Unit 8

L1 Introduction to Statistical Surveys

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 color, age, height, shoe size…)

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

1

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 get 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.)

2

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)
- 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)
- Bad Questions: Vague or misleading such as, "Are you a heavy drinker?" Define Heavy.
- 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)

3

Where and how can we gather the data?

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)

4

How and who to choose for a sample population?

1. **Simple Random Sampling**: Randomly choose the individuals (Choose names out of a hat)
2. **Systematic Random Sampling**: Choose a starting individual and then choose every n^{th} one after that. (Inspecting every 100^{th} TV on an assembly line for quality control)
3. **Cluster Sampling**: The population is made up of several similar clusters with a lot of variation within each cluster (ex: gr.9 homeroom), we randomly choose some clusters to survey.
4. **Stratified Sampling**: The population is divided into subgroups (strata) by personalities, gender, interests, grade level, age etc. We want to represent all proportions of the population in our sample.

5

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?

Practice:
8.1 # 1-4

8

Unit 8

L2 Qualitative Data

Tables, Diagrams & Analysis

Data may be:

1. Raw data
2. Tally & Frequency table
3. Relative frequency tables
4. **Grouped data tables**

Diagrams may include:

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

To analyze the data we may calculate:

Measures that stay close to the center of the ordered data:

1. **Mode (Mo):** the most frequent value in the data.
2. **Median (Md):** the middle value in the ordered data.
3. **Mean (average) ($\bar{x}$):** the sum of the data divided by the number of data entries.

Measures that show the dispersion of the ordered data:

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

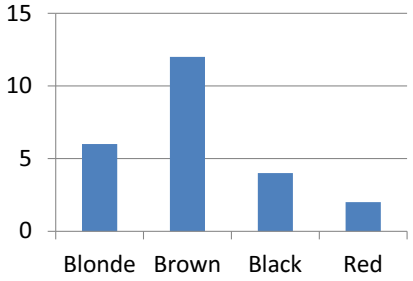
Measures that show position relative to the ordered data:

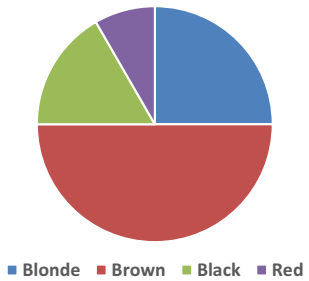
5. **Quartiles (Q1,Q2,Q3),** Quintiles, percentiles.

Qualitative data

Can be represented by a bar graph or a pie chart

color of hair	frequency	Relative Freq.	Angle
Blonde	6		
Brown	12		
Black	4		
Red	2		
Total	24		





Only Mode can be found:

Mo: Brown

Weighted Mean ($\bar{x}$): the average when some values have more weight than others

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: 8.2 # 1, 2and 8.6 # 1-7

Unit 8

L3 Discrete Quantitative Data

Discrete quantitative data is numerical data with values that are distinct and specific (often integers)

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

Ex 1 (Raw data):

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

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

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

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

4 entries 4 entries

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

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

Ex 5: (Raw data) find the mean of

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

n = 7

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

Ex 6: (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

4. Range (R): The difference between the highest and lowest value.

$R = X_{max} - X_{min}$

Ex 7 – 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

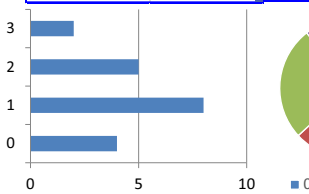
Mean Mode

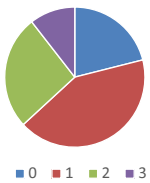
Median Range

Discrete Quantitative Data

Can be represented by a bar graph or a pie chart

# of pets	frequency	Relative Freq.	Angle	Total
0	4			
1	8			
2	5			
3	2			
Total	19			





Mo:

Md:

$\bar{x}$:

R :

Practice:

8.3 #1,2

Unit 8

L4 Continuous Quantitative Data

Continuous quantitative data is numerical data that can be any real number.

- **When there is a large amount of raw data and the numbers are distinct (very few repeating data values). We can't do a frequency table, instead we group the data into classes.**
- **Each class will be defined by an interval such as: [0,10[then [10,20[… etc.**
- **All classes should have equal amplitude (length) and leave no gaps.**
- **This is called a grouped data table.**

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

Ex 1:

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.

Ex 2:

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

n = 30 (even) so M_d is average of 15th and 16th entries

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

Ex 3: 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

Ex 4: Consider the marks that 15 students got on a quiz

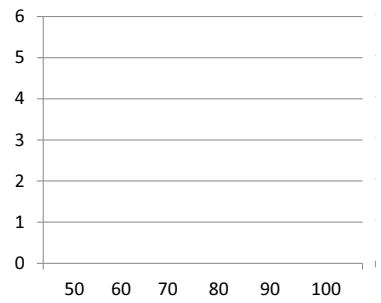
Raw data: ~~63, 72, 58, 80, 67, 63, 58, 92, 80, 80, 72, 76, 63, 63, 92~~

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

Continuous data can be represented by a **Histogram**

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



8.4 #1,2

Unit 8

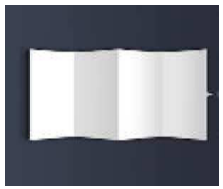
L5 Box and Whiskers plot

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.

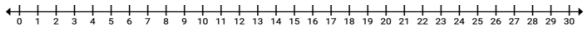
Ex 1: Consider the ordered set of n = 13 data

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

Q_2 =

Q_1 =

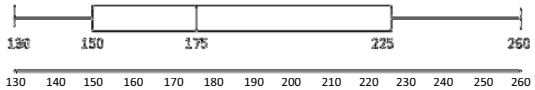
Q_3 =



A Box and Whiskers Plot:

- Shows only the min, Q1, Q2, Q3, and max.
- Vertical lines at Q_1, Q_2, Q_3 (Hinges) form a box with 50% of the data, lines to min and max are called whiskers
- There is the same number of data entries (25%) in each section, but they may be spread out (dispersed) in some sections more than others.
- Gives information on the concentration and dispersion of the data.

Ex 2: Weights of athletes on a men's' rowing team.



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

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 3: 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

n = min = max =

Q_1 = Q_2 = Q_3 =

Variation interval = R =

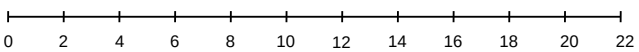
Interquartile interval = I =

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

Case 1 - Odd # of Items

Ex 4: # of Visits to Mrs. Botros website/ day

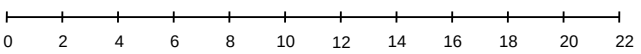
3 5 8 9 12 14 16 17 17 18 19 20 20



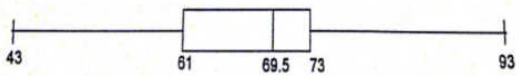
Case 2 - Even # of items

Ex 5: Number of _____ / day

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



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:
8.5 1-5