

GRADE 12 APPLIED UNIT D₂ – STATISTICS

CLASS NOTES

*“There are three kinds of lies:
lies, damned lies and statistics”*; Mark Twain.

INTRODUCTION

1. If you are adept at Statistics it is possible to credibly and convincingly fool many people about anything most of the time!
2. Statistics is one of the many branches of Mathematics. So far you have worked with **Arithmetic** and **Measurement** and **Geometry** and **Algebra** and **Trigonometry**; special branches of mathematics. The branch of **Probability** is very closely related to **Statistics** as well.

Definition. One simple definition of statistics is the ‘*collection, processing, and display of information*’.

3. **Objectives.** The objective of this unit is to learn about statistical measures of **Central Tendency and Variation** and how data is distributed.

Prior Knowledge. You would have done much of these same concepts in previous grades so much of this should be a refresher. In Grade 11 Essential you studied how to graph Statistical data and in Grade 9 the ways to collect information.

CENTRAL TENDENCY

4. Many students will be familiar with the ideas of central tendency from previous Grades. This part constitutes a comprehensive review.
5. We often want a single number to represent many numbers in a set. For example if the students in the class room have ages: 23, 25, 26, 23, 22, 23, 20, 34, 40 we might ask : ‘what **one** age is sort of the average or typical, or the normal, or the middle, or the central age that I can think about when I think about this class?’

6. A single number to represent a bunch of numbers in a set is called a **Central Tendency**. And there are several different ways to find a Central Tendency. And as you might expect the different ways do not necessarily give the same result.

MEAN

7. The **mean** (also called *average*) of data is a statistic that is a measure of the **central value** of the data. It is represented by the symbol: \bar{x} , 'x bar' or sometimes the Greek letter: μ ; 'mu'. A single number that somehow represents a bunch of numbers is called a '**statistic**'. There are many 'statistics' in common use. The mean is a very commonly used statistic, sometimes over-used.

8. There are several types of **mean**, however the one most common and that we will use exclusively (the '*arithmetic mean*') is the mean that is calculated by adding up all the data values and then dividing by the number of values.

9. For example; given the set of data values: {1, 2, 4, 6, 7, 9} the mean would be:

$$mean = \bar{x} = \frac{1 + 2 + 4 + 6 + 7 + 9}{6} = 4.83$$

Make sure you understand the order of operations! The entire top is summed up then the sum is divided by the size of the set of numbers.

10. You may want to remind yourself, and your calculator, that the top is calculated first by using **parentheses** () around the entire numerator:

$$mean = \bar{x} = \frac{(1+2+4+6+7+9)}{6} = 4.83 \text{ (especially if you plan on entering the entire expression into a calculator in one shot)}$$

MEAN UNITS

13. Means are made up of numbers that represent some sort of unit. So naturally a mean value of some data should have units attached to it.

Eg: Jeff runs the following daily amount of **km**'s: {5, 6, 7, 8, 9} The mean is **7 km**. So, don't forget to include units.

MEAN AS AN EQUAL SHARING

14. One way to think about a mean is what everyone would get if everything was equally shared. If Rob has \$4 and Janice has \$6 and Brandon has \$8 and they threw all their money in a pot ($\sum all \$$) and added it up it would be \$18, taking away \$3 at a time to divide up the money; each person would get \$6.

$$\frac{\sum all \$}{3} = 6$$

MEDIAN, \tilde{x}

15. The Median is *another* measure of **central tendency**. It is the **middle** datum in a set of data; like a police line-up, the central datum, ...the 'median' on a highway. It is calculated as follows:

- a. Arrange the data in numerical order by value;
- b. Find the middle data value; and
- c. *If* there are two middle data values, use their mean.

Eg: {**9, 2, 4, 6, 6, 7, 1**}. Putting in order: {**1, 2, 4, 6, 6, 7, 9**}. The median, \tilde{x} , is 6.

Eg: {**1, 2, 4, 6, 7, 9**}. The median, \tilde{x} is $\frac{4+6}{2} = 5$. Compare this to the mean. Notice that **means and medians are not necessarily** the same value; sometimes they are close, sometimes very different.

