Measures of Central Tendency: Part I – Ungrouped (Raw) Data

A. 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:

1. Median –

2. Mode –

3. Mean –

An extra element we need to consider is that some distributions contain _____. These are values that are *distant* from the majority of the values. 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)

B. Calculating & Using Mean Values

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*
$=$	$\bar{x} =$

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) Manually 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?

Ex. 2: Using Excel, find the mean, median and mode of the data for each physics class.

CW: p. 133 # 1, 4, 9, 10

Key Concepts

- The three principal measures of central tendency are the mean, median, and mode. The measures for a sample can differ from those for the whole population.
- The mean is the sum of the values in a set of data divided by the number of values in the set.
- The median is the middle value when the values are ranked in order. If there are two middle values, then the median is the mean of these two middle values.
- The mode is the most frequently occurring value.
- Outliers can have a dramatic effect on the mean if the sample size is small.
- A weighted mean can be a useful measure when all the data are not of equal significance.
- For data grouped into intervals, the mean and median can be estimated using the midpoints and frequencies of the intervals.

Communicate Your Understanding

1. Describe a situation in which the most useful measure of central tendency is
   a) the mean **b)** the median **c)** the mode
2. Explain why a weighted mean would be used to calculate an index such as the consumer price index.
3. Explain why the formula $\bar{x} \doteq \frac{\sum_i f_i m_i}{\sum_i f_i}$ gives only an approximate value for the mean for grouped data.

Practise

A

1. For each set of data, calculate the mean, median, and mode.
   a) 2.4 3.5 1.9 3.0 3.5 2.4 1.6 3.8 1.2 2.4 3.1 2.7 1.7 2.2 3.3
   b) 10 15 14 19 18 17 12 10 14 15 18 20 9 14 11 18
2. **a)** List a set of eight values that has no mode.
   b) List a set of eight values that has a median that is not one of the data values.
   c) List a set of eight values that has two modes.
   d) List a set of eight values that has a median that is one of the data values.

Apply, Solve, Communicate

3. Stacey got 87% on her term work in chemistry and 71% on the final examination. What will her final grade be if the term mark counts for 70% and the final examination counts for 30%?

4. Communication Determine which measure of central tendency is most appropriate for each of the following sets of data. Justify your choice in each case.
   a) baseball cap sizes
   b) standardized test scores for 2000 students
   c) final grades for a class of 18 students
   d) lifetimes of mass-produced items, such as batteries or light bulbs

B

5. An interviewer rates candidates out of 5 for each of three criteria: experience, education, and interview performance. If the first two criteria are each weighted twice as much as the interview, determine which of the following candidates should get the job.

Criterion	Nadia	Enzo	Stephan
Experience	4	5	5
Education	4	4	3
Interview	4	3	4

6. Determine the effect the two outliers have on the mean mark for all the students in Example 2. Explain why this effect is different from the effect the outliers had on the mean mark for class B.

7. Application The following table shows the grading system for Xabbu's calculus course.

Term Mark	Overall Mark
Knowledge and understanding (K/U) 35% Thinking, inquiry, problem solving (TIPS) 25% Communication (C) 15% Application (A) 25%	Term mark 70% Final examination 30%

   a) Determine Xabbu's term mark if he scored 82% in K/U, 71% in TIPS, 85% in C, and 75% in A.
   b) Determine Xabbu's overall mark if he scored 65% on the final examination.

8. Application An academic award is to be granted to the student with the highest overall score in four weighted categories. Here are the scores for the three finalists.

Criterion	Weighting	Paulo	Janet	Jamie
Academic achievement	3	4	3	5
Extra-curricular activities	2	4	4	4
Community service	2	2	5	3
Interview	1	5	5	4

   a) Calculate each student's mean score without considering the weighting factors.
   b) Calculate the weighted-mean score for each student.
   c) Who should win the award? Explain.

9. Al, a shoe salesman, needs to restock his best-selling sandal. Here is a list of the sizes of the pairs he sold last week. This sandal does not come in half-sizes.

