

Statistics 1013 Winter 2025

Readings & Exercises

Prep - Exercise : math review	pg. 3
<u>Unit 1: Foundations:</u>	pg. 5
Foundation 1: the research process	pg. 6
Foundation 2: Perspectives on Thinking	pg. 7
Foundation 3: Data (variable) types	pg. 9
Exercise 1a/b: data types & define/abstract actions	pg. 10
<u>Unit 2: Data Analysis – one variable</u>	pg. 13
Notes about Compute and Interpret actions	pg. 14
2a: one variable distribution characteristics	pg. 15
Exercise 2a-1: one categorical variable: visuals and computations	pg. 16
Exercise 2a-2: one measurement variable: visualize the distribution - dot plot	pg. 17
Exercise 2a-3: one measurement variable: computations	pg. 19
From dot plot to histogram (interpret centre, spread and shape)	pg. 21
Introduction to boxplot – another perspective	pg. 22
Exercise 2a-4: sketch boxplots from histograms	pg. 23
Review computations for one measurement variable	pg. 26
Exercise 2a-5: predict shapes of actual distributions	pg. 27
2b: The Normal Distribution – a mathematical model	pg. 29
Characteristics of a normal distribution: a mathematical tail	pg. 30
Exercise 2b-1: visualizing characteristics	pg. 31
Standardized normal distribution	pg. 32
Exercise 2b-2: z-score to areas under normal curve	pg. 33
The z-table: a useful tool	pg. 34
Normal distribution and percentiles – grounded in concrete	pg. 35
Exercise 2b-3,4,5	pg. 36
2c: introduction to SPSS as computation tool	pg. 39
Ppt: Introduction to SPSS (explore one variable)	pg. 40
Exercise 2c-1: Use SPSS to produce output for one variable	pg. 42
Exercise 2c-2,3: Use SPSS to produce output for one variable	pg. 43
Exercise 2c-4: SPSS boot camp one variable	pg. 45
<u>Unit 3: thinking tools when two variables are needed</u>	pg. 46
3a: variable relations: 3 concepts	pg. 47
Concept 1: Data type pairings	pg. 48
Concept 2: Independent and dependent variables	pg. 49
Exercise 3a-1: identify independent and dependent variables	pg. 50
Concept 3: measure strength of association	pg. 51
Computation: two categorical variables – comparing rates	pg. 52
Exercise 3a-2 relate visual to numerical two categorical variables	pg. 53
Strength of association: two categorical variables	pg. 54
Exercise 3a-3 calculate RR and RD	pg. 55

Unit 3a: two variable scenarios continued

Strength of association: 2 measurement pg. 56
 Exercise 3a-4 relate visual to numerical two measurement variables pg. 57
Strength of association: one categorical vs one measurement variables pg. 58
 Exercise 3a-5abc: relate visual to numerical - one categorical vs one measurement pg. 59
Compute Strength of association: one categorical vs one measurement pg. 62
 Exercise 3a-6: calculate and interpret practical significance – 1 cat vs 1 measurement pg. 63
Data analysis chart pg. 65

3b: Data analysis with SPSS as calculator

Exercise 3b-1ab: use spss to analyse scenarios with two categorical variables pg. 68
Ppt: solutions to exercise 3b-aabc with practical significance pg. 71
 Exercise 3b-2abc: choose correct approach pg. 74
 Exercise 3b-3: Data Analysis Boot camp exercises pg. 77
 Exercise 3b-4: More Boot camp questions pg. 83

Unit 4: Tools for Making Inferences (Quantitative Methods)

Exercise 4: from idea to interpretation – time studying pg. 88
4a: Inferential Statistics – a practical introduction (tossing the salad) pg. 91
 Exercise 4a-1 sampling simulation #1 pg. 92
 Ppt. Inferential statistics and Confidence intervals pg. 93
 Exercise 4a-2 sampling simulation #2 pg. 97
4b: compute inferences in scenarios with one measurement variable pg. 99
 Exercise 4b-1,2: calculate and interpret C.I. for one mean pg. 100
4c: compute inferences in scenarios with one categorical variable pg. 103
 Exercise 4c-1,2: C.I. for one proportion pg. 104

Unit 5: Interpreting the work of others

Template for reading research reports pg. 107
 Exercise 5-1: practice with reading research pg. 109
 Exercise 5-2: practice with ‘no data set’ scenario analysis pg. 110

Z-table pg. 111
t-table will be distributed separately

©2025 Taras Gula except for exercises 2a-2 and 3a-5abc, which were adapted from materials created as part of Rossman & Chance (2002) - see reference below.

Rossman, A., & Chance, B. (2002). A data-oriented, active-learning, post-calculus introduction to statistical concepts, applications, and theory. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/1/3i2_ross.pdf

Exercise 0: Math Review

1. When rolling 2 fair dice what is the probability that the sum is 11?

SS = n(SS)= E = n(E)= P(E)=

2. The probability of winning a game is $\frac{2}{3}$. Find the odds of winning the game.

3. Find a number that is approximately half-way between 0.05 and 0.025. _____

4. Victor is 147 cm tall and Victoria is 159cm tall. Victoria is _____% taller than Victor?

5. What makes a well-defined set well-defined?

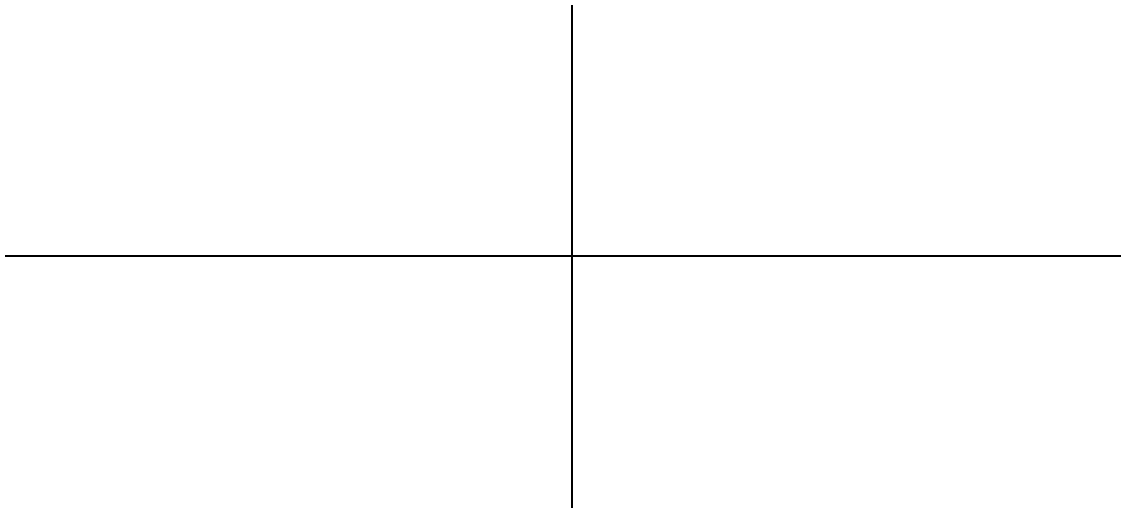
6. Take a good look at the contingency table below and answer the given questions.

		Arthritis		Total
		YES	NO	
Gender	Females	136	551	687
	Males	53	370	423
Total		189	921	1110

Who has a higher probability of arthritis, males or females (justify your answer).

7. Draw 3 lines on a Cartesian table and label them with the corresponding letter.

- a. Slope = 5 and y-intercept = -3 b. Slope = 0.01 and y-intercept = 2 c. Slope = -1 and y-intercept = 0



Unit 1: Foundations

Statistics – a term that everyone in this data drenched world of ours is familiar with, yet it is one that needs shaping as its meaning is context dependent. In this course you will be exposed to ‘statistics’ as a *formal* set of mathematical thinking tools that are a part of a systematic approach used in health sciences research. Quantitative information collected from a set of individuals will be used to build knowledge about them as a group and generalized to the *population* that they represent.

Thus, to know something about a population we must first know something about its individual members. There is a lot of information packed into a statistic like ‘*the prevalence of obesity in Canadian males has increased by 20% over the last 10 years.*’ The more we understand how research in health sciences works, and the more experience we have with (non-inferential and inferential) data analysis the more meaningful those statements become, and the better we will be able to discern their importance and usefulness.

Data – comes from an ancient term in Latin (pronounced *dar-eh*, and meaning *gift or to give*) and fits nicely with the idea that data comes from information *given* to the researcher by individuals about themselves - perhaps without informed consent. (e.g. *I am married with three children.*) Each bit of information (data point) about an individual member of a population tells us one thing about that individual (*I have three children*) but we need more information (about other aspects of the individual – i.e., more variables) in order to get to know more about the individual. More importantly for health science researchers, we need information from many individuals (more data) so that we can learn about a population. Statistical methods help us do that formally.

Variable – each variable is the name of a characteristic of a participant in research that we are interested in. (e.g. *marital status; number of children*). The term is rooted in the word vary which involves change. Every participant in a research project provides one bit of information per variable, the variation is rooted in the potential for a variety of responses from all participants, from within the population. In the real world nothing is perfectly constant. For example, blood pressure readings can change over a few minutes (*within subject variation*) while head circumference ranges from 35cm to 62cm (*between subject variation*). Variation may be *explainable by external factors* (e.g., by genetics), *by measurement error* (e.g., biases in data collection) or *unexplained/random variation* (there may be a reason, but we don’t know it).

Mathematical Models: are abstracted representations of concrete phenomena (e.g. *distribution of head circumference in population is Gaussian (normal – bell curve, uncertainty is modeled by a variety of probability distributions for example $P(E) = \frac{n(E)}{n(SS)}$*). Uncertainty and variation are core to health sciences research.

Reverse engineering - to learn how something works by taking it apart and studying its components. This is not a course in which you will learn how to build mathematical models for statistics, instead we will be taking apart existing models to see how they work so that you can have a grounded understanding of how researchers use them to build knowledge.

Foundation 1: the research process – building knowledge

A visual representation of the quantitative research process is pictured below. As you can see the analysis of data (#6) where we do statistical calculations and reasoning is but one step out of 8. The goal is to build knowledge in a community of scholars.

Figure 1: visualization of the research process.

Idea: Ideas come from the creative and curious parts of our brain activity and are the easy part of the research process. The challenge is to take ideas and refine them into a clear research question (RQ) or set of research questions.

Lit Review: Ideas need to be checked against previous research. The means reading previous research on the topic you are interested in studying (i.e. find out what is already known) or methods you wish to use.

Design: (abstraction) describe a research approach that will help answer the research question(s); include list of variables, description of target population, and description of sampling, recruitment, data collection and analysis methods.

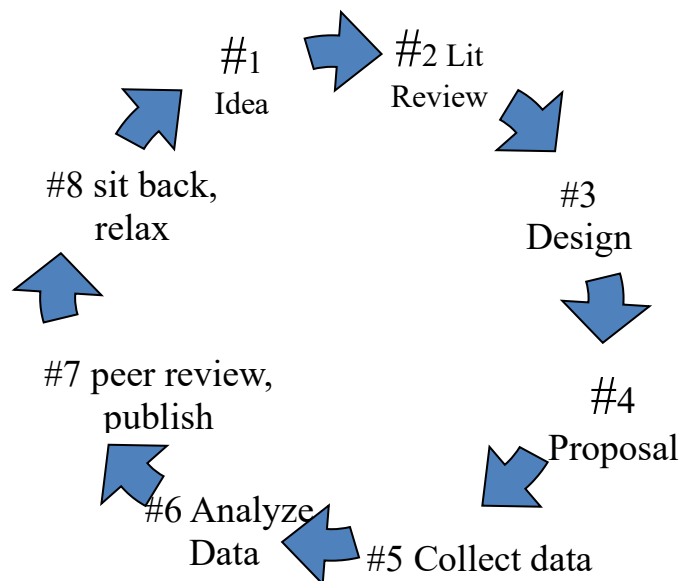
Proposal: write a proposal for funding (and ethics clearance) which describes the research design and expertise of personnel in detail for the committees who will review it.

Collect data: Must be done systematically and follow the design (#3) that was proposed (#4) to the funders of the research. There are two ways participants give their data – by answering questions (survey or interview) or by allowing themselves to be measured (observation).

Analyse Data: (calculate and interpret) look for patterns in data by building visual and numerical representations of the population (and fits them to mathematical models?) using quantitative statistical tools. Interpretat results through a search for patterns and evidence used to help answer research question(s).

Peer Review, Publish and present: Peer review is a formal process in which experts in the field assess the validity of the methods and conclusions of the research that was conducted.

Sit Back and Relax: this is even easier than the Idea part, but no less important.

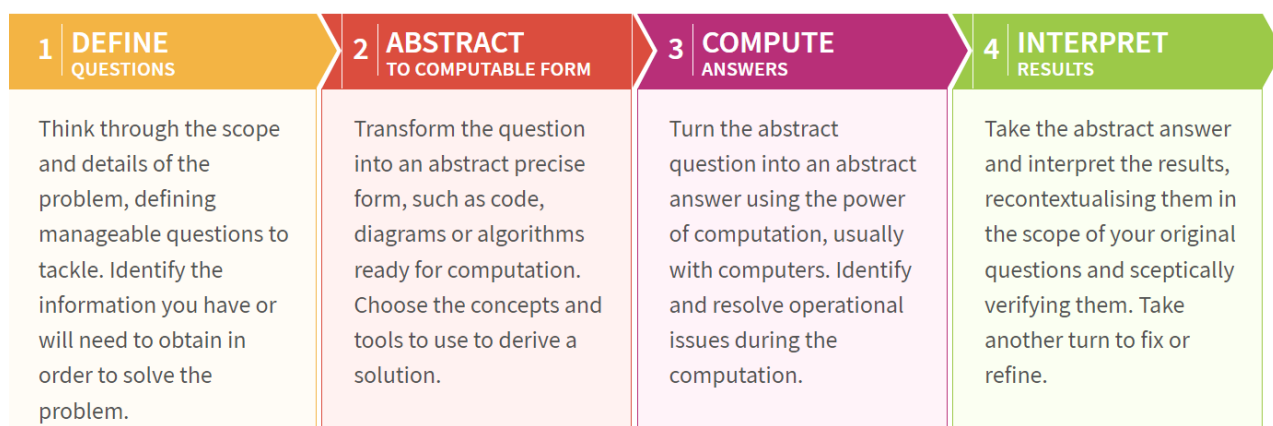


Foundation 2: Perspectives on Thinking. Four Thinking Actions – a proposal.

Each of us is exposed to a lot of information in the world around us. Some of that information is noise (*do I really need to know that the Canadian dollar went down \$0.0005 today vs the US dollar?*) and some of that information is useful to me (*garbage truck took out TTC wires at Spadina and King meaning that my commute will be 30 minutes longer*).

The noise around us often asks us to react quickly - to ‘think fast’ - and sometimes that is necessary, but we do have the ability to slow things down and I hope that this course will help you do that.

When we solve ‘real world problems’ we spend a lot of time jiggling between concrete and abstract thinking spaces (evidence shows that we actually activate different parts of the brain for each). Conrad Wolfram’s four part model of computational thinking recognizes the need for this jiggling and suggests the following actions we need to pay attention to. Note that they fit rather closely with the research process described on the previous page. I’d like to propose that we take the time to think about each activity/exercise/test and which of the following thinking actions they provoke.



Wolfram’s computational thinking model: [Computational Thinking: Be Empowered for the AI Age](#)

Extend Wolfram’s model to intro statistics for health sciences context:

Define Question - create a Research Question (RQ) that is grounded in the health science phenomenon that is being studied. A good RQ points to abstraction (*i.e., to the variable(s) and data that are to represent them*) and to the population that will be studied.

Abstract: What data will be needed? What variables? What will be their types? Quantified information (variables) will be collected as data coded as numbers. Abstraction may force a reconsideration of the RQ. By defining these you are defining the *mathematical model* you will use to generate evidence that can support your answer to the RQ.

Compute: ask (prompt?) a machine to produce the numerical and visual output (*depending on the mathematical model you chose in abstraction*) that will describe the patterns in the data as distribution (or calculate by hand 😊). The mathematical models you will use in this course have been packaged into functions that you choose from a menu, but you need to choose the right one.

Interpret: reconnect the abstract to the concrete by directly answering the research question using the results from Statistical computations to make predictions about populations and individuals in the future, while thinking carefully about the reasonableness of the results as a way of double checking the computations.

Foundation 2: perspectives on thinking (continued)

The 4 thinking actions researchers (and designers of computational models of various human systems) engage in are important to consider, but we will not be engaging fully in all of them in this course.

We will be spending more time deciphering the work others have completed – that is, analysing data sets and output from data sets. They did the Defining and Abstracting (and data collection and entry) work and we'll be picking up at that point and helping make sense of what story the data is telling.

From the list (Define, Abstract, Compute and Interpret) you will be spending a lot of time practicing with computation and interpretation – but hidden in those is figuring out the work that the researchers must have done to get to the completed data set.

Reverse Engineering: the idea of reverse engineering is to take something apart to see how it works e.g., a bicycle (How do they not fall over?) or a claim about vaccine effectiveness (How are claims made? What does the phrase '95% effective' mean? Where do these claims come from?)

When ideas, models or concepts are reverse engineered they are brought down to ground level – down from the abstract concepts, mathematical formulas to the practical day to day considerations.

It is hard work to do that, as the machinations of researchers are harder to see than the machinations of the bicycle – it requires going into a lot of detail that our intuitive brains typically avoid – preferring the 'good enough – I get it now' approach.

The goal is not only to take things apart, but to practice putting them back together again

Foundation 3: Data (variable) types: key to the abstraction thinking action (and interpretation) is deciding on (or figuring out) what the data ‘looks like’. There are 2 types of data representing two different ways human characteristics vary. Getting comfortable with that fact and how to use it will really help with all statistical work you may be asked to do in this course.

Data (variable) types: Human characteristics vary in different ways, thus there are different data types. In this course we’ll use the following two: categorical (e.g. *classify individuals by marital status: single, married, divorced...etc.*) and measurement (e.g. *head circumference is measured as a distance around in cm or mm – or inches*).

Categorical data: each individual participant is *classified* as belonging to one of a series of categories. Information about each participant is turned into information about a group through a *count* of individuals in each category; subtypes are nominal and ordinal.

nominal: classify by description; order is irrelevant. (e.g. *diagnosis of disease, political party preference, or marital status*)

ordinal: classify by order, no need for consistent difference between ranks. (e.g., *letter grades, birth order, Likert scale on a survey: agree, disagree, strongly disagree*)

Examples of research questions requiring the collection of data for one categorical variable.

RQ1. What is the sense of belonging in local community in ‘town X’?



RQ2. What is the prevalence of Covid in the city I live in?

RQ3. Which political party would win the election if one were held today?

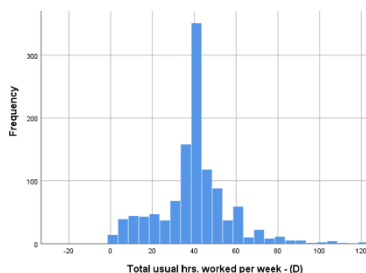
Measurement data: Often measured by mechanical means – called observation (e.g. head circumference, earnings per hour, grip strength), but could also be survey; subtypes: discrete and continuous.

discrete: values correspond to isolated (discrete) points along a line i.e. there is a gap between any two values. (e.g. age in years)

continuous: measurement variable that assumes any value, including every possible value between any two values. (e.g. weight on a super precise scale)

Examples of research question requiring the collection of data for one measurement variable.

RQ1. What is the total usual hours worked per week of 2000 Canadians?



RQ2. What is the duration of infectiousness for those who tested positive for Covid?

RQ3. What is the number of ICU beds available in Ontario hospitals?

Exercise 1a: recognize data types

You should have worked on the HNP webapp to develop your skills with recognizing data types before the first day of class. The following will be discussed in class, but the discussion will be much more meaningful after practicing with the app. After completing this exercise practice with statcat.ca as well.

Your task: Recognize the following research scenarios as having one or two variables. Name the variable(s) and the corresponding data type in each case.

Which type of hospital (teaching, community, research) has the highest occupancy rate?

What is the mean age of retirement of CHIMA members?

What is the typical LOS of covid patients who are hospitalized?

Is there a relation between income and Length of Stay of covid patients?

Exercise 1b: Define and abstract. (define and abstract)

Before you start: What does ‘quantification through classification and counting’ look like?

Idea: Are students who exercise regularly more successful in their college courses than those who do not?

Step 1: The individuals being studied are college students – most likely by survey as it would be too much work to try to follow them over a longer period of time and track exercise habits.

Step 2 (Abstraction): What characteristics (variables) are needed? Exercise as a categorical variable (yes/no is simplest set up as a comparison of those who exercise (>30 minutes per day to account for volume?) to those who do not; can also make it a measurement variable as something like minutes per week of exercise – categorical is simpler. The second variable is ‘successful in their college courses’ which could be a categorical variable (like graduated yes/no) or a measurement variable (like GPA). I will go with GPA!

Step 3: Reframe the research question: Do students who exercise more than 30 minutes per day have higher GPAs than those who do not?

Step 4: Independent variable ‘*Exercise*’ categorical; 2 categories: <30min per day; >=30min per day. Dependent variable ‘*GPA*’ measurement.

The task: Use abstraction to formalize the given research idea into a variable (or variables?), and reframe it into a clear research question.

Research idea: Students have been relegated to online learning over the past 2 years. What sorts of strategies have they employed to stay strong physically throughout the pandemic

List the variable name(s) and type(s)

Unit 2

Compute and Interpret

Scenarios where one variable is enough

In this section of the course, you will use a series of statistical tools to search for patterns in data one variable at a time.

Compute and Interpret are often referred to collectively as data analysis: the act of making sense of information (about human or other beings) that has been collected and coded, through the search for patterns. Recall that each variable (presented to you in a data set) is an abstracted representation of one characteristic of the individual beings, events, or other objects that are being studied. Researchers have already defined the question and abstracted for you.