16. Try these; find the median and compare with the mean previously calculated:

<p>a. {4, 8, 1, 1, 3, 5, 7} Median, \tilde{x} = Mean, \bar{x} =</p> <p style="text-align: right;">Ans: 4 and 4.14</p>	<p>b. {1.2, 3.5, 3.3, 3.3, 3.4, 0.9, 6.1} Median, \tilde{x} = Mean, \bar{x} =</p> <p style="text-align: right;">Ans: 3.3 and 3.1</p>
<p>c. {70, 80, 90, 100, 110, 120, 130} \tilde{x} = \bar{x} =</p> <p style="text-align: right;">Ans: 100 and 100</p>	<p>d. {1, 2, 4, 8, 16, 7, 3, 2, 1} \tilde{x} = \bar{x} =</p> <p style="text-align: right;">Ans: 3 and 4.88</p>

DIFFERENCE BETWEEN MEAN AND MEDIAN

17. You likely noticed that the mean and the median do not necessarily give the same '*statistic*' of central tendency! If you are sneaky you can use that to your advantage.

20. **Example - The Pay Difference.** John is on a crew of seven workers who shingle roofs for a company. The company is owned by a family; two brothers (Tom and Jerry) and an aunt Brenda. John and his crew claim they do not get paid enough! John does some research and gets the annual income of the ten people in the company.

21. Here is the data that John found: John \$24K, Kyle \$27K, Jeff \$26K, Bev \$21K, Terry \$24K, Mike \$24K, Kevin \$27K, and owners: Tom \$120K, Jerry \$110K, Brenda \$93K.

Calculate the **mean income** of the company

Calculate the **median income** of the company _____

What is the most common pay? _____

**Advanced concept.* Notice that when John researched company pay he was just rounding to the nearest 'K' or thousand. How do you think using rounded numbers might change the 'statistics'?*

22. Explain the big difference in the central tendency statistics:

Do you think that some companies (or politicians) might '*cherry pick*' and use the statistic that makes them look better? (Y / N)

MODE

23. The **mode** is yet *another* measure of a central tendency. It is the number value or category that happens the most often; the most popular.

24. Eg: {**2, 2, 4, 5, 6, 6, 7, 7, 7, 8, 9, 3**}. The data value '**7**' occurs the most often, so the mode is **7**. It helps to put the data in numerical order to better notice the most frequent value. Sometimes there may be two values that are equally common, a **bi-modal** distribution of data. If there are two or more most frequent values, list them all as the mode. The mode of {3, 4, 3, 1, 2, 4} is **3 and 4**

The mode is not an overly useful statistic for the central tendency. Although it may be useful in some context as below.

25. Carla sells bannock at her bannock shop. Here is a list of how many bannock each of her regular 12 customers ordered today:

{1, 2, 5, 5, 5, 2, 5, 3, 5, 1, 1, 2}.

Carla is thinking that 5 bannock seems to be a popular size quantity to order and maybe she should pre-package some of her bannock into packages of 5. The mode of her order quantities is that order size that occurs the most often. Determine the mean, median, mode, and range of her normal daily orders?

Mean: _____; **Median:** _____; **Mode:** _____

RANGE (SPREAD OF DATA)

26. The **range** of your data is another simple statistic. It tells how **spread out** your data are, a rather simple calculation. To calculate the range of your data, find the difference between the highest value in the set of data and lowest value in the set of data.

$$\text{Range} = x_{\max} - x_{\min}$$

where x_{\max} is the highest (maximum) value of the data and x_{\min} is the lowest (minimum) value of the data. Range will have units of that which is being measured.

30. Determine the Mean, Median, and Range of the following student course averages in a Math class:

{60%; 45%, 88%, 23%, 76%}

would the teacher like these marks? Explain_____

SPREAD OF DATA

31. The range of course is not a measure of the Central Tendency; it is a statistic that tells you how **spread out** your data is. It is known as a measure of **Variability**, or sometimes called **Dispersion**, which is not covered on this course. If you see resources that talk about Variance or Deviation or Inter-quartile Range of data then they will be referencing different ways of measuring how **spread out** the data is.

32. The importance of the **Range** to Grade 12 studies is that the **Mean** value, the **Median** value and the **Mode** of the data must fall within the x_{\max} and x_{\min} values of data. In other words; a Mean (average) Math mark cannot be more or less than the highest or lowest of any mark. You cannot have all your marks in the range from 35 to 72 and have a Central Tendency that says you have an 84!

FREQUENCY TABLES

35. Frequency tables just record how frequently different data values occur. You would normally make a Frequency Table when conducting a survey. It is usually easiest to make a stroke tally as you count values and then just count the tally for each value.

