First Name: _______________________ Last Name: _______________________ Student ID: __________

Chapter 5: Measures of Central Tendency, Measures of Spread

➢ The Basics

We now know how to collect data efficiently and put the data into frequency tables and diagrams. The next step in summarizing data is to use *measures of central tendency* to begin to draw simple conclusions. You should already be familiar with the most common three:

- **Median:**
- **Mode:**
- **Mean:**

An extra element we need to consider is that some distributions contain ____________________. These are values that are *distant* from the majority of the data. They have a greater effect on the _______________________ than they do on the __________________.

Determining which measure to use can sometimes be tricky. The following guidelines should be of assistance:

- ✓ Outliers will affect the mean the most, especially if the sample size is small. In general, use the median if the data contains outliers.
- ✓ If the data is mainly symmetric, the mean and median will be close so either is appropriate
- ✓ Use the mode when the frequency of the data is more important than the calculated value (e.g. shoe size) or when the data is non-numeric (e.g. hair colour)

➢ Calculating Mean, Median & Mode for Ungrouped Data

We can distinguish the difference between the mean of a *population* and the mean of a *sample* of that population.

For an entire population of size N	For a sample of size n
$\mu =$	$\bar{x} =$

If you choose a wise sampling protocol, and your survey contains no bias, the sample mean will *approximate* the true mean.

Ex.1: Two classes wrote the same physics exam and had the following results:

Class A	71	82	55	76	66	71	90	84	95	64	71	70	83	45	73	51	68	
Class B	54	80	12	61	73	69	92	81	80	61	75	74	15	44	91	63	50	84

a) Calculate the mean, median and mode test results for each class.

b) Use the measures of central tendency to compare the performances of both classes.

c) What is the effect of any outliers on the mean and median?

Weighted Means

Sometimes, certain data within a set are more significant than others. A *weighted mean* gives a measure of central tendency that reflects the relative importance of the data. Weighted means are extremely popular, especially with course mark breakdowns, university admissions, job interviews, etc.

We can use the following formula for **weighted means**:

Ex.2: The personnel manager for Hillside Marketing Limited considers five criteria when interviewing a job applicant. The manager gives each applicant a score between 1 and 5 in each category, with 5 as the highest score. Each category has a weighting between 1 and 3. The following table lists a recent applicant's scores and the company's weighting factors.

Criterion	Score, x_i	Weighting Factor, w_i
Education	4	2
Job Experience	2	2
Interpersonal Skills	5	3
Communication Skills	5	3
References	4	1

a) Determine the weighted mean score for this job applicant.
b) How does this weighted mean differ from the unweighted mean?
c) What do the weighting factors indicate about the company's hiring priorities?

➢ Means and Medians for Grouped Data

When a set of data has been grouped into k intervals (i.e. you do not have access to the original data), you can *approximate* the **mean** using the formula for **grouped data**:

For an entire population of size N	For a sample of size n
$\mu \approx$	$\bar{x} \approx$

Where m_i

and f_i

The **median** can be estimated by taking the **midpoint** of the **interval** within which the median datum is found.

Ex.3: A simple random sample of car owners were asked how old they were when they got their first car.

Age	$(15, 20]$	$(20, 25]$	$(25, 30]$	$(30, 35]$	$(35, 40]$
Frequency	7	8	15	7	2

a) Determine the mean and median age of first car ownership from this sample.

Age	Midpoint (m_i)	Frequency (f_i)	$f_i\, m_i$	Cumulative Frequency
$(15, 20]$		7		
$(20, 25]$		8		
$(25, 30]$		15		
$(30, 35]$		7		
$(35, 40]$		2		

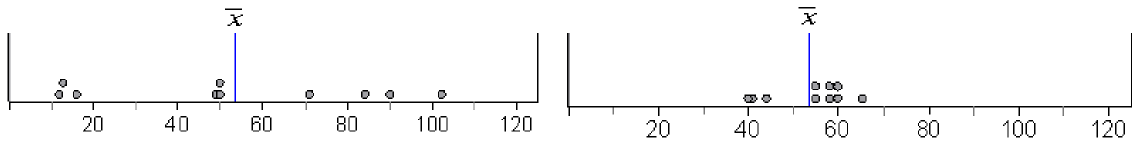
b) Why are these values approximations?

Often, we will also want to know how closely the data *cluster* around the centre of the data set.

