Math 1112
Unit 5: Reason with Quantities

5.1 Quantification of Population health:

1. Incidence/prevalence & other miscellaneous rates
2. Exercises with population based rates
3. continued
4. continued

5.2 The Nature of Data

5. Recognize quantitative data
6. Quantitative data vs variable
7. Categorical vs measurement data/variable
8. Recognize one vs two variable scenarios

5.3-5.4 Read graphs and charts (most of this work to be done in HNP webapp):

9. Read Histogram
10. Read Pie chart

5.5 Summarizing data (sets of data) with statistic al tools

11. Using a Few Numbers to Represent Many (notation)
12. Using a Few Numbers to Represent Many (centre and spread)
13. Exercises (continued)
14. Percentiles

Quantification of Population health below you will find the formulas for a series of statistics (all type III) which are used to gauge the health of a population with regards to a particular disease. The list is not exhaustive, i.e. there are many statistics calculated by epidemiologists that are not on the list. Note that most of these are presented usually 'per 100,000 population'

Incidence Rate: Refers to the number of *newly reported* cases of a disease in a specified time period divided by the population at that time.

$$\frac{\text{Total number of newly reported cases of a disease for a period}}{\text{Total number of population for the period}}$$

Prevalence Rate: Refers to the number of *existing* cases of a disease in a specified time period divided by the population at that time. This rate is very complicated to compute realistically for a period of time of more than one day since people die from diseases and get diagnosed with diseases every day.

$$\frac{\text{Total number of existing cases of a disease for a period}}{\text{Total number of population for the period}}$$

Case Fatality Rate: Refers to the rate of death from a disease; i.e. are a large proportion of sick people dying from the disease – how serious is it?

$$\frac{\text{Number of deaths due to specific disease in a time period}}{\text{Number of diagnosed cases during the same time period}}$$

Specific Mortality Rates: rate of death in a specific population: could be geographic (e.g. mortality rate in Canada, or disease specific – e.g. cancer mortality). The disease specific mortality rate is the same as Case fatality.

$$\frac{\text{Number of deaths a time period}}{\text{Total population at risk}}$$

Fertility Rates:

$$\text{Crude Birthrate} = \frac{\text{Number of live births}}{\text{Total number in population}}$$

$$\text{General fertility rate} = \frac{\text{Number of live births}}{\text{Number of women aged 15-44 in population}}$$

Other Rates: When you want to calculate something and don't have a specific formula, you can use the "other rates" formula.

$$\frac{\text{Number of times something occurred}}{\text{Number of times something could have occurred}}$$

Exercises: Population based Rates

Exercise 1: Population of Port Perry in the mid-year of 2005 = 16, 000
The number of new cases of cancer between January and December 1st 2005 = 8
The number of cases of cancer according to the registry on December 1st 2005 = 75
Calculate:

a. The incidence rate of new cases of cancer for the time Jan.-Dec. 1st 2005.

b. The prevalence rate of cancer cases as of Dec. 1st 2005.

Exercise 2: In 2008 the population of Oshawa was 100, 000 people.
There were 3 new cases of stomach cancer reported in 2008.
The number of stomach cancer cases according to the registry on December 31st 2008 was 145.
Calculate:

a. The incidence rate of new cases of stomach cancer for 2008.

b. The prevalence of stomach cancer cases as of December 31st 008.

Exercise 3: Population of Elizabethville in 2007 was 1200 people.
The number of new cases of cervix cancer reported was 2 in 2007.
The number of cervix cancer cases according to the registry at the end of 2007 was 15.
Calculate:

a. The incidence rate of new cases of cervix cancer for 2007.

b. The prevalence rate of cervix cancer cases at the end of 2007.

Exercise 4: In 2004, there were 10 new cases of liver cancer reported in NFLD for males. Given that there were 260,000 males in NFLD in 2004, calculate the incidence rate based on 100, 000 people.

Exercises: Population based Rates - continued

Exercise 5: In 2008, there were 5 new cases of larynx cancer reported for males in PEI. The population in 2008 for PEI was 250,000. Calculate the incidence rate for 2008 per 100 000 population.

Exercise 6: In 2007, according to the cancer registry, there were 25 cases of breast cancer in Ontario. The population at the time was 257, 095. Calculate the prevalence rate per 100 000 population.

Exercise 7: In 2005, according to the cancer registry, there were 38 cases of lung cancer in Alberta. The population at the time was 224, 590. Calculate the prevalence rate per 100 000 population.

Exercise 8: In 2003, there were 38 deaths reported due to lung cancer. According to the cancer registry there were 25, 897 cases of lung cancer. Calculate the case fatality rate per 100 000 population.

