T6.1 – Introduction to Statistics

IB Math SL1 - Santowski

Example to start with

- Given the following data set for 3 of my shot put athletes, determine their:
- (a) mean
- (b) median
- (c) mode
- (d) inter-quartile range
- (e) range
- Prepare a histograph of the data

Thrower 1	Thrower 2	Thrower 3
8.74 m	10.39 m	8.79 m
8.94 m	10.86 m	9.39 m
9.66 m	10.94 m	9.94 m
10.01 m	9.00 m	10.97 m
10.01 m	9.15 m	9.72 m
8.43 m	9.35 m	8.49 m
10.25 m	9.35 m	9.63 m
10.14 m	8.45 m	9.83 m
9.04 m	8.85 m	9.49 m
9.30 m	8.95 m	8.82 m
8.69 m	9.10 m	9.24 m
8.85 m	10.20 m	9.13 m

Statistics - Definition

- The scientific study of numerical data based on variation in nature.
- A set of procedures and rules for reducing large masses of data into manageable proportions allowing us to draw conclusions from those data.
- Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

(A) Review of Key Terms

- Measurement – assignment of a number to something
- Data – collection of measurements
- Population – all possible data
- Sample – collected data

(A) Review of Key Terms - Variables

- The variables are the characteristics or information collected about each individual in the study, i.e. what was measured
- Continuous Variable ➔ data that can take on ANY value between a minimum value and a maximum value
- Discrete Variable ➔ data result when the number of possible values is either a finite number or a 'countable' number of possible values

(A) Review of Key Terms - Variables

- Discrete
- The number of eggs that hens lay; for example, 3 eggs a day.
- Students' raw score on the last quiz
- Continuous
- The amounts of milk that cows produce; for example, 2.343115 gallons a day.
- The time spent study for the last quiz.

(B) Presentation of Data – Tables

- We can use frequency tables to show/organize our data
- - a table showing data values or categories and some measure of how often each value or category occurs
- - the number of times a value occurs is known as the **frequency**
- - if the frequency is divided by the total number of responses, the result is the **relative frequency** of that value

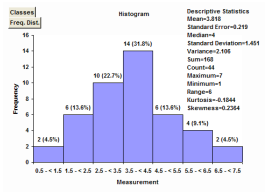
Table-I: Frequency of the patients in different age groups

Age Group (years)	Absolute Frequency	Relative Frequency (%)
≤20	8	1
21-30	150	18.8
31-40	465	58.1
41-50	118	14.8
51-60	34	4.2
61-70	13	1.6
71-80	10	1.3
>80	2	0.2

Day	# of Golf Balls Sold Frequency
Monday	17
Tuesday	13
Wednesday	15
Thursday	20
Friday	22

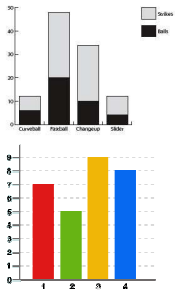
(C) Presentation of Data - Graphs

- We use histograms to visually represent **continuous data**
- Data values are grouped by class intervals and presented on the x-axis as a number line
- The frequency of the data in the class intervals appears on the y-axis



(C) Presentation of Data - Graphs

- We use column graphs to visually represent **discrete data**
- Data values are presented on the x-axis as a number line or descriptively
- Columns are separate
- The frequency of the data values appears on the y-axis



(D) Two Branches of Statistics

- (a) Descriptive Statistics : Involves organizing, summarizing, and displaying data of a **POPULATION**
- (b) Inferential Statistics : Involves using sample data to draw conclusions about a population.
- Statistics of **SAMPLES** from a population.
- Assumptions are made that the sample reflects the population in an unbiased form

(E) Measures of Central Tendencies

- A way of summarising the data using a single value that is in some way representative of the entire data set
  - It is not always possible to follow the same procedure in producing a central representative value: this changes with the shape of the distribution
- Include determination of mean, median, mode

(E) Measures of Central Tendencies

- The "**mean**" of some data is the average score or value, such as the average age of an IB1 student or average weight of track and field athletes that wish they were shot putters.
- Inferential mean of a sample: $X=(\Sigma X)/n$
- Mean of a population: $\mu=(\Sigma X)/N$
- The mean is the preferred measure of central tendency, except when:
  - There are extreme scores or skewed distributions
  - Non interval data
  - Discrete variables

(E) Measures of Central Tendencies

- The main problem associated with the mean value of some data is that **it is sensitive to outliers.**
- Example, the average weight of track and field athletes might be affected if there was one shot put thrower on the team that weighed 400 pounds.

