Chapter 2 - Workbook

Data Collection

MDM4U

2.1 - Developing a Thesis

MDM4U
Jensen

Part 1: ISU Intro

This chapter will prepare you to begin your ISU that is worth 10% of your final grade. For the ISU you will be required to choose a topic that interests you and conduct a study that analyses large amounts of data using:

- one-variable statistics tools (chapter 3)
- two variable statistics tools (chapter 1)
- probability (chapter 4/5)

Part 2: Mind-Map

Before you can begin your project, you must create a thesis:

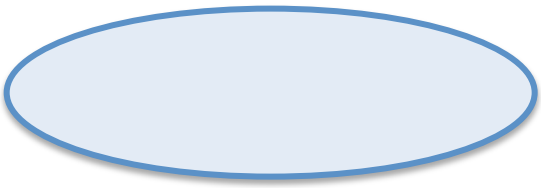
thesis: a formal statement or question that your project will answer or discuss

To begin creating a thesis, you must first determine what topics interest you and then determine what concepts related to that topic you want to study. A useful brainstorming tool that can illustrate how a topic relates to other concepts is a *mind map*.

mind map: a visual display used in brainstorming to illustrate relationships

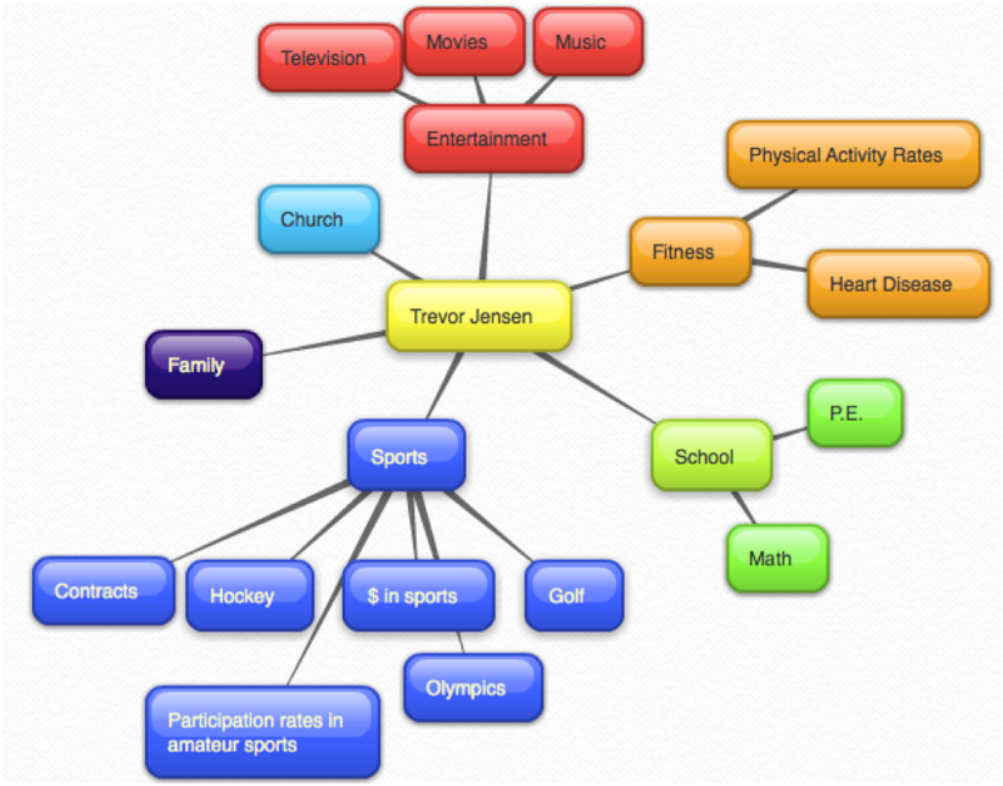
Constructing a Mind Map

1. Start by making a mind map of your interests with you at the center. Start off as simple as possible and draw arrows to show how topics are connected. Work from the inside out.



Extended Mind Map

2. Pick one of the topics from your mind map and extend it with sub-topics.



Part 3: Thesis Question Development

Once you have narrowed down your topic, you will need to pose a problem that you plan to investigate.