10	7	6	8	7	10	5	10	7	9
11	4	6	7	10	10	7	8	10	7
9	7	10	4	7	7	10	11		

   a) Determine the three measures of central tendency for these sandals.
   b) Which measure has the greatest significance for Al? Explain.
   c) What other value is also significant?
   d) Construct a histogram for the data. What might account for the shape of this histogram?

10. Communication Last year, the mean number of goals scored by a player on Statsville's soccer team was 6.
   a) How many goals did the team score last year if there were 15 players on the team?
   b) Explain how you arrived at the answer for part a) and show why your method works.

11. Inquiry/Problem Solving The following table shows the salary structure of Statsville Plush Toys, Inc. Assume that salaries exactly on an interval boundary have been placed in the higher interval.

Salary Range ($000)	Number of Employees
20–30	12
30–40	24
40–50	32
50–60	19
60–70	9
70–80	3
80–90	0
90–100	1

a) Determine the approximate mean salary for an employee of this firm.

b) Determine the approximate median salary.

c) How much does the outlier influence the mean and median salaries? Use calculations to justify your answer.

12. Inquiry/Problem Solving A group of friends and relatives get together every Sunday for a little pick-up hockey. The ages of the 30 regulars are shown below.

22	28	32	45	48	19	20	52	50	21
30	46	21	38	45	49	18	25	23	46
51	24	39	48	28	20	50	33	17	48

a) Determine the mean, median, and mode for this distribution.

b) Which measure best describes these data? Explain your choice.

c) Group these data into six intervals and produce a frequency table.

d) Illustrate the grouped data with a frequency diagram. Explain why the shape of this frequency diagram could be typical for such groups of hockey players.

e) Produce a cumulative-frequency diagram.

f) Determine a mean, median, and mode for the grouped data. Explain any differences between these measures and the ones you calculated in part a).

13. The **modal interval** for grouped data is the interval that contains more data than any other interval.

a) Determine the modal interval(s) for your data in part d) of question 12.

b) Is the modal interval a useful measure of central tendency for this particular distribution? Why or why not?

14. a) Explain the effect outliers have on the median of a distribution. Use examples to support your explanation.

b) Explain the effect outliers have on the mode of a distribution. Consider different cases and give examples of each.

C

15. The harmonic mean is defined as $\left(\sum_i \frac{1}{nx_i}\right)^{-1}$, where n is the number of values in the set of data.

a) Use a harmonic mean to find the average price of gasoline for a driver who bought $20 worth at 65¢/L last week and another $20 worth at 70¢/L this week.

b) Describe the types of calculations for which the harmonic mean is useful.

16. The geometric mean is defined as $\sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$, where n is the number of values in the set of data.

a) Use the geometric mean to find the average annual increase in a labour contract that gives a 4% raise the first year and a 2% raise for the next three years.

b) Describe the types of calculations for which the geometric mean is useful.

Measures of Central Tendency: Part II – Weighted & Grouped Data

A. 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**:

$$\overline{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Ex. 1: The MDM 4U course has a mark breakdown of: 3 summative assignments worth 10% each, 35% for tests, 15% for problem sets and a 20% final exam. Calculate a student's final grade if their breakdown was:

Summative 1 = 78%	Summative 2 = 83%	Summative 3 = 56%
Problem Sets = 86%	Tests = 75%	Exam = 63%

B. Means and Medians for Grouped Data

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

Population Grouped Mean	**Sample Grouped Mean**
$\mu \approx \frac{\sum f_i \cdot m_i}{\sum f_i}$	$\overline{x} \approx \frac{\sum f_i \cdot m_i}{\sum f_i}$

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

Ex. 2: A group of high school students were asked how long they spend on homework a day.

Amount of Time (hours)	[0, 0.5)	[0.5, 1)	[1, 1.5)	[1.5, 2)	[2, 2.5)
Number of Students	7	8	15	7	2

a) Determine the mean and median number of hours students spend on homework.

