

Chapter 4, Choosing Whom or What to Study

Intro. This is stuff that will be in the next exam, plus stuff you can use for making changes to your own questionnaire

- What do we do if we want to know if professor x is a good one? Sampling p.79
- But it is only one opinion, may be for personal reasons there was an issue between the friend and the professor, so the sample is small p.79
- A student who is motivated will give a different opinion than one whose purpose is only to pass the course p.79
- So sampling is a tricky business

- Population and samples

- (The Gallup poll and the election of Roosevelt in 1936)
- Box 4.3 about the 1936 elections and sampling error made by The Literary Digest p.93 (a lab about sampling?)
- So how do we get a good sample for a scientific study?
- First, you need to understand what it is that you are studying?
- You need to have in mind clear questions of what you want to know plus
- You need to know the range of empirical observations you are going to be making when exploring or researching these questions and ideas p.81
- You need to decide what or whom you are studying p.81 and get representative sample
- Box 4.1 on the referendum and the poll results p.80 (put it on the board)
- Do we want to study smoking habits for all students, for Cegep students, first year, second year, females, males, etc. p. 81
- The manner in which you select the sample will determine to which extent you can generalize the findings of your study to the entire population p.81
- (Doing a survey at the cafeteria or at the library, or in front of the college, on 300 students, what differences would that make?), on the importance of having the best possible grades

- What about if it had been conducted on every 5th students of various class last semester.
- What about every 20th student who registers in class? P.81
- It is more difficult than asking a couple of friends, but also more revealing p.81

The concept of a population

- What is a population in a study. It can tv programs, or radio ones, it can be an animal or human population, trees, buses, the population can be anything
- Some times the study of a population is difficult, for historians or anthropologists, they have incomplete data, to which extend can they rely on them? P.82
- The sampling of a census is not an issue because it involves everyone p.82
- History of the census, p.82-83
- So how do you select a sample? P.83
- By formulating a research question or framing a hypothesis, you already do part of the job p.83
- If you are looking for the most popular recreational activity among college students, than your population is college students.
- But is it in Canada, in North America, in Quebec, francophones, Anglophones, at Dawson you still need to be more specific p.83
- If it is Dawson, then you need to get a sample from that population
- **selecting the sample**
- there is the random method, or probability, or the non-random p.84
- The random one ensures that every member of the population studied has an equal chance of being selected p.84
- The non-random is used when it is impossible to make sure everyone has that chance p.84
- random does not mean you can select them in any way p.84
- the procedure ensures that no specific group of the population has a bigger chance to be selected in the sample, thus making it less representative p.84

- So first you need to have a complete picture of the population you want to study, that is a sampling frame p.84

- It can be the number of drug addicted people in Montreal, the number of clinics to treat them, veterans of Afghanistan, etc.

- So first there is no list of drug addicted people, so that is a problem, there is a list of veterans though, and a list of clinics

- So if you cannot get an entire list of the population, find something simpler to do.

- Once you have the list, you use a technique that ensures everyone in the population can be on it

- (speaking with the next person who will celebrate his or her anniversary)

- If it is a population of 500 students, write their names on a piece of paper of equal size and proceed with a draw to get 100

- Or you can proceed with a table of random numbers p.85, created by a computer

- From one number picked on the list, you move in one direction, and this is your sample p.85-86

- You decide on pattern to select the numbers, upwards, left to right, etc, p.87

- If it does not work for a number skip it p.87

- You can also use a systematic random sample, the next birthday, every 15th member of an association on a list p.87

- If you have the list of students at Dawson then

- You divide the population on the list by the number you want in the sample and this will give you the interval you need

- For instance with a list of 250 people and you need 75, then the interval is three, but you start by randomly selecting the first one, and then you proceed

- Example p.87

- **stratified random sample**

- suppose you want to compare the aspirations of males and females in a health science program, you have the list of students but you want to make sure your sample reflects the

proportion of the population in terms of gender p.88, this population is 60% female and 40% male.

- Some of you said I will get a 50-50 sample of students, why?
- You will use the stratified random sample, a variation of the stratified sample
- With the complete list of the population (ex. The census) social scientists will stratify their sample using a number of variables such as gender, age, ethnicity, occupation or income p.88
- In the health program there are 500 students, you need a sample of 100.
- One possibility is to divide the group between males and females, so 200 and 300 p.88
- Then you select randomly 60 females out of the 300 and 40 males out of the 200.
- So you have a stratified random sample

Multi cluster sampling

- What do you do with when you can't get a list of the population you want to study?
- Say you want to do a study with the graduating students of the province
- You will proceed by selecting a sample of schools, or a cluster and then survey the students who graduated from these schools p.89
- you can do the multistage cluster sampling by getting first a random sample of schools, then a random sample of students from each of these schools p.89
- It is faster and cheaper, but you may well end up with a biased sample
- Let's assume that you want to study something related to first year student in social science, and there is no student list
- But they all have to take Research Methods, and there are 40 sections,
- You randomly select 10 sections and survey all students in the 10 sections
- Or you do a random selection of the students in the 10 sections based on the list you did of them
- Random assignment involves assigning the population of your sample in one group or another for the purpose of the experiment p.90

- Non-random samples

- When you lack time or resources, you proceed with a non-random sample
- In history, you work with accidental samples, what is left of the past because of hazard, because of the accidents of history p.91
- You try to put it into perspective, how this info came to us, etc. so that you can establish the limits of your sample
- The convenience sampling, you go on and ask whoever suits you (le micro trottoir, the coalition and the interview at UQAM)
- You can get anything by asking people (like in ask a silly question)
- You want to carry on a study on nutrition and then you ask the students in your class p.91
- Is it reflective of the student population, you have no way of knowing
- The Shere Hite report on female sexuality, thousands of questionnaires distributed.
- One such example of sampling bias is discussed in Discovering Psychology with Philip Zimbardo (Episode 2: Understanding Research). Zimbardo explained that women in Hite's study were given a survey about marriage satisfaction, where 98% reported dissatisfaction, and 75% reported having had extra-marital affairs (after 5 years
- However, only 4% of women given the survey responded. Zimbardo notes that the women who had dissatisfaction may have been more motivated to respond than women who were satisfied. ABC News & The Washington Post later gave surveys on marriage satisfaction to women randomly (also known as a random sample). 93% reported being satisfied with their relationships, and only 7% reported having had affairs. Zimbardo's final note is to “beware of science-coded journalism.”
- Researchers do that to pre-test questionnaires p.91
- Or non-random are used in situations for which you can't do otherwise, that is for instance, you study people who were once homeless, getting a random sample of that population would be impossible
- non-random are used in psychology because they use the functioning of cognitive brain functions in healthy persons (students)

