Review

Organization of Data for Analysis

Page 314 #s 6 - 8

- 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.

- A population is the entire group of a set of people or things. A sample is a smaller portion of that population.
- You can learn a lot about a population by examining samples of that population, as long as all members of the population are equally likely to be part of the sample.
- When multiple samples are taken from the same population, they are different from each other. This is called variability of samples. The smaller the differences in the samples, the more likely the sample closely represents the population.
- There are many types of sampling techniques. Some types of samples work better in certain situations. A good sample is random, and each person in the population has an equally likely chance to be chosen.

Type of Sample	Example
Simple Random • randomly choose a specific number of people • examples: stratified samples and systematic samples	Put all the names in a population into a hat and draw one or several names. Each person has an equal chance of being chosen.
Systematic • put the population in an ordered list and choose people at regular intervals	Order all the patients of a doctor in some way (e.g., alphabetically) and choose one randomly. Select the rest of the data at regular intervals from the original starting point (e.g., every tenth name after the original).
Stratified • divide the sample into groups with the same proportions as those groups in the population • time- and cost-efficient to conduct	Survey factory employees about new safety initiatives. There are 1000 employees in the factory, of which 633 are women and 367 are men. Randomly select 63 women and 37 men to take the survey.
Cluster • divide the population into groups, randomly choose a number of the groups, and sample each member of the chosen groups	Survey Little League Canada baseball players. Randomly select five districts in each province and give the survey to every player in those districts.
Multistage • divide the population into a hierarchy and choose a random sample at each level	Conduct an employee wellness survey by randomly selecting 10 stores. Randomly select three departments in each store, and randomly select 10 employees in each of those departments.
Convenience • choose individuals from the population who are easy to access • can yield unreliable results since it inadvertently omits large portions of the population • often very inexpensive to conduct	To get the public's input on a new pet by-law, a local politician goes to a local park and asks people their opinion.
Voluntary • allow participants to choose whether or not to participate • often the only people who respond are either heavily in favour or heavily against what the survey is about	Conduct an online poll asking people whether banning junk food in schools will fight obesity.

- In an observational study, the researcher records behaviour and tries to draw conclusions based on the observations.
- Experimental studies try to determine the cause and effect relationship between two variables by controlling for one variable to see what effect it has on the other variable.
- Effective experiments have good control, randomize the members of the treatment and control groups, and try to have a similar demographic make-up in each group.
- Surveys are a powerful way to gain information about a group of people.
- Surveys should be anonymous but can ask for precise demographic information.
- Items on surveys should be clear, concise, and ask only one question that is free of bias.
- Rating scales on a survey should be evenly distributed between good and bad outcomes.
- Data from surveys can be efficiently collected using technology.

- Primary sources of data are collected directly from the source and are not manipulated or summarized in any way.
- Microdata are the individual pieces of data that make up all of the primary data.
- Secondary sources of data are used by someone who did not collect them. Often these data have been manipulated and summarized. Data found in the media often are secondary data.
- Data that are summarized in some way are called aggregate data.
- Large sources of data are available for analysis on the Internet.
- Sources of data also are hidden in digital items like songs and photos.

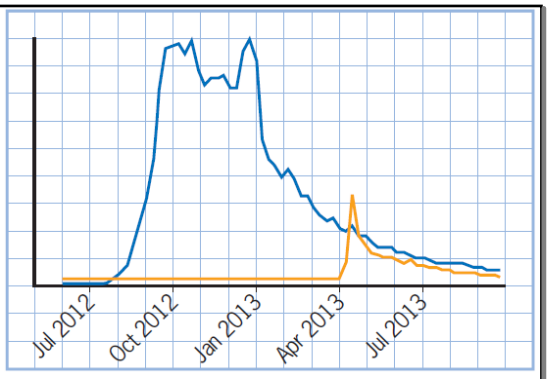
- For data to be valid, collection methods must be free from bias.
- The data can be affected if the collection methods suffer from sampling, measurement, response, or non-response bias.
- Different ways of displaying data can distort it and make it biased.
- Large numbers should always be put in context.
- Infographics can be dense with information or convey an idea with unique methods.

Type of Bias	Example
response bias	A teacher asks students to raise their hand if they cheated on last week's test. Students will not want to admit to cheating on a test so it is unlikely that many will raise their hand.
sampling bias	A politician goes back to the farming community she grew up in to ask for opinions on her latest initiative for the agriculture industry. It is likely that a larger proportion of the people she speaks to would support the initiative, both because it would benefit them and because she grew up in the area. This would not accurately represent the entire population.
measurement bias	A survey question asks, "A lot of people do not like math. How would you feel being referred to as a math geek?" This is a leading question; the wording of the question can affect the outcome by influencing someone's answer. Other types of measurement bias can occur when the collection method affects the results, for example when the options in a multiple choice question are too limited for an honest response.
non-response bias	A mail-in survey asked respondents about their drinking habits. Only 3% of the surveys were returned. Such a small return rate would likely not yield a representative sample. In fact, those who respond often have very strong opinions about the subject matter and so the results could easily over- or under-estimate the feelings of the population.