Amount of Time	Midpoint (m_i)	Number of Students (f_i)	$f_i m_i$	Cumulative Frequency
[0, 0.5)		7		
[0.5, 1)		8		
[1, 1.5)		15		
[1.5, 2)		7		
[2, 2.5)		2		

b) Why are these values *approximations*?

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

Age	16-20	21-25	26-30	31-35	36-40
Frequency	10	18	12	8	2

Use Excel to approximate the mean and median age for this sample.

CW: p. 134 # 5, 8, 11, 12

Measures of Dispersion for Ungrouped (Raw) Data

A. Recall Measures of Central Tendency

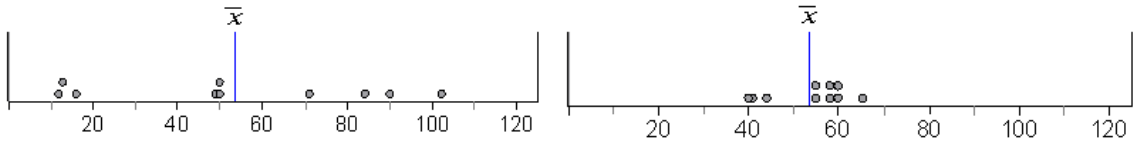
The measures of central tendency indicate the **central value** or **centre point** of a data set (the mean or the median) or the value that is **repeated** most often (the mode).

Often, however, you will also want to know how closely the data ***cluster*** around the centre…

B. Measures of Spread or Dispersion for Ungrouped (Raw) 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 today: **range**, **variance**, and **standard deviation**.

Ex. 1: 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?



1) 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.

2) Variance:

The variance is a measure of dispersion that describes the *relative distance* between the data points and the mean of the data set. It is 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, and the mean of the sample, $\bar{x}$ (or the mean of the population, μ).

Population Variance	**Sample Variance***
$\sigma^2 = \dfrac{\sum(x-\mu)^2}{N}$	$s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$

** 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.*

3) 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**
$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

NOTE: ALL FORMULAS GIVEN SO FAR ARE FOR UNGROUPED RAW DATA ONLY (NO INTERVALS)!

Ex. 2: Both Ms. Hughes and Mr. Jackson were hired to wrangle dinosaurs from the field so that they could be tagged with laser beams. Team managers tracked how many dinosaurs we were able to collect in each of the first 10 days on the job. At the end of the 10 days, the managers hired Mr. Jackson on full-time, as he collected an average of 53.7 dinosaurs per day, and unfortunately let Ms. Hughes go because she only collected 53.6 dinosaurs on average. Did the managers make the right call?

Mr Jackson's record for catching dinosaurs is shown in the table below. The sample mean is 53.7.

a) Find the **range**:

b) Find the **sample variance, s^2**:

i) *Calculate the* ***deviation*** *and the* ***square deviation****.*

ii) Sum *the square deviation.*

iii) *Divide the square deviation by* ***(n-1).***

c) Find the **sample standard deviation, *s***:

Take the square root of the variance.

Day	**Dinosaurs Caught** (x)	**Deviation** $(x-\bar{x})$	**Square Deviation** $(x-\bar{x})^2$
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:	**9734.1**

A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean. Before we can judge how *large* Mr. Jackson's standard deviation is, we should compare it against Ms. Hughes'.

Ms. Hughes' record for catching dinosaurs is shown in the table below. The sample mean is 53.6.

a) Find the **range**:

b) Find the **sample variance, s^2**:

i) *Calculate the* ***deviation*** *and the* ***square deviation****.*

ii) Sum *the square deviation.*

iii) *Divide the square deviation by* ***(n-1).***

c) Find the **sample standard deviation, *s***:

Take the square root of the variance.

Day	**Dinosaurs Caught** (x)	**Deviation** $(x-\bar{x})$	**Square Deviation** $(x-\bar{x})^2$
1	41		
2	55		
3	55		
4	58		
5	60		
6	65		
7	44		
8	40		
9	58		
10	60		
		SUM:	

Did the managers hire the right person for the job? Why or why not?