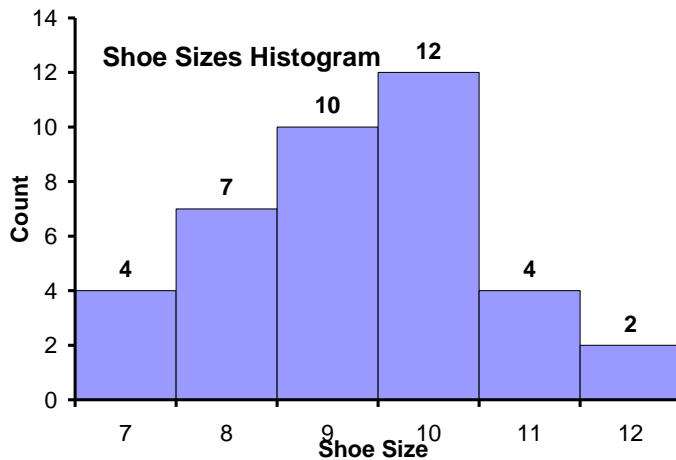
Eg: a frequency table from a survey of shoe sizes at our school:

Shoe Size	Tally	Count
7	////	4
8	### //	7
9	### ###	10
10	### ## //	12
11	////	4
12	//	2
	Total sample size, n:	39

HISTOGRAMS

36. You will recall from Grade 11 that Histograms are just a picture of the data. The human eye usually sees patterns better in such a graph or picture. A picture is worth a thousand numbers.

A histogram of the above shoe size frequency table would look like this.



Try doing this histogram in a graphing tool! You know how to use a spreadsheet (?) or even DESMOS can graph statistics too!

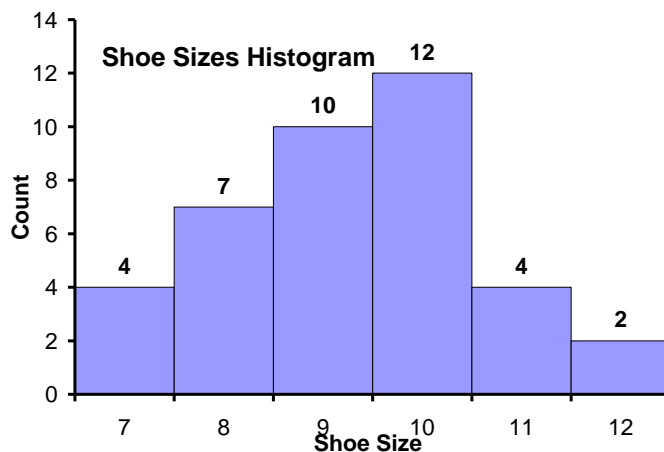
37. Notice that in a histogram that the area of each bar represents a percentage of the entire sample. Each category is the same width (one shoe size in this case). In the above case there were **39 samples**. So the category for shoe size **9** takes up **10 out of 39** parts of the graph, or **25.64%** of the data. In a histogram, there should not normally be any gaps between the bars in the categories (unlike a 'bar graph').

38. A histogram can also show percentages or probabilities (instead of a raw count) of categories along the vertical axis, simply by taking the count and dividing by the total number of the sample. We could have said, for example, that $4/39$ of shoes are size 7, so 10.26%, and graphed the percentage along the vertical axis instead.

In the above survey on shoe size the histogram tells us that there is a 46% probability that next random person to walk in the door will have a size 10 or bigger shoe. Histograms are useful to study probability (another unit in Grade 12)

FINDING CENTRAL TENDENCIES FROM A HISTOGRAM

39. Given the shoe size histogram it is not difficult to find the important statistics of **mean**, **median**, and **mode** (almost just by looking) directly from the histogram!



40. **Find the mean from the histogram.** Add up all the data values, divide by the total number of values. There are 39 data values, their total is $(4*7) + (7*8) + (10*9) + (12*10) + (4*11) + (2*12) = 362$.