Purposive samples

- The sample involves selecting whoever the researcher judges has characteristics that meet the purpose of the study p.92

- If you want to do a research on female students who are single parents, there is no list and you may find it easier to simply take any female students single parent you know p.92

Quota sampling resemble stratified samples

- It is a non-random sample that is stratified according to some known characteristics of the population you are studying p.92

- Females, males, first year and second year student, etc. p.92-94 and their voting patterns

Snowball sampling

- When you find people that meet the characteristics of what you study, say anti-racist militants, and they give you names of other people like them p.94

- But of course it is a biased approach p.94

- But if you study a small group, white supremacists, it is difficult to get people, so it might be the only way of knowing p.94

- Type of samples on page 95-96

- Determining the size of the sample

- The size depends on the purpose of your studies, time, money, etc.

- For non-random it is not a big deal since you cannot generalize the results p.96

- The greater the scope, the greater the sample

- If you study Dawson Cegep population as a whole, you would need students in social science, commerce and other profiles p.96

- Other studies may not, say those who are practicing sport, how they manage it with their studies, etc. p.96

- The bigger a random sample is, the better it is p.97

- There is always a margin of error, or sampling error, because of the differences between the sample and the entire population it represents p.97

- We can establish the **confidence interval**, the value is accurate + or - 1.2,

- And there is the **confidence level**, it is accurate 19 times out of 20
- (he then talks about confidence level related to the size of the sample)
- Box. 4.5 the Margin of Error, examples of the above with the polls p.98
- It is not only the size of the sample nor the way you selected it that will determine your confidence, but the way you prepared the questions, how many refused to answer, etc.
- Examples, the probability sample established to study prostitutes in LA, complicated but it worked p.100
- (based on information given by informants, where prostitution was taking place, etc).
- The non probability is the most common sample, for reasons of cost p.100
- The internal conclusion are still valid but cannot be generalized

Various web sites, including a sample size calculator p.102, and a video on how polls are really carried out p.103

Chapter 5 Social Survey

- According to D. Abowitz and D. Knox, who did a study on college students, women more than men ranked goals related to individual happiness, interpersonal happiness and fulfillment higher than career goals p.104
- Then surveys on alcohol and blackouts, etc.
- The public is generally unaware of what makes a survey trustworthy or not p.105
- The dial-in polls are not scientific p.105 (although they do show something)

Casual and scientific inquiry

- The example of the casual inquiry about students who work part time and who do not, where the sample is 10, 5 who work and 5 who do not p.106 and the impact of that on academic performance
- This approach is flawed but it resembles the social survey p.106

The scientific approach

- Instead of the casual approach, you select randomly a sample, every 20th student on the student list p.106, systematic random sampling

- First, how do you define academic performance? P.106, is it only the grade average or other things?

- Then you construct a questionnaire. You can put it online, send it by mail, conduct direct interviews, etc. p.107

The aim and purpose of a survey

- The main purpose of these surveys is to describe the characteristics of a given population p.107 such as age and income

- In the case of a random sample, the only interest of that sample is the extent to which you can make inferences about the population

- They can be used to study attitudes and opinions about issues, consumption, etc.

- To examine relationship between 2 variables, say age or revenue and voting patterns

- Or to test a theory, human capital theory, the more you invest in your education, the more money you will earn p.107

- Descriptive research are concerned about how and who, and explanatory one about why

- You can find out who are the students who are working part-time, then you may try to find why p.107

- They serve to see if governmental intervention is needed in terms of health or social problems p.107

- If a national campaign against drinking and driving is achieving its goals

- Most surveys though are about attitudes p.107

- The trend study in the longitudinal study is a comparative study of the same population over time, at least at 2 different moments, but with a different sample each time but measuring the same thing, attitudes and opinion towards this and that

- The panel study does the same thing but with the same sample

- These studies are long and costly p.108

The questionnaire

- You need to think a lot about how you prepare the questionnaire, it not merely a matter of writing down a few questions p.108
- Should you ask for instance 1st year high school students if they own a computer?
- If it is their family, is it theirs, if it is something to play videogames, is it theirs?
- How are you going to organize and collect the data p.109

Guidelines for asking questions

- Box 5.1 on how to conduct a survey
- Respondents should understand the questions and it should measure what you are studying p.110
- (the Yes Minister Episode), why not?
- Suppose you want to know who is using the community center, in terms of gender and age,
- Your hypothesis should be measured, (women are and old people are using more the community center)
- Another theory to be measured, the unemployed vote less in a federal election p.110
- In order to properly measure what you want to measure, you need operational definition, in that case operational definition of the employed and not employed p. 110
- Who should we consider employed or unemployed?
- So you need a definition for employed and unemployed, it may or may not be the governmental definition
- Say that the important point is that the person works at the moment, period
- So the question to measure that can be “Do you presently work at a job or business, for pay or profit at the moment?”
- Each question measure a specific variable, gender, age, schooling, etc.
- This will enable the researcher to measure the relationship between unemployment and schooling, for instance p.110
- Box 5.2 p.111-112 on who are the unemployed

- Aim for clarity in your questions
- You can go for open-ended questions or closed-ended questions
- Open-ended, no specific choices
- Closed-ended, specific choices
- Closed ended is easier to analyse to an extent.
- Between these 2 options you have to consider what is more convenient for your research p.113
- Closed-ended will simplify the data analysis
- They should be mutually exclusive, that is no double answer possible, so the answers should be conceived accordingly p.113
- All the possible answers should be included, the example p.113 of the number of hours worked per week p.113 (10 hours and less, between 10 and 20 or 20 and more)