Ex. 3: Reuben and Dana are laying patio stones. They record how many stones they put in place each hour in an organized table:

Hour #	1	2	3	4	5	6	**TOTAL**
Reuben	34	41	40	38	38	45	
Dana	51	28	36	44	41	46	

a) Which worker gets more stones put down during the 6-hour day? _____

b) Which worker is more consistent on a per-hour basis? **Hint**: *Consistency* can be evaluated by comparing the standard deviation of the two sets of data. Use Excel to calculate their standard deviations and draw your conclusion.

CW: p. 148-149 #1, 6a, 10, 15a

Practise

A

1. Determine the mean, standard deviation, and variance for the following samples.
   a) Scores on a data management quiz (out of 10 with a bonus question):

5	7	9	6	5	10	8	2
11	8	7	7	6	9	5	8

   b) Costs for books purchased including taxes (in dollars):

12.55	15.31	21.98	45.35	19.81
33.89	29.53	30.19	38.20	

2. Determine the median, Q_1, Q_3, the interquartile range, and semi-interquartile range for the following sets of data.
   a) Number of home runs hit by players on the Statsville little league team:

6	4	3	8	9	11	6	5	15

   b) Final grades in a geography class:

88	56	72	67	59	48	81	62
90	75	75	43	71	64	78	84

3. For a recent standardized test, the median was 88, Q_1 was 67, and Q_3 was 105. Describe the following scores in terms of quartiles.
   a) 8
   b) 81
   c) 103

4. What percentile corresponds to
   a) the first quartile?
   b) the median?
   c) the third quartile?

5. Convert these raw scores to z-scores.

18	15	26	20	21

Apply, Solve, Communicate

B

6. The board members of a provincial organization receive a car allowance for travel to meetings. Here are the distances the board logged last year (in kilometres).

44	18	125	80	63	42	35	68	52
75	260	96	110	72	51			

   a) Determine the mean, standard deviation, and variance for these data.
   b) Determine the median, interquartile range, and semi-interquartile range.
   c) Illustrate these data using a box-and-whisker plot.
   d) Identify any outliers.

7. The nurses' union collects data on the hours worked by operating-room nurses at the Statsville General Hospital.

Hours Per Week	Number of Employees
12	1
32	5
35	7
38	8
42	5

   a) Determine the mean, variance, and standard deviation for the nurses' hours.
   b) Determine the median, interquartile range, and semi-interquartile range.
   c) Illustrate these data using a box-and-whisker plot.

8. Application
   a) Predict the changes in the standard deviation and the box-and-whisker plot if the outlier were removed from the data in question 7.
   b) Remove the outlier and compare the new results to your original results.
   c) Account for any differences between your prediction and your results in part b).

9. Application Here are the current salaries for François' team.

Salary ($)	Number of Players
300 000	2
500 000	3
750 000	8
900 000	6
1 000 000	2
1 500 000	1
3 000 000	1
4 000 000	1

a) Determine the standard deviation, variance, interquartile range, and semi-interquartile range for these data.

b) Illustrate the data with a modified box-and-whisker plot.

c) Determine the z-score of François' current salary of $300 000.

d) What will the new z-score be if François' agent does get him a million-dollar contract?

10. Communication Carol's golf drives have a mean of 185 m with a standard deviation of 25 m, while her friend Chi-Yan shoots a mean distance of 170 m with a standard deviation of 10 m. Explain which of the two friends is likely to have a better score in a round of golf. What assumptions do you have to make for your answer?

11. Under what conditions will Q_1 equal one of the data points in a distribution?

12. a) Construct a set of data in which $Q_1 = Q_3$ and describe a situation in which this equality might occur.

b) Will such data sets always have a median equal to Q_1 and Q_3? Explain your reasoning.

13. Is it possible for a set of data to have a standard deviation much smaller than its semi-interquartile range? Give an example or explain why one is not possible.

14. Inquiry/Problem Solving A business-travellers' association rates hotels on a variety of factors including price, cleanliness, services, and amenities to produce an overall score out of 100 for each hotel. Here are the ratings for 50 hotels in a major city.