The information collected will arrange itself into a distribution with the help of statistical tools. Computers need to be asked to produce appropriate visualizations and numerical output for the research question to be answered. Before giving the computer a task, it is important to know what computations we want it to do. To make the right choice you will need to reverse engineer the scenario (see next page)

In this unit all scenarios will involve only one variable, meaning that you'll only have to figure out the data type for one variable at a time. Thus, you will have only one decision to make as the analysis approach (computations) is dependent on the data type: take a look at visuals below.

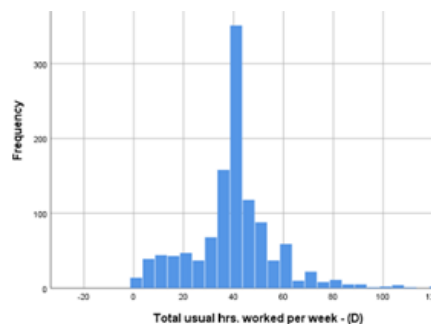
pie chart (for categorical variable)

Sense of belonging - local community

- VERY STRONG
- SOMEWHAT STRONG
- SOMEWHAT WEAK
- VERY WEAK



histogram (for measurement variables)



Compute and Interpret actions: formal process for producing a valid answer to the RQ being considered.

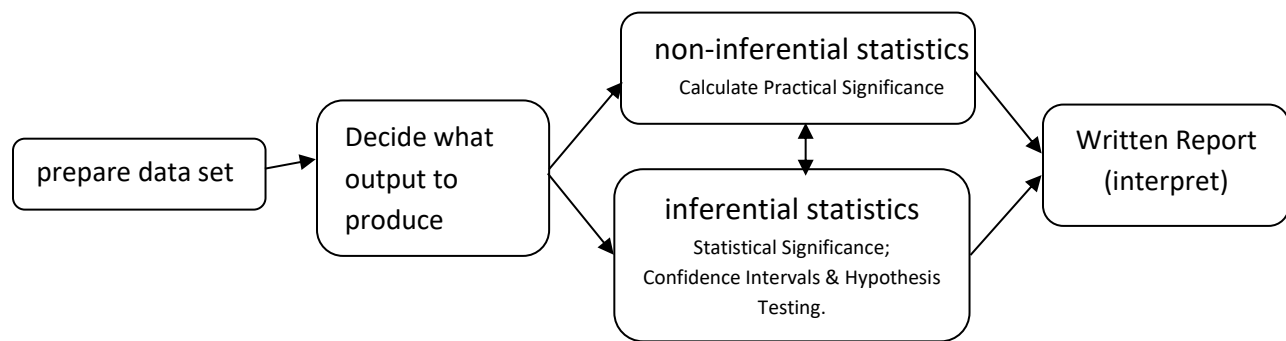


Figure 2: visualization of elements of data analysis

Learn this visual by heart!!

Data Preparation (you will not learn the task of preparing data for analysis in this course)

process: enter data into software, clean up data entry errors, and deal with missing data and/or outliers. If data set is large and/or ‘messy’ then this can take days or weeks.

product: a nice clean data set that can be used to help answer the RQ.

Decide what output to produce

process: how many variables are you working with? what types of variables are they? What are the statistics needed for a valid answer to the research question?

product: summary statistics (visual and numerical output) of the variable(s) relevant to the RQ.

Non-inferential Statistics (calculations of practical significance) (for scenarios with 2 variables)

process: calculate, with support from SPSS, appropriate standardized measures relevant to the data types given by RQ. (Pearson’s r, difference between means, relative risk).

product: non-inferential answer to the RQ. How strong is relation/association between variables? Practical significance measures strength of association between 2 variables and is called *clinical significance* in medical research, or *measure of effect* by statisticians.

Inferential Statistics (calculations of statistical significance)

process: calculate confidence intervals, hypothesis testing (calculate the infamous *p*-value).

product: Inferential statistics help us convince the sceptical audience that our findings (e.g. this vaccine was effective for 950 out of 1000 individuals) will be true for individuals beyond the participants in the study. More formally, we get an estimate of what the results may look like in the population beyond the data set and hopefully a confirmation that the observed relations (practical significance) indicate that there is something more than chance involved as we generalize results to the population.

Written report (Interpret)

process: step 1: explore the output to look for any surprises or think about how results from A, B, and C help answer the RQ and compose into a story.

product: a written report that answers the RQ and provides inferential and non-inferential evidence to support the answer. (In some studies non-inferential tools could be sufficient.)

2a Compute and Interpret scenarios with one variable as forwards/backwards thinking.

Goal is to become comfortable with the characteristics of categorical and measurement variables. Given that the nature of measurement data is more complex there will be many more exercises with measurement variables.

You may need to think in a ***forward*** direction (design) or in a ***backward*** direction (reverse engineer to figure out what the researchers that did the design were thinking when they did the design).

In this unit you will start from the middle and work forward: the research question will be set for you and the data set will be ready. You will need to be a reverse engineer to figure out what the researcher was doing (work backwards) then based on what you find work forwards and get SPSS or other tools to compute numerical and visual output, then interpret what you found to answer the research question (RQ).

Your tasks will involve choosing the right computations, making them, and then making valid interpretations.

You will also learn about the various tools created by mathematicians and other scientists to help us make valid interpretations when investigating one variable at a time.

Exercise 2a-1: Computation in scenarios with one categorical variable

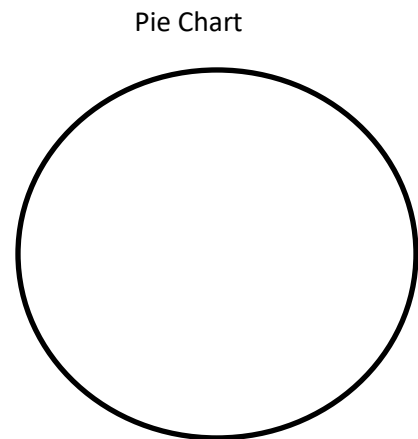
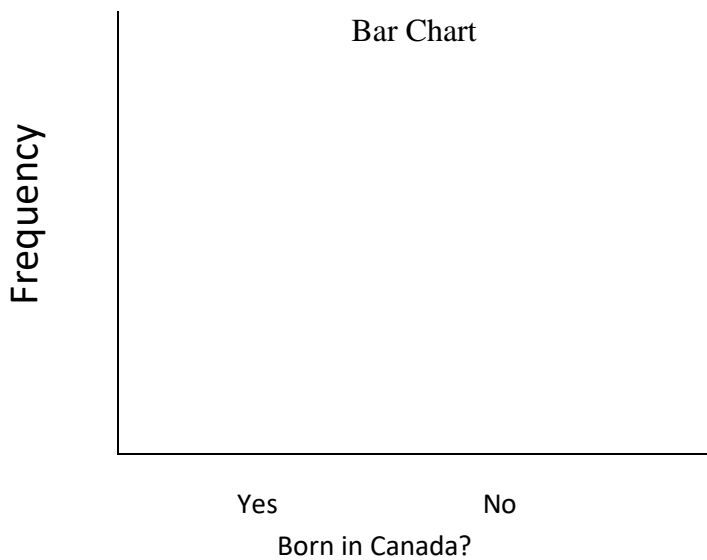
The following is a data set of a categorical variable which comprises a set of answers to the question: Were you born in Canada? 1 means 'yes' and 0 means 'no'

Data set {1,0,0,1,1,1,1,0,0,0,0,1,1,1,0,0,1,0,0}

1. Fill in the frequency table (numerical and visual output below):

Frequency Table:

Born in Canada?	Number	%	Cumulative %
Yes			
No			
Total			



2. What is being quantified in this scenario? How is the quantification happening?

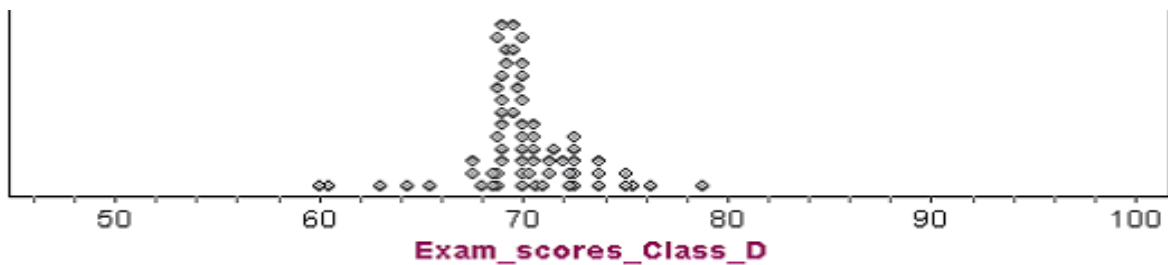
Exercise 2a-2: Interpret dot plot distributions in scenarios with one measurement variable

Dot plots are a close representation of the raw data as each dot is an individual raw score (one case).

We can explore them particularly (i.e. investigate how one point is positioned in the distribution relative to the others) or generally (describe characteristics of the distribution as a whole).

1. Take a look at the dot plot of Class D exam marks below.
 - a. Describe particularly: Is a mark of 75 a high mark? Would you be happy with 75?

- b. Describe generally: If you wanted to summarize how the students did on this exam in general what would you say?

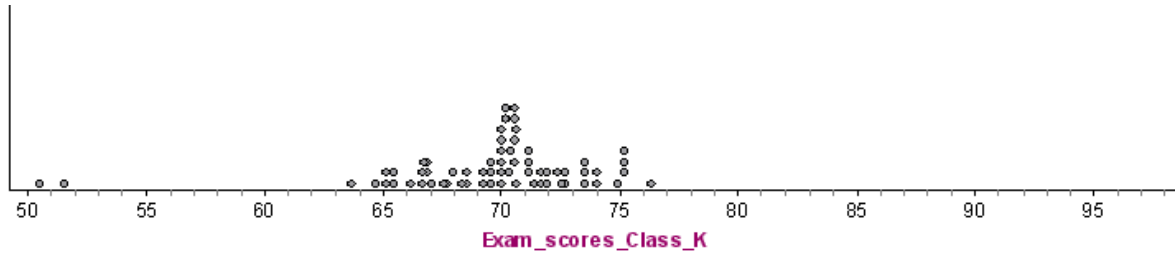


Exercise 2a-2: continued

2. Take a look at the dot plot of Class K exam marks below.

a. Describe particularly: Is a mark of 75 a high mark? Would you be happy with 75?

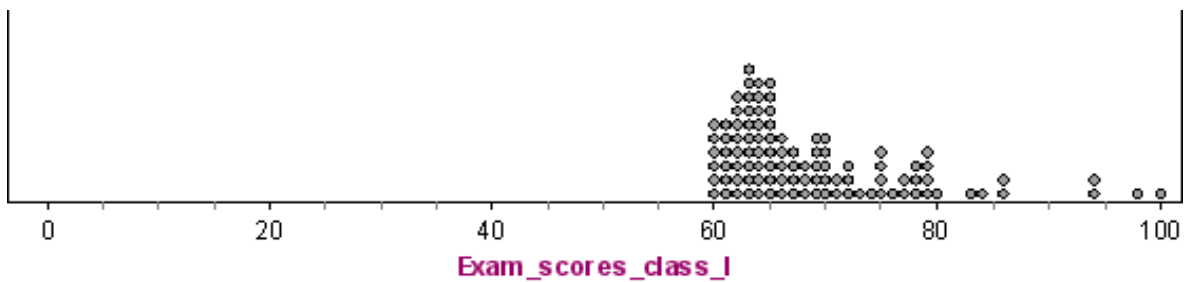
b. Describe generally: If you wanted to summarize how the students did on this exam in general what would you say?



3. Take a look at the dot plot of Class I marks below.

a. Describe particularly: What is a 'good mark' in this class?

b. Describe generally: If you wanted to summarize how the students did on this exam in general what would you say? What is unusual in the distribution of marks in class I?

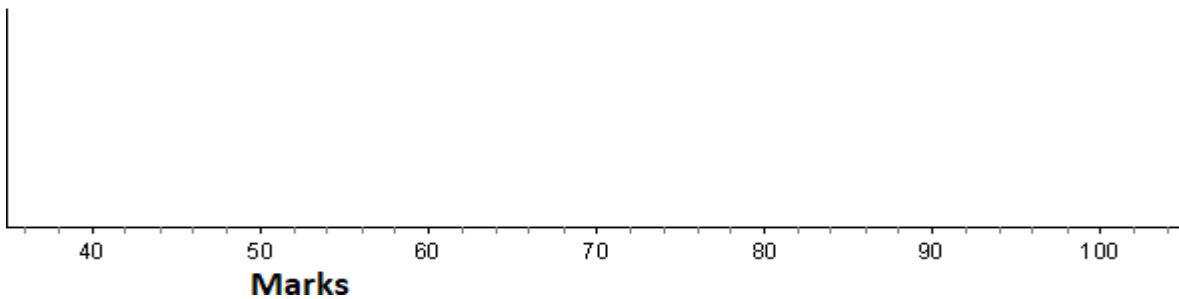


Exercise 2a-3: Calculations with one measurement variable.

Answer the questions posed for each of the distributions pictured as dot plots below. Each dot placed on the graph will be a separate participant (student) in the study and the number of dots rising above a particular mark represents the number of students who had that particular mark on their exam (the frequency).

- 4. Take a look at the list of marks of 15 students, then plot them individually in the space below with one dot for each individual mark.

Marks = {38, 43, 45, 51, 51, 51, 64, 67, 72, 74, 77, 77, 85, 85, 85, 90, 91}



- 5. Recall summary statistics that you learned to calculate for measurement variables; We'll add one more here to the list

Calculations of centre:

mean= median (50th percentile) = mode =

Calculations of dispersion:

standard deviation = mean deviation = $\sum_1^n \frac{|x_i - \mu|}{n}$

Estimate the following:

25th percentile = 75th percentile =

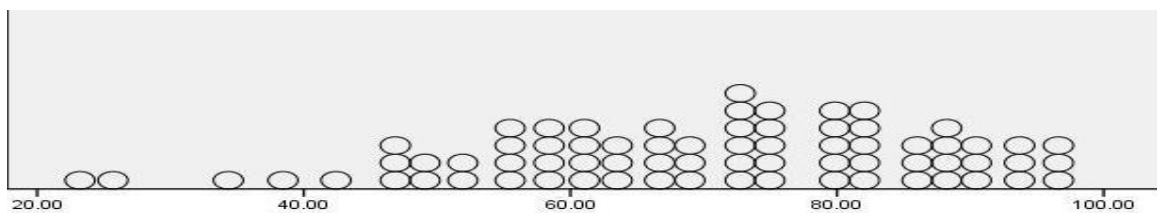
From dot plot to histogram (shape, centre and spread)

In Exercise 2a-2 and 2a-3, you investigated the distributions of marks in a class by looking at the *dot-plots*. They allow you to use *particular* individual scores, and to start looking at *general* trends, (by using **centre, shape and spread** to describe the distributions).

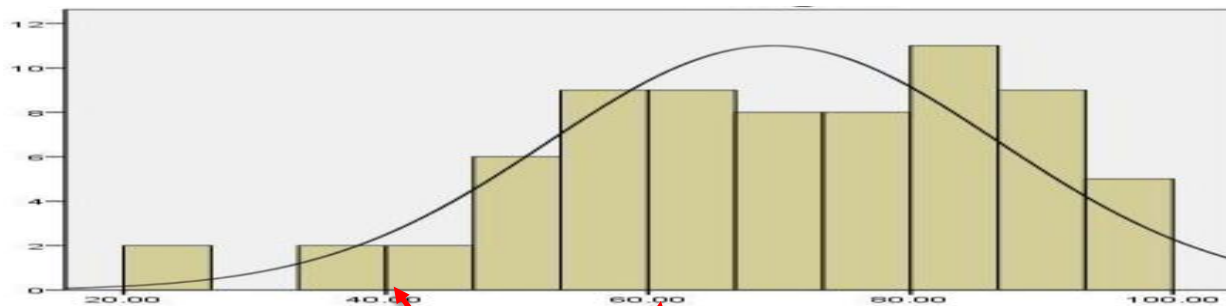
Dot plots are rarely used in visualizing distributions of measurement variables, we typically use the histogram and boxplot instead as the variation is smoothed out, while with the dot plot you see all of the noisy individual values.

The *histogram* is a very close relative of the dot plot in which the horizontal axis is chunked in to classes with a range of values, and the vertical represents a frequency. This usually gives a smoother flow to the shape of the distribution, especially with larger data sets

Below you can take a look at a dot-plot and histogram of the distribution of math1112 class marks from many years ago.



Dot plot of distribution of math 1112 marks



Histogram of distribution of math 1112 marks

The dot plot shows a much more precise view of the distribution as you can see each particular score. SPSS has generated the histogram using classes of width of $20/3 = 6.66667$. (I can tell that because there are three bars between 40 and 60.) Both tell us something about the distribution of marks in the population because we can see which values (marks) were possible, and how often they appeared. In a histogram we get a better sense of the general shape (or lack thereof) of the distribution without all the noisy detail the dot plot.

Visualize one measurement variable: shape, centre and spread.

Distribution: How is the data distributed? We will answer this question by using three tools: shape, *centre and spread*. Numerical values and visualizations will both be used to help us recognize patterns.

The distributions pictured to the right are interesting in their **shapes**. Shape can best be assessed through a visual representation of the data (histograms come closest to the curves presented to the right). The ‘*rectangular*’ distribution is also known as the *uniform* distribution. *Skewed* distributions are defined by the direction of the stretch (or tail) not the side on which the hump is found.

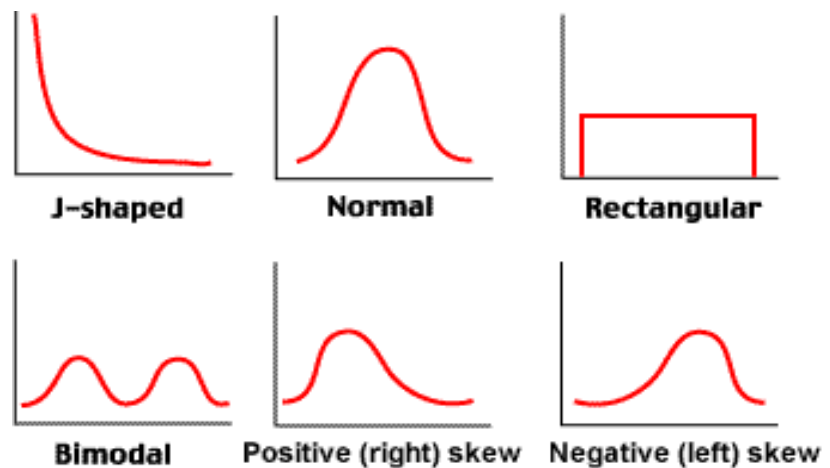


Figure 3: distribution types by shape

Centre: Where is it? Knowing the centre of the distribution tells us one characteristic of the population that we are studying, and allows us to compare two or more groups. (e.g. on average males are taller than females) without focusing on the variation.

In figure 3 above finding the centre is easy for the Normal and Rectangular(Uniform) distributions, but is much more tricky for the J-shaped, Bimodal, and Skewed distributions. Having three different measures of centre (mean, median, mode) allows us to capture different kinds of centre.

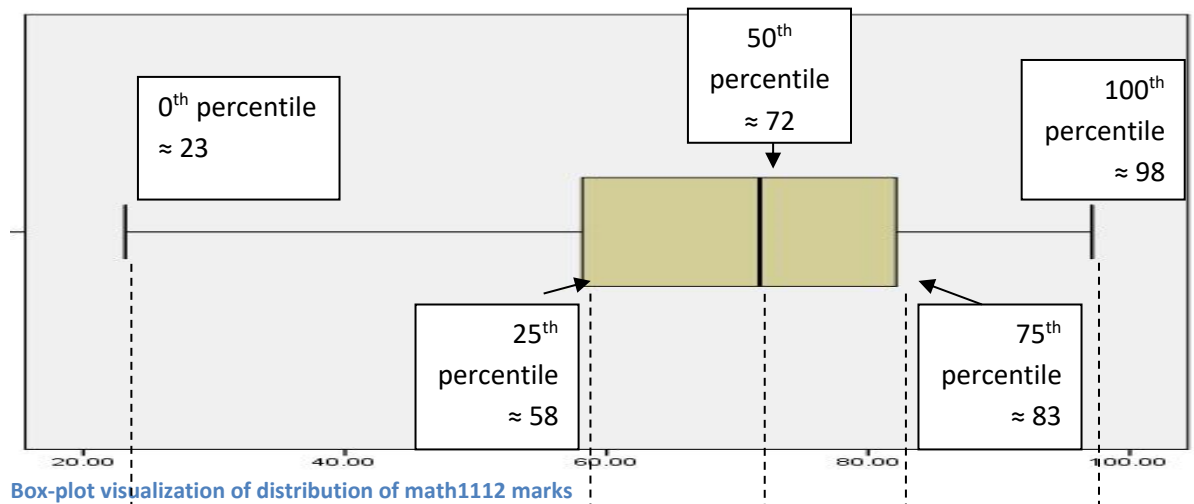
Spread: How spread out are the values? Measures of spread help tell us how widely the values of the distribution are dispersed. The *range* is very simple as it tells you the difference between the highest and lowest value, which is ok, but I prefer just looking at the minimum and maximum values to get a sense of the spread. The *standard deviation* is the average distance of each point from the mean. This statistic works well if the distribution is normal, but not for skewed, bimodal or j-shaped distributions as the mean does not clearly represent the centre in those distributions.

Outliers: Are there any extreme (particular) values that may be distorting our attempt to find patterns through numbers? If so, should they be removed? When looking at centre, spread and shape it is important to consider these outliers – also known as extreme values – they can easily be spotted in a detailed examination of visual output. Outliers need to be investigated individually during data preparation as they may be data entry or collection errors. If you go back to exercise 3 and take a look at the dot-plot for class K, you will see that there are two marks far to the left of the others. These two may be considered outliers, however, in this case they are likely valid and should not be removed from the data. If there were scores far to the right (e.g. at 120) then I would be concerned that these are not valid as the maximum possible mark should be 100.