Problems with closed ended

- The problem is that, if you ask what is the most important problem facing Canada, the respondent will select a problem among the list, but he or she may have another answer in mind,
- Also the respondent who answers national unity will not be able to elaborate p.113
- And you are suggesting answers
- But open-ended questions are difficult to interpret (the example of the dropouts going back to school) p.114
- Examples of unclear questions: do you regularly go to the library? (what do we mean by regularly)
- What was your total income last year? (what do we mean by total income)
- Avoid double-barrelled question, example, “does the community center have a library and a daycare?”
- Avoid leading questions or biased questions, or a sequences of questions that will lead the respondent into one direction p.115 (the Yes Prime Minister Episode)
- After the Yes Prime Minister, give the examples on p.115 and 116

- Example of questions that encourage one answer, p.115, and the remedy
- Example of faulty sequence of questions p.116 and the remedy p.116
- Ask questions that the respondent is likely to be able to answer p.116
- Faulty example, when did you learn to count from 1 to 5, how much did you spend on magazines last year?
- Questions too technical that the respondent is unlikely to know should also be avoided p.117
- Negations in a question are grammatically incorrect and they make the question more difficult to answer

Guidelines for questionnaire construction

- When you prepare the questions, the emphasis is on the wording, for the questionnaire it is the sequence of the questions
- Prepare introduction for the questionnaire
- Then some instructions on how to complete the questionnaire p.118 and then perhaps specific instructions for some of the questions p.118
- spread the questions on each sheet p.119
- You may set up the answers with boxes and the respondent put an X or a check mark, or you put a number, which makes it easy for coding
 1. Yes
 2. No
- In certain cases, one answer will lead to contingency questions, only for some respondents, say if a student answers that he is a part time student p.119, then you put an arrow or you indicate skip to questions no. X, etc. p.119
- Matrix questions, same kind of answers to measure opinions, for this and that do you strongly agree, somewhat agree, somewhat disagree or strongly disagree p.120-121
- The opening questions should be easy and non-threatening p.121
- Difficult questions and open ended questions should be at the end p.121

- Then you pre-test the questionnaire p.121
- Note the problems and make the necessary changes p.121

Administering the questionnaire

- There are 2 ways, either the respondents fill in the questionnaire and give it back to the researcher or the researcher read out the questions and record their responses
- The respondents are doing you a favour p.122
- If it is a self-administered questionnaire, the response rate will be higher if the researcher is there
- By mail it will be less p.123
- The group administered has the advantage of the researcher being present but it is less time consuming
- But if someone makes a comment it might influence the other people in the group p.123
- If other people are conducting the study for you, you should train them p.123

Interview surveys

- Time consuming but efficient, better rate of response
- But the interviewers might influence the respondent p.130
- The telephone interviews, you get hung up, but you can enter the answers right away p.130

Lecture on analysing the data, populations and samples

Explain the acetate on populations and samples

Chapter 9: What are the results?

Analysing quantitative data

Coding is the first step.

1. single
2. cohabiting
3. married
4. separated, etc
5. widowed

6. no answer

- You should always leave the data in a more detailed state than you need, you can later simplify it p.217

- Say that you put separated and divorced into the same category, but later you find that there is a difference between respondent who are divorced and those are separated, than you have to start all over again p.217

Data summary

Box 9.1, p.219

- You can use excel to do that, you place the data in rows and columns p.219

- You can find the mean and the standard deviation when the data are in the spreadsheet p.219

- You can also create line graphs, bar charts and pie charts, see box 9.1 p.219

- Once the data is coded and the categories set up, you may start to summarize the data, something you can do by using a computer spreadsheet program p.217

- Let's say that it is a spreadsheet with gender, marital status and the type of music p.218
(acetate to be made)

Respondent number	Gender	Marital status	Music
01	2	3	1
02	1	5	4
03	2	1	1

- Row number 1 becomes 01 for that respondent

- Then you organize the data, and say that you want to carry out a univariate analysis (analyse univari e) with one variable, the music your respondents are listening

- Simple analysis, you add up the number of respondents for each musical category

- This is called the frequency distribution, you add up the numbers and get the % for each type of music

- See table 9.1 for that example, under f, the total number in the sample listening to this or that type of music, under %, what it represents in terms of % **(acetate to be made)**

- Another example of a frequency distribution taken from QM

- In this case, the categories do not have to be presented from the lowest to the highest, but sometimes it does matter, say with income, grades or level of education p.218

- In that case there is the cumulative frequency distribution, example p.220, box 9.2

- You want to present grades, but some grades have no frequency, the spreadsheet might be pretty spread, so you regroup the grades, say 91-100, 81 to 90, etc. and that is the grouped frequency distribution p.220

(acetate to be made)

- Box 9.2, example of a cumulative frequency distribution p.220

- The cumulative frequency allows you to locate the proportion of observations above or below a given point very easily

- All those who got between 71-80 and below, cf is 26 students in box 9.2

- You can add in another column the % along with the 26, that is 74.3%

Levels of measurement

- All data are not measured the same way

- On top of marital status, gender and type of music, we may want to include information on age, how much is spent on CDs and the respondents' favourite radio station p.222

- One is either male or female, but income, age amount spent on CDs vary much more p.222

- These data differ in terms of their levels of measurement

- Measurements refers to sorting out or making distinctions between observations in terms of differences in some specific characteristics

- Making categories of measurement in other words, typification

- In figure 9.1 p.223, we see the nominal scale, a name for each type of music, then ordinal scale, first choice, second choice, interval scale, -5, 5, 15, 25 degrees C, an interval of 10, then the ratio scale with income, 10 000, 20 000, etc.

- Explanation of various types of music and of the nominal level of measurement
- With the example of music, the data is measure with a nominal level of measurement, respondents are identified by their main musical preferences and put into the corresponding category, same thing with gender p.222
- Nominal scale also means that the categories are mutually exclusive p.222
- It allows the researcher to count the numbers in each group p.222
- Nominal level of measurement provides with a simple information, the number of people in each category, their %, the ordinal level allows you divide them in group but the researcher is also able to rank the categories p.222
- Which type of music is the favourite type of people
- The ordinal level puts things in order, first choice, 2nd, somewhat agree, strongly agree
- But it lacks precision, what is the difference between somewhat agree and strongly agree, difficult to tell p.222
- If you use an interval level of measurement, you can do that p.22, you can rank observations or data on scale with regular intervals

(The scale of agreement and disagreement from 1 to 10)