39	50	56	60	65	68	73	77	81	87
41	50	56	60	65	68	74	78	81	89
42	51	57	60	66	70	74	78	84	91
44	53	58	62	67	71	75	79	85	94
48	55	59	63	68	73	76	80	86	96

a) What score represents

i) the 50th percentile?

ii) the 95th percentile?

b) What percentile corresponds to a rating of 50?

c) The travellers' association lists hotels above the 90th percentile as "highly recommended" and hotels between the 75th and 90th percentiles as "recommended." What are the minimum scores for the two levels of recommended hotels?

ACHIEVEMENT CHECK

Knowledge/ Understanding	Thinking/Inquiry/ Problem Solving	Communication	Application

15. a) A data-management teacher has two classes whose midterm marks have identical means. However, the standard deviations for each class are significantly different. Describe what these measures tell you about the two classes.

b) If two sets of data have the same mean, can one of them have a larger standard deviation and a smaller interquartile range than the other? Give an example or explain why one is not possible.

C

16. Show that $\Sigma(x - \bar{x}) = 0$ for any distribution.

17. **a)** Show that $s = \sqrt{\dfrac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}}$.

(Hint: Use the fact that $\Sigma x = n\bar{x}$.)

b) What are two advantages of using the formula in part a) for calculating standard deviations?

18. Communication The **midrange** of a set of data is defined as half of the sum of the highest value and the lowest value. The incomes for the employees of Statsville Lawn Ornaments Limited are listed below (in thousands of dollars).

28	34	49	22	50	31	55	32	73	21
63	112	35	19	44	28	59	85	47	39

a) Determine the midrange and interquartile range for these data.

b) What are the similarities and differences between these two measures of spread?

19. The **mean absolute deviation** of a set of data is defined as $\dfrac{\Sigma|x - \bar{x}|}{n}$, where $|x - \bar{x}|$ is the absolute value of the difference between each data point and the mean.

a) Calculate the mean absolute deviation and the standard deviation for the data in question 18.

b) What are the similarities and differences between these two measures of spread?

Career Connection

Statistician

Use of statistics today is so widespread that there are numerous career opportunities for statisticians in a broad range of fields. Governments, medical-research laboratories, sports agencies, financial groups, and universities are just a few of the many organizations that employ statisticians. Current trends suggest an ongoing need for statisticians in many areas.

A statistician is engaged in the collection, analysis, presentation, and interpretation of data in a variety of forms. Statisticians provide insight into which data are likely to be reliable and whether valid conclusions or predictions can be drawn from them. A research statistician might develop new statistical techniques or applications.

Because computers are essential for analysing large amounts of data, a statistician should possess a strong background in computers as well as mathematics. Many positions call for a minimum of a bachelor's or master's degree. Research at a university or work for a consulting firm usually requires a doctorate.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

For more information about a career as a statistician and other careers related to mathematics, visit the above web site and follow the links.

Measures of Dispersion for Grouped Data

A. Recall Measures of Dispersion for *Ungrouped* Data

Measures of dispersion or spread for *ungrouped* data describe how far <u>each individual data value</u> has *strayed* from the **mean** (also described as how closely the individual data values *cluster* around the mean). We use the formulas learned in the last lesson when we have all of the **raw data** and can compare each data point to the mean.

But what if we don't have access to the raw data? What if we ONLY have *grouped* data (i.e. a frequency table with intervals and/or midpoints)? The best we can do is ESTIMATE the measures of dispersion by treating the midpoints as though they are the data points. This requires slightly different formulas!

B. 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 standard deviation. To find the variance, simply square the standard deviation!

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

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

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

Ex. 1: Use Excel to calculate the standard deviation for the salaries listed below:

Midpoint Salary, ($1000)	28	30	32	43	36	38
Frequency	4	6	7	4	2	1

To find the standard deviation for grouped data:

i) Find the grouped mean using the appropriate formula from Lesson 2.

ii) Find the midpoint deviations.

iii) Square the midpoint deviations.

iv) Multiply the squared midpoint deviations by their frequency, so that the values are weighted appropriately.

v) Find the sum of the weighted midpoint deviations and divide by N or *n* – 1.

vi) Square root the result!