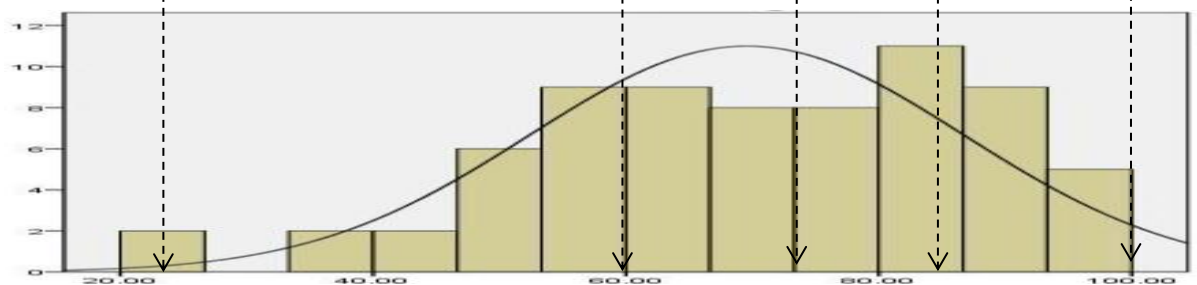
Introduction to boxplot – another perspective

The box-plot is another way of visualizing the distribution of a measurement variable. It can be presented in vertical or horizontal orientation. The one below is presented in a horizontal orientation. The centre ‘boxed’ area is surrounded by two capped lines called whiskers. The boxplot is sketched using a few key percentiles listed below and mared on the visuals below.

- the left-most (or bottom) of the whisker being the 0th percentile
- the left side (or bottom) of the box is the 25th percentile
- the centre bar is the 50th percentile (also known as the median)
- the right side (or top) of the box is the 75th percentile
- the right-most (or top) of the whisker is the 100th percentile



Compare the boxplot above to the histogram of the same data below - dotted lines are percentile markers.

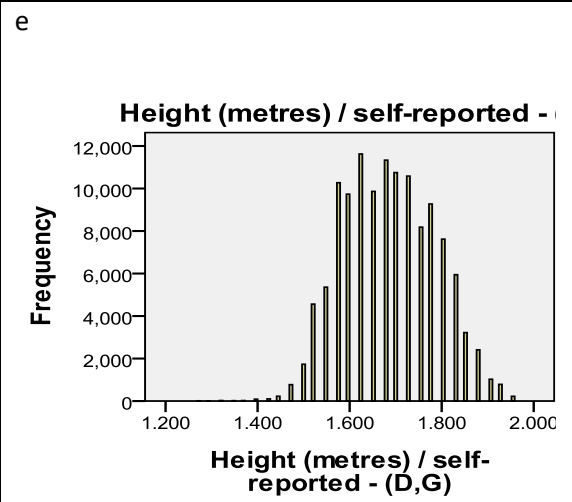
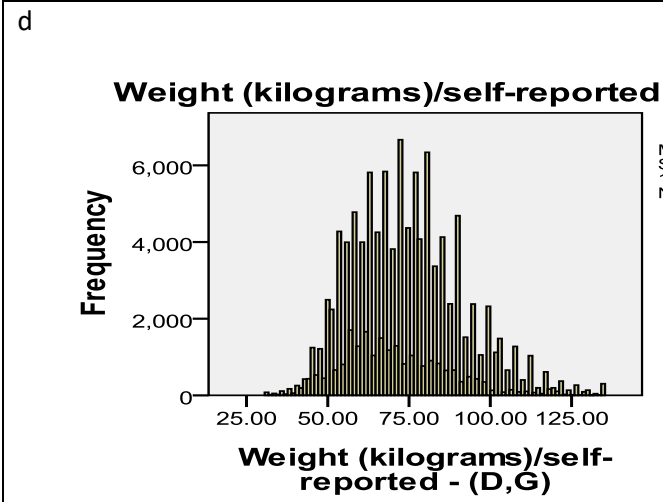


You will get a chance to practice with creating box plots from given histograms in exercise 4.

Exercise 2a-4: Visualizing boxplots from histograms: Sketch a box plot in the space that more or less matches percentiles with the corresponding histogram. Remember that you can draw box plots horizontally or vertically. Make sure to label the 0th, 25th, 50th, 75th, and 100th percentiles.

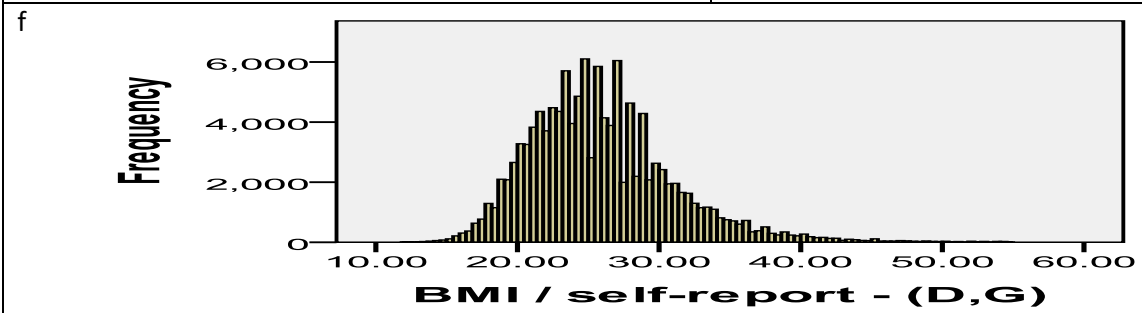
<p>a</p> <p>Histogram</p> <p>Frequency</p> <p>Total usual hrs. ...</p> <p>This histogram shows a distribution of 'Total usual hrs. ...' with a frequency axis from 0 to 20,000. The data is highly right-skewed, with a peak frequency of approximately 18,000 occurring between 30 and 40 hours. The x-axis ranges from 0 to 150.</p>	<p>b</p> <p>Number of hours spent sleeping per night</p> <p>Frequency</p> <p>Number of hours spent sleeping per night</p> <p>This histogram shows the frequency of hours spent sleeping per night. The frequency axis ranges from 0 to 1,250. The distribution is roughly bell-shaped and centered around 7 hours, with a peak frequency of about 1,300. The x-axis ranges from 0 to 12.</p>
<p>a</p>	<p>b</p>
<p>c</p> <p>Histogram</p> <p>Frequency</p> <p>Energy expenditure (kcal/kg/day)</p> <p>This histogram shows the frequency of energy expenditure in kcal/kg/day. The frequency axis ranges from 0 to 1,000. The distribution is highly right-skewed, with a peak frequency of about 800 occurring at very low energy expenditure values (around 0.5 kcal/kg/day). The x-axis ranges from 0 to 30.0.</p>	
<p>c</p>	

Exercise 2a-4 continued



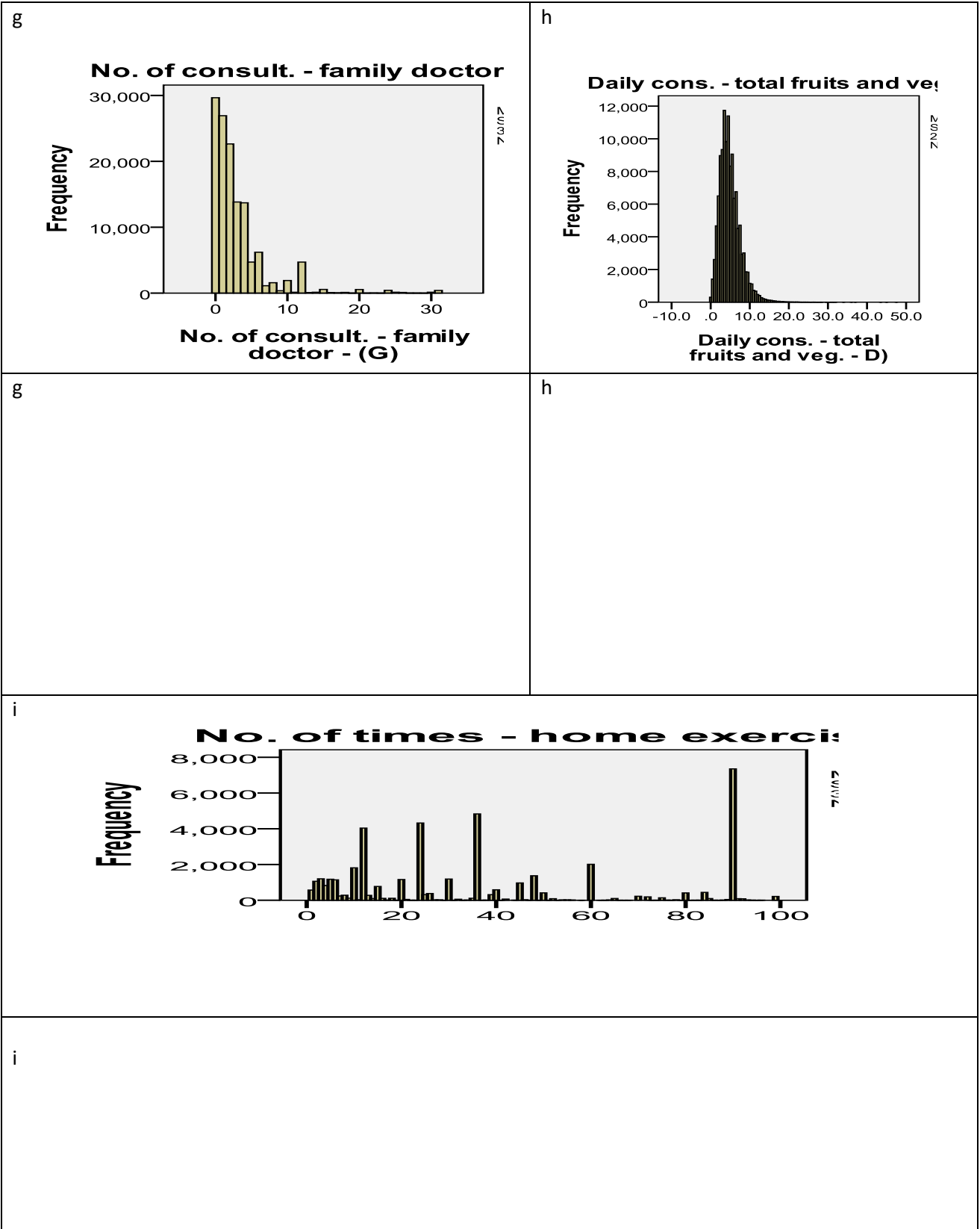
d

e



f

Exercise 2a-4 continued



Review of computations - one measurement variable

Distributions of Measurement variables can be quite complex. Mathematicians have developed many tools that capture the wide variety of their characteristics. We will focus on reviewing calculations of centre and spread (even though a software package will do all the work for us later in the course).

Example: Heights (in cm) of a family = {75, 89, 111, 144, 178, 179, 186}

N: cardinality of the set. In 'Heights' $N = 7$; Max: highest value; $\max \text{Height} = 186$; Min: lowest value; $\min \text{Height} = 75$;

x_i points to the i^{th} individual case. $x_5 = 178$

$\sum_{i=1}^n x_i$ - means the sum of all x_i - starting from $i = 1$ all the way up to n

$$\sum_{i=1}^3 x_i = 75 + 89 + 111 = 275 \text{ (using the Heights data set above)}$$

Centre is found by calculating the mean or median or finding the mode.

Median: a measure of centre found by choosing the centre number of an ordered distribution. *The median height of the heights distribution above is 144.*

Mean (μ or \bar{x}): is a measure of centre; $\frac{\sum_{i=1}^N x_i}{N}$ sum of all numbers in the set divided by N (the cardinality)

$$\mu_{\text{Height}} = (75 + 89 + \dots + 186)/7 = 962/7 = 137.43$$

Mode – appears the most often in the data set; In Heights data set there is no single mode as each height appears only once

Spread is measured by calculating the range, or standard deviation.

Range – a measure of spread = max – min {in the example range = $186 - 75 = 111$ }

Absolute deviation from mean: a nicely intuitive measure of dispersion. $\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$

Standard deviation provides a result that is close to the MAD, and relevant in specific types of distributions of single continuous variables, and rather challenging to conceptualize but it is preferred by mathematicians

$$(\text{std dev}) = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{\frac{\sum (75 - 137.43)^2 + (89 - 137.43)^2 + \dots}{7}} = \sqrt{2119.62} = 46.04$$

Shape – we will use only 2 numerical descriptions of shape skewness (stretching of bell curve) and kurtosis (pointiness of bell curve) without introducing the formulas for them.

Exercise 2a-5: predict shapes of distributions from the world around us.

Predict the shapes of the following distributions by either by sketching an outline of a histogram, and entering the following values: min, max, mean, median, mode, standard deviation, shape, or if the variable is categorical create an appropriate pie chart and mark in frequencies by %. All N = 5000.

Length of stay for acute appendicitis patients

Length of stay for fractured skull patients

Country of Birth of CHIMA members

Mean = 45; std dev = 5; shape = bell curve

Unit 2b: The Normal Distribution

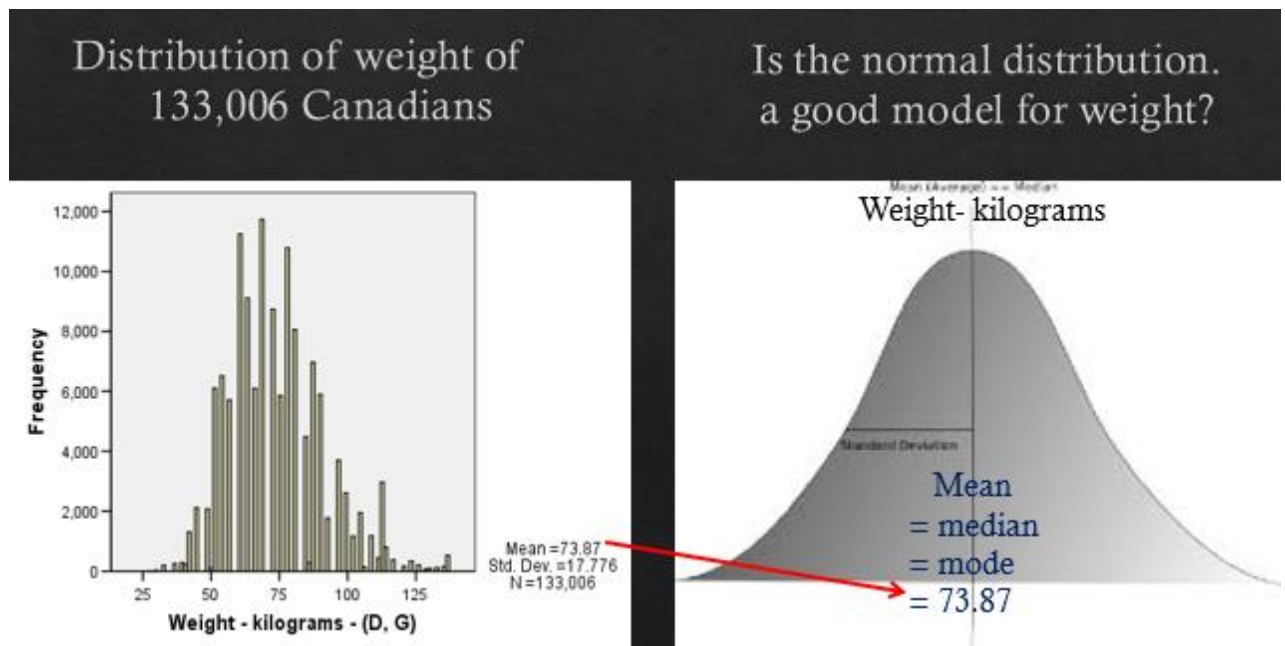
a mathematical model (abstraction)

and

foundation for inferential statistics

Over 100 years ago, humans noticed that data from many measurements of biological phenomena (thumb lengths, head circumferences etc.) have distributions that are similar to a bell in shape. There is a lot of data around a particular value at the centre (mean=median=mode) and the frequency drops off as one moves farther and farther above or below the centre. We now call the shape of this distribution normal.

Mathematicians have created a very useful mathematical model for these kinds of distributions called the normal (or Gaussian) distribution (see next page).



Numbers are abstract objects that can represent quantities of concrete objects, but also have a mathematical life of their own. In the same way, mathematical models, which are abstract and have a mathematical life of their own, can represent actual concrete distributions. They can be used to solve abstract mathematical problems, or as a tool to help make sense of (make inferences/predictions about) concrete situations that we don't know a lot about (e.g. how a virus will spread in a population?) Mathematical models provide a foundation for thinking beyond the data - i.e. for inferential statistics.

In this unit you will learn about the normal distribution, and practice working with it abstractly and as an abstraction of concrete situations.

Characteristics of a normal distribution:

There are a few keys to identifying whether the normal distribution fits the distribution of a variable you are investigating:

- the data type of the variable must be a measurement
- calculating values: Mean=median=mode (or close enough); maximum value is about 3 standard deviations above the mean and minimum is about 3 standard deviations below the mean.
- Visual representation is shaped more or less like a bell curve which means that it is symmetrical (has a line that divides it vertically in half along the mean/median/mode).

Mathematical aspects: The Normal distribution is a mathematical function that can be plotted on a 2 dimensional cartesian plane with x = individual bit of data, μ = *mean* and σ = *standard deviation*;

$$\text{General equation for Normal distribution: } f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note that the exact placement of the curve is dependent on the mean (μ) and the exact shape is dependant on the standard deviation (σ). Take a look at the 3 normal curves below and note how the change in μ shifts the graph along the x-axis and how the change in σ (from 1 to 2) flattens the shape of the bell curve.

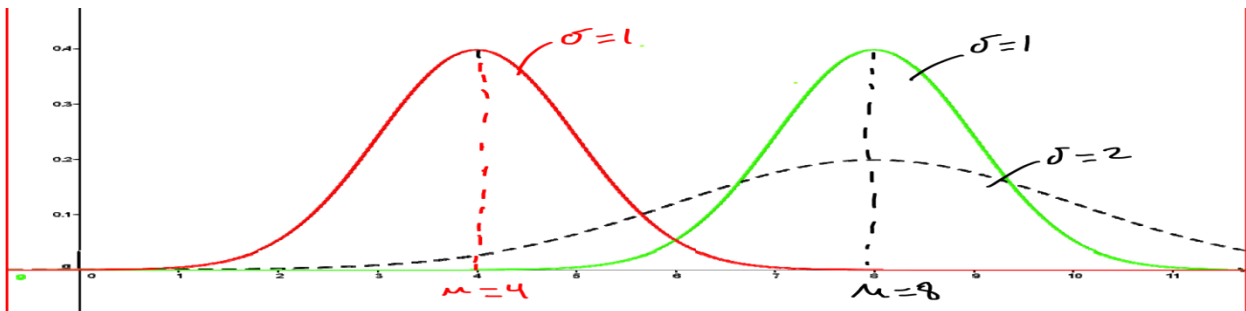
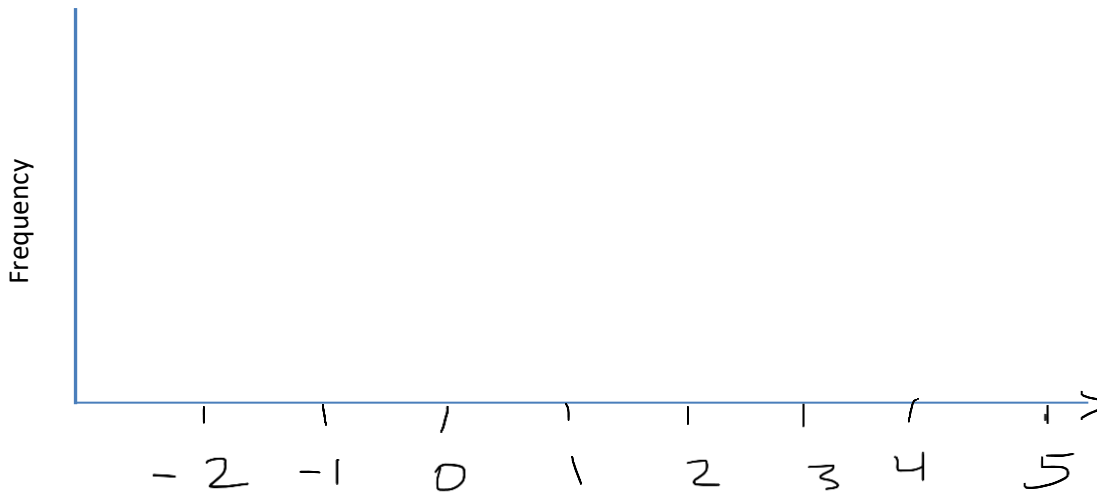


Figure 1: shifting and flattening a normal distribution

Exercise 2b-1: plot the following distribution of X on the graph provided below

- A. $\mu = 3$, and $\sigma=1$; B. $\mu = 2$, and $\sigma=0.5$; C. $\mu = 0$, and $\sigma=1$;



Exercise 2b-1b: plot a normal curve with the following characteristics in the space below.

- $\mu = 37$, and $\sigma =4$, and an outlier at 3.5σ above the mean.



Standardized normal distribution: (i.e. with $\mu = 0$ and $\sigma = 1$) is mathematically beautiful, and has many useful characteristics.

