9.1 Statistical Surveys —Intro and review of grades 7,8

Population: the complete set of objects or persons that are being considered in a survey

Sample: a subset of the population.

Variables (data): What is being surveyed. (annual income, eye colour, age, height, shoe size…)

- Quantitative variables: Ones that can be measured numerically. Ex: Age, height, weight.

  i) Discrete variables: Variables that have separate values (often integer values)

  (# of goals in a hockey game, Shoe size 9, 9 ½ , 10…)

  ii) Continuous variable: Ones that can be any real number within an interval

  (time it takes to run 100m)

- Qualitative variables: Ones that don't have a numerical value. Ex: Eye colour

Census: Statistical survey where **all** of the population is being surveyed.

Canada has a census every ten years…. Every person in Canada is counted.

Poll: Survey where a **sample** is studied to infer information about the population being studied.

A group of students at a university are asked which night of the week they party the most often

Study: A statistical survey where **experts** in the field being studied are surveyed

("4 out of 5 Dentists prefer X brand of toothpaste" is a study.)

Bias: Any error involved in a statistical survey. There are several sources… basically… why is it not a good survey….

Types of Bias:

- Choosing a sample: Don't poll a Seniors retirement home about what activities Senior Citizens partake in. (Seniors in a home are probably less active than those who still live at home)
- Bad Questions: Vague or misleading such as, "Are you a heavy drinker?" Define Heavy.
- Non Random Sampling: Time and place. Asking random spectators at the Olympic Games in Vancouver if they smoke. (perhaps people interested in sports are less likely to smoke)
- Errors in processing data: tabulating results can lead to an error.
- Errors in Analysis of data: Making assumptions, (like discounting undecided voters before an election … maybe they are decided… just don't want to tell you)

Look over Activity 1 on P 256. Check your answers.

Practice: P 257 # 1-4, P 259 # 5

9.2 Statistical Tables and Diagrams

Diagrams may include:

1. Bar Graphs
2. Pie charts/circle graphs
3. Broken line graphs
4. Histogram
5. Box and whiskers plots

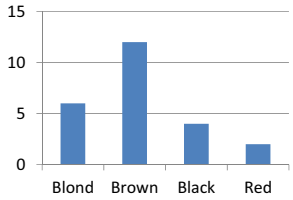
Tables may include:

1. Condensed frequency tables
2. Relative frequency tables
3. Grouped data tables

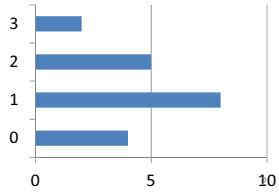
1. Bar graphs – studied in grade 6/7

For representing qualitative data or discrete quantitative data

color of hair	frequency
Blond	6
Brown	12
Black	4
Red	2



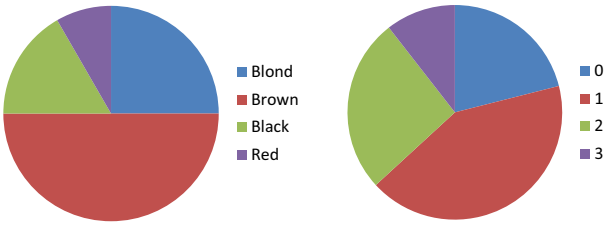
# of pets	frequency
0	4
1	8
2	5
3	2



2. Pie/circle Charts – studied in grade 7/8

For representing qualitative data or discrete quantitative data

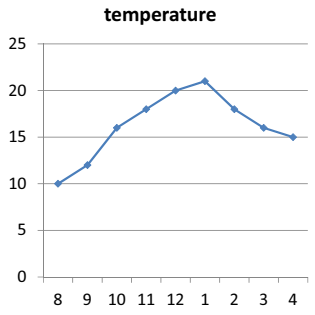
We represent the percentage of each category as a percentage relative to the total. (part of the whole)



3. Broken line graph - studied in grade 6

For representing data that continually changes over time.

hour	temperature
8	10
9	12
10	16
11	18
12	20
1	21
2	18
3	16
4	15



4. Histograms

For representing quantitative data grouped in classes

But first we need to look at different types of tables

- **Sometimes we will want to group raw data into classes. We will do this when there is a large amount of raw data and if the numbers are distinct (very few repeating data values).**
- **Each class will be defined by an interval such as: [0 , 10 [then [10,20[… etc.**
- **This is called a grouped data table.**

4. Histograms

Ex: Consider the marks that 15 students got on a quiz

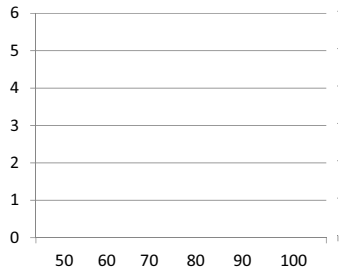
Raw data: **63, 72, 58, 80, 67, 63, 58, 92, 80, 80, 72, 76, 63, 63, 92**

Marks	Tally	Frequency	Relative frequency
[50,60[			
[60,70[			
[70,80[			
[80,90[			
[90,100]			
Total	**15**	**15**	**100**