Money in Sports

3. Brainstorm and create number of questions that can be explored with the use of statistical information

a) How do people at my school feel about high salaries in professional sports?

b) How have salaries paid to professional hockey players changed from 1960 to present?

c) Is there a relationship between a very large salary increase to an athlete and his or her subsequent performance?

d) Does the amount a country spends to prepare its athletes for the Olympics correspond to the country's success at the games?

Thesis Question Analysis

Questions to ask of your Thesis:

i. What are the main variables in my question?

ii. Can these variables be measured statistically?

iii. Is there enough data to make an interesting analysis

4. Once you have chosen your thesis, analyze it using the three questions above to make sure your study will be able to provide an insightful answer.

Thesis: Is there a relationship between a very large salary increase to an athlete and his or her subsequent performance?

Analysis:

i. player salaries, performance statistics (goals, home-runs, etc.)

ii. yes; however it may be difficult to choose which performance statistics to use

iii. yes there would be lots of available data for professional athletes and their salaries and performance.

Project tips:

One way of posing a problem is to generate questions from data. For example, once a topic has been identified, do a preliminary data search. The type and quantity of available data may indicate some possible questions. Data from print sources, the Internet, and E-Stat are some resources that may be used.

2.2 Worksheet Characteristics of Data – Worksheet

MDM4U
Jensen

1) Identify each of the following variables as quantitative or qualitative. For each quantitative variable, identify whether it is continuous or discrete.

a) age

b) favourite meal

c) television viewing preferences

d) speed of car

e) colour of hair

f) fabric texture

g) pH of water samples

h) seating capacity

i) test mark

j) paint colours

2) Identify the variables and their types, as well as the population for the following thesis questions. Also, would you collect a sample or conduct a census? Would each question require a cross-sectional study or a longitudinal study?

a) Is there a relationship between weather conditions and absenteeism in Grade 9 at your school?

b) Is there a relationship between the amount of television watched and the level of physical fitness among adult females?

c) Are teenage drivers who have been issued speeding tickets more likely to be males?

3) Consider this thesis question: *In North America, do foreign cars depreciate in value faster than domestic cars?* Now answer the questions that follow:

a) What is the population?

b) What are the key variables that must be considered? Are these quantitative or qualitative? If quantitative, are they discrete or continuous?

c) Should a census or a sample be used to collect data?

d) Is a cross-sectional or a longitudinal study more appropriate for drawing conclusions?

4) Explain the differences between each pair of terms.

a) population/sample

b) cross-sectional study/longitudinal study

c) quantitative variable/qualitative variable

d) discrete data/continuous data

2.3 Worksheet - Collecting Samples

MDM4U
Jensen

1) Identify the type of random sampling in each of the following scenarios.

a) The principal randomly selects four classes and surveys each student in those classes

b) William picks names out of a hat

c) A hockey card collector opens a drawer of sorted cards and, after selecting a random starting point, takes out every fifth card.

d) The Ministry of Education randomly selects your school for testing, and 40 student names are randomly selected from a student list.

e) Your class submits solutions to a problem and your teacher divides the work into four piles by achievement levels. She then randomly picks three examples from each.

f) A farmer brings a juice company several crates of oranges each week. A company inspector looks at 10 oranges from the top of each crate before deciding whether to buy all the oranges.

g) The ABC program Nightline once asked whether the United Nations should continue to have its headquarters in the United States. Viewers were invited to call one telephone number to respond 'yes' and another for 'no.' More than 186 000 callers responded.

2) A textbook has 600 pages and 6 chapters. Describe how to you could design and carry out the following samples of its pages.

a) Select 6 pages from the textbook using simple random sampling

b) Select 10 pages using systematic random sampling

c) Select 12 pages using stratified random sampling

d) Select 10 pages using multi-stage random sampling.

3) Based on the following groups of names, identify a sampling method that may have been used to collect the samples listed in parts (a) through (e).