Formula for Standardized Normal curve:
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The z-score: A key to standardization is the z-score formula which turns a raw score from any normal distribution into a raw score from the standardize distribution with $\mu=0$ and $\sigma=1$ by use of the z-score formula. If we know μ and σ for a distribution, we can translate each raw score 'x' into a z-score and use the z-score to find its percentile in a table of values called the z-table (see appendix)

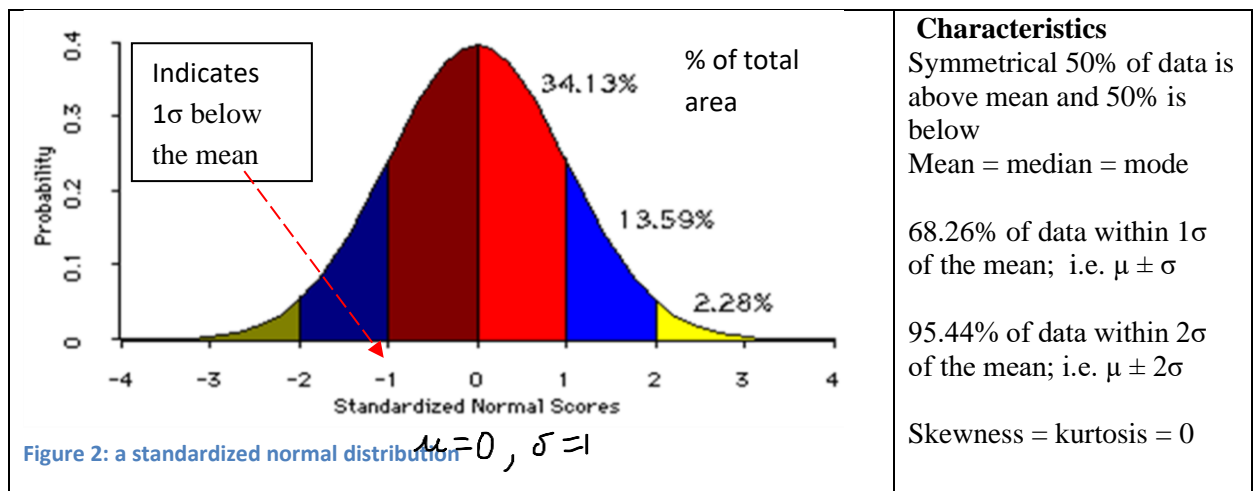
The z-score equation:
$$z = \frac{x-\mu}{\sigma}$$

The z-score tells us how many standard deviations a particular raw score ('x') is above or below the mean in a standardized normal distribution.

Area under the curve: a key concept used to find percentiles of raw score 'x' in the distribution. Because the normal distribution is used as a valid representation of the distribution of concrete information (e.g. head circumference of a population) the area under the curve (as a portion of the total) represents the portion of the actual data in that population.

For example, all normal distributions are symmetrical in nature, which means that 50% of the data lies below the mean and 50% above the mean (or put another way mean = median). If head circumferences were normally distributed with mean = 48cm then I would automatically know that 50% of the population has a head circumference > 48cm.

The graph of the standardized normal distribution below shows some other markers of area by z-score.

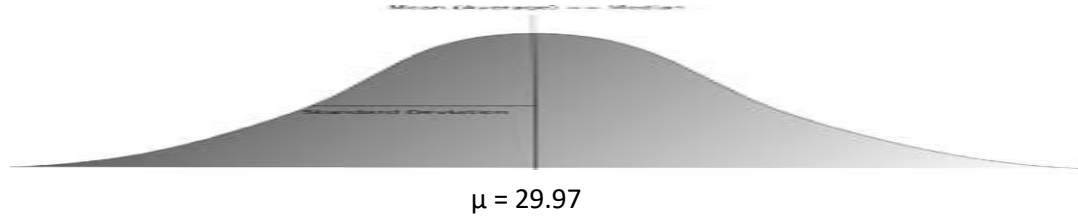


Exercise 2b-2: from z-score to areas under normal curve

A normal distribution with $\mu=29.96$ and $\sigma=8.977$.

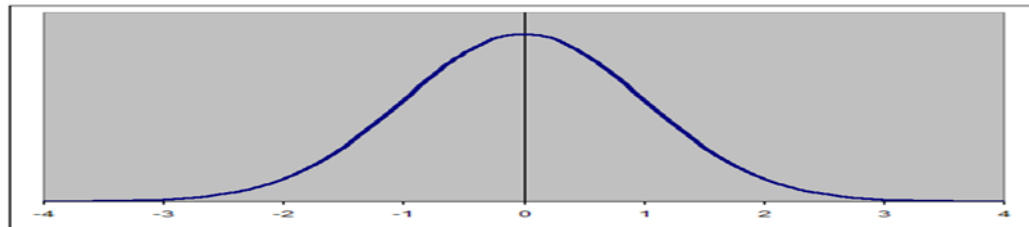
Q1. Plot the following marks onto the normal curve model of the above distribution.

- a. $X = 44$ b. $x = 17$ c. $x = 2$ d. $x = 174$



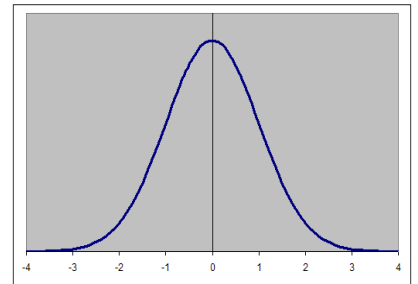
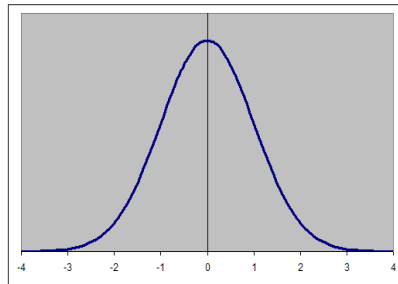
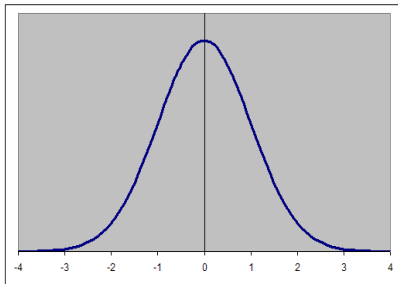
Q2. Standardization: calculate z-scores for each and plot them on the graph below.

- a. $X = 44$; $z = \underline{\hspace{1cm}}$ b. $x = 17$; $z = \underline{\hspace{1cm}}$ c. $x = 2$; $z = \underline{\hspace{1cm}}$ d. $x = 174$; $z = \underline{\hspace{1cm}}$



Q3. Shade the following areas using the z-score information from above

- a. $X < 44$ b. $x > 17$ c. $2 < x < 44$



Q4. Estimate the portion of the total area you shaded for each as a %.

- a. b. c.

Q4. Use the z-scores from Q2 and the z-table (see next page for details) to find each of the following probabilities (some calculation may be needed)

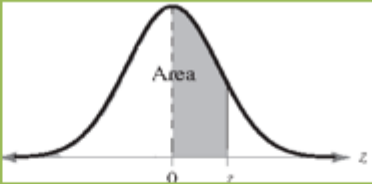
- a. $P(x < 44) =$ b. $P(x > 17) =$ c. $P(2 < x < 44) =$

The z-table: someone did the calculations for us...

The z-score tells us how many standard deviations a particular raw score ('x') is above or below the mean in a standardized normal distribution.

The z-table is a document that tells us the portion of the area under the standard normal distribution (see full table in appendix) for every z-score.

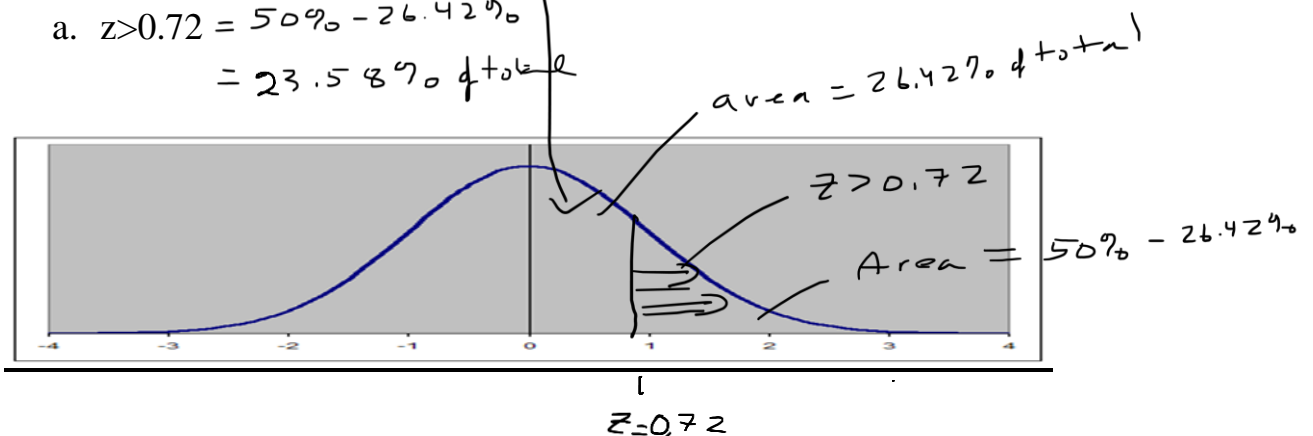
Z table Area between mean and z-score.
For example $z = 0.72$ corresponds to area = 0.2642 (or 26.42% of total area).



Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015

2 examples: Find the area under the curve for the following:

a. $z > 0.72 = 50\% - 26.42\%$
 $= 23.58\%$ of total



There is a full z-table in the appendix which you can use to help with exercise 12

Normal Distribution as an abstraction of concrete situation

Once it is established that the characteristic of interest (e.g. timestudying) in a population is normally distributed and we know μ and σ we can use those values to 'standardize the distribution' and make predictions (make inferences) about the population.

Example: Let's take our data from 'time studying' exercise as a starting point: Time studying (measured in hours per week) in semester 1 of HIM was normally distributed with $\mu = 17.2$ and $\sigma = 8.1$. Imagine that we know that Arika studied 25 hours per week, but we don't know where she stood in relation to the HIM student population. We frame this as: What % of students had a lower time studying than Arika? and use characteristics of the normal distribution to estimate an answer which we could get from the data set.

a. Calculate the z-score for Arika's raw score: $x = 25$.

$$z = \frac{x - \mu}{\sigma}$$

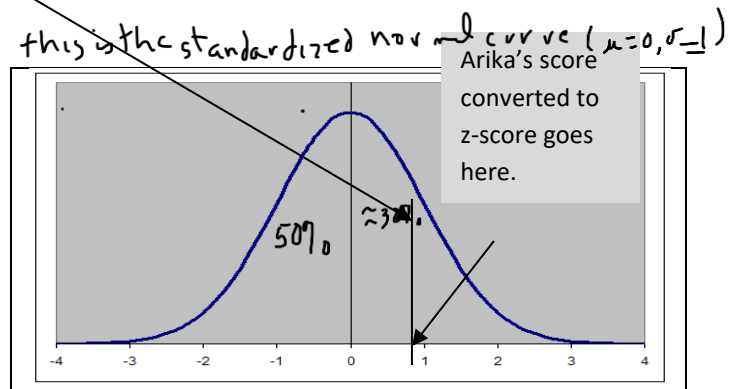
$$z = \frac{25 - 17.2}{8.1} = \frac{7.8}{8.1} = 0.963. \text{ (+0.963 means that Arika's time studying was 0.963 standard deviations above the mean)}$$

b. Use the z-score and normal curve to estimate her position. What % of students studied less than Arika? (i.e. less than 25 hours per week) This question can be stated as a probability question. What is $P(x < 25)$?

Use the z-score from a. ($z = 0.963$) and the normal distribution to find out about Arika's time spent studying in relation to her peers.

Given that time spent studying is normally distributed we can plot Arika's timestudying as a point on the standardized normal distribution and see where she sits compared to her classmates. We know she is above average (since $25 > 17.2$), but how much above is she?

Estimate answer to $P(x < 25)$ by estimating area below $z = 0.963$ using visual only:



We want to know what % of students studied less than Arika. We know that the area between the mean and $\sigma = 1$ is about 34%. Use that to estimate that the area below the curve as about $50\% + 30\% = 80\%$. Thus by rough estimate $P(x < 25) \approx 0.8$.

c. Calculate exact answer by using table of values:

Step 1: look up $z = 0.963$ in z-table to get 0.3315 (33.15%). This means that the area between the mean and $z = 0.963$ is equivalent to 0.3315 of the area under the normal curve.

Step 2. 33.15% is the % of students studied less than 25 hours and more than 17.2 (the mean). We need to add on all the students who studied less than 17.2 hours too (that is 50%)

The predicted proportion of students who studied less than 25 hours: $P(x < 25) = 50\% + 33.15\% = 83.15\%$

Exercise 2b-3: relating individual to the group (finding percentiles) when you have the data.

Below is an ordered array of the length of stay data for the Cardiac Wing of Dunchurch Hospital.

2 4 5 7 9 9 14 14 14 16 16 17 18
 19 19 19 20 20 22 22 22 22 22 22 25 25
 26 26 26 28 29 29 30 31 32 33 35 37 37
 38 39 41

LengthofStay	Statistics	
N	Valid	42
	Missing	0
Mean		22.4048
Median		22.0000
Mode		22.00
Std. Deviation		9.87744
Skewness		-.109
Std. Error of Skewness		.365
Kurtosis		-.484
Std. Error of Kurtosis		.717

- open the LOS data set in www.stataras.com
- verify the numerical output above and take a look at the distribution visually.
- Calculate percentiles for particular values using SPSS – see SPSS instructions booklet
 - e.g. Find percentile for LOS =37 using LOS data set and SPSS software
 - Step 1: using SPSS generate a frequency table for 'LOS data'.
 - Step 2: look at the cumulative frequency for x = 37. You should get 92.9 percentile, which means that 92.9% of the Length of Stays are below 37.

Based on the LOS data set provided what percentile corresponds to an LOS of

a) 29 b) 7 c) 38 d) 22.4048
- Based on the LOS data set provided what % of Cardiac patients had an LOS
 - a) > 29 b) < 7 c) < 38 d) > 22.4048

Exercise 2b-4: use normal model to relate individual to group (finding percentiles). Assume that you don't have the data.

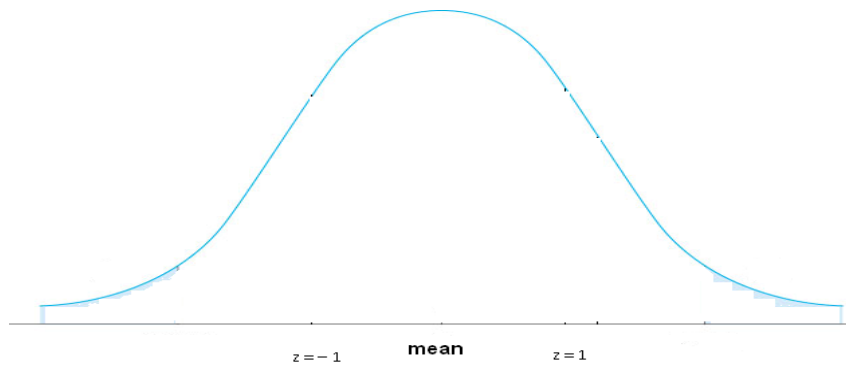
Now imagine that the only LOS information you had about the Cardiac Wing of Dunchurch hospital was that the mean was $\mu = 22.4048$ and $\sigma = 9.8774$.

Complete the following:

- 1 a) Find the z-score for $x = 29$? _____ How many σ s is 29 from the mean? _____
- b) Find the z-score for $x = 7$? _____ How many σ 's is 7 from the mean? _____
- c) Find the z-score for $x = 38$? _____ How many σ 's is 38 from the mean? _____
- d) Find the z-score for $x = 22.4048$? _____ How many σ 's is 22.4048 from the mean? _____

2 Place the x-values on the horizontal axis in the normal curve below.

- a) 29 b) 7 c) 38 d) 22.4048



3. Use the z-score values from #1 and the z-table to calculate the percentage of values between the mean and each of the following Length of Stays:

- a) 29 b) 7 c) 38 d) 22.4048

4 Based on the z-table predict the percentage of individual you expect to lie below 29 a) b) 7 c) 38 d) 22.4048

5. If you were to pick a value (x) randomly from the LOS distribution with $\mu=22.4048$ and $\sigma=9.87744$ calculate

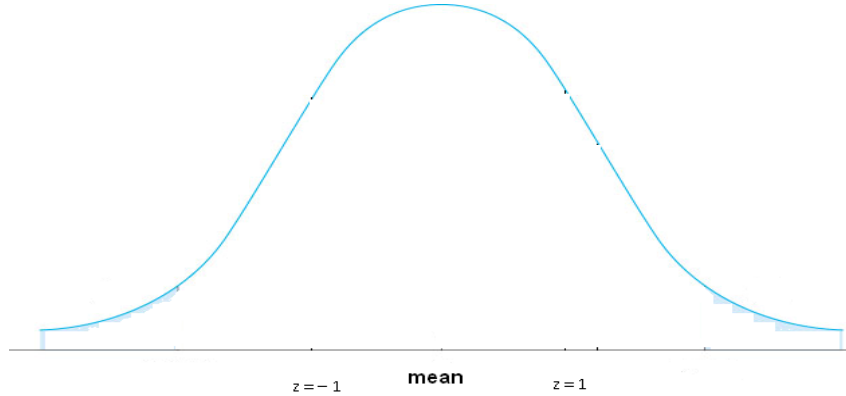
- a) $P(x < 29)$
- b) $P(x < 7)$
- c) $P(7 < x < 38)$
- d) $P(x > 22.4048)$

Exercise 2b-5: shift from actual data to normal model. Open the marks1112 data set from www.stataras.com. Use only the final mark variable to answer the following.

- 1
- | | |
|--|---|
| a) Find the z-score for $x = 77$? _____ | How many σ 's is 77 from the mean? _____ |
| b) Find the z-score for $x = 20$? _____ | How many σ 's is 20 from the mean? _____ |
| c) Find the z-score for $x = 50$? _____ | How many σ 's is 50 from the mean? _____ |
| d) Find the z-score for $x = 95$? _____ | How many σ 's is 95 from the mean? _____ |

2 Place each X-value on the horizontal axis above/below the mean.

- a) 77 b) 20 c) 50 d) 95



3. Based on the z-table and using the z-score values from #1, what percentage of values would you expect to lie between the mean and

- a) 77 b) 20 c) 50 d) 95

4 Use the z-table to find what percentage of values would be expected to lie below

- a) 77 b) 20 c) 50 d) 95

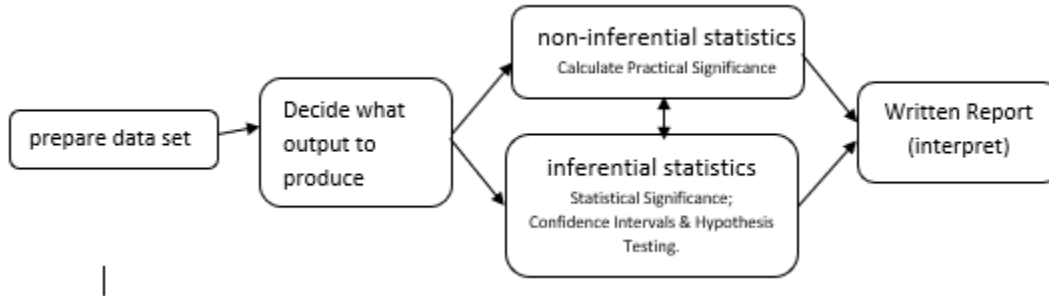
5 If you were to pick a value randomly from the actual distribution (without using normal) find

- a) $P(x < 77)$
 b) $P(x < 20)$
 c) $P(50 < x < 95)$
 d) $P(x > 95)$

2c introduction to SPSS – a computational tool

In this unit you will practice getting SPSS to produce output for scenarios with one variable, then to use that information to answer the research question posed.

Recall the following elements of data analysis (calculate and interpret actions):



The tasks you will need to figure out how to do:

1. Decide whether the scenario requires computations for measurement or categorical variable by reading the research question and looking at the data set.
2. Figure out (with support of the instructions booklet) how to get SPSS to produce what you need, and how to enter each type of data.
3. Compute: get spss to produce visual and numerical output
4. Figure out how to transfer output from SPSS output file to OneNote
5. Interpret the output: use it to support appropriate answer to the research question posed including the degree of confidence you have in your answer.

Introduction to SPSS (calculate and interpret: one variable) pg 1

Calculate and interpret with one variable

numerically

33% of students ...
The average mark in the class was ...

visually

1

Elements of Data Analysis

Learn this by heart!!!

2

Compute and interpret with one measurement variable:
exercise ~~20~~ 20-1

Marks for test1 are available for a series of math 1112 HIM students.
Decide which output is needed and get spss to do produce it.

RQ: How did the students on the test?

3

Numerical and visual output for marks 1112 (with notes in red)

test1	
1	84.00
2	88.00
3	95.00
4	92.00
5	78.00
6	84.00
7	91.00
8	89.00
9	41.00
10	93.00
11	71.00
12	78.00
13	72.00
14	95.00
15	72.00
16	88.00
17	80.00
18	90.00

Statistics	
N	Valid 402
	Missing 0
Mean	80.1913
Median	84.3750
Mode	100.00
Std. Deviation	18.84906
Minimum	20.00
Maximum	100.00

4

Interpret visual and numerical output.

RQ: How did the students do on the test?

Statistics	
N	Valid 402
	Missing 0
Mean	80.1913
Median	84.3750
Mode	100.00
Std. Deviation	18.84906
Minimum	20.00
Maximum	100.00

Note the left skew in histogram (visual) is echoed by the mean being lower than the median

5

Compute and Interpret one Categorical variable

Exercise ~~20~~ 20-2

A box of smarties was opened up and the colour of each smartie was recorded.

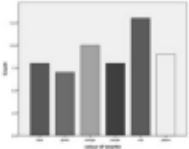

Research question: Are there more red smarties than other colours in the opened boxes?

6

Introduction to SPSS (calculate and interpret: one variable) pg 2

Exercise 8b: Use SPSS instructions to reproduce the following output for the *smarties* data set

Colour of smarties				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
blue	8	14.5	14.5	14.5
green	7	12.7	12.7	27.3
orange	10	18.2	18.2	45.5
purple	8	14.5	14.5	60.0
red	13	23.6	23.6	83.6
yellow	6	10.8	10.8	100.0
Total	55	100.0	100.0	

interpretation: there seem to be more red smarties, but it would be hard to call the difference dramatic.

7

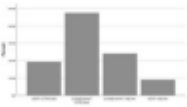
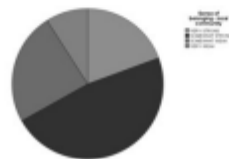
20-3
Exercise ~~8b~~

Statistics-Canada asked a random sample of 2000 Canadians the following question:
Research question: Do a majority of Canadians have a strong sense of belonging to their local community?