Now we can draw the histogram for the marks of the 15 students

It looks almost like the bar graph, but the intervals are connected

Marks	Frequency
[50,60[	2
[60,70[	5
[70,80[	3
[80,90[	3
[90,100]	2
Total	15



Practice: Page 266 # 1,2

9.3 Sampling Techniques

When conducting a survey, we can gather the data by:

1. **Direct observation** (record events or behaviours of people)
2. **Face to face interview** (in a public place or shopping center)
3. **Telephone interview** (usually faster, efficient)
4. **Written questionnaire** (to be filled out and returned)
5. **Documentary observation** (pull from existing data base)
6. **Mechanical or electronic instrumentation** (scanner in supermarkets, electronic counters)

To choose the sample of our population:

You have already seen 2 Techniques in grades 7/8 :

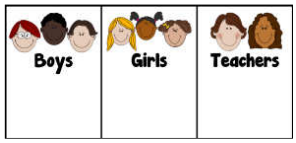
1. **Simple Random Sampling**: Method where we randomly choose the individuals belonging to the sample. (Choose numbers out of a hat)

2. **Systematic Random Sampling**: Method where we choose a starting individual and then choose every n^{th} individual after that. (Inspecting every 100^{th} TV on an assembly line for quality control)

There are two other techniques:

3. Stratified Sampling: we use it when the population is divided into subgroups called Strata (by personalities, sex, interests, grade level, age groups)

We want all proportions of the population to be represented in the sample.



Ex 1: A small school with 400 students has 160 grade 9's, 140 grade 10's, and 100 grade 11's .

A sample of 60 students are to be chosen for a survey. How many of each grade should be included?

Ex 2: The following table shows the distribution of the 1200 students in a school.

	# of girls	# of boys
First cycle	360	345
Second cycle	240	255

A sample of 180 students is required, it must be representative of the population. How many girls from the second cycle should be in the sample?

Another sampling technique:

4. Cluster sampling: When the population is made up of several similar clusters where there is a lot of variation inside each cluster. (such as grade 9 homerooms)

We then randomly choose some clusters.

Each individual in the cluster is surveyed.

Practice Page 272 # 1-3

9.5 Measures of Central Tendency

Measures that tend to be close to the center of the data are:

1. **Mode**
2. **Median**
3. **Mean (average)**

1. Mode (M_o): the most frequent value in the data.

Ex 1 (Raw data):

22, 24, 24, 25, 26, 26, 26, 31, 32

Ex 2: Mode is the only measure of central tendency that we can use for qualitative variables.

Hair color	Frequency
Red	1
Blond	5
Brown	7
Black	3

1. Mode (M_o): the most frequent value in the data.

Ex 3 (Grouped data):

Age	Frequency
[10,14[	7
[14,18[	4
[18,22[	9
[22,26[	4

2. Median (M_d): the middle value in an ordered group of data.

Case 1: n(number of data entries) is odd:

22, 24, 24, 25, 26, 28, 29, 31, 32

2. Median (M_d): the middle value in an ordered group of data.

Case 2: n(number of data entries) is even:

22, 24, 24, 25, 28, 29, 31, 32

2. Median (M_d): the middle value in an ordered group of data.

Ex 3:

Value	Frequency
2	10
4	6
5	2
6	9
9	5
Total	32

Ex 4:

Value	Frequency
20	2
25	8
30	4
40	9
50	6
Total	29

2. Median (M_d): the middle value in an ordered group of data.

Ex 5: (Grouped data)

Class of marks	Frequency
[0,20[	4
[20,40[	8
[40,60[	10
[60,80[	7
[80,100]	1
Total	30

3. Mean ($\overline{x}$): /average the sum of the data divided by the number of data.

Ex 1: (Raw data) find the mean of

21, 31, 24, 26, 32, 42, 25

3. Mean ($\overline{x}$): /average the sum of the data divided by the number of data.

Ex 2: (Frequency table) find the mean age

Age	Frequency	Total age
12	2	2(12)=24
13	4	4(13)=52
14	3	3(14)=42
Total	9	118

3. Mean ($\overline{x}$): /average the sum of the data divided by the number of data.

Ex 3: (Grouped data) find the mean height

Height	Frequency	Midpoint	Total height
[100,110[	8	105	8(105)=840
[110,120[	2	115	2(115)=230
[120,130[	7	125	7(125)=875
Total	17		1945

3. Mean ($\overline{x}$): /average the sum of the data divided by the number of data.

Ex 4: (Weighted data) find John's mean mark

Term	Weighting	John's mark
1	0.14	65%
2	0.14	70%
Mid exam	0.075	70%
3	0.285	85%
Final exam	.36	63%

Practice: page 281 # 1-4
page 283 # 5,6
page 284 # 7

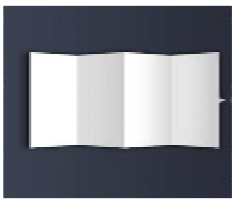
9.6 Measures of Position: Quartiles

How do you fold a paper into 4 equal parts?

---------- Q_1 ---------- Q_2 ---------- Q_3 ----------

25% of data | 25% of data | 25% of data | 25% of data

Q_2 = Median



Quartiles Q_1, Q_2, Q_3, divide a set of ordered data into four groups with an equal amount of data in each.