Shaggy	Paul	Joey	Susan
Fred	John	Monica	Elmo
Scooby	George	Rachel	Ernie
Thelma	Ringo	Ross	Oscar
Daphne		Chandler	Zoe
		Phoebe	Maria

a) Joey, Monica, Fred, Paul, Daphne

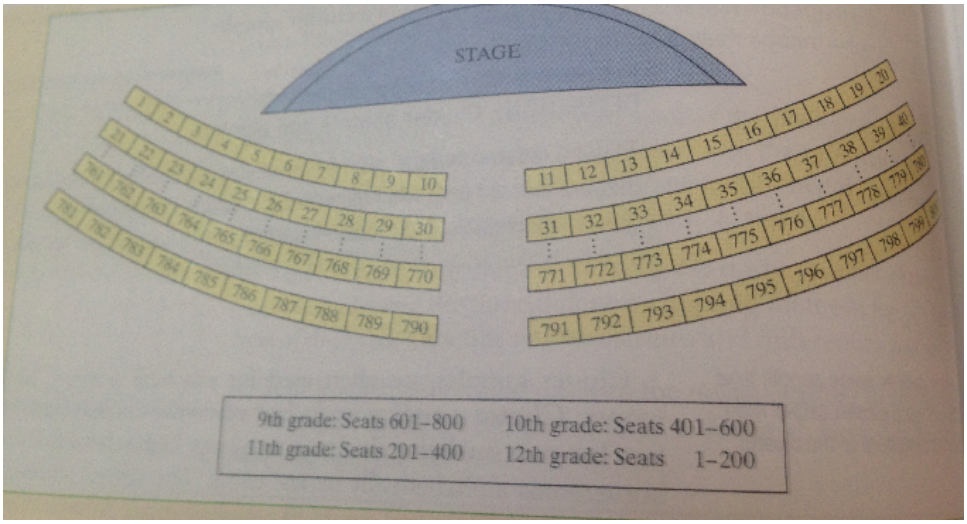
b) Susan, Elmo, Ernie, Oscar, Zoe, Maria

c) Shaggy Scooby, Daphne

d) John, George, Ringo

e) Shaggy, Fred, George, John, Joey, Chandler, Susan, Ernie

4) Student council wants to conduct a survey during the first five minutes of an assembly. There are 800 students at the assembly. A map of the auditorium is shown below. Note that the students are seated by grade level and the seats are numbered from 1 to 800. Describe how you would use your calculatr to select 80 students to complete the survey with each of the following methods:



a) Simple Random Sample

b) Stratified Random Sample

c) Cluster Random Sample

d) Systematic Random Sample

2.4 Worksheet – Survey Design and Bias

MDM4U
Jensen

1) For each of the following questions, state if it is an open question, rating question, ranking question, checklist question, or information question.

i) Please provide the following information:

Gender: _____ Grade: _____

ii) Please provide the following information

Gender: ☐M ☐F Grade: ☐9 ☐10 ☐11 ☐12

iii) With 1 meaning most helpful and 10 meaning not at all helpful, rate each of the chapters of the textbook.

_____ Chapter 1 _____ Chapter 2 _____ Chapter 3

_____ Chapter 4 _____ Chapter 5 _____ Chapter 6

iv) Rank the chapters of this textbook in order from 1-6 (1 being most useful).

_____ Chapter 1 _____ Chapter 2 _____ Chapter 3

_____ Chapter 4 _____ Chapter 5 _____ Chapter 6

2) Describe the characteristics of a good question and what good questions must avoid.

3) Using your criteria from the previous question, evaluate the following survey questions and re-write the question to meet all of the criteria of a good question.

a) Should the OMB be funded to initiate waste audits across the province?

☐Y ☐N

b) Given the large amount of sulfur dioxide that is spewed out of smelters, should mining companies be forced to clean up their act? Please comment.

c) On a 5-point scale, do you agree with the bleeding-heart Liberals that all corporations should pay higher taxes? (1: totally agree, 5: totally disagree)

d) Why shouldn't forestry companies clear cut? Please explain.

4) For each of the following questions, state if it is an open question, rating question, ranking question, checklist question, or information question.

a) You are presently in grade (circle the appropriate answer):

9 10 11 12

b) I find mathematics stimulating because:

_____

c) Rank the following foods from favourite (1) to least favourite (4).