- It provides us with a difference, a more or less, and a quantitative way of measuring it p.222
- There is no true 0 for the temperature, for the time
- The problem is that no true 0 exist with the interval measure of measurement, so we cannot say that Ian, whose IQ score is 100, is 80% as intelligent as Ellen, whose IQ is 125.
- We cannot say that Monday was twice as warm as Sunday, because there is no 0 value with either IQs or temperature
- Where a true 0 exists, we can use the ratio level of measurement p.222, or ratio scale, same thing
- It specifies the difference in rank and the ratio numerically
- So we can say, if John spent 25\$ on CD and Mary 50\$, that she spent twice as much
- The level of measurement affects how we organize the data

- Certain statistical techniques require that we meet certain minimum levels of measurement p.224

Measures of central tendency

- At times researcher may want to have a single number that summarizes certain information, the measure of central tendency, usually located toward the center, around a value where most of the data are concentrated
- The class average is a measure of central tendency p.224
- We get a sense of where we stand compared with the rest of the class
- But this is not enough for researchers, they need something more specific
- There are three measures of central tendency, the mode, the median and the mean p.224
- The mode is the most frequent value in a distribution, like rock music in table 9.1, p.218

Isabelle's acetate on the mode

- If most respondents answered that they listen to music about 1 hour a week, that is the mode p.225
- The median is the middle point in frequency distribution, the point where there are half the cases below, and half above p.225
- Say you have the following distribution 1,1,1,2,5,5,6 the number of hours people listen to music, the median is 2
- But what if there is an even number of cases, say 8. You need to calculate the 2 median value and then calculate their average p.225
- 1,1,1,2,5,5,6,7
- 2 and 5 are the median value, their average is 3.5 and that is the median p.225
- Then there is the mean, symbolized by \bar{X} and above it, X is the raw score in the set of score and N the total number of scores, and ΣE , the sum of all the raw scores p.225-226

Isabelle's acetate with the mean

- Why is the mean sometimes insufficient, let's say that the amount of money spent on music is this: 10, 15, 20, 20, 10, 10, 15, 300
- The mean is 50 but it does not reflect the central tendency because of the 300, without it the average is 14,29
- But the median is 15, and it is a more reliable measurement p.226
- Because the mean is influenced by extremely large or small scores p.226
- Which one should we use, the mode, median or mean, it depends of the type of research one is conducting p.226
- The mode can be useful with ordinal, interval and ration levels of measurements,
- For instance you want to know which radio station is mostly listened to
- Median require that we rank the categories, so it makes no sense with nominal data such as gender p.226m, preferred radio station or type of music
- The mean is used with interval and ratio data, such as the average age of respondents
- When the mode, the median and the mean are equal, it means that the scores have a normal distribution p.226
- **Figure 9-2 p.227**
- Many physical or psychological characteristics are normally distributed, few people are very smart, very stupid, large, short, tall etc.
- Most people are around a certain average p.226
- For social characteristics, many are not following the normal curve in their distribution patterns p.226
- The proportion of population in certain age group does not, the distribution of income does not
- So the mode, the median and the mean are not equal p.226, these asymmetrical distribution, or skewed distribution p.226
- A negatively skewed distribution has a much larger tail to the left and the bulge to the right, and the positively one it is the reverse p.226
- The positive one would probably be like income distribution in society p.226

Measures of dispersion

- The mode, the median and the mean are useful summaries of the data but they do not, however, tell us how widely spread the values are p.226 and this information can be useful

- Let's say we have 2 different set of scores,

A. 10, 10, 14, 16, 17, 20, 25

B. 12, 13, 14, 16, 18, 19, 20

- Both of them have a \bar{X} (mean) of 16, but it is not the entire story p.227 since the scores in the first test are more spread out than in the 2nd one

- As a first measure of dispersion there is the range, it is simply the difference between the lowest and the highest scores

- The range for set A is $25-10=15$ and 8 for the second one

- Since the mean is 16 for both, we can now tell that this mean is more reflective in the 2nd set p.228

- The problem with the range is that it takes into account only the 2 most extreme values p.228

- In the CD example, 10, 15, 20, 20, 10, 10, 15, 300 the range is 290 whereas in all but one case the range is 10\$ p.228

Variance and standard deviation

- We use the mean to know how much we did compared with others for an exam for instance
- The mean is 70 and you have 78, what does it mean you can be proud of being 8% above the average p.228
- It does not mean that you have scored higher than most students in the class p.228
- A few extreme scores may have a big impact on the mean p.228
- There is a way of taking into account the difference from the mean of every score in the distribution and coming up with a measure that reflects the variability of the scores
- This can be done by calculating the variance p.228
- You calculate the deviation of each score with the mean and you square each one, you add them up and then you divide the sum by the number of scores p.228

S^2 means the variance

We have 8, 9, 11, 5, 2 (say money spent on cds per month in a sample, is 8 a value far typically close or far from the mean)

Sigma E of all the X (35) divided by N(5) = 7

So the mean = 7

Then you do $(X - X_{\bar{}})^2$ for each score to calculate the deviation and then

$$8 - 7 = 1 \text{ square} = 1$$

$$9 - 7 = 2 \text{ square} = 4$$

$$11 - 7 = 4 \text{ square} = 16$$

$$5 - 7 = -2 \text{ square} = 4$$

$$2 - 7 = -5 \text{ square} = 25$$

Then the means of all the square $1+4+16+4+25=50$

Then 50 divided by 5 (N)=10

The variance $S^2 = 10$

- But the variation of money spent on CDs was not squared, so we take the square root of the variance, of 10, = 3.2

- This is called the standard deviation p.228, SD
 - SD is equal to the square root of the sum of squared deviations from the mean p.229
 - The SD tells us what is the typical distance of the values from the mean, the average deviation if you want.
 - Excel has a function to do that automatically
 - In order to know if 8 above the average is good or not at the exam, we are asking how spread are the scores p.229
 - SD helps answering that
 - Say that 2 sections had the same exam
 - The mean is the same 16, but for section A the range is 12 and it is 10 for section B, but the standard deviation is 3.2 for A and 4 for B
- A. 10, 10, 14, 16, 17, 20, 25
 B. 12, 13, 14, 16, 18, 19, 20
- Without knowing the SD for B, you may have concluded that the grades were more clustered around the mean, but it is not the case
 - Despite a bigger range in section A, the spread is bigger in section B
 - The overall performance of the class, and the comparative performance of one individual student, is better assessed using the standard deviation than the only the mean p.230
 - This information is especially useful when the sample is representative and normally distributed p.230, when the mean, the median and the mode are equal
 - If the mean is 65% for the exam, then you know that there are 50% of the students above it, and 50% below it p.230
 - If SD is large, scores are largely spread around the mean, if it is small, they are close the mean **Acetate of p.231 with the normal curve**
 - Mathematicians have established that if it is a normal distribution, we are able to tell the proportions of scores within certain limits
 - Let's say that there is an IQ test, the mean is 100 and it is normally distributed and SD is 15 p. 230