- **Measures of Spread or Dispersion for Ungrouped Data**

Measures of dispersion describe how far the individual data values have *strayed* from the **mean** (also described as how closely the data values *cluster* around the mean). There are three measures of dispersion we will investigate: **range**, **variance**, and **standard deviation**.

Ex.4: The two dot-plots below each have a sample mean of approximately 54. How would you describe the similarities and differences between these two samples? Why is it important to note the differences?



- **Range:** The range is the simplest measure of dispersion, calculated by finding the difference between the ________________ value and the ________________ value of a data set. It is a quick way to get a feel for the ______________ of the data, but it relies on only ________ data points to describe the variation in a sample, as no other values between the highest and the lowest are involved in the calculation.

- **Variance:** The variance is a measure of dispersion calculated by **squaring each deviation** for an entire set of data, and then finding the mean of these values. A **deviation** is the difference between a data value, x_i, and the mean of the sample, $\bar{x}$ (or the mean of the population, μ).

Population Variance	**Sample Variance★**
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

★*Note the denominator of n – 1 for the sample variance. This compensates for the fact that a sample from a population tends to underestimate the deviations in the population.*

- **Standard Deviation:** The standard deviation is simply the **square root of the variance**. It is a more useful measure than the variance because the standard deviation is in the *same units* as the data set (while the variance is in units *squared*).

Population Standard Deviation	**Sample Standard Deviation★**
$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Note: *These formulae are only for ungrouped data sets.*

Ex.5: Both Hughes and Jackson were hired to wrangle dinosaurs from the field so that they could be tagged with laser beams. Team managers tracked how many dinosaurs they were able to collect in each of the first 10 days on the job. At the end of the 10 days, the managers hired Jackson on full-time, as he collected an average of 53.7 dinosaurs per day, and unfortunately let Hughes go because he only collected 53.6 dinosaurs on average. Did the managers make the right call? Jackson's and Hughes' record for catching dinosaurs is shown in the table below. Use the range, the variance and standard deviation of the data for both Hughes and Jackson to help you decide.

Jackson:

Day	Dinosaurs Caught	Deviation	Square Deviation
1	12	-41.7	1738.89
2	49	-4.7	22.09
3	102	48.3	2332.89
4	16	-37.7	1421.29
5	50	-3.7	13.69
6	71	17.3	299.29
7	84	30.3	918.09
8	50	-3.7	13.69
9	90	36.3	1317.69
10	13	-40.7	1656.49
		SUM:	

Hughes:

Day	Dinosaurs Caught	Deviation	Square Deviation
1	41		
2	55		
3	55		
4	58		
5	60		
6	65		
7	44		
8	40		
9	58		
10	60		
		SUM:	

Ex.6: A veterinarian has collected data on the life spans of a rare breed of cats. Determine the mean, standard deviation, and the variance for these data.

16 18 19 12 11 15 20 21 18 15 16 13 16 22 18 19 17 14 9 14 15 19 20 15 15

- **Measures of Dispersion for Grouped Data**

We will now consider the same measures of spread for grouped data (i.e. data that has been organized into intervals and frequency tables). The formulas below are for calculating standard deviation of a data set grouped into k intervals. To find the variance, simply square the standard deviation.

Population Standard Deviation(Grouped)	**Sample Standard Deviation (Grouped)**
$\sigma \approx \sqrt{\frac{\sum_{i=1}^{k} f_i (m_i - \mu)^2}{N}}$	$s \approx \sqrt{\frac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{n-1}}$

Recall that f_i is the frequency for a given interval and m_i is the midpoint of the interval. It should be noted that calculating standard deviations from ungrouped data will give more accurate results, and that measures of dispersion from grouped data are only *estimates*.

Ex.7: Calculate the standard deviation for the sample of salaries listed below:

Salary ($1000)	$(47,49]$	$(49,51]$	$(51,53]$	$(53,55]$	$(55,57]$	$(57,59]$
Frequency	4	6	7	4	2	1

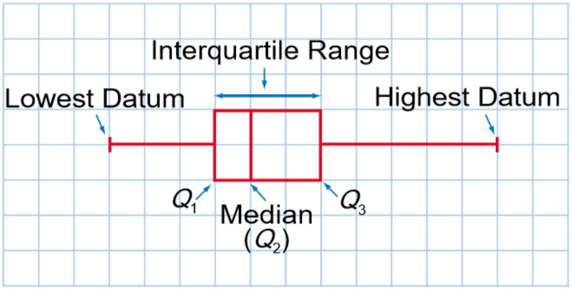
There are two other important measures of dispersion, called **measures of relative position**, which describe the portion of data below a certain data point.