_____ pizza _____ hamburgers _____ tacos _____ watermelon

d) Do you wear a wristwatch?

_____ always _____ sometimes _____ seldom _____ never

e) How much do you like math on a scale of 1 to 5 (1 being the lowest)

f) Estimate your net income:

_____ $15000-$19999 _____$20000-$39999 _____ $40000-$59999 _____ $60000 +

5) Explain the difference between primary and secondary data.

6) Identify the type(s) of bias that might result from each of the following data collection methods.

a) You hand out surveys to your classmates to be returned to you next week.

b) You are interested in the study habits of grade 12 students, so you interview students from your class.

c) You ask students about their recycling habits on behalf of the Greenteam, the school environment club.

d) You take a random sample of 5 students from each block A class to determine their attitudes toward the new school attendance policy.

7) When a phone questionnaire is conducted, many people with call display will not answer their phone. What kind of bias does this represent? What can be done to minimize this kind of bias?

8) Suppose you want to know the average amount of money spent by the fans attending opening day for the Toronto Blue Jays. You get permission from the team's management to conduct a survey at the stadium, but they will not allow you to bother the fans in the club seating or box seats (the most expensive seating). Using a computer, you randomly select 500 seats from the rest o the stadium. During the game, you ask the fans in those seats how much they spent that day. What type of bias is present in this survey method?

2.5 Worksheet– Experiment Design

MDM4U
Jensen

Refer to part 2 of the lesson for help with the following questions

1) An educator wants to compare the effectiveness of computer software for teaching biology with that of a textbook presentation. She gives a biology pretest to each of a group of high school juniors, then randomly divides them into two groups. One group uses the computer, and the other studies the text. At the end of the year, she tests all the students again and compares the increase in biology test scores in the two groups. Is this an observational study or an experiment? Justify your answer.

2) One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same age, gender, and race who did not have brain cancer, then asked about the use of cell phones. The results suggested that the use of cell phones is not associated with risk of brain cancer. Is this an observational study or an experiment? Justify your answer.

3) Do smaller classes in elementary school really benefit students in areas such as scores on standardized tests, staying in school, and going on to college? We might do an observational study that compares students who happened to be in smaller and larger classes in their early school years. Identify a potential variable that may be confounding with the effects of small classes.

4) Ability to grow in shade may help pines found in the dry forests of Arizona to resist drought. How well do these pines grow in shade? Investigators planted pine seedlings in a greenhouse in either full light, light reduced to 25% of normal by shade cloth, or light reduced to 5% of normal. At the end of the study, they dried the young trees and weighed them.

a) Identify the experimental units.

b) What are the explanatory and response variables?

c) What are the treatments used?

5) You can use Skype to make long-distance calls over the Internet. How will the appearance of ads during calls affect the use of this service? Researchers design an experiment to find out. They recruit 300 people who have not used Skype before to participate. Some people get the current version of Skype with no ads. Others see ads whenever they make calls. The researchers are interested in frequency and length of phone calls.

a) Identify the experimental units.

b) What are the explanatory and response variables?

c) What are the treatments used?

Refer to part 3 of the lesson for help with the following questions

6) Dr. Linda Stern and her colleagues recruited 132 obese adults at the Philadelphia Veterans Affairs Medical Center in Pennsylvania. Half the participants were randomly assigned to a low-carbohydrate diet and the other half to a low-fat diet. Researchers measured each participant's change in weight and cholesterol level after six months and again after one year. Explain how each of the four principles of experimental design was used in this study.

7) Does day care help low-in-come children stay in school and hold good jobs later in life? Carolina Abecedarian Project has followed a group of 111 children since 1972. Back then, these individuals were all healthy but low-income infants in Chapel Hill, North Carolina. All the infants received nutritional supplements and help from social workers. Half were also assigned at random to an intensive preschool program. Explain how each of the four principles of experimental design was used in this study.