41. So the mean, when we count the frequency of data values, is given by:

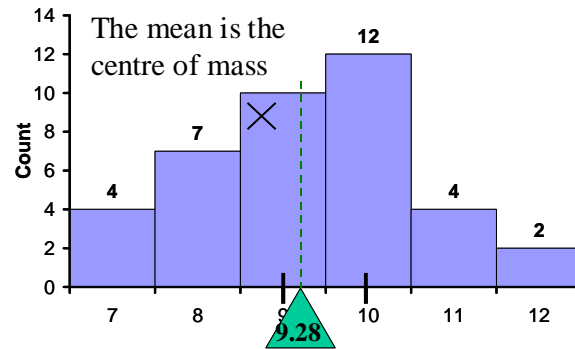
$\bar{x} = \frac{\sum(f_i x_i)}{N} = \frac{(4*7)+(7*8)+(10*9)+(12*10)+(4*11)+(2*12)}{39}$ which is the mathematical way of saying “take the frequency count, f_i , of each data value category, x_i , multiply them together, add them all up, and divide by the total number”.

The Greek letter Σ (capital ‘sigma’), is used almost exclusively to mean ‘add up all this crap’.

(But I don’t think crap is a Greek word.)

42. So we have $362 \div 39 = 9.28$. The **mean** shoe size is **9.28**. If you went to the shoe store and asked for a size **9.28**, then you would have a ‘mean’ or ‘average’ foot size. Obviously, stores don’t actually sell that size, you have a choice between discrete shoe size values only: either size **9** or **10**.

43. One rather important interpretation of the mean is that it is the '*centre of balance*' of the histogram, the histogram would balance at the mean if we stuck a knife edge there along the bottom.



You may know a bit about 'Centre of Mass' from science.

45. Find the median from the Histogram. The median is the exact half-way piece of data when you line the data up in order of value. There are **39** data values. If we knock off **19** values from the *left*, and **19** values from the *right*, the **20th** value from either end would be the central one. The first 4 values from the left were sevens, the next 7 were eights, so the 20th would occur in the size nine category. The **median** is 9. I put an X in the median category where the 20th person was. A median will generally be one of the categories, unless it is exactly half way between two categories with an even number of samples.

Find the Mode from the Histogram

46. Find the mode from the histogram. This is the easiest statistic of the histogram! It is readily obvious that the most frequent shoe size is size **10**. So the **mode** is **10**. Occasionally you might notice two peaks (or more) that are the same frequency. It is possible to be *multi-modal* and have more than one most common value.

Notice this histogram is not a nice perfect (normal) symmetrical shape. We say it is '**skewed**'. The data favours the lower data by a bit. There is a statistic to measure 'skewedness' too which we do not cover.

47. *The Three Different Measures of Central Tendency Generally Give Different Answers.* The wonder of statistics! So, as you can see..., the **three different measures** of central tendency give similar but slightly different answers. For our shoe sizes the **mean**, \bar{x} , was 9.28, the **median**, \tilde{x} , was 9 and the **mode** was 10. You could say the average was 9.28 (if we all had our feet surgically adjusted so we were all the same shoe size), or a median: where half the people have a size 9 or more and half have a size 9 or less. The mode, the most common size, is 10.

If you can master statistics, you would be a great politician; you can give three different answers for the same question!!! Pick the one that best makes you look good!

A good example of a misleading measure of central tendency.

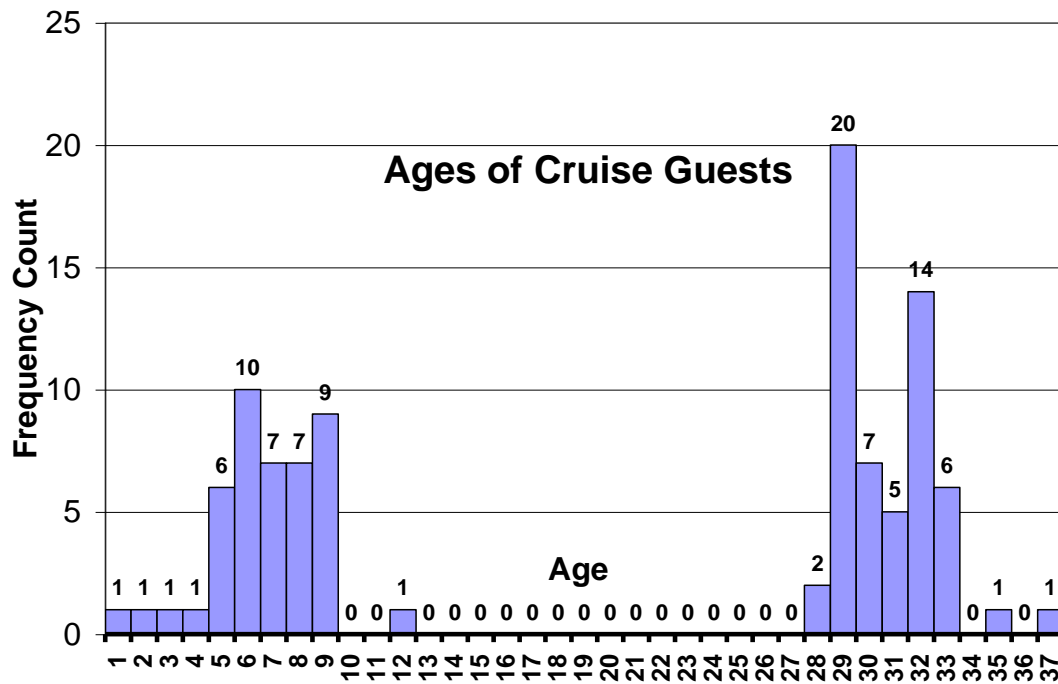
48. We get the sense that measures of central tendency are a good representation of the measure of the 'centre' of data. But be careful!! Statistics can be misleading, especially in certain hands! If you really understand statistics, you can pretty well rule the world! (ie: **BS** baffles brains!)