Solutions

6. Thinking In 2012, the musician Psy brought Korean pop music (K-Pop) to the world with his hit song "Gangnam Style." In 2013, he released another big hit, "Gentleman."

a) The graph shows the number of Google searches related to "Gangnam Style" and "Gentleman." Which colour represents which search? Give reasons for your answer.

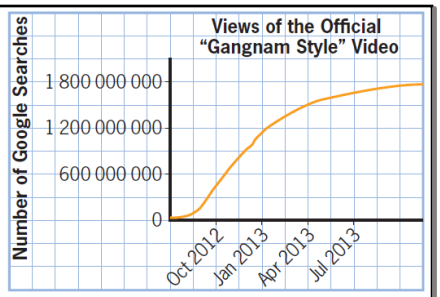


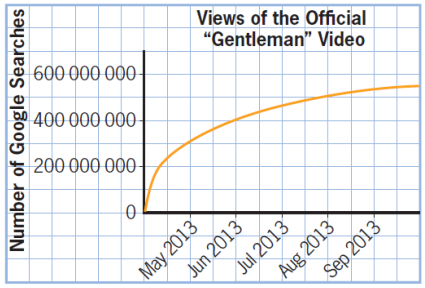
The blue line represents Gangnam Style as this shows a large spike in searches during the second half of 2012.

There is a smaller spike in mid 2013 which represents Gentleman.

6. Thinking In 2012, the musician Psy brought Korean pop music (K-Pop) to the world with his hit song "Gangnam Style." In 2013, he released another big hit, "Gentleman."

b) The graphs below show the number of YouTube views of the videos for each song as of September 2013. Based on these data, do you think that "Gentleman" is doing better or worse than "Gangnam Style"? Justify your answer.





As the graphs have similar shapes, it is tempting to say that they are doing as well as each other. In fact, Gentleman might be doing better as it starts off with a steeper rise.

However, when we look at the scales on the y-axis we can see that Gangnam Style is doing better because each square represents 300,000,000 searches where as for Gentleman each square only represents 100,000,000 searches. If we also look at the x-axis we can see that in the first 5 months Gentleman had about 540,000,000 searches where as Gangnam Style had over 1,700,000,000 searches.

Summary - Gangnam Style is doing better than Gentleman.

7. You are conducting a survey about one of the following topics: entertainment, sports, the environment, school, or technology.

a) Discuss the methods you will use to conduct your survey.

b) What questions will you ask?

c) What types of data will you collect?

d) How will you keep the data free of bias?

a) Use a sample that is representative of the population. The sample should be random, so that it is free from bias. This way the sample is reliable. The best ways of achieving this is to use a systematic or stratified sample.

b) The questions should be clear, concise and bias free. The survey should also be anonymous.

c) Depending on your questions, that data that you collect can be continuous, discrete, nominal, and/or ordinal.

d) To keep the data free from bias, the sampling method used should be free from response, non-respnse and measurement bias. The data should also be displayed in forms that are also free from bias.

8. The table shows some data from *Romeo and Juliet*.

Character	Number of Words Spoken	Appearances
Romeo	4690	163
Juliet	4314	118
Friar	2624	52
Nurse	2223	91
Capulet	2156	50
Mercutio	2112	62
Benvolio	1157	64
Lady Capulet	874	45
Prince	590	16
Paris	542	23
Montague	319	10
Tybalt	263	17
Sampson	256	20
Peter	248	13
Balthasar	233	12

a) How is the title of the play reflected in the data?

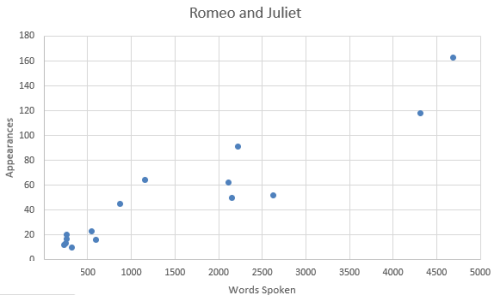
b) Does it appear that the number of words spoken is related to the number of appearances? Justify your answer using the data.

c) Who has the most spoken words in relation to their number of appearances?

d) Create a graphical summary of the data. Your teacher may provide you with a file called **RomeoAndJuliet.csv**.

a) The title reflects that Romeo and Juliet are the two central characters and are likely to have the most lines and be on stage the most.

b) There seems to be a positive in correlation in that the more words that are spoken, the more appearances a character has.



c) Friar has the most number of words spoken per appearance at a rate 50.5 (2624 ÷ 52).

d)