8) Researchers in Japan conducted an experiment on 13 individuals who were extremely allergic to poison ivy. On one arm, each subject was rubbed with a poison ivy leaf and told the leaf was harmless. On the other arm, each subject was rubbed with a harmless leaf and told it was poison ivy. All the subjects developed a rash on the arm where the harmless leaf was rubbed. Of the 13 subjects, 11 did not have any reaction to the real poison ivy leaf. Explain how the results of this study support the idea of a placebo effect.

9) The progress of a type of cancer differs in women and men. Researchers want to design an experiment to compare tree therapies for this cancer. They recruit 500 male and 300 female patients who are willing to serve as subjects. Which are the block in this experiment: the cancer therapies or the two genders? Why?

10) A nutrition experimenter intends to compare the weight gain of newly weaned male rats fed Diet A with that of rats fed Diet B. To do this, she will feed each diet to 10 rats. She has available 10 rats from one litter and 10 rats from a second litter. Rats in the first litter appear to be slightly healthier.

a) Why would it be poor design to have the 10 rats from Litter 1 be fed Diet A, and the 10 rats from Litter 2 be fed Diet B?

b) Describe a better design for this experiment

CHAPTER 2 ANSWER KEY

SECTION 2.2

1) a) quantitative, usually discrete
b) qualitative
c) qualitative
d) quantitative, continuous
e) qualitative
f) qualitative
g) quantitative, continuous
h) quantitative, discrete
i) quantitative, discrete
j) qualitative

2) a) Weather condition is a qualitative variable (can be quantitative and continuous if looking at temperature). Absenteeism is a quantitative and discrete. The population is grade 9 students in our school. Sample is collected. Longitudinal study would be required to track attendance records over a period of time with different weather conditions.
b) Amount of television is a quantitative and continuous variable (measured in minutes). Physical fitness is a quantitative and continuous variable (using BMI). The population is adult females. Sample is collected. Cross sectional study would be easiest to do but longitudinal is an option.
c) Gender is a qualitative variable. Number of female students with speeding tickets is a quantitative and discrete variable. Number of male students with speeding tickets is a quantitative and discrete variable. The population is teenagers who have been issued speeding tickets. Sample is collected. Cross-sectional study is required.

3) a) The population is cars in North America.
b) Type of car is a qualitative variable. Value of car is a quantitative discrete variable.
c) A sample should be used.
d) Cross-sectional

4) a) Population is the group being studied when sample is a selection of individual taken from the population.
b) Cross-sectional study is a study that considers individuals from different groups at the same time.
Longitudinal study is a study of a single group (or sample) over a long period of time.
c) Quantitative data are numerical and qualitative data are non-numerical.
d) Discrete data is data that can only take on a finite number of values within a given range. For example, number of vehicles is a discrete data. Continuous data is data that are measurable with all real numbers and therefore can take on an infinite number of values within a given range.

SECTION 2.3

1) a) Cluster random sampling
b) Simple random sampling
c) Systematic random sampling
d) Multi-stage random sampling
e) Stratified random sampling
f) Convenience non-random sampling
g) Voluntary non-random sampling

2)a) Use randint(1, 600, 6) to randomly select 6 pages.
b) select random starting point using randint(1, 600, 1) and then select every 60th page (sampling interval = 600/10 = 60)
c) Divide pages into groups based on chapters. Take a simple random sample of 2% (12/600 = 0.02) of the pages from each chapter.
d) Divide pages into groups based on chapter. Do a simple random sample of chapters and then do a simple random sample of the pages within the chosen chapters.

3) a) simple **b)** cluster **c)** systematic **d)** multi-stage **e)** stratified

4) a) Use randint(1, 800, 80) to choose which students to give the survey to.
b) Use the grade levels at the strata. Within each grade's seating area, we'll select 10% (20) of the seats.
For 9th grade use randint(601, 800, 20)
For 10th grade use randint(401, 600, 20)
For 11th grade use randint(201, 400, 20)
For 12th grade use randint(1, 200, 20)

c) When using cluster random sampling, it is best if each cluster has the same characteristics as the population. For this reason, it would be best to use each column of seats as a cluster because that will ensure there are students of each grade level in each cluster. Because there are 20 columns (clusters), each with 40 seats, we need to randomly choose 2 clusters to get our sample of 80. Use randint(1, 20, 2) to select two clusters and then give the surveys to ALL of the students in those clusters.
d) Use randint(1, 800, 1) to determine a random starting point. Then give the survey to every 10th student (sampling interval = 800/80).