- We can then that 34% or about of those who took the IQ test will be between the mean and above within one standard deviation
- In this example, it means that 34% of the students will be between 100 and 115
- Another 34% will be below the mean within the standard deviation, that is to say between 85 and a 100 p.230
- So 68% fall within the SD, above or below the mean, between +1 and -1 standard deviation p.230, figure 9.3 B p.231
- Then if you add 1 more standard deviation, you add 13.59% of the results, above and below the mean
- That means that 95.44% of the results fall between +2 or -2 standard deviation, below and above the mean p.230
- In terms of the IQ score, it means that 95.44% are between 70 and 130
- So there is only 5% outside of these values p.230
- The 1-2-3 rules

Tabular analysis

- Measures of central tendency are useful in understanding particular variables
- Researchers are in general interested in understanding the relationship between 2 variables p.230, that is what tabular analysis is
- Let's that a given research found that males and females enjoy different types of music p.230, you want to analyse the data taking this reality into account p.231
- So you set up a table with the independent variable on the X axis, above, p.231
- And the dependent variable on the Y axis, from upward downward p.231
- This is cross tabulation, with 2 variables, which is like setting up more than one frequency distribution
- Using 2 variables is to carry out a bivariate analysis
- See table 9.3 for cross-tab of music enjoyed by gender of respondent p.232
- Acetate of p.232 and 233

- There are 3 ways of presenting the percentages, calculating dividing the result of one cell divided by the total number of observations in the sample, like table 9.4, males listening to rock music represent 14.5% of the total sample
- Or you can compare the number of male rock listeners to the total number of male listeners, not the entire sample, 9.4 b
- Or you can compare it with the total number of rock listeners 9.4 c p.233
- So each of the table on p.233 is telling a different story, it depends of what you are looking for
- Since the purpose was to examine the relationship between music and gender, the % of music listener for each gender and for each type of music, table b, is what is interesting for us

Correlation

- The question is, what is the statistical relationships between the variables p.234
- When looking at table 9.4b, one can get to the conclusion that there is a relationship between gender and the type of music listened, since females listen much more to rock music
- **Acetate of table 9-4 to be made**
- This is what the table indicates, if it had shown similar results for males and females, then the statistical relationship would have been nil
- So a careful examination of the percentage allows us to discover any statistical relationship that might be present
- But the technique of visual presentation can also be used p.234
- You can put the data on a scattergram or scattergraph
- Let say that we assume that there is a relationship between the number of hours students spend studying and the results they get at the exam p.234
- Each point on the scattergram represents a score on 2 variables p.234
- The scattergram for class 1, p.235 shows a strong positive correlation, that is to say that grades increase with an increase on time spent studying p.234
- **Acetate of p.235**

- There is a negative correlation for class 2 and no correlation for class 3
- The more or less straight lines of the points, in class 1 and 2 indicates the correlation
- We may calculate the correlation coefficient, or the numerical value of the more or less straight line p.235, the more straight it is, the higher the correlation coefficient, represented by the letter r
- r varies from -1 to $+1$
- Do not forget that correlation is different than causation p.235
- (nominal variables means that the variables are in categories, exclusive, male, female, rock, classic)
- r tells us the amount of relationship between 2 variables (and to an extent, enables us to make some prediction) p.235
- With 2 nominal variables, the strength of the relationship is all that r can tell us
(Example, gender and the likeliness of listening to rock music)
- r can tell us more with ordinal, interval or ratio level variables p.235, like income, years of schooling, IQ, age, it can tell us the correlation but also the direction, negative or positive
- Computer programs easily calculate the correlations p.236

Visual summaries

- Data summarized by graphic forms are easier to read than tabular forms
- Computer spreadsheet will help build a graph but you need to be careful when deciding which type you want p.236
- The pie chart graph would be appropriate for table 9.1, the frequency distribution of types of music
- The number of sections should be limited or else it loses its usefulness
- There is also the bar graph, figure 9.6 p.238
- Correlation and causation, the example of the price of gas and the price of bus fares p.237,
- The bar graph is very good for comparison p.238

- There is also the line graph (used for polling information) p.239
- It is good for displaying values that change across time p.239

Lecture 3

Statistics and variables

- It is a branch of applied mathematics that is concerned with 2 areas of application
it p.240
- It is a kind of a language, or a way to think about information,
- It allows us to say complicated things briefly and with precisions

Statistics in our daily life

- We use statistics everyday

Descriptive statistics

- The first is descriptive statistics and it deals with the collection and classification of information as numbers
- You have already been using statistics in your life
- You had 10 courses over the academic year, you can look at your result 1 by 1 or calculate
- You can create a data distribution of your scores
- You can calculate the mean to give you a more complete picture
- The mean allows you to summarize or describe the data, so it belongs to the realm of descriptive statistics p.12

Examples of descriptive statistics

- Examples: my car's gas consumption is 20% less per full-up since its last tune-up, that is a saving of 15dollars a week
- Government revenue from tobacco are 10% lower, so are my grades, etc.