8

Interpret categorical data output
numerical & graphical output is used to help answer the research question.

Sense of belonging - local community				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
VERY STRONG	376	18.8	18.8	18.8
SOMEWHAT STRONG	638	31.9	31.9	50.7
SOMEWHAT WEAK	487	24.3	24.3	75.0
VERY WEAK	176	8.8	8.8	83.8
Total	1687	83.8	100.0	
Missing	DO NOT KNOW	24	1.2	
	REFUSAL	2	.1	
	NOT STATED	37	1.9	
Total	63	3.2		
Total	2000	100.0		

interpretation:

9

Relating Frequency to percent and probability

Sense of belonging - local community				
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
VERY STRONG	376	18.8	18.8	18.8
SOMEWHAT STRONG	638	31.9	31.9	50.7
SOMEWHAT WEAK	487	24.3	24.3	75.0
VERY WEAK	176	8.8	8.8	83.8
Total	1687	83.8	100.0	
Missing	DO NOT KNOW	24	1.2	
	REFUSAL	2	.1	
	NOT STATED	37	1.9	
Total	63	3.2		
Total	2000	100.0		

Is the probability that a Canadian respondent thinks they should do something about their health 18.8% or 19.4%?

10

Exercise 2c-1 SPSS output for one measurement variable and interpretation

Data set: 'marks 1112' from www.stataras.com; variable name: test1 (test 1 mark)

Research Question: How did students do (generally) on the test?

1. Numerical output: Write out the following:

What type of variable is test1 mark? _____

μ = median = mode = σ = Range =

min = max = N = skewness = kurtosis =

12th percentile = 25th percentile = 50th percentile = 75th percentile = 95th percentile =

2. Visual output:

- a. produce dot plot, histogram and box plot using SPSS and sketch them below. Mark the 25th 50th and 75th percentiles on each.

dot plot

histogram (describe shape)

box plot

--	--	--

Discussion:

- b. Are there any outliers? If so, does it seem that the outlier(s) may be errors in data entry?

3. Answer the research question by describing the distribution of marks using elements from numerical and visual output.

Exercise 2c-3: Use SPSS to produce output for one categorical variable (likert scale).

Data set *cchs condensed* data set, variable: sense of belonging

Research Question: Do a majority of Canadians have a strong sense of belonging to their local community?

1. What type of variable is sense of belonging? _____

Numerical output. Recopy information from SPSS output:

List of categories	Frequency	Valid percent	Cumulative percent
Total			

2. Visual Output: Produce a bar chart and pie chart of the data and sketch them below

Bar Chart	Pie Chart
-----------	-----------

3. Is it accurate to say that a majority respondents have a strong sense of belonging to their local community?

Explain your conclusion referring to numerical and visual output from above.

4. Is cumulative frequency appropriate and/or useful here? What % of respondents had a sense of belonging that was more than somewhat weak.

Exercise 2c-4: boot camp (compute and interpret) scenarios with one variable

Follow the 5 tasks from page 39 to complete the following:

Q1. Research Question: Are most Canadians who responded to the Canadian Community Health Survey born in Canada?

Data set: *'cchs condensed 24 variables'*

Q2. Research question: How satisfied are Canadians who responded to the Canadian Community Health Survey with the way their body looks?

Data set: *'cchs condensed 24 variables'*

Q3. Research Question: How much did Canadians who responded to the Canadian Community Health Survey work per week?

Data set: *'cchs condensed 24 variables'*

Q4 Research question: How much fruits and vegetables do Canadians who responded to the Canadian Community Health Survey say they eat?

Data set: *'cchs condensed 24 variables'*

Unit 3

Forwards and backwards thinking tools for variable relations.

Up to now you have concentrated your efforts on thinking about one variable at a time. This is important foundational work, but for the most part does not help us answer more interesting questions in the health sciences. We want to know answers to questions like: Does wearing a mask protect us from testing positive with covid-19? (two variables: mask wearing *yes/no* and testing positive *yes/no*)

Recall: your goal is to master thinking both ways through the ‘define, abstract, compute and interpret’ process (*forwards*... and *backwards*).

Forwards means that you will start with an idea for research then conduct the following steps

- define a research question (RQ),
- abstract by quantifying the question into components (unit of analysis, variables) then collect data.
- compute a variety of numerical and visual descriptions that allow you to study the relations and measure the strength of association between the two variables.
- interpret the numbers: Provide an answer to the research question (RQ) by interpreting what the numbers and visualizations tell us about the relation between the two variables.

Backwards means that you will be involved in reverse engineering (figuring out what the researchers who collected data were thinking as they moved through the define, abstract, compute and interpret process)

Your task will involve choosing the right computations, making them, and then making valid interpretations.

In this unit you will start from the middle: the research question will be set for you and the data set will be ready. You will need to be a reverse engineer to figure out what the researcher was doing (working backwards) then based on what you find get SPSS or other tools to compute numerical and visual output, and interpret what you found to answer the research question (RQ).

Data analysis involves both reverse engineering the RQ (backwards through abstraction and definition) and working forwards in computing (including choosing the correct computations) and interpreting results.

3a Computational tools when two variables are needed

In unit 1 you were introduced to data types as a foundation for this statistics course. The extension to this foundation will add 3 concepts to help investigate the relation between 2 variables.

Concept 1: data type pairings,

Concept 2: independent (exposure) vs dependent (outcome) variables,

Concept 3: the relation between variables in the abstract can represent relations between characteristics of individuals in the concrete world. The strength and direction of these relations can be measured. Measures of strength of association/relation between variables will be called practical significance.

Remember this diagram?

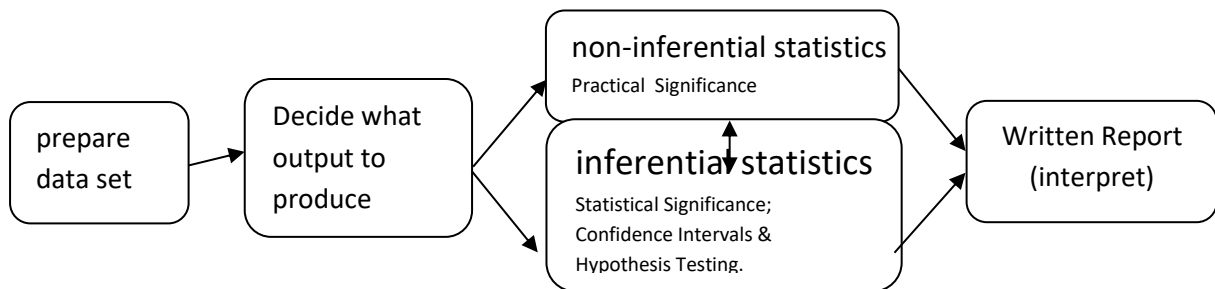


Figure 2: visualization of elements of data analysis

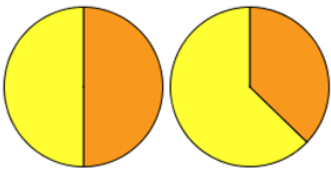
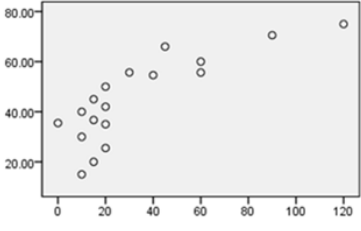
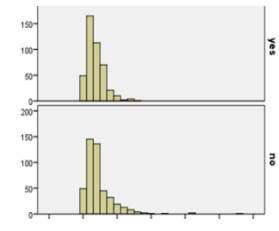
If the concrete situation has been competently quantified (into relevant variables) and crafted into a good research question and the data collected competently, then your task is to do reverse engineer the abstraction, decide on and conduct the calculations (or to get SPSS to do the calculations) and do the interpretation to justify your answer the research question posed.

Concept 1: data type pairings.

Given that there are two types of variables [categorical (c) and measurement (m)], we know there can only be three ways they can be paired: cc, mm, cm.

Knowing the three options for data pairing will help you design research (abstraction) or to reverse engineer the abstract structure (the data set) created by researchers. The three options are named below, with a short name for type of analysis in brackets followed by an example of research question (RQ) and visualization:

Three options for data type pairings with two variables.

both categorical (cc) (compare rates)	both measurement (mm) (correlation)	1 categorical & 1 measurement (cm) (compare means or medians)
RQ: Are people who smoke a pack a day more likely to have emphysema?	RQ: Is time spent studying correlated with course mark?	RQ: Do people who smoke have a higher number of siblings than those who don't smoke?
<p style="text-align: center;">yes no</p> 		

Different **types of pairings** of variables necessitate different analysis approaches and differentiating them and using them will comprise the bulk of the work in this section of the course.

Before investigating computations that help investigate the relations that happen between the different pairings there is one more concept to spend a bit of time with: independent and dependent variables.

Concept 2: Independent and dependent variables:

In each type of pairing there will typically be a variable that is thought of as leading to, predicting, or having an effect on the other. (e.g., vitamin C use reduced duration of a cold). The variable that is thought to have an effect (vitamin C use) is called the **independent** (*exposure*) variable. The variable that may be affected (duration of a cold) is called the **dependent** (*outcome*) variable.

If the independent variable 'x' (vitamin c dose) is strongly associated with the dependent variable 'y' (duration of cold in days) then knowing the value of or the category 'x' we can **predict** the value of 'y', and perhaps find that x has an effect on y. (e.g. $y = -0.5x + 5$ - For every dose of vitamin c your duration goes down by half a day.)

The goal of data analysis is to describe the relation and measure the strength of association between the two variables (also called measure of practical significance, or clinical significance) Each type of pairing requires a unique calculations and visualizations.

Part of the reverse engineering process involves reading the research question (RQ) and figuring out what the variables are, which of them is the independent and dependent variables, and what their types are. With those bits of information calculation/computation are quite simple as they will be done by machines.

Example: fill in the blanks (answers below)

Research Question: Are those individuals who are singers more likely to enjoy musicals?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Short form for analysis that would be appropriate: _____

Independent variable: singer (yes vs no)	categorical
Dependent variable: enjoy musicals (yes vs no)	categorical
Short form for analysis: compare rates (since both variables are categorical)	

Exercise 3a-1: identify independent and dependent variables

fill in the blanks for the following scenario (answers are below)

1. Research Question: Is the number of hours students study per week associated with higher marks?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Short form for analysis that would be appropriate: _____

2. Research Question: Do smokers have a higher likelihood of having diabetes?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Short form for analysis that would be appropriate: _____

3. Research Question: Do smokers have a lower lung capacity (in cm³)?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Short form for analysis that would be appropriate: _____

4. Research Question: Does 15 minutes of exercise per day reduce one's stress score?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Short form for analysis that would be appropriate: _____

Practice with recognizing data types of variables is available on www.statcat.ca.

Concept 3: measures of strength of association with two variables

The goal of every data analyst is to help the researcher answer a research question i.e. to show the degree to which a particular treatment cures a disease, or that a teaching approach helps students learn math better, or that a particular chemical in the water is harmful (or helpful). This is done by presenting numbers that measure the extent to which the independent (exposure) variable ‘x’ contributes to, or is related to changes in the dependent (outcome) variable ‘y’.

Each type of pairing (cc, mm, cm) will require you to use that type of pairing as a lens in choosing the right action to take, (the correct visual and numerical output to produce and correct measure of strength of association to calculate). Only then can you answer the research question using interpretations based on the information you have.

Data types pairing	Short name for action	Measure of strength of association (calculation of practical significance)
Both categorical (cc)	Compare rates	Ratio of rates = likelihood ratio = RR difference between rates = RD Calculate by hand
Both measurement (mm)	Correlation	Pearson’s r and r^2 ; equation of line of best fit Calculated by SPSS
one categorical & one measurement (cm)	Compare means	raw difference between means % difference between means & Cohen’s d Calculate by hand

Interpretation challenges: The evidence can sometimes provide a clear direction (e.g. 100% of smokers get lung cancer, and 0% of non-smokers get lung cancer.), however, in most cases the associations (relations) between the independent variable and dependent variable are not as clear. Do not expect to have obvious answers to research questions, do expect to have to make tough decisions, or state that the data is not providing evidence for a definitive answer to the RQ when results are unclear.

Computation: two categorical variables – comparing rates:

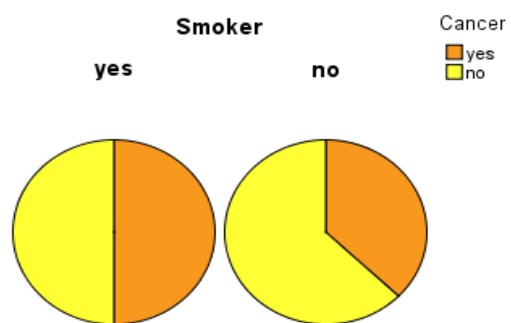
Numerically that is done by the creation of a crosstabulation (contingency table). This is much easier to do when both variables have only two categories. Comparing rates is difficult to do when there are more than 2 categories in either of the variables. Most research questions you will confront in this course are set up to investigate the comparison of two rates/ratios.

RQ: Are smokers more likely to have cancer? 63 individuals were asked whether they smoked and whether they had cancer of any kind.

Name of independent categorical variable: _____ #of categories _____

Name of dependent categorical variable: _____ #of categories _____

Smoker * Cancer Crosstabulation		Cancer		Total
		yes	no	
Smoker yes	Count	10	10	20
	%	50%	50%	100%
no	Count	16	27	43
	%	37.2%	62.8%	100%
Total	Count	26	37	63



Numerical output: use the contingency table. Exposure (independent) variable is set up in rows and outcome (dependent) in columns.

Calculate rates of outcome (in this case ‘cancer’) in those exposed (smoker = yes) vs those in unexposed (smoker = no)

Example: The grey bands show calculations of rate of cancer in smokers ($10/20 = 50\%$) and rate of cancer in non-smokers ($16/43 = 37.2\%$) We can see that the rate of cancer in smokers is higher.

Visual output the paneled pie chart shows that the portion of smokers with cancer is larger than the portion of non-smokers with cancer but the details and meaning are hard to interpret.

Exercise 3a-2 relate visual to numerical with two categorical variables

Relating pie charts to likelihood ratio; Create a comparison pie chart to match the contingency table to the right, or fill in the contingency table to match the pie chart to the left. Remember that accuracy is more important than precision. This exercise is meant to build connections between numerical and visual representations of 2 categorical variable scenarios.

Comparison pie chart	Contingency table																								
<p>Immigrant status - (F)</p> <p>YES NO</p> <p>Dwelling - owned by a member of hslid</p> <p>■ YES ■ NO</p>	<table border="1"> <thead> <tr> <th colspan="2"></th> <th colspan="2">Dwelling owned by family</th> <th>Total</th> </tr> <tr> <th colspan="2"></th> <th>yes</th> <th>no</th> <th></th> </tr> </thead> <tbody> <tr> <th rowspan="2">Immigrant</th> <th>yes</th> <td></td> <td></td> <td>100</td> </tr> <tr> <th>no</th> <td></td> <td></td> <td>100</td> </tr> <tr> <th colspan="2">Total</th> <td></td> <td></td> <td>200</td> </tr> </tbody> </table>			Dwelling owned by family		Total			yes	no		Immigrant	yes			100	no			100	Total				200
		Dwelling owned by family		Total																					
		yes	no																						
Immigrant	yes			100																					
	no			100																					
Total				200																					
<p>YES NO</p> <p>■ YES ■ NO</p>	<table border="1"> <thead> <tr> <th colspan="2"></th> <th colspan="2">Played video</th> <th>Total</th> </tr> <tr> <th colspan="2"></th> <th>yes</th> <th>no</th> <th></th> </tr> </thead> <tbody> <tr> <th rowspan="2">Played sports</th> <th>yes</th> <td>20</td> <td>80</td> <td>200</td> </tr> <tr> <th>no</th> <td>220</td> <td>80</td> <td>400</td> </tr> <tr> <th colspan="2">Total</th> <td>240</td> <td>60</td> <td>600</td> </tr> </tbody> </table>			Played video		Total			yes	no		Played sports	yes	20	80	200	no	220	80	400	Total		240	60	600
		Played video		Total																					
		yes	no																						
Played sports	yes	20	80	200																					
	no	220	80	400																					
Total		240	60	600																					
<p>Smoker</p> <p>yes no</p> <p>Cancer</p> <p>■ yes ■ no</p>	<table border="1"> <thead> <tr> <th colspan="2"></th> <th colspan="2">Cancer</th> <th>Total</th> </tr> <tr> <th colspan="2"></th> <th>yes</th> <th>no</th> <th></th> </tr> </thead> <tbody> <tr> <th rowspan="2">Smoker</th> <th>yes</th> <td></td> <td></td> <td>500</td> </tr> <tr> <th>no</th> <td></td> <td></td> <td>500</td> </tr> <tr> <th colspan="2">Total</th> <td></td> <td></td> <td>1000</td> </tr> </tbody> </table>			Cancer		Total			yes	no		Smoker	yes			500	no			500	Total				1000
		Cancer		Total																					
		yes	no																						
Smoker	yes			500																					
	no			500																					
Total				1000																					
<p>YES NO</p> <p>■ YES ■ NO</p>	<table border="1"> <thead> <tr> <th colspan="2"></th> <th colspan="2">passed</th> <th>Total</th> </tr> <tr> <th colspan="2"></th> <th>yes</th> <th>no</th> <th></th> </tr> </thead> <tbody> <tr> <th rowspan="2">Studied?</th> <th>yes</th> <td>64</td> <td>36</td> <td>100</td> </tr> <tr> <th>no</th> <td>120</td> <td>180</td> <td>300</td> </tr> <tr> <th colspan="2">Total</th> <td>184</td> <td>216</td> <td>400</td> </tr> </tbody> </table>			passed		Total			yes	no		Studied?	yes	64	36	100	no	120	180	300	Total		184	216	400
		passed		Total																					
		yes	no																						
Studied?	yes	64	36	100																					
	no	120	180	300																					
Total		184	216	400																					

Strength of association: two categorical variables

Calculations for two categorical variables: In order to standardize comparisons of rates when analyzing 2 categorical variables 2 approaches have been used (both based on the 2 by 2 contingency table).

The independent variable is called ‘exposure’, and the dependent variable is called ‘outcome’.

		Outcome		Totals
		Yes	No	
Exposure (or treatment)	Yes	a	b	a+b
	No	c	d	c+d
	Totals	a+c	b+d	N

Ratio of Rates (a.k.a. the likelihood ratio or RR – relative risk)	Difference between rates (a.k.a. RD – risk difference)
$\frac{P(\text{outcome} = \text{yes}, \text{given exposure} = \text{yes})}{P(\text{outcome} = \text{yes}, \text{given exposure} = \text{no})} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	$\begin{aligned} &P(\text{outcome} = \text{yes}, \text{given exposure} = \text{yes}) \\ &P(\text{outcome} = \text{yes}, \text{given exposure} = \text{no}) \\ &= \frac{a}{a+b} - \frac{c}{c+d} \end{aligned}$

Example: Research question: Are those exposed to smoking more likely to get cancer? Use the results in the contingency table to calculate RR and RD and answer the question.

		Outcome		Totals	Rates of illness:
		Yes	No		
Exposure (smoking)	Yes	37	22	59	37/59 = 62.71% of those who were exposed to smoking got sick.
	No	6	18	24	6/24 = 25% of those who were not exposed to smoking got sick
	Totals	43	34	83 (=n)	

To answer this question we would compare the rates – it is clear that the rate of illness is higher in those that were exposed (62.71%) when compared to the rate of illness in those that were not exposed (25%)

Ratio of Rates (RR) tells us how many times more likely one is to get the disease if one is exposed to the risk factor {or how much more likely one is to get better if one is exposed to the treatment}. Also known as likelihood ratio, relative risk, ratio of proportions, ratio of probabilities

Calculation: Ratio of Rates = 62.71%/25% = 2.5084.

Statement: People exposed to the risk factor are 2.51 times as likely to get sick as those not exposed.

Difference of Rates (RD) easy to calculate, but can be tricky to conceptualize. It is simply the absolute (raw) difference between the two rates.

Calculation: Risk difference (RD) = 62.71% – 25% = 37.71%.

Statement: Given that there are 100 exposed people, and 100 unexposed, the exposed group will have 37.71 more individuals with the disease.

Exercise 3a-3 calculate RR

In this exercise you will have a 3 tasks: come up with an appropriate research question (using reverse engineering) calculate RR, use RR to answer the research question and interpret result.

Contingency table					Reverse Engineer Research Question and respond using Relative Risk
		Dwelling owned by family		Total	RQ RR= Statement
		yes	no		
Immigrant	yes	77	23	100	
	no	74	26	100	
Total		151	49	200	
		Played video		Total	RQ RR= Statement
		yes	no		
Played sports	yes	140	60	200	
	no	220	180	400	
Total		3600	240	600	
		cancer		Total	RQ RR= Statement
		yes	no		
smoker	yes	350	150	500	
	no	70	630	700	
Total		420	780	1200	
		passed		Total	RQ RR= Statement
		yes	no		
Studied?	yes	64	36	100	
	no	120	180	300	
Total		184	216	400	

Strength and direction of association: two measurement variables

Key phrase: **strength and direction** of relation between the two variables.

Direction: When a higher value of the independent variable (e.g. years of schooling) is related to a higher value in the dependent variable (e.g. income) the direction is **positive**.

When a higher value of the independent variable (e.g. years of schooling) is related to a lower value in the dependent variable (e.g. happiness score) the direction is **negative**.

Strength: The relation between years of schooling and income will be very strong and positive if the more schooling one has means that one will have a higher income.

Calculation of practical significance: the **strength** and **direction** of correlation between 2 measurement variables is measured by Pearson's r. SPSS will do the work for us!