Exercise 9: 12, 356 people have a life threatening illness known as the HIM flu. 50 people die from this disease within 1 year. What is the case-fatality rate per 100 000 population, as a result of the HIM flu?

Exercise 10: 20 people are sick with Exam disease in 1 year, 11 people die from this disease. What is the case-fatality rate as a result of Exam disease? Calculate per 100 000 and per 1000 – which makes more sense to report?

Exercises: continued

Exercise 11: Hospital 1 has 978 beds of which 888 are occupied. Hospital 2 has 274 beds. How many of beds in hospital 2 would have to be occupied in order to have an equivalent bed occupancy rate?

Exercise 12: The prevalence rate of Multiple Sclerosis (MS) in Canada is 240 per 100 000. Given that the population of Canada is 33,739,900. How many people in Canada have MS?

Exercise 13: There are 756 existing cases in a population of 345,000. Find prevalence per 100,000.

Exercise 14: There are 12 new cases in a population of 145,000,000. Find incidence per 100,000

Exercise 15: There are 690,442 existing cases in a population of 145,000,000. Find prevalence per 100,000.

Recognize Quantitative Data

In unit 4 you learned that there are 2 types of quantities: counts of concrete objects (10,000 protesters, one and a half onions in the fridge) vs counts of units of measure of one object (my head circumference is 59 cm).

These characteristics of quantities can be used to decide whether claims about the effectiveness of treatments, or other health related questions require the collection of quantitative data.

Claims such about the effectiveness of medications require quantitative data since these claims need to be verified through experiments in which 1 group gets a treatment and the other gets a placebo, or another treatment. A count of improved (% of those improved), or another measure of effectiveness (e.g. pain score reduction) are used to compare the 2 groups indicate that the data collected must be quantitative in nature.

Information (evidence?) about questions like 'What is it like to be homeless in Toronto?' or How does it feel to win a million dollars? are more difficult to quantify. This kind of data is usually collected through longer stories, focus groups.

Exercise 1: Would quantitative data Be appropriate to investigate the truth of each claim below. Justify your choice with a description of the data (as count of concrete vs abstract objects) you suggest be collected.

a. Canadians are willing to sacrifice some of their comforts to help reduce climate change.

b. Most Canadian families played board games to preserve their mental health during the early weeks of the Covid-19 pandemic.

c. Echinacea is not effective in reducing the duration of the common cold.

d. Residents of Canada more likely to be tea drinkers than residents of France.

e. Electric cars will improve the lives of the poor.

f. Residents of Canada drink more tea per day than residents of France.

Quantitative Data vs Variable: A variable is the name of the characteristic of interest for an object being studied, while data is the information about the characteristic that has been (or will be) collected.

Exercise 1: Is the following a good variable name for the data that has been collected? If you think it is a good name describe what makes it so, if not, come up with a better one.

a. 50 students were asked the extent to which they were satisfied with the mathematics course they were taking. They were offered the following options: very satisfied, somewhat satisfied, neutral, unsatisfied.
Variable name 'Satisfaction level'.

b. The heart rate of 50 students after walking up 3 flights of stairs at Waterfront Campus was taken just when they got to the top.
Variable name 'heart rate'.

c. Researchers were interested in what portion of students preferred online studies after a year of enforced online learning due to the pandemic.
Variable name 'rate of preference'.

d. Researchers were interested in the mean GPA of students studying online so that they could compare to previous years.
Variable name 'mean GPA'

Categorical vs Measurement Data: Once data is collected it will typically be entered as a set of numbers, one row or column for each variable. The numbers can be measurements of the objects being studied (head circumference = 45cm), or codes representing categories (like shirt size: small = 1, medium=2, large=3, etc) } Thus, we define two types of quantitative data: measurement and categorical.

Exercise 1: For each of the data sets pictured decide whether the data is categorical or measurement.

a.

Total number of cigarettes smoked per month
4
5
2
9
5
3
1
6
3
5

Data Type_____

b.

Caffeine_mg
136.00
.00
170.00
204.00
136.00
255.00
238.00
272.00

Data Type_____

c.

selfhealth
poor health
poor health
ok
ok
very healthy
very healthy
poor health

Data Type_____

d.

TwoDays OfStudy
no
yes
yes
no
yes
no
no
no
yes

Data Type_____

e.

treatmentgroup
cryotherapy
cryotherapy
cryotherapy
cryotherapy
duct tape
duct tape
duct tape
duct tape
duct tape
duct tape

Data Type_____

f.

timestudying
10.00
7.00
15.00
25.00
15.00
15.00
15.00
10.00
15.00
25.00

Data Type_____

Recognize one vs two variable scenarios: There is very little research that is done or reported where only one variable is being studied especially in health research where comparing rates, comparing means and correlations are the building blocks to understanding how to prevent and treat diseases. In the exercises below you will practice thinking slowly and deciding whether the scenario description is based on a single variable, or on two variables.