Inferential statistics

- 2nd branch, inferential statistics
- Let's first make a difference between sample statistics and population parameters p.13
- From a representative sample of 2000, of US adult population, we calculate that they watched 15.4 hours of TV last week,
- This is a summary measure of the sample, a summary characteristics p.13
- It is a summary characteristic of the sample p.13
- If you calculate the difference between the highest value and the lowest, from 0 to 38 hours, than again it is a summary characteristic of the sample p.13
- As opposed to that, when we talk about characteristics of the entire population, we talk about population parameter p.13
- There is an average of hours spent watching TV but it is too costly to calculate
- So we rely on the sample of 2000 to make some inferences on the number of hours that is watched by the entire population
- That is the branch of statistics called inferential statistics p.14
- The referendum polls for the 1995 October 30th referendum, 3 polls conducted within 5 days before the referendum

Angus Reid	52% No,	48% Yes	1029 respondents
Som	50.5% No	49.5 Yes	1115 respondents
LL	50.2% No	49.8 Yes	1003 respondents

- The study of the sample in itself is not interesting, it is interesting only to the extent that it lets us make inferences for the entire population
- But watch out for sampling errors p.14
- So the second area is called inferential statistics, it applies mathematical ideas to the organization of numerical data (numbers) in order to draw conclusions or inferences from it p.240, from a sample
- Inferential statistics also involves interpreting data to see what they mean in relation to a hypothesis

- Examples: what are the chances of my passing the next test if I postpone studying until the night before?
- Are the outline and the first three lectures of course enough to conclude enough to conclude that the course is interesting or not?
- Inferential statistics involve interpreting data to see what they mean in relation to a question or a hypothesis
- 2 questions then, do the data support our hypothesis and, 2nd, are they the result of chance and coincidence p.241
- Statistics are a toolbox and we have to use the proper tool p.241

In the real life (Youri and the documentary)

- People do different things with statistics, some work on the level income of some communities, others collect information about how voters plan to vote p.5
- These cases are similar to the extent that they all involve collecting information on a particular variable, be that voting intention or level of income p.5, etc.

Populations and samples, definitions

- What is the Canadian population and what does it include?
- (transparency with diagram in Isabelle's notes)
- A population is all possible cases that meet certain criteria, the total collection of all cases
- If you are interested in the grade point averages of students enrolled for 6 hours or more at a particular college, then all the students who met the criteria (all those enrolled for 6 hours or more), would constitute your population p.10
- The average income of the people working in Quebec, my population is...

(You are interested in all the male college students who have a girlfriend because you want to know how much money they spend on flowers), so your population is the male students with girlfriend

- Are there many studies conducted on an entire population?
- Some students drop others register late, so this population is in a state of constant flux (couples break up, new couples are formed)

- So there is always a margin of error regarding the data you can collect about a given population, because first it varies
- Second, it almost impossible to have access to an entire population
- And that is why we do sampling
- Box 4.1 on the referendum and the poll results p.80, as an example (yes why not)
- A sample is a portion of that population p.10
- The issue is, is this sample representative? p.11
- If in a poll they speak to the person who answers the phone, is it representative?
- A random sample is, how do pollsters do that?
- So how do we get a good sample for a scientific study?
- You cannot draw many conclusions from a non-random sample
- (example of a non random sample, the cafeteria)
- Did you conduct a small survey in RM?

More definitions of concepts

- For our purposes, a variable is anything that can and that can take on a different quality or quantity (age, weight, number of car accidents, etc.)
- Qualitative variables yield categorical responses
- Quantitative variables yield numerical responses
- Discrete quantitative variables are numerical responses which arise from a counting process
- Continuous quantitative variables are numerical responses which arise from a measuring process

Examples, write if it qualitative or quantitative and if quantitative, if it is discrete or continuous (take a sheet of paper?)

- A. Ownership of a ipod. (qualitative)
- B. The number of CDs purchased in the world in 2008. Quantitative and discrete. A number, a numerical value, and something you count. 1-2-3, etc.
- C. The playing length of the last Coldplay album. Quantitative and continuous. How long, we measure it.
- D. Favourite type of food. Qualitative.

- The data we collect in statistics is information about variables, the weight of students at Dawson, 130 pounds, 125 pounds, etc,

- The data altogether about the weight of the Dawson students is the data set

- The individual pieces of information are the data points p.5

- The amount of money male college students spend on buying flowers to their girlfriend, the amount of money, p.5,

- The amount of money on flowers spent by each male college students are the data points

- The overall bundle of information is referred to as the data set

- The example of a study in this class with 10 guys, and the amount they spend on flowers and where you collect information on 10 individuals

- So the data set consists of the 10 observations, or 10 cases

- A specific piece of information about one guy, the amount spent on flowers, would be a data point p.5

- The data distribution, is like a list of the values or responses associated with a particular variable in a data set p.5 (the variable is the amount of money, a discrete quantitative variable)

- In our case, the list of the various amount of money spent by the 10 male students

Lecture on experimental research

- Which brand of cola do you prefer?, Most people will say they prefer one brand over another p.138

- But can you really distinguish between two types of Cola? (exercise: the Pepsi challenge?)
- A simple taste test has many of the features of a real experimental study p.138
- Experimental research is a technique that aims at demonstrating and specifying or clarifying cause and effect relationships p.139
- In chemistry, experimentation will determine how much a mixture will need to be heated before its components will react and form a new compound p.139
- Experiment is viewed as the prime technique in scientific research, the basic logic of experimental research is also very much a part of everyday problem solving p.139
- For example, on a cold damp day, your car would not start.
- You do not want to spend money uselessly by calling a tow truck
- So you open the hood and you look for any loose electrical leads,
- Then you check the battery by turning on the lights
- You also clean the spark plugs or dry off the distributor cap
- Then you try to start the engine once again p.139
- Similarly when you cook each meal is an experiment p.139
- With every repetition you learn more precisely how much seasoning to use, how long to cook, and so on p.139
- Learning to cook better or trying to fix your car are 2 activities that involve the search for causes p.139
- What is causing problems with the car or the vacuum cleaner?
- What makes the sauce taste more spicy or less, or the cake to not rise or to rise p.139
- In scientific research, we manipulate one variable to assess its effect on another variable
- Experiment in social science poses different problems p.139
- First, it involves human beings, so there are ethical and legal considerations p.139