SECTION 2.4

1) i) info **ii)** info **iii)** rating **iv)** ranking

2) Good questions are specific, simple, relevant, readable. Good questions avoid jargon, abbreviations, negatives, being leading, and insensitivity

3) a) abbreviation; Should the Ontario Municipal Board be funded to initiate waste audits across the province? ☐Y ☐N
b) leading, insensitivity; Should mining companies be forced to decrease the amount of sulphur dioxide being emitted at smelters? Please comment.
c) leading, insensitivity; On a 5–point scale, do you agree that all corporations should pay higher taxes? (1: totally agree, 5: totally disagree)
d) jargon, negatives; Should forestry companies be able to cut down all trees in certain areas? Explain.

4) a) info **b)** open **c)** ranking **d)** info **e)** rating **f)** info

5) If you conduct a survey and collect information firsthand, this is called primary data. This type of data is easy to work with because you control how it is collected.

Information obtained from similar studies conducted by OTHER researchers is called secondary data.

6) a) NON-RESPONSE – some students will not return the survey
b) SAMPLING BIAS – the students in this one particular class may not represent all grade 12's. Students with similar traits tend to take the same types of classes. This is a convenience sample which will always lead to sample bias.
c) RESPONSE BIAS – since your are asking on behalf of the Greenteam, students may feel pressured to give answers that they know the Greenteam would like to hear instead of giving honest answers. Anything in the survey method that causes people to give incorrect answers creates a response bias.

d) HOUSEHOLD BIAS – not all classes are the same size. Classes should be surveyed proportionately, not equally. Smaller classes are over represented in this scenario.

7) NON-RESPONSE BIAS – company can block name from call display to reduce this bias

8) SAMPLING BIAS – because you are sampling only from the lower priced ticket holders, this will likely produce an estimate that is too small and not representative of the entire population.

SECTION 2.5

1) Experiment, because students were randomly assigned to the different teaching methods.

2) Observational study, because the researchers did not assign people to either use or not use of cell phones.

3) Type of school and socioeconomic status are possible confounding variables. Private schools tend to have smaller class sizes and students that come from families with higher socioeconomic status. If these students do better in the future, we wouldn't know if it was due to smaller class sizes or type of school or socioeconomic status.

4) a) Pine seedlings
b) Explanatory variable: light intensity, Response variable: weight of tree
c) Full light, 25% light, and 5% light

5) a) 300 people who haven't used Skype before
b) Explanatory variable: whether ads are present or not, Response variable: length and frequency of calls
c) No ads shown during calls and ads shown during calls

6) Comparison: researchers used a design that compares low-carb diets with low-fat diets.
Random Assignment: Subjects were randomly assigned to one of the two diets.
Control: The experiment used subjects who were all obese at the beginning of the study and who all lived in the same area.
Replication: There were 66 subjects in each treatment group

7) Comparison: Researchers used a design that compared children who were assigned to an intensive pre-school program to children who were not enrolled in an intensive preschool program.
Random Assignment: Subjects were randomly assigned to be enrolled in the intensive program or not.
Control: All subjects were healthy, low-income, and from the same area. Also, all subjects received nutritional supplements and help from social workers.
Replication: Over 50 subjects in each group.

8) The subjects developed rashes on the arm exposed to the placebo (a harmless leaf) simply because they thought they were being exposed to a poison ivy leaf. Likewise, most of the subjects didn't develop rashes on the arm that was exposed to poison ivy because they didn't believe they were being exposed to the real thing.

9) The genders, because researchers will randomly assign all three therapies within each gender.

10) a) If one of the groups gained more weight, we would not know if this was because of the diet or because of genetics and initial health. Genetics and diet would be confounded.
b) Use a randomized block design with the litters as blocks. For each of the litters, randomly assign half of the rats to receive Diet A and the other half to receive Diet B. This will allow researchers to account for differences in weight gain caused by differences in genetics.