Exercise 1: Do the following research questions require data to be collected on one or two variables? Name the variable(s) and whether they are categorical or measurement.

a. Distribution of head circumference in Canadian adult males.

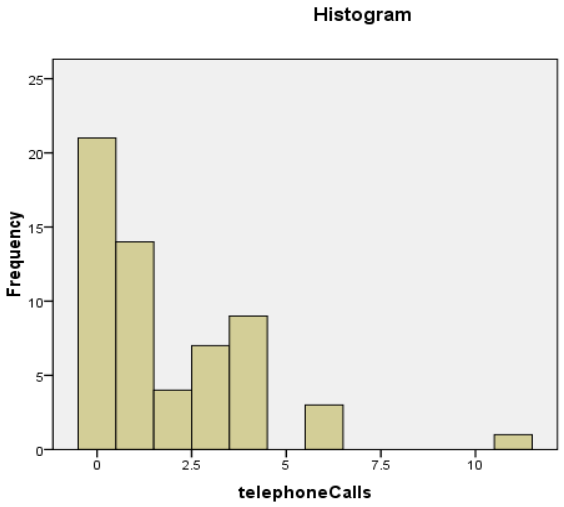
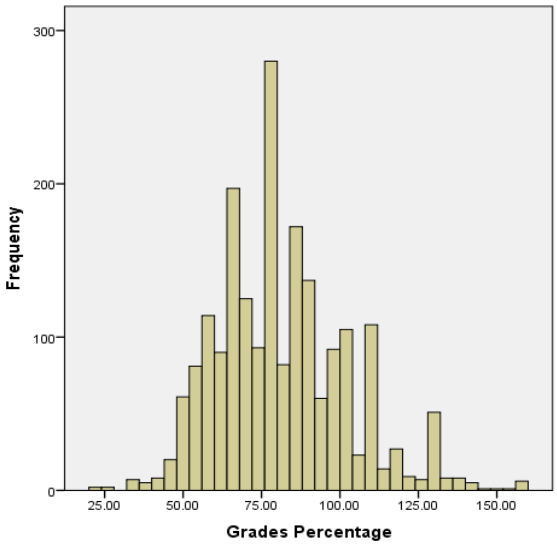
b. Comparison of rates of bicycle ownership in urban vs rural teens in Ontario.

c. Change in GPA scores of HIM students before to after online due to covid-19 pandemic.

d. Does number of minutes student exercises per week predict their GPA score?

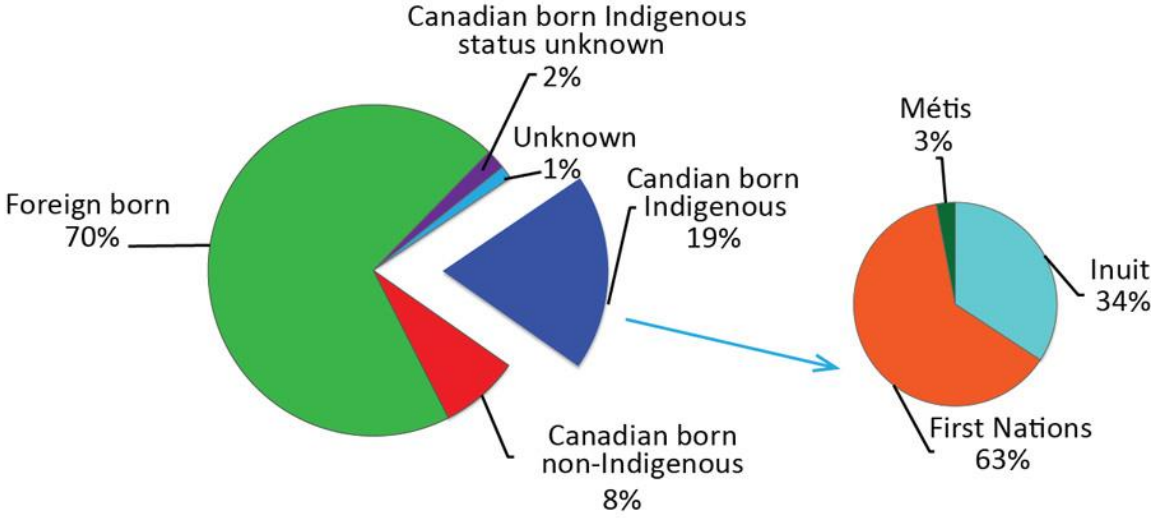
e. Do hospitals with bed occupancy over 99% have a higher rate of doctor burnout?