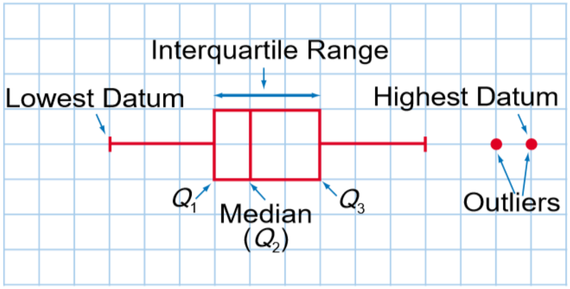
- **Quartiles and Interquartile Range**

The ____________ divide a set of *ordered data* into four groups with equal numbers of data, after it has been arranged in ascending order, just as the median divides data into two equally sized groups. The three dividing points are ______________, _____________, and the ______________.

The ______________________________________ (IQR) is the range of the middle half of the data. It has a value of $Q3 - Q1$. ***Note***: the **semi-interquartile range** is one half of the IQR. Both these ranges indicate how closely the data are clustered around the median.

A **box-and-whisker plot** and a **modified box-and-whisker plot** illustrate quartiles and interquartile ranges. The only difference between the two is that a *modified* box-and-whisker plot shows *outliers*. If a point is *outside* of $\{Q1 - 1.5(IQR)\}$ or $\{Q3 + 1.5(IQR)\}$, it is considered an outlier.





Ex.8: A random survey of 13 people walking into the school were asked how many times they have attended a live concert. The results were as follows.

3 2 1 10 4 7 35 12 0 1 4 4 3

Determine the median, the first and third quartiles, the interquartile range, and any outliers. Draw a modified box-and-whisker plot.

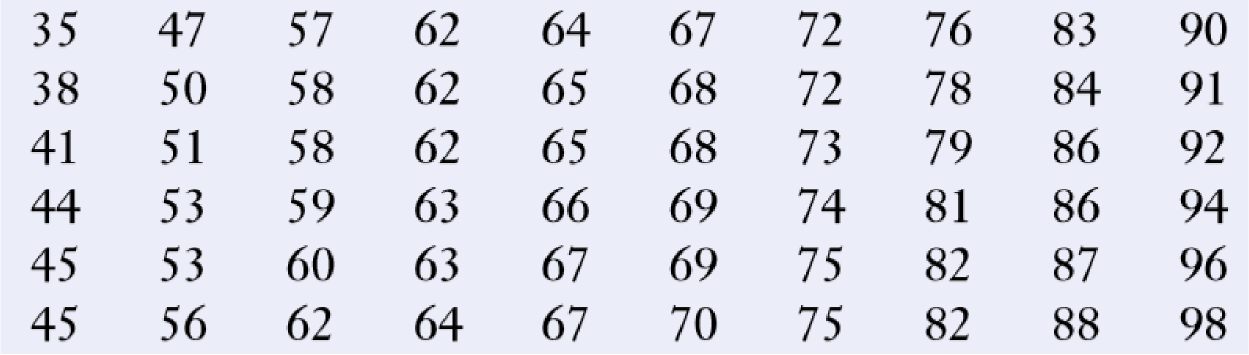
- **Percentiles**

Percentiles are similar to quartiles, except that they divide the sets of data into 100 intervals with equal numbers of values (Therefore, percentiles are usually applied to larger data sets).

- ✓ Percentiles are labeled as P_k. This value is called the k^{th} percentile.
- ✓ In a dataset, $k\%$ of the data is less than or equal to the value P_k.
- ✓ <u>ALWAYS</u> place the data in ascending order when working with percentiles.

For example, P_{80} means that 80% of the data is *less than or equal* to the value of P_{80} and 20% of the data is *greater than or equal to* the value of P_{80}.
Note: k^{th} percentile may or may not be a value in the data set.

Ex.9: The given set of data summarizes exam scores for one section of STAT101 at the University of Waterloo.



35	47	57	62	64	67	72	76	83	90
38	50	58	62	65	68	72	78	84	91
41	51	58	62	65	68	73	79	86	92
44	53	59	63	66	69	74	81	86	94
45	53	60	63	67	69	75	82	87	96
45	56	62	64	67	70	75	82	88	98

a) Find the 90th percentile.

b) Does a certain student's score of 75% place the student at the 70th percentile?