- When people are volunteering for an experiment, many things might be happening in their lives which will affect the experiment p .139 and over which the experimenter has no influence
- Furthermore, human beings may well change their behaviours just because they know that they are part of an experiment p.139
- Social scientists have developed various types of experiment to overcome these problems
- One type is the laboratory experiment, which is an artificial situation that the researchers create and can control
- Another type is the field experiment, which is conducted in real life settings where researchers give up some of their control over the situation p.139
- In both of these types of experiment, the researcher retains some control over the experiment p.139
- In contrast, the natural experiment involves no such manipulation, researchers observe changes in social behaviour as they occur after natural events, such as floods or earthquakes, or social events, such as plant closures or riots p.139
- Laboratory experiment is the model, the least expensive and the one in which the researcher has more control, so this is the one that will be mostly discussed in the chapter p.139
- **The logic behind experiments is causation**
- As we saw, one of the underlying assumptions of science is that it believes in patterns,
- For many researchers, the research process is one that moves from discovering and describing trends and patterns, and then understanding them and explaining them p.141
- Experimental research is mainly explanatory, it is designed to ascertain the degree to which one thing affects or causes another p.141
- So you go for a taste test of colas with friends who all have stated a difference, but then the taste test is inconclusive. So you still keep searching, why do they prefer one brand or another p.141
- Asking why is looking for a cause p.141
- The experimental method is about providing a means to find causes p.141

- However, neither nature nor social life comes already labelled as a sets of causes and effects
- Consequently it is difficult to discover and measure causal relationships
- The search for causes combines logical analysis and empirical research,
- So now we are going to focus on the 3 conditions to establish causality p.141
- First there is the temporal order, 2nd the consistent association and 3rd the elimination of plausible alternatives p.141
- So first the temporal order
- There must be a clear time sequence, such that the cause consistently appears before the effect p.141
- So if you want to know causal relationship between time spent studying and exam results p.141, obviously the time spent studying has to be before the exam p.141
- Say you want to know the effect a certain film on the dangers of smoking for students. So you give a group of students a questionnaire on the dangers of smoking. You show them the movie and then you give them a post-questionnaire movie p.141
- Before you can determine the effect of the movie, the students really need to have seen the film p.141
- In research, it can often be quite difficult to establish a temporal order p.141
- Let's take the example of what some social scientists have called the culture of poverty p.141
- It is characterized by, among other things, negative attitudes toward regular work and undisciplined work behaviour p.141
- But this theory has been plagued by the chicken-and-egg controversy p.141
- Some researchers claim that this is the result of the absence of work, people get discouraged and they develop negative views about working p.141
- Therefore, an independent variable, bad working conditions, cause a "cultural pattern of negative worker attitudes p.142
- Other researcher argue that the culture of poverty is learned, it is transmitted from generation to generation and moulds people so that they are incapable from the outset of fitting into stable routines (so they are the victims of circumstances, or of the system) p.142

- So what is the cause and the effect also depends on how you analyse the situation p.142
- (there is an ideological dimension here also), conservatives, it is because of the individual, left, it is because of the system p.142
- Condition 2, consistent association
- There must be a consistent pattern of association between cause and effect.
- In other words, whenever and wherever you see the effect, you will, at least in most cases, find the cause
- In the example of the film about the dangers of smoking, there is an association between the film, the cause, which appeared before their change in knowledge of the dangers of smoking p.142
- We presume that what the students learnt about smoking that made them change their attitudes, they got it from the film p.142
- But is there a vast change of knowledge and attitude after they watched the film, a slight one, are we sure that it is because of the film p.142
- To measure the scope of the change, we use statistical methods, for the question, is it really because of the film, we move to conditions no.3
- Condition 3, elimination of plausible alternatives p.142
- Before accepting the idea that there is a causal connection, researchers have to satisfy themselves that the connection is the best possible explanation p.142
- To make sure that their attitude has changed because of the film, the researchers must rule out any alternative explanations
- Watching the movie must come to be the best possible explanation p.142
- So how do we rule out that between the moment that they watched the film and the moment they answered the 2nd questionnaire, they did not participate in some activity that may have changed their attitude p.142
- Or maybe the pre-film questionnaire got the students to think about the dangers of smoking and it introduced a bias
- So we need to rule out other explanations, alternative explanations

- The experiment must enable us to do that by isolating various independent variables that might affect our dependent variables p.142

- One possibility is to repeat the experiment several times to see if the association between possible cause and effect is consistently found p.142

- Box 6-2 on how to conduct an experiment p.143

- The various steps in the experimental research

- What are the various steps in the experimental research, in the case of the laboratory experiment p.144

- Essentially, the laboratory experiment involves manipulating one variable while attempting to either control the influence of other variables or hold them constant p.144

- The focus is on the effect of the manipulated variable, the independent variable p.144

- In the example of smoking, showing the film is the manipulated variable p.144

- Step 1 formulating the experimental hypothesis

- The point of the experiment is to expose causation at work p.144

- Therefore it follows that you have to be clear about the relationship you expect to find between two variables in a setting over which you have much control p.144

- For example, what did you expect to find in showing the film on the dangers of smoking p.144

- Without a clear hypothesis, no well-designed experiment can emerge p.144

- One example to follow is Stanley Schachter, when he developed an experiment on anxiety, fear and affiliation p.144

- He had read a number of articles on religious hermits, convicts, and prisoners of war who had been socially isolated for long periods of time

- These accounts revealed high level of anxiety and fear p.144

- He asked himself, is it the isolation from others that creates these distressing reactions p.144

- His hypothesis was that when people experience anxiety, they turn to others p.144, something they cannot do in isolation p.144

- So this was his hypothesis, anxiety and fear lead to affiliation with others p.144
- So the first step in experimental research is to assert the relationship that exists between 2 variables p.144
- But then the next challenge is to devise a way to measure the relationship between the variables p.144
- So in the case of Schachter it was anxiety-provoking experience (independent) and affiliation (dependent variable) p.144
- One of the major means for the researcher to gather information is by asking them questions
- With children it is also done but the answers have to be yes or no p.145, to which extent can we ask the question
- **Step 2. Operationalizing**
- Designing an experiment requires considerable imagination and creativity p.146
- You have to make sure that those who are subject of the experiment are not subjected to unnecessary or excessively upsetting experience p.146
- Schachter developed a 2 group approach, the experimental group and the control group p.146
- The experimental group was to be subjected to a strong anxiety-provoking experience and its reactions were to be measured p.146
- The control group was to be placed in an identical situation or environment as the experimental group but without being subjected to such a strong anxiety-provoking experience p.146
- The disturbing or anxiety provoking experience is the independent variable p.146
- If circumstances like that provoke people to turn to each other (affiliation) then there should be a significant difference between the 2 groups p.146
- The anxiety experience consisted of telling people in the experimental group that they were going to be submitted to electrical shock p.146, and at that point an assistant brought a formidable machine to that effect p.146
- The assistant acted in a cold and distant way and said the electrical shock were going to be quite uncomfortable p.146