Example - Track & Field Athletes

Athlete	Weight	Weight
Schmuggles	165	165
Bopsey	213	213
Pallitto	189	410
Homer	187	610
Schnickerson	165	165
Levin	148	148
Honkey-Doorey	251	251
Zingers	308	308
Boehmer	151	151
Queenie	132	132
Googles-Boop	199	199
Calzone	227	227
AVERAGE	194.6	248.3

(E) Measures of Central Tendencies

- **The Median** (not the cement in the middle of the road)
- Because the mean average can be sensitive to extreme values, the median is sometimes useful and more accurate.
- The median is simply the middle value among some scores of a variable. (no standard formula for its computation)
- The values that falls exactly in the midpoint of a ranked distribution
- Does not take into account exact scores
- Unaffected by extreme scores
- In a small set it can be unrepresentative

(E) Measures of Central Tendencies

Athlete	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227
	194.6

Rank order and choose middle value.

If even then average between two in the middle

i.e. in this case (187 + 189)/2 = 188

Weight
132
148
151
165
165
187
189
199
213
227
251
308

(E) Measures of Central Tendencies

- **The Mode**
- The most frequent response or value for a variable.
- Multiple modes are possible: bimodal or multimodal.

(E) Measures of Central Tendencies

Athlete	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227

Figuring the Mode

What is the mode? ➔ Most frequent value

Answer: 165

➔ Does not take into account exact scores
➔ Unaffected by extreme scores
➔ Not useful when there are several values that occur equally often in a set

(F) Measuring the Spread of Data - Dispersion

- Measures of dispersion tell us about variability in the data.
- Basic question: how much do values differ for a variable from the min to max, and distance among scores in between.
- Variability is usually defined in terms of distance
  - How far apart scores are from each other
  - How far apart scores are from the mean
  - How representative a score is of the data set as a whole
- We use:
  - Range
  - Standard Deviation
  - Variance

(F) Measuring the Spread of Data - Dispersion

- Measures of dispersion give us information about how much our variables **vary from the mean**, because if they don't it makes it difficult infer anything from the data.
- Dispersion is also known as the spread or range of variability.
- Describes in an exact quantitative measure, how spread out/clustered together the scores are

(F) Measuring the Spread of Data - Range

- **The Range** (no Buffalo roaming!!)
- $r = h - l$ ➔ Where h is high and l is low
- In other words, the range gives us the value between the minimum and maximum values of a variable.
- Understanding this statistic is important in understanding your data, especially for management and diagnostic purposes.

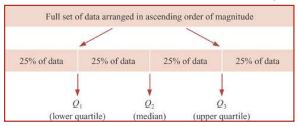
(G) Measuring the Spread of Data - Quartiles

- Quartiles
- Three quartiles approximately divide an ordered data set into four equal parts
- First Quartile is about one quarter of the data ➔ 1st quartile [Q_1] is the score at the 25th percentile
- Second Quartile is about one half of the data ➔ 2nd quartile [Q_2] is the score at the 50th percentile—the median
- Third Quartile is about three quarters of the data ➔ 3rd quartile [Q_3] is the score at the 75th percentile

(G) Measuring the Spread of Data - Quartiles

Inter-quartile Range

When the data is arranged in ascending order of magnitude, the **quartiles** divide the data into four parts. There are a total of three quartiles which are usually denoted by Q_1, Q_2 and Q_3.

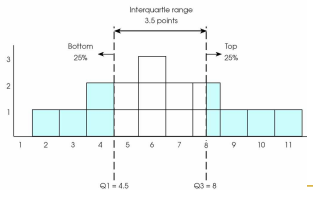


The **inter-quartile range** is defined as the difference between the upper quartile and the lower quartile of a set of data.

Inter-quartile range = $Q_3 - Q_1$

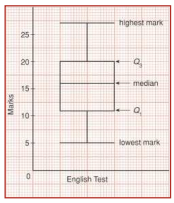
(G) Measuring the Spread of Data - Quartiles

- IQR provides information about how much distance on the X scale covers or contains the middle 50% of the distribution.



(G) Measuring the Spread of Data - Quartiles

A **box-and-whisker diagram** illustrates the spread of a set of data. It provides a graphical summary of the set of data by showing the quartiles and the extreme values of the data.



The difference between the two end-points of the line (represented by the highest and lowest marks) is the range.
The length of the box is the inter-quartile range.

From the above diagram, we know that the range of the data is 22 and the inter-quartile range is 9.

Homework

- HW

- Ex 18A #3 (no graphs), 4, 5;
- Ex 18B.1 #1b, 5, 11, 15;
- Ex 18B.2 #1, 4ab, 6b, 10ab;
- Ex 18D.1 #1c, 3;
- Ex 18D.2 #3ac