79.3
Spain	3080	81.8
Sweden	3703	81.5
Switzerland	5157	82.3
United Kingdom	3456	80.4
United States	8006	78.5

b) Spending per Capita: Highest, USA, lowest, Mexico. Life Expectancy: Highest, Japan, lowest Mexico. Canada sits in the top 33% for life expectancy and the top 25% on spending on healthcare per capita. There appears to be a trend that the more spent on healthcare, the higher the life expectancy.



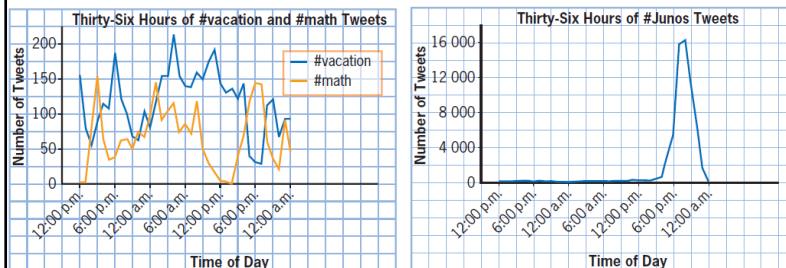
The country that is most outside of the trend is the USA. Other countries are Chile, Korea, Israel, and Japan. These four seem to have a higher life expectancy compared to their spending on healthcare.

d) The preferred area of the graph would be the top left, where life expectancy is high, but spending on healthcare per capita is low.

e) Against: The USA already spends nearly double compared to Canada and has lower life expectancy. Why bother spending more? For: The trend seems to indicate a link, but maybe there are other factors at play? It could be to do with how the money is being spent, or diet, or exercise, or lifestyle choices for example.

Your Turn

The graphs show the number of tweets at various times of day that contain the hashtags #vacation, #math, and #Junos.



c) Likely broadcast at 8:00pm. The hashtag was used less often after that.

- a) What do these graphs tell you about the frequency of tweets with each type of hashtag?
- b) What do the data seem to indicate about #math and #vacation as Twitter topics?
- c) The Juno Awards is the Canadian music industry's yearly award show. When do you think the show was broadcast?
- d) Could these graphs be compared to each other as they are? Explain why or why not.

b) The volume of tweets would suggest that #vacation and #math are not popular hashtags.

a) For #vacation, tweets go both up and down, peaking at 6:00pm, 4:00am, and 11:00am during the first 24 hours. For #math, tweets go both up and down, peaking at 3:00pm, 1:00am, and 8:00am during the first 24 hours. For #Junos, there are minimal tweets until 3:00pm when there is then a huge surge and decline over the next 9 hours with the peak occurring at 8:00pm.

d) The graphs should not be compared to each other. There are significantly more tweets on the third graph. We don't even know if these data are from the same day.

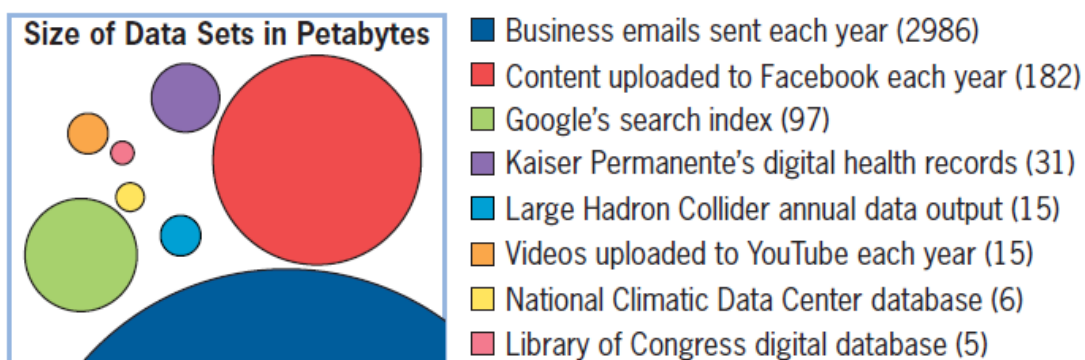
Key Concepts

- Depending on who is analysing the data and their intention, the information taken from the data can be very different.
- Variability in data exists due to errors in measurement or varying conditions in experiments.
- Different people can interpret data in different ways.
- There are two main types of data: numerical and categorical. Numerical data may be classified as continuous or discrete. Categorical data may be classified as ordinal or nominal.
- When researchers collect data on more than one variable, they can compare the data to see if there is a relationship.

R1. Studies show that political experts' predictions are correct or mostly correct 46% of the time and incorrect or mostly incorrect 47% of the time. The remaining 7% are a mix of correct and incorrect predictions. What does this suggest about how reliable experts might be when making predictions?

This suggests that predictions made by political experts are not very reliable. They are correct only about half the time.

R2. In the last 20 years, the amount of data being moved and stored online has become staggeringly large. In 2012 alone, almost 3 zetabytes of information moved online. Discuss the positives and negatives of collecting this large amount of data each year.



Positive: Availability, ease of back-up, lower risk of loss, less susceptible to viruses

Negative: Access speed (although this is improving), need an internet connection, environmental impact of the data farms (emissions and electricity usage), could get hacked