Ex 1: Consider the ordered set of n = 13 data

1 4 7 8 9 9 11 16 17 19 25 30 30

Q_2 =

Q_1 =

Q_3 =

Q3 =

Ex 2: Consider the following frequency table for the number of pets that students have.

# of pets	frequency
0	4
1	8
2	5
3	2

n =

M_o =

$\bar{x}$ =

$M_d = Q_2$ =

Q_1 =

Q_3 =

Ex 3: (p286#4) In a class of 40 students, if the 1st quartile is 64, the median is 70 and the 3rd quartile is 78. What is the maximum number of students that have a mark **less** than you if you got…

a) 62% b) 69% c) 76%

We recreate the data:

1) 40 students means 10 in each quarter

2) Even n=40 means none of the medians is an actual mark

Practice: page 286 # 1, 2, 3

9.7 Measures of Dispersion

Variation Interval : The interval with the lowest and highest data: $[X_{min}, X_{max}]$

Interquartile Interval: The interval $[Q_1, Q_3]$
50% of the data lies in this interval

Range (R): The difference between the highest and lowest value. $R = X_{max} - X_{min}$

Interquartile Range (I): The difference between Q_3 and Q_1 $I = Q_3 - Q_1$

Note that **range** and **interquartile range** refer to single numerical values

Ex 1: A group of 11 friends are playing a game of bowling. Here are their scores

123 99 139 100 88 86 133 100 153 112 93

We first need to rearrange the data

86 88 93 99 100 100 112 123 133 139 153

n = min = max =

Q_1 = Q_2 = Q_3 =

Variation interval = R =

Interquartile interval = I =

Which score is less than the median but more than Q1?

Ex 2 – Ruler Reaction Time (cm)

Boys

6 9 9 9 11 12 14 14 14 14 15 15 17

Girls

9 9 10 11 11 11 12 13 14 14 15 15 18

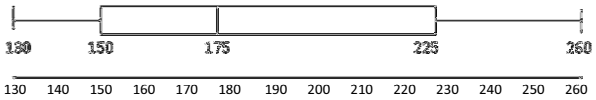
Mean Mode

Median Range

Practice:
page 288 # 1-4

9.8 Box and Whisker Plots

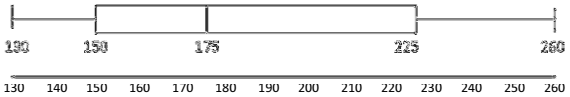
Box and Whisker Plots are a visual way to look at Quartiles.



The Box and Whisker plot gives us information on the concentration and dispersion of the data.

There is exactly the same number of data entries in each section (25%) but they may be spread out (dispersed) in some sections more than others.

Ex 1: Weights of athletes on a men's' rowing team.



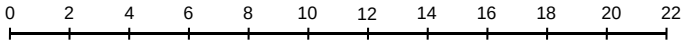
X_{min} = 130 Q_1= 150 Q_2 = 175 Q_3 = 225 X_{max} = 260

- The rectangle = Interquartile interval (where 50% of the data lies… 50% of the rowers are between 150 and 225 lbs
- Vertical lines = Hinges = Q_1, Q_2, Q_3
- Left whisker= distance between X_{min} and Q_1 (where the lower 25% of the data lies.. between 130 and 150 lbs)
- Right whisker= distance between Q_3 and X_{max} (where the upper 25% of the data lies.. between 225 and 260 lbs)

Case 1 - Odd # of Items

Ex 2: # of Visits to Mrs. Botros website/ day

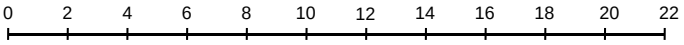
3 5 8 // 9 12 14 16 17 17 18 // 19 20 20



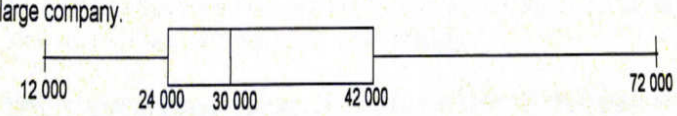
Case 2 - Even # of items

Ex 3: Number of _____/ day

1 1 2 // 4 5 7 // 9 10 11 // 14 16 17



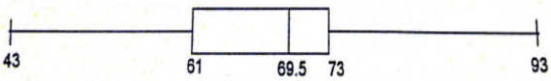
1. The following box-and-whisker plot represents the annual salary of employees in a large company.



Which of the following statements is definitely TRUE?

A) The mean salary is $30 000.
B) Half the employees earn between $24 000 and $42 000.
C) The salaries are more densely concentrated between the second and third quartiles.
D) There are more employees who earn over $42 000 than employees who earn under $24 000.

2. The marks on a history test for a class of 30 students are represented by the following box-and-whisker plot.



The following are the marks for 26 of the 30 students:

43	54	55	56	58	60	61	65	66	66	66	67	68
70	71	71	71	71	71	71	74	77	77	79	86	88

Find the four missing marks. Explain why you chose each mark.

Practice P. 291 # 1-5