- The control group was not placed in such situations, the assistant was friendly and informal and went out of his way to emphasize the mild nature of the shocks p.146
- Before the shocks were going to be administered the assistant told each group that there was a delay in the proceedings while the machinery was set up
- Then he told them they could wait either together or individually in different rooms
- Before doing so they had to fill out a questionnaire p.146
- The questionnaire probed how anxious they were p.146
- When they finished completing the questionnaire, both groups were told that the experiment was over p.146
- The threat of physical pain is a stressful situation, as very few people are masochistic p.146
- This is almost universal, most of the time that is going to be the reaction, there is a trend here

- Step 3 Designing the study

- In setting up the experiment, the researcher attempts to rule out possible explanations other than the one intended p.147
- So that is why we use control group, who is not submitted to the experimental variable p.147
- By comparing the experimental group to the control group, we can actually rule out many other possible explanations than the experimental variable p.147
- That is why the use of a control group is so common p.147
- But other experimental design also exists p.147
- You can decide not to use any control group, as in the film movie,
- Sometimes it is simply impossible to set up a control group (too complicated) p.147
- There are different possibilities, based on the circumstances, the resources of the researcher, etc. p.147

- Step 4, selecting the subjects

- If the research design is effective, any differences between the control group and the experimental group will be the result of the experimental manipulations, not other factors p.147

- So it is vitally important to ensure that both groups are identical and that the only difference is that one experiences the experimental stimulus p.147

- So when you set up the experimental group and the control group, you have to make sure that both groups are identical p.149

- That is what the researcher has to do when he assigns individuals to each group p.149

- The most common method to do that is random assignment p.149

- It is like selecting a random sample, every individual must have an equal chance to end up in the experimental or the control group p.149

- So perhaps you select a number from a random table after you have given a number to each person taking part in the experiment p.149

- Step 5, determining internal and external validity

- If researcher could choose who is going to which group, it would bias the experiment

- A similar would occur if participants were allowed to choose in which group they want to go p.149

- We must eliminate all possible source of bias we can think of, so that we are confident that the experimental results (the difference between experimental group and control group) are purely the result of the experiment

- In such a case the experiment is said to have internal validity p.149

- We must also check the experiment design to see to which extent it is possible to generalise the conclusions, the generalizability of the experiment p.149

- If the experiment is repeated several times, are we going to come up with the same results p.149

- In the case of Schachter, we have to see if the results are going to be the same with men and women, old and young, etc. p.149

- If we can do that, the experiment is said to have external validity p.149

- Field experiment

- We have focused so far on the laboratory experiment, the environment is artificially controlled by the researcher p.150
- The disadvantage of this approach is that it does not recreate real life setting p.150
- Obviously things are more complex in the real life p.151
- So the field experiment is conducted in a real life setting
- Say you want to know if students would help a drunken colleague p.151
- Obviously this cannot be done with a lab. Experiment

Lecture on Ethical considerations

- There a number of criteria that we should bear in mind when we conduct research
- Although it is not always obvious to get a clear sense of the course to follow in some occasions
- Some criteria are more obvious than others
- **1. No physical harm, but also no psychological or moral harm**
- Obviously, everyone agrees with physical harm
- So the experiment should not be harmful for the subject physically, psychologically, and morally
- What about the order to crash the chalk, and the obedience experiment in class. Perhaps those who were asked felt demeaned after the experiment, perhaps their feelings were hurt, who knows.
- So the issue here is more tricky
- **2nd What about consent?**
- If individuals were asked to be part of the experiment they would probably behave differently
- So what do we do? The questionnaire on sexual behaviour in class.
- **3. Confidentiality**
- The result must be kept confidential, which makes sense if you want to convince people to give you information (this one is rather easy)
- **4. Invasion of privacy**
- To which extent can you move into the private life of individuals, trying to get the information you need. The answer is you cannot, not without their consent.

- 5. Deception

To which extent can you lie to say that you are doing a study on this and that, whereas you are actually doing something else.

6. Pressure

One should not put pressure on the subject if they are reluctant to carry on

- Ethical considerations, their origins

- The results of individual rights, authorities have to guarantee almost happiness to individuals

The case of Laud Humphreys

Background

- Laud Humphreys was a student sociologist in the 1960's and attempting to earn his Ph.D. at Washington University at the time of his research.

His experiment

- As a doctoral candidate at Washington University, Laud Humphreys began researching what he referred to as "tea room trade" or the act of fellatio between two anonymous men in public restroom.

- Humphreys spent time learning about this practice and determined to become an insider so that he could study these behaviours.

- Focusing primarily on the restrooms in public parks, Humphreys made himself a regular where these activities were displayed. He offered to be a look-out and warn the participants about unexpected visitors such as police officers by making noise to interrupt the sexual act before the participants could be caught and possibly arrested.

- Humphreys wanted to track those he observed so he would note the license plate on the vehicle of the participant.

- He observed and interviewed participants without their full disclosure regarding his intent to conduct research.

- He also manipulated information to obtain home addresses and used these later to interview participants again.

Results

- In total, Humphreys observed a time and place representative sample of 134 men and finally reported on 100 men due to attrition.

- He conducted 50 interviews in the tearooms and another 50 interviews with participants in their homes one year later, posing as a social health surveyor.

- In order for participants to open their homes to him, Humphrey's realized that he must change his appearance, demeanor and his automobile, so he did.

- None of the participants seemed to make the connection to their previous interactions with him.
- As Humphreys gained entry into this private system of instant sex, he learned that many (over 50%) of these men did not consider themselves homosexuals
- They were in fact mostly happily married, carried on very important roles in their respective communities and preferred to have quick sexual encounters, with few words and with men, all the while maintaining the appearance of being typical heterosexual males.

Commentaries

- Humphreys was committed to the research of this behaviour and took considerable risks to conduct this single study.
- He got arrested at one point with other homosexuals and so he kept silent about his research to make sure they would not know what he was doing.
- Due to the perceived dishonest nature of his study, his Ph.D. was later rescinded by the university.
- While the members of the Sociology Department at Washington University did not award Humphreys his Ph.D., many individuals today see his work of pioneering significance.