CW: p. 148 #7a, 9a

Measures of Relative Position

A. Recall Standard Deviation

Last class, we began our discussion on *dispersion* by introducing variance and standard deviation:

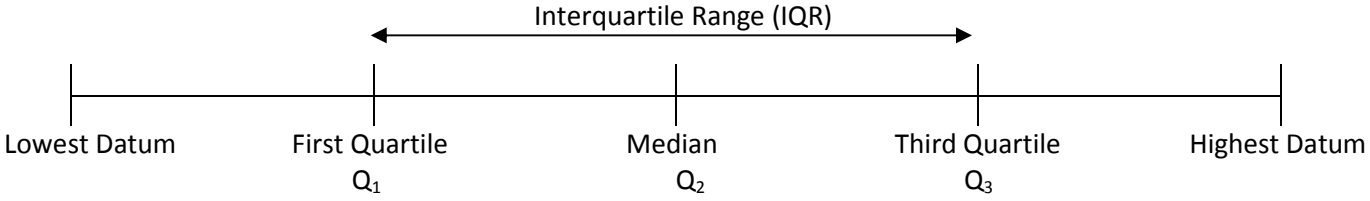
Population Standard Deviation	Sample Standard Deviation
$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
Population Standard Deviation (Grouped)	Sample Standard Deviation (Grouped)
$\sigma \approx \sqrt{\frac{\sum f_i(m_i-\mu)^2}{N}}$	$s \approx \sqrt{\frac{\sum f_i(m_i-\bar{x})^2}{n-1}}$

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

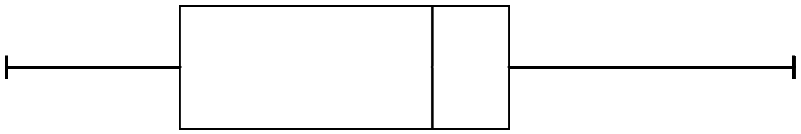
B. Quartiles and Interquartile Range

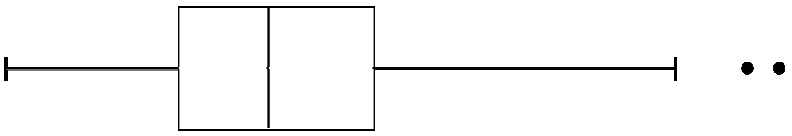
- The _____ divide a set of *ordered data* into four equal segments after it has been arranged in ascending order. (This is similar to how the median divides data in two equally sized groups!)

- The _____(IQR) is the range of the middle half of the data. It has a value of $Q_3 - Q_1$. The larger the interquartile range, the larger the spread. *Note*: the **semi-interquartile range** is one half of the IQR.



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 $\{Q_1 - 1.5(IQR)\}$ or $\{Q_3 + 1.5(IQR)\}$, it is considered an outlier.





Ex. 1: A random survey of 15 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	1	8	

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

C. Percentiles

Percentiles are similar to quartiles, except that they divide the sets of data into 100 intervals with equal numbers of values. Percentiles bring some confusing notation with them, so we need to define a few variables:

- ✓ Percentiles are labeled as P_k. As such, P_k is an actual item of data. This value is called the k^{th} percentile.
- ✓ In any particular set of data, k% of the data is less than or equal to the value P_k.
- ✓ **ALWAYS** place the data in numerical 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}. To find the value of P_{80} first multiply **0.8** by ***n***: this datum and the one above it contain *between them* the 80^{th} percentile. Thus, it is the *average* of these two pieces of data.

Ex. 2: The given set of data summarizes exam scores for one section of STAT 101 at the University of Waterloo.

34	50	58	62	65	68	72	78	84	91
41	51	58	62	65	68	73	79	86	92
45	53	60	63	67	69	75	82	87	96
48	56	62	64	67	70	76	82	89	99

a) Find the 90^{th} percentile.