Pearson's r is close to 0: very weak (no) correlation, (inequality in a country has no relation to life expectancy)

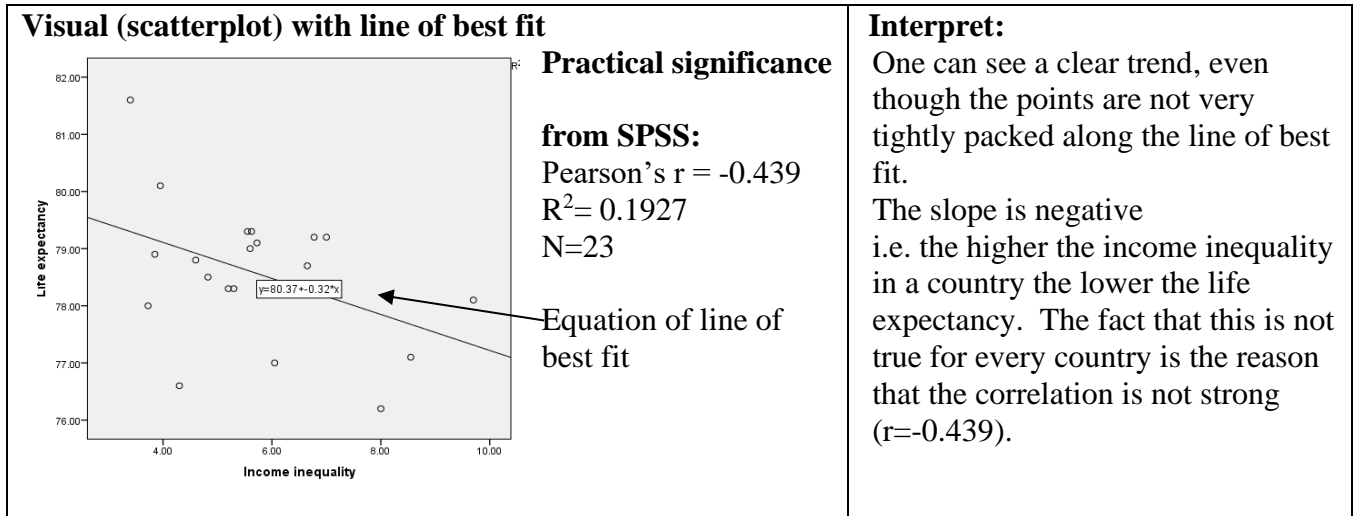
Pearson's r is close to +1: very strong positive correlation, (Higher inequality related to higher life expectancy)

Pearson's r is close to -1: very strong negative correlation, (Higher inequality related to lower life expectancy)

Note: The **correlation coefficient**, r^2 tells you what portion (%) of the variability in the dependent variable is directly related to changes in the independent (predictor) variable.

Visual output: A **scatter plot** is investigated to get a visual estimate of the strength and direction of correlation. It also helps identify any **outliers**. Look for the **direction** of association (positive slope or negative slope), and for the **strength** of association (how tightly are the points packed together in a line – or curve).

Adding a **line of best fit** really helps identify trends.



Exercise 3a-4: relate visual to numerical with two measurement variables

Reverse engineering: Researchers study characteristics of countries not just people to see if policy can affect health outcomes. One researcher proposed that high income inequality (measured by a number representing the relative worth of top 20% divided by relative worth of bottom 20% of population) may have something to do with a variety of health outcomes including life expectancy (quantified/abstracted as age at death)? Take a look at the previous page for what was found.

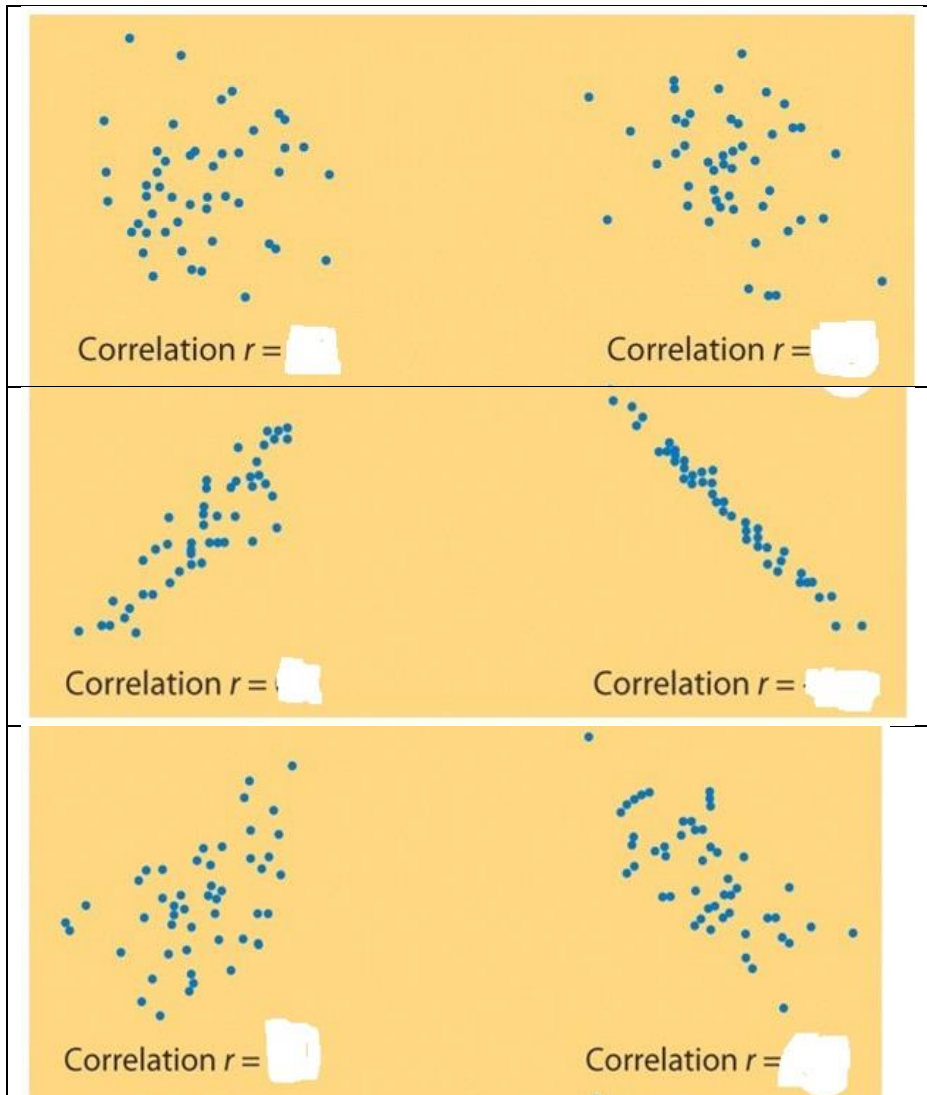
Your job is to reverse engineer a research question for the scenario.

Name of independent variable: _____ and type: _____

Name of dependent variable: _____ and type: _____

Research Question:

Fill in estimates for Pearson's r for the scatterplots below.



Strength of association between one categorical & one measurement variable

Very important! Make sure to keep in mind that the categorical variable is the independent (exposure) variable and the measurement variable is the dependent (outcome) variable. There is no other option.

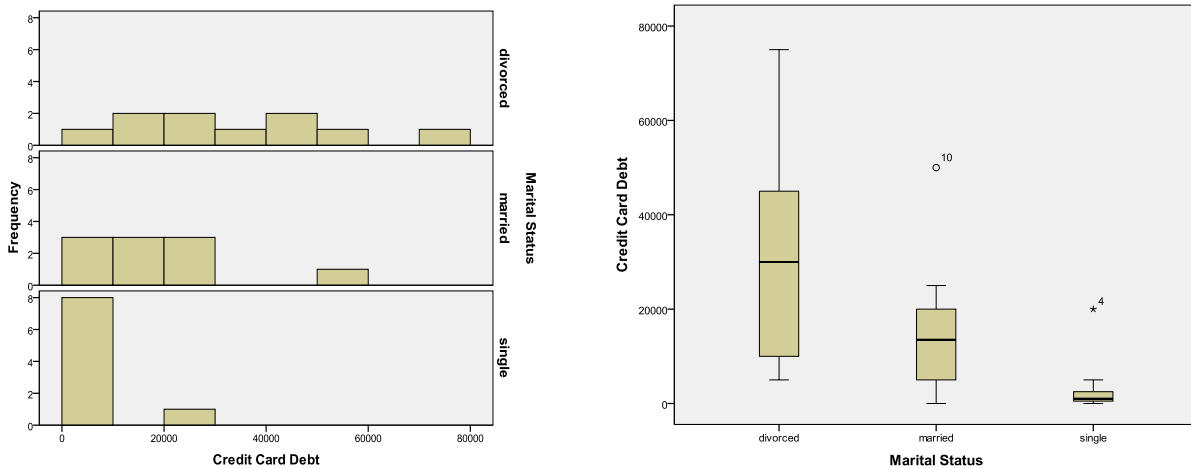
Categories are often called groups as statements of comparison (usually of mean or median)

For example: **RQ:** Which marital status carries the most credit card debt?

Name of independent variable: _____ type: _____

Name of dependent variable: _____ type: _____

Visual output, compare distributions by comparing centre, shape and spread in box plots and paneled histograms for each marital status.



Numerical output: compare distributions by comparing mean, median, and other statistics that you can get SPSS to produce in the table:

Credit Card Debt						
Marital Status	Mean	N	Std. Deviation	Median	Minimum	Maximum
divorced	31500.00	10	21864.482	30000.00	5000	75000
married	15900.00	10	14578.904	13500.00	0	50000
single	3472.22	9	6396.342	1000.00	0	20000
Total	17422.41	29	19178.800	11000.00	0	75000

Quick sample statement of interpretation: The compare means table and boxplot show that those who are single have less debt than others. The mean and median for singles is much lower and also the spread within the singles group is very small, while the spread in the divorced group is especially high (see the box plot and the values of the standard deviation). There is no real shape to any of the distributions, but that may be because we only have a few respondents from each group.

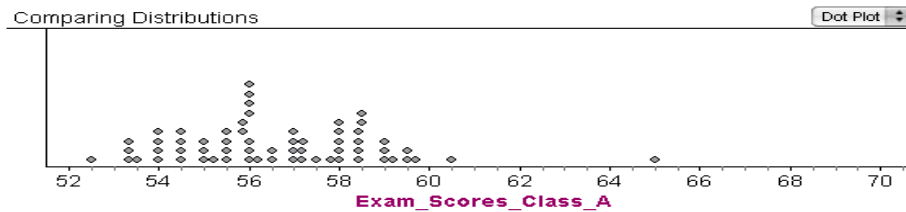
Exercise 3a-5a: *relate visual to numerical with one categorical vs one measurement variable*

Take a look at marks from exams of 3 classes (A, B and C) and use centre, spread and shape to compare them.

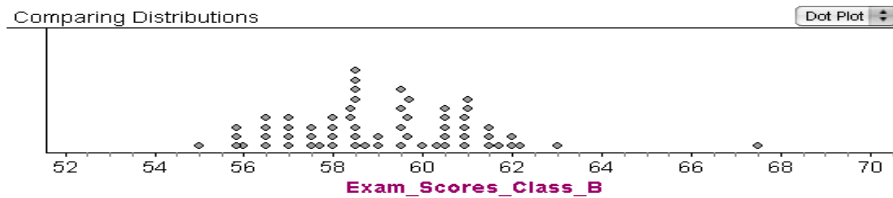
Research question: Which class did best on the exam?

Independent variable: _____ type _____

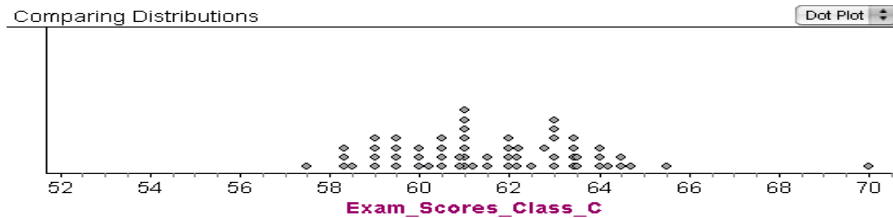
Dependent variable: _____ type _____



Q1. In what ways are the class marks different?



Q2. In what ways are the class marks similar?



Q3. Which class would you prefer to be in? Explain why.

Q4 Fill in the table below with estimates for the three classes

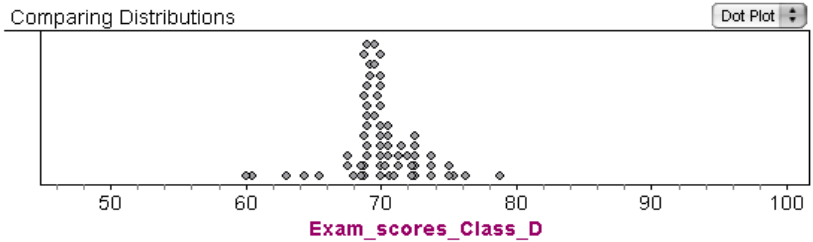
Class name	N	Mean	Mean Deviation	Min,	Max
Class A					
Class B					
Class C					

Exercise 3a-5b: Take a look at marks from exams of 3 classes (D, E and F) and use centre, spread and shape to compare them.

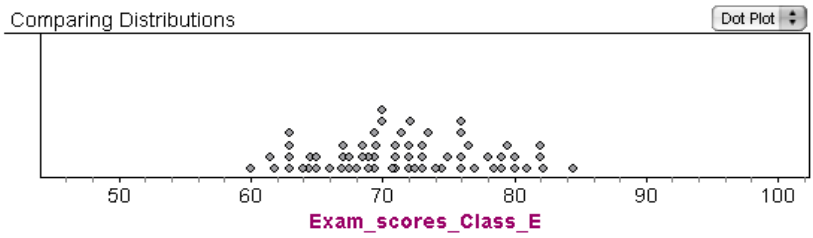
Research question: Which class did best on the exam?

Independent variable: _____ type _____

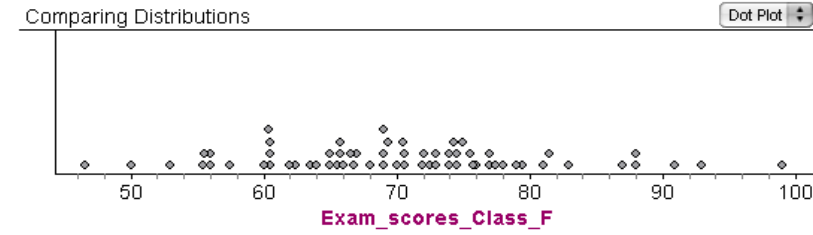
Dependent variable: _____ type _____



Q1. In what ways are the class marks different?



Q2. In what ways are the class marks similar?



Q3. Which class would you prefer to be in? Explain why.

Q4 Fill in the table below with estimates for the three classes.

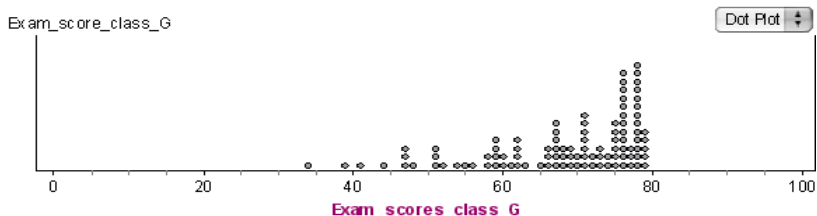
Class name	N	Mean	Mean Deviation	Min,	Max
Class D					
Class E					
Class F					

Exercise 3a-5c: Take a look at marks from exams of 3 classes (G, H and I) and use centre, spread and shape to compare them.

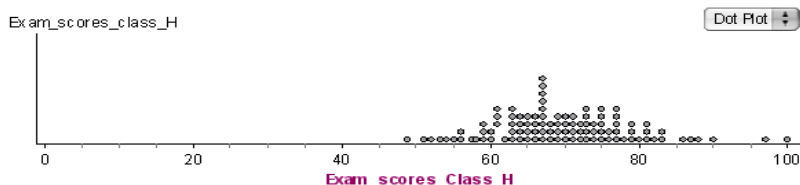
Research question: Which class did best on the exam?

Independent variable: _____ type _____

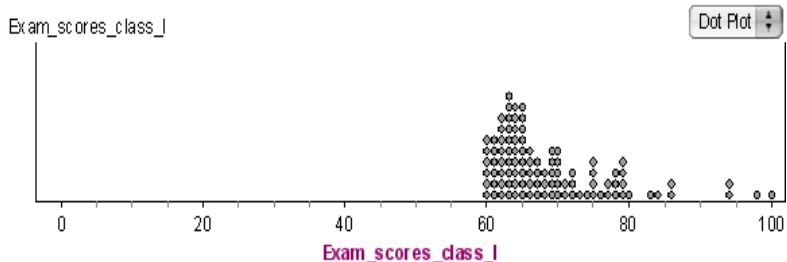
Dependent variable: _____ type _____



Q1. In what ways are the class marks different?



Q2. In what ways are the class marks similar?



Q3. Which class would you prefer to be in? Explain why.

Q4 Fill in the table below with estimates for the three classes.

Class name	N	Mean	Mean Deviation	Min,	Max
Class G					
Class H					
Class I					

Compute strength of association: one categorical and one measurement

Calculations for One Categorical and One Measurement = comparing means: The higher the difference between means the stronger the association. In this course you will use 3 methods to compare means of two groups (categories) at a time. The statement that ‘the higher the difference... the stronger the association’ sounds simple but without context a difference of 2, or 20, or 200 is meaningless. Simply comparing the means does not give us a full picture of the differences between the distribution of class marks in different classes.

Example: the calculations:

All information will come from the compare means table. (after numerical and visual output). Take a look at a simple compare means table of marks in 2 schools below.

School	mean	std dev	min	max
A ₁	74.3	10.1	42	95
B ₁	72.1	9.7	47	93
total	73.4	9.8	42	95

Below are the three calculations you will work with. The first two are ones that you should be comfortable with even without the formulas. The third is like the second except with standard deviation in denominator.

Raw difference , in absolute units	Percent difference: start with the larger mean	Cohen’s d: captures the difference in terms of standard deviation (σ)
$\mu_1 - \mu_2$	$\frac{\mu_1 - \mu_2}{\mu_2} \times 100$ or $\frac{\mu_2 - \mu_1}{\mu_1} \times 100$	$\frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ ($\sigma_{pooled} = \text{total } \sigma$)

Raw difference: B₁ has a mean mark that is 2.2 higher than A₁
 %difference: B₁ has a mean mark that is 3.05% higher than A₁
 Cohen’s d: B₁ has a mean mark that is 0.22 σ higher than A₁

Exercise 3a-6: calculate practical significance for comparing means:

4 research studies were conducted in which 2 teaching approaches are compared. School A is using one teaching approach to math, while school B another. Results from the 2 approaches were quantified as results on a standardized math test. In studies 1 & 2 test marks were calculated out of 100. In studies 3 and 4 the marks were calculated out of 20.

Research Question: Is the teaching of math in School A more effective than in School B? Do Students from school A score higher on a math test than those in school B?

Independent variable and type: _____

Dependent variable and type: _____

Numerical Results - compare means tables.

Study 1

School	mean	std dev	min	max
A ₁	74.3	10.1	42	95
B ₁	72.1	9.7	47	93
total	73.4	9.8	42	95

Study 2

School	mean	std dev	min	max
A ₂	77.5	11.5	44	95
B ₂	71.3	9.3	40	95
total	74.7	10.7	40	95

Study 3

School	mean	std dev	min	max
A ₃	14.3	4.1	8	19
B ₃	12.1	3.7	9	20
total	13.4	3.8	8	20

Study 4

School	mean	std dev	min	max
A ₄	17.5	4.5	7	19
B ₄	11.3	3.3	8	20
total	14.7	3.7	7	20

Raw differences of means (remember: raw differences is a straightforward subtraction $\mu_1 - \mu_2$)

Study 1: Students in School A's mean scores were ____ marks higher than school B's mean scores

Study 2: Students in School A's mean scores were ____ marks higher than school B's mean scores

Study 3: Students in School A's mean scores were ____ marks higher than school B's mean scores

Study 4: Students in School A's mean scores were ____ marks higher than school B's mean scores

Because the maximum mark achievable was different in the different studies. Looking at the raw differences alone does not help answer the research question. At what point are raw differences worth announcing and publishing? How many marks of a difference would get the researcher to cry: "Eureka – I found something" ?

Exercise 3a-6 (continued) calculate practical significance for comparison of means

In study 2 the raw difference between means (6.2) is much higher than in study 1 (2.2).

In study 4 the raw difference between means (6.2) is the same as in study 2, but is a much more dramatic difference because the test mark is out of 20 (instead of out of 100).

In study 3 the raw difference between means (2.2 marks) may be significant given that there were only 20 marks possible.

Results of the 4 studies from the previous page are reproduced below.

Study 1			Study 2		Study 3		Study 4	
School	mean	std dev	mean	std dev	mean	std dev	mean	std dev
A	74.3	10.1	77.5	11.5	14.3	4.1	17.5	4.5
B	72.1	9.7	71.3	9.3	12.1	3.7	11.3	3.3
total	73.4	9.8	74.7	10.7	13.4	3.8	14.7	3.7

Your task: Fill in the chart below by calculating the standardized comparisons for each of the studies above. In each case the analysis is looking at the difference School A – School B. Results for study 1 are done for you.

	Study1	Study 2	Study3	Study4
Maximum marks	100	100	20	20
Raw score difference	2.2	6.2	2.2	6.2
% difference	3.05%			
Cohen's d	0.2245 std dev			

Calculations for study 1 and statement

% difference = $100 * (74.3 - 72.1) / 72.1 = 3.05\%$ School A scored 3.05% higher than school B.

Cohen's d = $(74.3 - 72.1) / 9.8 = 0.2245$ School A scored 0.2245 std devs higher than school B.

Do the Standardized comparisons help answer the Research Question? Use the calculations of practical significance to decide whether the teaching of math in School A is more effective than in School B for each by entering 'yes', 'no' or 'still unsure' and a short justification in the spaces below.

	Study1	Study 2	Study3	Study4
Decision about superior effectiveness of school A vs B				

Data Analysis Chart.