Read Histogram: Take a look at the histograms reproduced below and 'read' them carefully. Answer all questions below based on what you see in the histograms and the titles given.

<table>
<tr><td>1a. Number of telephone calls made by nurses to a group of 59 discharged patients.
 
</td><td>1b. Teachers marks (in %) for a class of 2000 physics students.
 
</td></tr>
<tr><td>What is the typical number of calls made? Explain your reasoning.</td><td>What is the typical mark a student got in this rather large class?</td></tr>
<tr><td>Describe the distribution of #of telephone calls made by nurses by highlighting information that you found interesting or that may be important to the researcher.</td><td>Describe the distribution of marks for this class and highlight information that you found interesting or that may be important to the researcher.</td></tr>
<tr><td colspan="2">Describe 3 differences that you noticed between the two histograms.</td></tr>
</table>

Read Pie Graph: Pie charts are used for categorical data. Take a look at the chart below and think about the story that it is trying to tell. This is a tricky one to read. Make sure to play with 'graphs and charts' in the HNP web-app for more practice.

Distribution of active tuberculosis cases (new and re-treatment) by patient origin, Canada, 2016



Source: https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2018-44/issue-3-4-march-1-2018/article-1-tuberculosis-2016.html

a. What are the categories of 'patient origin' used for this variable? How many are there?

b. Write out a sentence or two indicating what the 70% (foreign born) refers to.

c. Write out a statement about the second – smaller - pie chart. The categories represent the 3 sub-classifications of indigenous communities in Canada.

d. For the Canadian born non-Indigenous category is the 8% value low, high or just about what would be expected? Explain your answer.

e. Does the graph indicate that there is a tuberculosis crisis in Canada? Use numbers to explain why or why not?

Working with data – using few numbers to represent many (notation)

When we have data in front of us (see data set 1a and 1b below), this is usually an indication that we are trying to learn something about a group of individuals by collecting characteristics of some or all of the individuals in the group. (could be humans, mice, cars or other objects)

Looking at a data set is usually not much fun, and it is pretty hard to describe the group by talking about the individuals in the group. Mathematicians have attempted to capture the characteristics of the group by devising various numerical and graphical tools.

Numerical tools: Measures of centre (mean, median & mode) variability/spread (standard deviation & range) and position (min/max & percentiles). Percentiles will be dealt with separately.

Examine Data sets 1a & 1b below – same data, but 1b has been ordered and nicely presented.

Data set 1a: math 1112 marks: {*This set of marks looks chaotic – it is hard to say anything about them*.}

66 45 37 52 44 55 75 93 36 91 61 83 61 45 48 94 83 80 00 62 97 71 72 88 67 61 76 61 54 61 59 47 82 77 55 40 57 65 84 49 65 63 68 70 88 57 21 66 75 33 70 50 44 73 86 70 58

Data set 1b (ordered): distribution of math 1112 marks (low to high): {*much easier on the eye!*}

0	21	33	36	37	40	44	44	45	45	47	48	49
50	52	54	55	55	57	57	58	59	61	61	61	61
61	62	63	65	65	66	66	67	68	70	70	70	71
72	73	75	75	76	77	80	82	83	83	84	86	88
88	91	93	94	97								

Notation: Before getting to the calculations of centre and spread it is important that you get comfortable with algebraic notation that is used to represent elements of a data set.

N – number of individuals in the data set. In data set 1b there are 57 individuals {$N = 57$}

Max: highest value {*max = 97- Easy to find in ordered data set};* Min: lowest value {*min = 0}*

x_i : points to the i^{th} individual element. {*in 1b $x_7 = 44$. It is the 7^{th} data point starting from 0 as the first. }*

$\sum_{i=1}^{n} x_i$ - the sum of elements 'x_i' from $i = 1$ to $i = n$ {*when n = 3 this means the sum of first 3 elements* $\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 0+21+33=54$}; sum of all elements= $\sum_{i=1}^{57} x_i = 3591$

Notation - exercises:

1. Using the ordered data set of math 1112 marks 1b find the following:
   a. x_4 b. x_{12} c. x_1 d. x_{56} e. x_0

2. Using the ordered data set of math 1112 marks 1b calculate
   a. $\sum_{i=3}^{5} x_i$ b. $\sum_{i=51}^{57} x_i$ c. $(\sum_{i=1}^{5} x_i)$ d. $\sum_{i=1}^{4} (x_i - 2)^2$

Working with data – using few numbers to represent many (calculations)