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

CW: p. 148 #2 – 4, 6bcd, 7, 14

Unit 6: One Variable Stats – Review

You should be able to complete the following problems both in Excel and by hand. Answer the questions and check your solutions against those provided.

1. The table below shows the lengths of ski poles sold by a sporting goods store last December. Create a frequency polygon to display this information.

Length (cm)	**Frequency**
125–129	6
130–134	18
135–139	44
140–144	10
145–149	5

2. The following data summarize the attendance records for a class of mathematics students.

Classes Missed	**Number of Students**
0	0
1–3	7
4–6	12
7–9	5
10–12	3
13–15	2

   a) Estimate the mean number of classes missed by students in this course.
   b) What is the median interval for the numbers of classes missed?

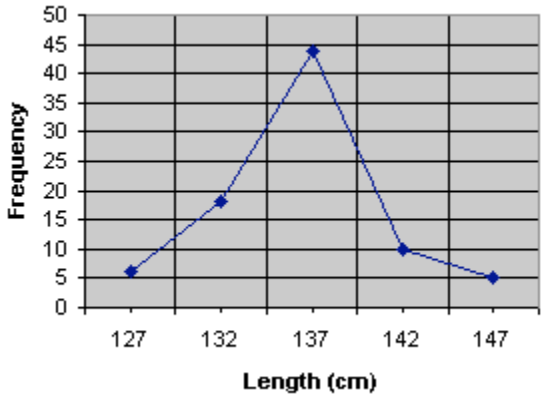
3. As of 2002, the top ten movies in terms of gross earnings were as follows (figures rounded to the nearest million dollars).

Rank	**Movie**	**Gross ($millions)**
1	*Titanic*	1836
2	*Harry Potter and the Sorcerer's Stone*	962
3	*Star Wars: The Phantom Menace*	924
4	*Jurassic Park*	921
5	*The Lord of the Rings: The Fellowship of the Ring*	836
6	*Independence Day*	813
7	*Star Wars*	798
8	*The Lion King*	767
9	*E.T. the Extra-Terrestrial*	757
10	*Forrest Gump*	680

a) Find the median, first quartile, and third quartile for the gross earnings for these ten movies.
b) Calculate the range and interquartile range.
c) Calculate the mean, standard deviation, and variance.

Unit 6: One Variable Stats – Review – Solutions

1.



2. a) $\mu = \dfrac{\sum\limits_i f_i m_i}{\sum\limits_i f_i}$

$= \dfrac{7(2) + 12(5) + 5(8) + 3(11) + 2(14)}{7 + 12 + 5 + 3 + 2}$

$= 6.03$

Students in this course missed an average of about six classes each.

b) The total number of students is 7 + 12 + 5 + 3 + 2 = 29, so the median interval is the one that includes the 15th largest value. Thus, the interval for 4–6 classes missed is the median interval.

3. a) The median has a value halfway between the fifth and sixth highest gross earnings. Thus, the median is $\dfrac{813 + 836}{2} = 824.5$ million dollars. The first quartile is the median of the lower half of the data and the third quartile is the median of the upper half, so Q_1 = 767 million dollars and Q_3 = 924 million dollars.

b) The range is 1836 – 680 = 1156 million dollars. The interquartile range is

$Q_3 - Q_1 = 924 - 767$

$= 157$ million dollars

c) The ten top-grossing films are a separate population of unusually successful movies rather than a representative sample of all movies. Therefore, use the population formulas to calculate the mean, standard deviation, and variance:

$\mu = \dfrac{\sum x}{N}$

$= \dfrac{1836 + 962 + 924 + 921 + 836 + 813 + 798 + 767 + 757 + 680}{10}$

$= 929.4$

$\sigma = \sqrt{\dfrac{\sum (x - \mu)^2}{N}}$ and $\sigma^2 = \dfrac{\sum (x - \mu)^2}{N}$

$= 313.2$ $= 98\,092$