Data Type(s)	Visual output	Numerical output	Strength of association (Non-inferential - Practical significance)	Inference (statistical significance)
one measurement m	- dot plot - box plot - histogram	- shape, centre, spread - mean, median/mode std. dev./ variance, - check for normality	N/A	Confidence Interval for mean
one categorical c	- bar - pie	Rate/frequency Frequency table	N/A	Confidence interval for proportion
Two Measurement (correlation) mm	- scatter plot - look for tightly packed points along a line	Correlation Pearson's 'r' - Strength and Direction	Strength and Direction Pearson's 'r' + 'r-squared' equation of regression line	Confidence Interval for r
Two categorical (compare rates) cc	- paneled pie	Compare rates or %/probabilities through Contingency table	Ratio of rates (RR- Relative Risk) Difference between rates (RD)	Confidence Interval for RR
One categorical & one Measurement (compare means) cm	- comparison box plot - paneled histogram	Compare means (difference of means/medians/std. deviations etc)	Raw difference between means, % difference between means Cohen's d	Confidence Interval for raw difference between means

3b: Data analysis with SPSS as calculator

Use of SPSS for computation and calculation in producing output for relations between two variables and calculations of strength of association.

Your job is to reverse engineer the connection between the research question and the variables (and data) then calculate (and visualize) relevant numerical/quantitative relations between independent and dependent variables, then prepare for, and interpret the numbers and how they help you answer the research question.

This work is a mix between rigid numeric technical work and creative playful interpretation.

Exercise 3b-1a: Use SPSS to help analyse scenarios with two categorical variables (two categories)

A survey was carried out by Stats Canada in which they asked people to identify whether they were born in Canada or not, and whether they lived in a home owned by a member of their family. (two very well defined sets!) Both variables have 2 categories (yes/no). It was thought that those born in Canada would have a higher rate of living in a dwelling owned by member of their family.

Research Question: Are Immigrants to Canada more likely to live in a dwelling owned by a member of their household than non-immigrant Canadians?

The data set '*cchs condensed*' can be found on www.statcan.ca. The two variables of interest are *Country of Birth* and *dwelling owned by member of household*.

Reverse engineer the abstractions:

Independent variable (type):

Dependent variable (type):

Numerical and visual output

Calculations of practical significance will have to be done by hand

Short written report: answer the research question directly and provide relevant evidence from Output and Calculations of Practical significance.

Exercise 3b-1b: Use SPSS to help analyse scenarios with two measurement variables.

HIM faculty at GBC were interested in finding out whether the second semester medical terminology grades were good predictors of grades in coding. They collected data for a few years thinking that higher medterm2 marks would be a good predictor of coding mark.

Research Question: Is the second semester medical terminology mark a good predictor of HIM students' mark in coding?

The data set '*med terminology vs coding marks*' can be found on www.stataras.com. Your task is to use the skills you learned to build evidence for an answer to the Research Question.

Reverse engineer the abstractions:

Independent variable (type):

Dependent variable (type):

numerical and visual output.

Calculations of practical significance will come from output and from SPSS producing the equation of line of best fit.

Short written report: answer the research question directly and provide relevant evidence from output and Calculations of Practical significance.

Exercise 3b-1c: Use SPSS to help analyse scenarios with one categorical and one measurement variable

A survey of friends and family was carried out by HIM students in their stat1013 course. It was thought that given that the birth rate in Canada is very low the number of siblings of those individuals born in Canada would be lower.

Research Question: Do Canadians not born in Canada have more siblings than those born in Canada?

The data set '*survey stat1013 w21*' can be found on the Onenote page where this exercise is posted. Two variables are of interest to us: *born in Canada* and *number of siblings*. Your task is to use the skills you learned to build evidence for an answer to the Research Question.

Reverse engineer the abstractions:

Independent variable (type):

Dependent variable (type):

numerical and visual output

Calculations of practical significance will have to be done by hand

Short written report: answer the research question directly and provide relevant evidence from output and Calculations of Practical significance.

Solutions to exercise 3b-1a

Exercise 3b-1a:

Research Question: Are Immigrants to Canada as likely to live in a dwelling owned by a member of their household in as non-immigrant Canadians?

I expect:
Immigrants will be less likely than non-immigrants to live in a dwelling owned by a family member.

7

Variables:
Country of Birth:
'Canada' 'other'

Dwelling owned by member of household:
'yes', 'no'

Data set *cchs condensed*

Independent variable is:

8

**A. Numerical exploration:
Crosstab with % (by row)**

Independent variable is better in rows

Country of birth - (G) * Dwelling - owned by a member of hold Crosstabulation

Country of birth - (G)		Count	Dwelling - owned by a member of hold		Total
			YES	NO	
CANADA	Count	1260	423	1683	
	% within Country of birth - (G)		74.9%	25.1%	100.0%
OTHER	Count	104	73	257	
	% within Country of birth - (G)		71.6%	28.4%	100.0%
Total	Count	1444	496	1940	
	% within Country of birth - (G)		74.4%	25.6%	100.0%

With % by rows we can see that a higher % of non immigrants (74.9%) live in a dwelling owned by member of household vs. immigrants (71.6%). That does not seem like a big difference.

9

**A. Visual exploration:
paneled pie chart**

I can see that a smaller 'slice' of immigrants lived in a dwelling owned by family, but it seems pretty close.

10

**B. Practical significance
Comparing rates (likelihood)**

Country of birth - (G) * Dwelling - owned by a member of hold Crosstabulation

Country of birth - (G)		Count	Dwelling - owned by a member of hold		Total
			YES	NO	
CANADA	Count	1260	423	1683	
	% within Country of birth - (G)		74.9%	25.1%	100.0%
OTHER	Count	104	73	257	
	% within Country of birth - (G)		71.6%	28.4%	100.0%
Total	Count	1444	496	1940	
	% within Country of birth - (G)		74.4%	25.6%	100.0%

$P(\text{non-immigrant lives in family dwelling}) = 0.749$

$P(\text{imm lives in family dwelling}) = 0.716$

$RR = 0.749/0.716 = 1.048$

$RD = 0.033$

Compare $P(\text{immigrant lives in family dwelling})$ to $P(\text{non-immigrant lives in family dwelling})$
Non immigrants are 1.048 times more likely to live in a dwelling owned by a member of their household. The two are equivalent.

I wanted to know:
•Are Immigrants to Canada as likely to live in a dwelling owned by a member of their household as Canadians born in Canada?

I found:
Immigrants are slightly less likely to live in a dwelling owned by a member of their household (0.716 vs 0.749). Non-immigrants are 1.05 times more likely to live in a dwelling owned by a member of their household.

Decision: Immigrants are equally likely as non-immigrants to own the home that they live in (RR = 1.05). There does not seem to be much of a difference between them.

Solutions to exercise 3b-1b

Exercise 3b

Research Question: Is the second semester medical terminology mark a good predictor of ones mark in coding?

The marks of a sample HIM students at a large urban college were analysed to help answer the question.

15

The data:

'medical terminology vs coding marks'

	medterm1	medterm2	coding	codingvar
1	88.77	76.39	61.34	2007
2	80.33	75.13	66.54	2007
3	87.33	73.96	78.84	2007
4	89.00	87.19	82.16	2007
5	84.84	86.26	89.13	2007
6	92.52	88.17	79.29	2007
7	91.67	86.65	70.72	2007
8	63.65	56.26	63.31	2007
9	85.91	72.10	77.17	2007
10	68.94	51.40	61.86	2007
11	92.66	89.24	82.29	2007
12	89.67	75.76	79.51	2007
13	78.11	68.43	42.66	2007
14	77.98	64.83	62.38	2007
15	76.36	75.95	57.63	2007
16	78.97	71.31	76.70	2007
17	78.36	80.25	37.78	2007
18	79.13	71.49	72.18	2007
19	90.43	74.84	74.00	2007
20	78.36	64.25	63.52	2007
21	79.82	65.90	71.60	2007
22	87.01	80.73	71.32	2007
23	87.78	79.21	78.27	2007
24	60.21	36.31	62.75	2007

16

Exploration & Practical Significance:
scatterplot with line of best fit + Pearson's r & correlation coefficient

$r^2 = 0.498$
 $r = 0.705$

	medterm2	coding
medterm2	Pearson Correlation	1
	Sig. (2-tailed)	.000
	N	88
coding	Pearson Correlation	.705
	Sig. (2-tailed)	.000
	N	87

** Correlation is significant at the 0.01 level (2-tailed).

17

Practical Significance: Regression
equation for line of best fit

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	12.370	6.189	1.999	.049
	medterm2	.733	.080	.705	.000

a. Dependent Variable: coding

Regression equation: $y = 12.370 + 0.733x$
y is 'coding mark'; x is 'medterm2 mark'
Predict your own coding mark...

18

D. What have we learned?

Strength and direction:
Relation between the 2 marks is strong and positive ($r = 0.705$)

$r^2 = 0.497$: 49.7% of the variability in the coding mark comes from the medterm2 mark

I can predict the coding mark if I know the students medterm2 mark

$y = 12.370 + 0.733x$

Answer to research question

There is evidence ($r = 0.705$), that an HIM students' medterm2 mark is strongly and positively related to (or is a strong predictor of) his or her coding mark.

49.7% of changes in the coding mark come from changes in the medterm2 mark (leaving over 50% of changes unaccounted for).

The predicted relationship is as follows:

Coding Mark = $12.370 + 0.733(\text{medterm2 mark}) + \text{error}$

Solutions to exercise 3b-1c

Exercise *ME*

Research question:
Do Canadians born outside of Canada have more siblings than those born in Canada?

data set: *surveystat1013 (older version)*

26

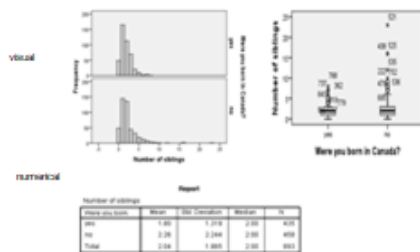
Do Canadians born outside of Canada have more siblings than those born in Canada?

Independent Variable: Birthplace (2 categories – Canada/other)

Dependent Variable: #of siblings (measurement)

27

Exercise *ME* visual and numerical exploration



28

B. Practical significance.

Report				
Number of siblings	Mean	Std. Deviation	Median	N
Were you born...				
yes	1.60	1.319	2.00	435
no	2.26	2.244	2.00	458
Total	2.04	1.665	2.00	893

There is not much difference between the 2 groups,

Raw difference: Those not born in Canada have 0.46 more siblings on average.

Percent difference: Those not born in Canada have 25.56% more siblings on average.

Cohen's $d = 0.247$. Those not born in Canada have 0.247 std devs more siblings.

29

Friends and family of HIM students not born in Canada have more siblings than those who are born in Canada

difference is not significant. Both groups have mean & median # of siblings ≈ 2 . (Raw difference 0.46 siblings; % difference = 25%)

The histograms and boxplot comparisons show that the distributions are similar except that those not born in Canada have a few outliers with ≥ 10 siblings.

It would be prudent to run the comparison without these outliers as they likely contributed to the difference.

Exercise 3b-4 – more boot camp questions I have left a little bit of space for some rough work, but please practice completing full solutions in a separate space.

Q1. George Brown HIM students collected data from their friends and family. One of the issues of interest was whether those born in Canada had a higher probability of being happy than those that were not born in Canada.

Research Question Were respondents to the friends of HIM survey who are born in Canada more likely to be happy than those respondents not born in Canada?

Data set: *'asst survey all'*

Q2. Research engineers were hired to help standardize brick production in 143 brick factories in Egypt. After 3 years data was collected on the number of bricks produced and gas consumption (m^3). Use data from the *'brick factory data'* to answer the research question below.

Research Question: Is there evidence to suggest that brick production has been standardized?

Q3. Answer the given research question using information from the *marks1112* data set:

Research Question: Did any of the classes between the fall of 2009 and fall 2016, do better in the math 1112 course than the others?

Q4. Answer the given research question using information from *modetrans vs numcolds* data set:

Research Question: How was the number of colds respondents got related to the mode of transportation they used?

Exercise 3b-4 – more boot camp questions - continued

Q5. Research Question: Were Canadian respondents who worked more likely to think they should do something to improve their health?

Data set: '*cchs condensed*'

Q6. Income inequality is calculated by dividing the incomes of the top 20% of a population by the incomes of the bottom 20% of the population. Researchers believe that income inequality in a country tells us about many other things in that country. One of the characteristics that researchers believe is dependent on income inequality is teenage births. Data set: '*income inequality*'

Research question: Did countries with high income inequality also have high rates of teenage births?

Q7. Those who sit on a couch all day and do not work are less likely to be injured. This is a speculation that many could agree with. But is there any truth to it?

Research Question: Were Canadian respondents who worked more likely to be injured in past 12 months?

Data set: '*cchs condensed*'

Q8 There is a general perception that taller people weigh more. Now you have a chance to see if there is any evidence to that perception in Canada

Research question: Do taller Canadian respondents really weigh more? Use the '*cchs2.1 data set*' to answer the question.

Exercise 3b-4 – more boot camp questions - continued

Q9. The LOS in Toronto data set is from a study in which LOS for appendicitis is compared in a group of 4 hospitals in the city of Toronto. (St. Joseph's, Mount Sinai, Toronto General, and Sunnybrook). The data is made up. Answer the research question assuming the data is real.

Research Question: Was the LOS at Sunnybrook lower than other hospitals in Toronto?

Q10. There is a general stereotype that men drink more than women, but some media reports indicate that this is not in fact true. Use the *cchs condensed* data set to help clear this up by answering the research question posed.

Research Question: Do Canadian males consume more drinks per week than Canadian females?

Q11. Many Canadians get injured every year, but little is known about the effect that has on productivity. Do those that get injured leave the workplace? Do they continue to work as effectively as those that did not get injured? Use data from the *cchs condensed* data set to answer the research question posed.

Research Question: Did Canadian respondent who were injured in the last 12 months work as many hours a week as Canadians who were not injured?

Q12. George Brown HIM students collected data from their friends and family. One of the issues of interest was whether happy people would have more siblings than unhappy people. Data was stored in the *asst.survey.all* data set:

Research Question: Who has a higher number of siblings GBC-HIM friends who are happy or those who are not happy?

Unit 4

Inferential statistics – a conceptual Introduction

Inferential statistics is made up of a series of thinking tools which help the researcher learn something about populations, systems, groups or events by studying a sample (a subset) of the population.

Statistical Inference around us:

Every time you take a headache pill, or some other medication you are trusting inferential statistics. Experiments were conducted on a small group of individuals somewhere in Canada (or other country) that found the medication to be effective and safe. By trusting the claims of the drug companies and trusting the Health Canada approval process, you are trusting the process of inferential statistics.

When food inspectors hired by the government to ensure the safety of our food system test cold cuts for bacteria that can make a person sick they also use inferential statistics as they don't go out and test every single cold cut produced. They take a sample of cold cuts, test them, and make inferences about the safety of a large population of cold cuts.

When there is an outbreak of disease and researchers try to find the cause for the outbreak, they do not have the time (nor the resources) to investigate every case. Instead they investigate a sample of individuals to see if exposure to a potential pathogen (bacteria in romaine lettuce?) is somehow related to diagnosis of disease.

Limits of inference:

One study on one sample gives researchers a fuzzy picture of the truth with a probability level attached. e.g. we have 95% confidence that 75% +/- 10% of individuals will feel reduced pain level after taking this medication. Many other steps need to take place (e.g. replication of results in another inferential study) since usually the confidence level is 95%, which means that we expect to be wrong 1 out of 20 times – that is not a probability that it is worth risking your life on... is it? In the research methods course you will get a more complex look at the limits (and power) of inference.

What you will learn:

In this unit you will experience inference through the action of calculation and interpretation. You will do in class exercises to help you develop a sense of how one of the inferential thinking tools (confidence intervals for single mean and proportion) works and learn how to interpret confidence intervals to answer inferential research questions.

Exercise 4: Thinking forward from idea to interpretation (using 4 actions: define question, abstract, compute, interpret)

Phase 1: Define question... I have done that for you.

RQ: How much time do HIM students spend studying in semester 1 of their program?

Phase 2: abstract (quantification) Make ‘time spent studying’ well defined & come up with a reasonable hypothesis (prepare for interpretation)

Step 1: We will come to a consensus as a class about which activities should be included in time spent studying. Write down the list in the space below:

Step 2: Think about how much time you spent studying in a typical week in your first semester. Make sure to take into account the busy weeks (end of semester) and the less busy weeks (start of term). Enter values below: these will be entered into OneNote by Taras:

My estimated typical time spent studying per week = _____

My minimum time studying = _____ My maximum time studying = _____

Step 3 (individual prediction): Hypothesis 1. Make a predictions (an inference based on your own values from step 2) about the distribution of ‘time spent studying’ by your classmates.

$\mu_{\text{timestud1}} \approx$ _____ $\min_{\text{timestud1}} \approx$ _____ $\max_{\text{timestud1}} \approx$ _____ $\text{Range}_{\text{timestud1}} \approx$ _____

Phase 3: compute

Step 1: (collect and calculate for sample of 4-5) Collect estimates (from step 2 above) for each group member and enter them below.

Write out ‘time spent studying’ for each group member: _____

Calculate the following and write them below and into OneNote.

$\bar{x}_{\text{timestud}} =$ _____ $\min_{\text{timestud}} =$ _____ $\max_{\text{timestud}} =$ _____ $\text{Range}_{\text{timestud}} =$ _____

Step 2: Hypothesis 2: Make individual prediction about the distribution of ‘time spent studying’ of your all your classmates based on information from your group results by predicting the following:

$\mu_{\text{timestud2}} \approx$ _____ $\min_{\text{timestud2}} \approx$ _____ $\max_{\text{timestud2}} \approx$ _____ $\text{Range}_{\text{timestud2}} \approx$ _____

Exercise 4: continued

Phase 3:

Step 3: (visual and numerical summaries of timestudying) Use the ‘timestudying’ data set (Taras will post it).

A. Use SPSS to generate a dot plot, histogram, box plot and descriptive statistics. Draw a rough sketch of the distribution below by hand.

Dot plot	Box plot	Histogram

Write out descriptive statistics: $\mu_{\text{timestud}} = \underline{\hspace{2cm}}$ $\text{median}_{\text{timestud}} = \underline{\hspace{2cm}}$

$\text{min} = \underline{\hspace{2cm}}$ $\text{max} = \underline{\hspace{2cm}}$ $\text{range} = \underline{\hspace{2cm}}$

Describe the distribution in terms of shape, center, and spread.

Phase 4: Interpret

Q1: Which of your predictions for μ was closer to the true value: individual prediction (phase 2 step3)... or group prediction (phase 3 step2)? Discuss any possible reasons for this?

Q2: Imagine if **four** classes of similar students were combined (i.e. to get a data set with approximately 200 students). In what ways would you expect to see this histogram differ from the one for your class of about 50 students?

Unit 4a: Inference concepts

Inference in practice. (to toss or not to toss the salad)

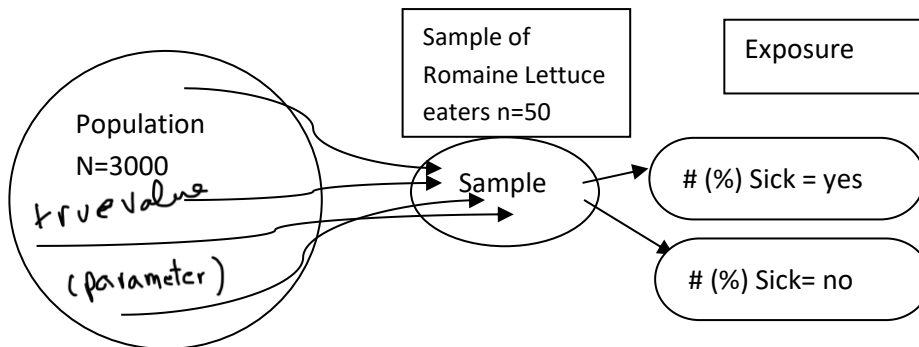
Imagine that there has been a claim made that romaine lettuce has been causing illnesses on a cruise ship. Should it be taken off the menu? This would be a bad idea as tonight's dinner is lasagna, and everyone wants Caesar salad with their lasagna. There is no 'salmonella' test on board, and it is your job to make the decision – toss the lettuce or serve it and maybe poison the passengers. The trick is to figure out a way to get information without bothering everyone.

Luckily you took this course and know all about inferential stats and a few tools of data analysis. Your plan is to assess the risk of getting sick from eating Romaine Lettuce by finding 50 individuals who ate romaine lettuce yesterday (there are 3000 total on the cruise) and finding out whether they felt sick within 24 hours after that.

Non-inferential research question: what % of Romaine Lettuce eaters got sick?

Inferential research question: What % of Romaine Lettuce eaters can we expect to get sick on the cruise after the lasagna dinner? What is the risk of eating Romaine Lettuce on that ship?

Flow Chart of sampling and analysis on cruise ship Romaine Lettuce problem.



Results after data collection and entry analysis:

Ate romaine		Total (n)
yes	no	
32	18	50

non-inferential: 32/50 (64%) of Romaine Lettuce eaters got sick.

Inferential (after Excel calculation):

I am 95% confident that the rate of illness in Romaine Lettuce eaters will be $64\% \pm 13\%$ (i.e. between 51% and 77%)

Note the differences between inferential and non-inferential analysis. In either case, the choice is pretty easy here! Toss the salad!!! (i.e. throw it out!)

Exercise 4a-1: sampling simulation with dice

Background: Recall that the SS for rolling one fair six-sided die is $SS = \{1,2,3,4,5,6\}$; based on this fact we would expect that in one roll of 6 dice we would get one of each facing up. Perhaps you are not sure about that... if so, then how about 6,000,000 rolls, here it is easier to imagine that there would be 1,000,000 1s, 2s, etc. Extending this imaginative journey we would then expect the ‘true’ mean (μ) of the act of rolling these 6 dice would be $(1+2+3+4+5+6)/6 = 3.5$; as it would be for 6 million dice.

Today’s exercise involves testing this expectation individually and collectively - as a class of 30+ students - and testing the ability of confidence intervals to capture ‘true means’.

Step 1: roll the 6 dice given (or one die 6 times) and record the 6 values ___ ___ ___ ___ ___ ___

Step 2: using SPSS or other software calculate $\bar{x} =$ ___ and $s =$ ___

Step 3: get your software of choice to calculate the 95% confidence interval for μ : LB = ___ UB = ___

Step 4: plot your values on the whiteboard under Taras’ guidance.