Measures of Centre of distribution:

Median: a measure of centre of the distribution. Once the numbers are in order look at the middle number. {*The median of the class marks (i.e. 1b) is x_{29} = 63.*}

Mean ($\bar{x}$ *or* μ): a measure of centre of the distribution $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ The sum of all the numbers in the data set divided by N {*sum of all terms is* $\sum_{i=1}^{57} x_i$ *=3591, mean = 3591/57 = 63*}

Mode: appears the most often in the data set – {*look through the ordered data set (1b) and find that 70 is repeated 3 times, but 61 is repeated 5 times, so 61 is the mode*}

Measures of variability (spread) in distribution:

Range: a measure of spread = max – min {*in the data set 1b the range = 97 – 0 = 97.* }

Standard deviation (std dev): a measure of the spread of distribution $\sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{354.1071} =$ 18.82. Below I will go through a calculation where N is not so large

Example of calculation : Data set Q = { 4, 6, 6, 8}; $\bar{x} = 6$,

$$\text{std dev} = \sqrt{\frac{(4-6)^2+(6-6)^2+(6-6)^2+(8-6)^2}{4}} = \sqrt{2} = 1.414$$

Exercises:

1. Below is a distribution of heights of the Gula family (in cms). Find N, min, max, x_4, median, mode, mean, range, and std.dev. {82 144 154 156 167 171 174 188 }

2. Below is a distribution of heights of the Matvienko family (in cms). Find N, min, max, x_4, median, mode, mean, range, and std.dev. {75 89 111 144 178 179 186 }

Working with data: exercises (continued)

3. Below is a distribution of heights of the Chalupa family (in cms). Find N, min, max, x_4, median, mode, mean, range, and std.dev. {155 166 177 62 128 76 166 }

4. Below is a distribution of heights of the Zinger family (in cms). Find N, min, max, x_4, median, mode, mean, range, and std.dev. {182 44 115 136 67 71 117 201 188 144}

5. Which family has the highest mean height? How much more is it than the shortest mean height as raw difference (in cms) and as a % difference.

6. Which family has the highest median height? How much more is it than the shortest mean height as raw difference (in cms) and as a % difference.

7. Which family is the tallest? Which numbers would you use as evidence for your answer?

8. Which family has the most variability in heights? Which numbers would you use as evidence for your answer?

Percentiles

In the previous pages you practiced thinking about data 'generally', using measurements of characteristics (centre, spread) usually through calculations. In this section we will examine the distribution of data set 1b by using measures of position – i.e. looking at where a particular value is in relation to the others as a percentile.

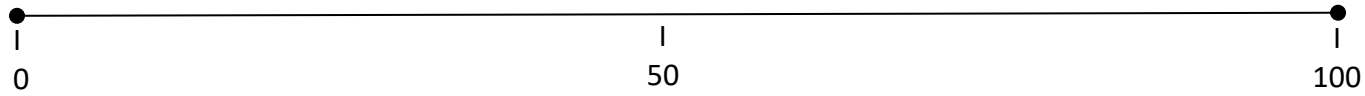
The median (a measure of the physical centre of a distribution) is also the 50^{th} percentile in a distribution. This means that 50% of the data is above the median and 50% is below.

Data set 1b (ordered): HIM math 1112 marks ordered lowest to highest: {*same as on page 19*}

0	21	33	36	37	40	44	44	45	45	47	48	49
50	52	54	55	55	57	57	58	59	61	61	61	61
61	62	63	65	65	66	66	67	68	70	70	70	71
72	73	75	75	76	77	80	82	83	83	84	86	88
88	91	93	94	97								

Below I have reformatted the data set to fit on one line from 0 to 100 representing %.

0 21 33 36 37 40 44 44 45 45 47 48 49 5052 54 55 55 57 57 58 59 61 61 61 61 61 62 **63** 65 65 66 66 67 68 70 70 70 71 72 73 75 75 76 77 80 82 83 83 84 86 88 88 91 93 94 **97**



You can see that the 10^{th} percentile (with 10% of the data below it) is approximately 37, and the 50^{th} percentile is 63 (the median). Approximately 50% of the students had a mark higher than 63, and 50% had a mark lower than 50%.

Exercises:

1. without doing any calculations, use the number line look for (estimate) the following percentiles in data set 1b.

a. 25^{th} b. 75^{th} c. 90^{th} d. 95^{th}

2 a. If you got a mark of 73 you would be (approximately) in what percentile of the class marks?

b. If you got a mark of 73, approximately what % of the class had a higher mark than you?

c. If you got a mark of 44, approximately what % of the class had a mark lower than yours?