Step 5: copy the plot representing every class members \bar{x} and confidence interval and mark those that did not capture the true population mean $\mu = 3.5$



Step 6: the 95% confidence interval claims to be correct 19 times out of 20 ($19/20 = 95\%$). i.e. it will not capture the ‘true mean (μ) 5% of the time. What % of the confidence intervals in the class missed the ‘true mean’ (μ)? _____

Example: if LB = 2.7 and UB = 3.4, then the confidence interval missed the true mean $\mu = 3.5$

Inferential Stats and Confidence Intervals

The Four Fundamental Concepts of Statistics

Principles underlying inferential statistics
(as described by Sally Caldwell)

Census vs sample?

Problem: In most research a census is not possible
e.g. Find the mean number of hours of sleep of GBC students

Solution: estimate the population mean (or other statistic) by taking a sample from the population.

Problem: we typically don't know the true population mean (or other characteristic) and thus have no idea if our sample is a 'good one'

Solution: Inferential statistics

symbols you will need to learn by heart

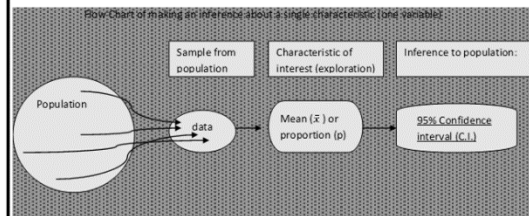
\bar{x} – sample mean

μ – population mean

s – sample standard deviation

σ – population standard deviation

Making inferences with one variable a flowchart:



What happened in exercise 4?

Concept #1 Random Sampling

In order to be able to make valid inferences about a population sampling must be done randomly.

To choose a sample randomly means that:

1. Every member of the population being studied has an equal chance of being chosen.
2. Each individual is chosen independently.
3. All combinations – even the weirdest ones are possible.

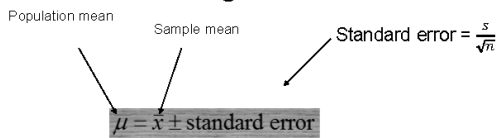
Concept #2 Sampling Error

Most sample means will approximate the population mean, some will be way off. Because each sample contains only a fraction of the population each one will contain some error.

this 'error' is called sampling error, and because of it, we cannot make automatic inferences from sample to population.

sampling error can be estimated and used to help make inferences.

Here is an idea for estimating μ using \bar{x} and s



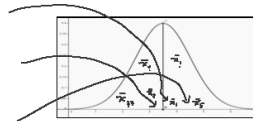
This is ok as an idea, but the standard error is too small...

Concept #3 Distribution of Sample Means

Imagine that the researcher is able to keep taking samples and calculating the mean for each one. This will create a set of sample means as follows: $\{\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots\}$

The collected and ordered data set containing sample means is called *the distribution of sample means*.

We could have created a distribution of sample means in exercise *Q15A 4f*



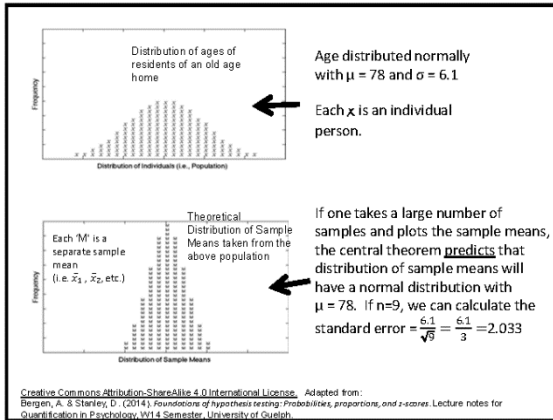
Each sample mean (randomly selected of course) is one sample from the infinite possible sample means – i.e., from the distribution of sample means.

Concept #4 Central Limit Theorem

If... the population (from which the samples are taken) has mean = μ , and standard deviation = σ

...then the distribution of sample means will have mean = μ and standard error = $\frac{\sigma}{\sqrt{n}}$

as n gets larger the standard error gets smaller



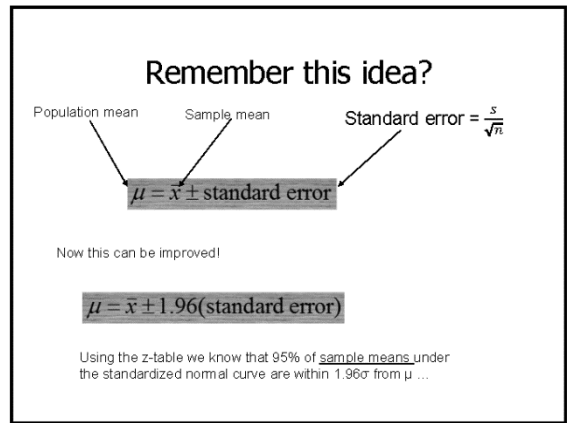
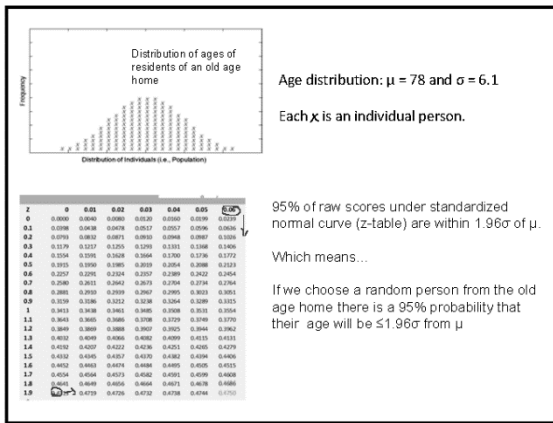
- the distribution of sample means is normal (even if the original distribution was not)

....and....

The standard error is the std deviation of the distribution of sample means...

...and..

we can use those two facts to find areas in the distribution of the sample means just like with raw scores (exercise 14)...



Problema.....

There is a different distribution of sample means for every sample size – which makes things a bit more complicated.

This was discovered by a math student working for Guinness breweries

It means that we can't blindly use 1.96 when building a confidence interval for means. Different sample sizes have different multipliers.

$df = \infty$ $df = 6$ $df = 3$

<http://www.psychstat.in.tusculostate.edu/Introbook/Isbi-24.htm>

Take a look at the t-tables on www.statcrs.com

Family of t Distributions (Two-Tailed Test)

Degrees of Freedom (df)	LEVEL OF SIGNIFICANCE				
	.20	.10	.05	.01	.001
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861

$Df = n - 1...$

For $n = 9$ we would use $t = 2.306$ to capture 95% of the area under the curve.

Shift from population distribution to distribution of sample means
(a.k.a. shift from z-score to t-score)

Population distribution	Distribution of sample means
Normal or other	Normal
Mean = μ	Mean = μ
Std deviation = σ	estimate for standard error $\left(\frac{s}{\sqrt{n}}\right)$
Raw score - x	Raw score - \bar{x}
For areas – use z-tables	For areas – use t-tables

The confidence interval for means

Population mean

Sample mean

Sample standard deviation

Sample size

The value in the Brackets is the Standard error

t-score, dependent on Confidence level and d.f. = n - 1

What is it for?

The confidence interval captures the population mean in a range.
After solving for the unknowns you will get a statement like the following :

We are 95% confident that $\mu = 167 \pm 4.5$

The interval $\left\{ \begin{array}{l} \text{---} \\ \text{---} \end{array} \right\}$

or.... The 95% Confidence interval for height is $162.5 \leq \mu \leq 171.5$

or... We are 95% Confident that the population mean height is between 162.5 and 171.5

LB UB

Z-score vs t-score

z-score

$$z = \frac{x - \mu}{\sigma}$$

t-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

Estimate for σ

n is the sample size;
s is the sample standard deviation
degrees of freedom = n - 1

The confidence interval for proportions

Population proportion

Sample proportion

Estimate of variation

z-score, dependent on Confidence level only

Sample size

LB = Lower Bound
UB = Upper Bound

Exercise 4a-2: sampling simulation 2 – (with SPSS)

Open up the *weights* data set found on www.stataras.com These are the actual weights of every single resident of a town of 2000 people. The true mean population mean ($\mu = 73.61$) and standard deviation ($\sigma = 17.598$) can be found from this census.

In this exercise you will be taking a sample ($n= 9$) from this population in order to simulate how well we can estimate the true mean using sampling and confidence intervals, and to learn a bit about sampling theory.

1. Use the SPSS instructions booklet to do the following: Set the random number seed then take a sample of size 9 from the townspeople’s weights,

Sampled weights: _____

3. Calculate the sample mean and sample standard deviation for your sample by hand or by using SPSS.

$\bar{x}_{individual}$ _____ $S_{individual} =$ _____

4. Use SPSS instructions to calculate the 95% confidence interval for the ‘true population mean’

Lower bound: _____ Upper bound: _____

5. Someone will collect the sample means from every individual in the class and enter them into a data set called ‘sample mean weights’. Open this new data set which is a ‘distribution of sample means’ i.e. \bar{x} from each person in the class and fill in the blanks below.

$\bar{x}_{distribution} =$ _____ $S_{distribution} =$ _____

is the distribution of sample means normally distributed? _____

6. Is your sample mean ($\bar{x}_{distribution}$) close to the true population mean ($\mu = 73.61$)? _____; The answer may be yes or may be no as predicted by sampling error (fundamental concept #2).

Is the mean of sample means ($\bar{x}_{distribution}$) close to the true population mean ($\mu = 73.61$)? _____; The answer should be yes as predicted by the central limit theorem (fundamental concept #4)

Unit 4b: Compute Confidence Interval for single mean (for one measurement variable)

Notation:

Population	Sample
μ	\bar{x}
σ	s

Goal: to capture the true population mean (μ) by using the sample mean (\bar{x})

(Up to now the only symbol we used for a mean was μ)

Confidence intervals are constructed when taking a sample from a population and estimating the true population mean (or proportion) based on the sample mean (or proportion).

For example: We want to know whether coders working in a particular LHIN are taking only 30 minutes per day for lunch as expected. What we need to do is to keep track of when they leave for lunch and when they come back, but that is very difficult to do for all coders in this LHIN. Instead we check a sample of coders by sampling a few locations and ‘shifts’.

Because we are taking a sample from the population of coder lunch times (in minutes) we know that there is a probability that the resulting mean \bar{x} time for lunch is not a perfect representation of the true population proportion μ time for lunch.

In order to get an estimate we build a confidence interval for the population mean μ (see formula to the below).

$$\mu = \bar{x} \pm t \left(\frac{s}{\sqrt{n}} \right)$$

Because we are working with a sample from a population – instead of using the z-table we need to use a different table for every sample size –

Each sample size has its own ‘normal’ distribution. The t-table is actually made up of many tables. (degrees of freedom) d.f. = sample size – 1 = $n - 1$.

Exercise 4b-1: calculate confidence interval for measurement variable

Confidence intervals for means	
$\mu = \bar{x} \pm t \left(\frac{s}{\sqrt{n}} \right)$	<p>You will get \bar{x} and 's' and the sample size 'n' from the sample data.</p>
<p>Practice: Find the 95% confidence interval for the lunch times (in minutes) in the following.</p> <ol style="list-style-type: none"> 1. Sample size = 16; $\bar{x} = 33.2$, $s = 2.12$. 2. Sample size = 73; $\bar{x} = 33.2$, $s = 2.12$. 3. Sample size = 100; $\bar{x} = 33.2$, $s = 2.12$ 	<p>'t' will come from the tables – you will need the following to find t</p> <p>d.f. = $n - 1$</p> <p>confidence level = 95% = 0.95 (always 95% in this course)</p> <p>significance level – 1 – confidence = 0.05 (100% - 95% = 5% = 0.05)</p> <p>Express your answer in one of two formats:</p> <p>The 95% C.I. for the mean time spent at lunch is between 27 and 45 minutes. Or...</p> <p>The 95% C.I. for time at lunch is $27 < \mu < 45$.</p>

Luckily, for most of the work in this course we will get SPSS or excel to calculate it for us.

Exercise 4b-2: Confidence intervals for means using SPSS

In this lesson we will go back to the ‘weights ’ data set available on www.stataras.com. The data set represents the responses of 2000 residents of the village of Magnetawan, Ontario. We can find the actual mean weight μ of this village, but we will go ahead and estimate the sample mean \bar{x} and by taking various samples and building various Confidence intervals around each.

- The actual population mean weight is: $\mu = 73.61$

Your tasks:

1. Take a random sample of 9, and build a 95% (and 99%)c.i. for the mean
2. Take a random sample of 49 and build a 95% (and 99%) c.i. for the mean
3. Take a random sample of 121 and build a 95% (and 99%) c.i. for the mean
 - Place all your results in the chart below.

sample	sample size	confidence level	\bar{x}	s	standard error	lower bound	upper bound	capture of actual mean yes/no
1a	9	95%						
1b	9	99%						
2a	49	95%						
2b	49	99%						
3a	121	95%						
3b	121	99%						

Unit 4c: Compute Confidence intervals for single proportion (for one categorical variable)

Notation:

Population Proportion(rate/percentage)	Sample Proportion/rate
π	p

Goal: to capture the true population proportion (π) by using the sample proportion (p)

Remember that confidence intervals are constructed when taking a sample of (size 'n') from a population (size 'N').

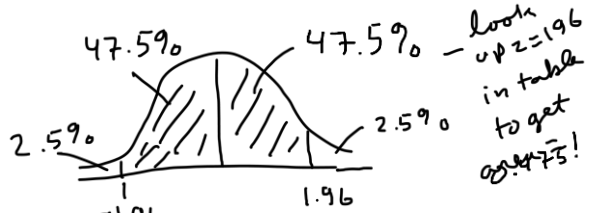
The equation representing the relation between π and p is as follows:

$$\pi = p \pm z \sqrt{\frac{p(1-p)}{n}}$$

For example: We want to know whether a group of coders is keeping their error rate down below a certain target. What we need to do is to compare their codes to the original charts. It would be impossible to check every single chart that each coder has coded, instead we check a random sample of charts over the last month. Because we are taking a sample from the population of charts that have been coded, we know that there is a probability that the resulting proportion of errors 'p' from the sample is not a perfect representation of the true population proportion ' π ' of errors. In order to get an estimate we build a confidence interval for the proportion (using the formula above).

Note that the population proportion (π) is estimated by adding a bit and subtracting a bit from the sample proportion (p) for 95% confidence we would use $z = 1.96$ as the area under z between -1.96 and +1.96 is 95% of the total area. We will be 95% sure that the true mean will be within the upper and lower bounds this equation produces.

Exercise 4c-1: calculate C.I. for proportion

<p>Formula</p> $\pi = p \pm z \sqrt{\frac{p(1-p)}{n}}$	<p>You will get the proportion (also known as rate) 'p' and the sample size 'n' from the sample data. 'z' will come from the tables – you will always use Z = 1.96 since that produces a confidence of 95% for a two tail test (see diagram below)</p>
<p>Practice: Find the 95% confidence interval for the proportions of errors in the following.</p> <ol style="list-style-type: none"> 1. Sample size = 100; $p = .73$ 2. Sample size = 500; $p = .73$ 3. Sample size = 2000; $p = .73$ 	 <p>When you add up the two tails you get $2.5\% + 2.5\% = 5\%$ (95% confidence)</p> <p>Express your answer in one of two formats: e.g. If $p = 0.36$ and $n = 109$</p> $\pi = 0.36 \pm 1.96 \left(\sqrt{\frac{0.36(0.64)}{109}} \right) = 0.36 \pm 0.0901$ <p>The 95% C.I. for the proportion of coder errors is between 27% and 45 %.</p> <p>Or $\pi = 36\% \pm 9\%$ which is more typical. 9% is called the margin of error</p>

Exercise 4c-2: Confidence Intervals. with one proportion

Q1. In a sample of 500 Canadians it was found that 46% were male. Calculate the 95% confidence interval for the proportion of males in Canada.

Q2. In a sample of 2000 Canadians it was found that 46% were male. Calculate the 95% confidence interval for the proportion of males in Canada.

Q3. Use the CCHS condensed data set to estimate the 95% Confidence interval for the following:

- a. The proportion of Canadians who think that they should do something to improve their health.
- b. The proportion of Canadians who were injured in the past 12 months.
- c. The proportion of Canadians who worked at a job in the last 12 months.
- d. The proportion of Canadians who are married.
- e. The proportion of Canadians who are single.
- f. The proportion of Canadians who are widowed/separated/divorced.

Unit 5

Interpreting the work of others

All of us will make many decisions about our health and health care or prevention using evidence from research studies.

In this last section of the course you will get a taste of reading a few research reports and looking for the thinking tools that they used to support their findings.

Reading Research Template: What to look for in research articles: a template.

When reading research it is important to pull out key pieces of information that will help you understand the goals of the study and be able to evaluate whether it achieves those goals.

Our reading of research will become more complex as we develop the language of research. To start off here are 5 questions that you need to be able to answer when reading a report on research.

1. What is the research question (RQ)?

You should be able to pull this out of the article word for word, unless the writers did not do a good job in presenting the information. The research question could also be framed as goals, topic etc.; typically presented in first paragraph or even in the title.

2. What data is being collected? Name of independent (exposure) variable – and type: Name of dependent (outcome) variable – and type:

This takes practice to figure out. Variable names and types are often coded in the research question.

Name(s) of other variable(s) being collected:

these are often demographic information or other characteristics of interest that are not part of the main research question – e.g. side effects of treatments.

3. Who or what is being studied? (unit of analysis)

Most of the time in health care research this is easy to answer, but occasionally this is a key to understanding a more complicated study.

4. What are the findings/conclusions according to the researchers?

What did the researchers describe as their findings? Were the findings definitive, or was the language hesitant?

5. Numerical and visual evidence that support the findings presented in the article.

Refer to all evidence even if you are not sure what it means.

Exercise 5-1: Practice with reading research

Read the article below and answer the 5 questions given in the template on the previous page.

Pregnancy Problems Tied to Caffeine Denise Grady (NY Times Jan 21, 2008)

Too much caffeine during pregnancy may increase the risk of miscarriage, a new study says, and the authors suggest that pregnant women may want to reduce their intake or cut it out entirely. Many obstetricians already advise women to limit caffeine, though the subject has long been contentious, with conflicting studies, fuzzy data and various recommendations given over the years.

The new study, being published Monday in the American Journal of Obstetrics & Gynecology, finds that pregnant women who consume 200 milligrams or more of caffeine a day — the amount in 10 ounces of coffee or 25 ounces of tea — may double their risk of miscarriage. Pregnant women should try to give up caffeine for at least the first three or four months, said the lead author of the study, Dr. De-Kun Li.

On Friday, the March of Dimes Web site said most experts agreed that the amount of caffeine found in 8 to 16 ounces of coffee a day was safe. It noted that some studies had linked higher amounts to miscarriage and low birth weight, but stated: “However, there is no solid proof that caffeine causes these problems. Until more is known, women should limit their caffeine intake during pregnancy.” Now, having reviewed the new study, the March of Dimes plans to change its message, to advise women who are pregnant or trying to conceive to limit their daily caffeine intake to 200 milligrams or less, said Janis Biermann, its senior vice president of education and health promotion.

Dr. Li’s study included 1,063 pregnant women who were interviewed once about their caffeine intake. Of 264 women who said they had used no caffeine, 12.5 percent had miscarriages. But the miscarriage rate was 24.5 percent in the 164 women who consumed 200 milligrams or more per day. The increased risk was associated with caffeine itself and not with other known risk factors like the mother’s age or smoking habits, the researchers said.

Dr. Li said the study answered an important question that previous research had left unresolved. Women who have morning sickness are less likely to miscarry than those who do not, possibly because the same hormonal changes that cause nausea and vomiting contribute to a healthy pregnancy. But some researchers said morning sickness could lead to confusing results in caffeine studies. These researchers argued that because they feel ill, some women may consume less caffeine. Dr. Li said he and his colleagues had determined that the risk from caffeine was real and could not be explained away by different rates of morning sickness.

Dr. Carolyn Westhoff, a professor of obstetrics and gynecology, and epidemiology, at Columbia University Medical Center, had reservations about the study, noting that miscarriage is difficult to study or explain. Dr. Westhoff said most miscarriages resulted from chromosomal abnormalities, and there was no evidence that caffeine could cause those problems. “Just interviewing women, over half of whom had already had their miscarriage, does not strike me as the best way to get at the real scientific question here,” she said. “But it is an excellent way to scare women.” She said that smoking, chlamydial infections and increasing maternal age were stronger risk factors for miscarriage, and ones that women could do something about. “Moderation in all things is still an excellent rule,” Dr. Westhoff said. “I think we tend to go overboard on saying expose your body to zero anything when pregnant. The human race wouldn’t have succeeded if the early pregnancy was so vulnerable to a little bit of anything. We’re more robust than that.”

Exercise 5-2: practice with ‘no data set’ scenario analysis:

In a recent survey, 695 Canadians were asked whether they had been injured in the last 12 months and whether they felt that they should do something to improve their health. Of the 108 who were injured, 71 felt that they should do something to improve their health, while out of the 587 who were not injured 367 felt they should do something to improve their health.

Research Question: Are Canadians who were injured more likely to feel that they should do something to improve their health?

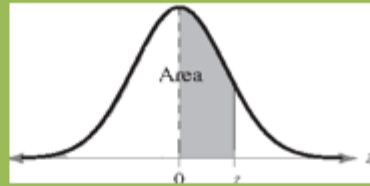
1. What is the unit of analysis? Who or what is being studied?
2. What data is being collected?
 Name of independent (exposure) variable:
 Name of dependent (outcome) variable:
3. Answer the research question using the information provided. First fill in the contingency table to the left - then calculate RR and answer the research question.

		Yes		No	Total	Answer to RQ
		Yes	No	Total		
Injured	Yes					
	No					
Total						

Calculate RR

Appendix 1: z-table

Z table Area between mean and z-score.
 For example $z = 0.72$ corresponds to area = 0.2642
 (or 26.42% of total area).



Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995