50. Let's say you decide to go on a holiday after grad! You really deserve a cruise in the Caribbean after all that Math! So, you book a holiday with Funtastic Cruise Lines that **guarantee** the average (ie: **mean**) age is about **20** years old!

Check out the graph below and calculate the mean, median, mode, and Range!

Aren't you surprised then when you get on board and find out that there are lots of married **30** year olds, many with their young toddlers!

Oh well, at least you can make some babysitting money while on-board!



You try calculating the mean, median, and mode of this data!

51. The 'bimodal' distribution in the histogram above shows the hazards of believing a single statistic. So *be very careful* which statistics you use. You might '*accidentally*' mislead people! Often **median** is a more relevant statistic. Further, any suspicious statistic that isn't supported by a histogram could (should) be readily challenged! In this particular case with the cruise ship I would certainly want to see the histogram of all the data, not just one number that summarizes everything!

52. The important rule to remember is that **using a single statistical number to represent a lot of numbers can be very dangerous** unless you are familiar with the 'distribution' of the data too! Data that is '*skewed*' generally has a large difference between the Mean, Median, and Mode. Data that is not '*skewed*' is called 'normal'. The best way to see if the data is distributed in a well-behaved manner is to ask to see the graph of the data.

Be very careful with statistical numbers!! **Ask to see a graph also.**

Finding the middle number of an ordered set of numbers (ordinals).

To find the middle position of any set of objects (including an ordered list of numbers); add one to the size of the set, n , and divide the total by 2. If the set size 'n' is an even number then the middle position will be between two objects (numbers).

Eg: the middle position between 7 objects is $(7 + 1)/2$ so the 4th position or 4th number is the centre object or number.

The middle position between 12 objects is: $(12 + 1)/2 = 6.5$; so the position is between the 6th position and the 7th position.

Can you think why this is true?

WEIGHTED MEANS

58. Teacher Rick has a class of 20 students with a class average of 74%. Teacher Courtney has an English class of 40 students with a class average of 86%. What is the average mark for both classes combined?

59. We cannot just add the two marks together and divide by two since Courtney had twice as many students, her average should actually count twice as much. Had we got all the students into one room and made an average of their individual marks their individual averages we get a different answer. The proper average mark combining both classes is:

$\frac{74*1+86*2}{3} = 82\%$; closer to the English class average since there were more students in the English class. The important point here is that you **cannot** generally **average averages** with considering the size of the sets of numbers involved.

60. Teachers often use weighted. Most teachers do weighted means in which tests count twice as much as quizzes and quizzes count twice as much as assignments for example.

61. **Example:** Teacher Mike has three quizzes and one test the first part of the course. Tests count twice as much as a quiz. Bev's quiz scores were 45, 68, and 77 and her test score was 85. What is Bev's mark?

$$\frac{45(1)+68(1)+77(1)+85(2)}{1+1+1+2}= 72.$$

Compare this to Bev's mark if the test **did not** have a heavier weight factor than the quiz:

$$\frac{45 + 68 + 77 + 85}{4} =$$

Example:

62. Last year the Jets won 35% of their games in the 80 game season. This year after 40 games they have won 25% of their games. Combine those two statistics into one average:

SOME PRACTICE PROBLEMS

SAMPLE QUESTIONS

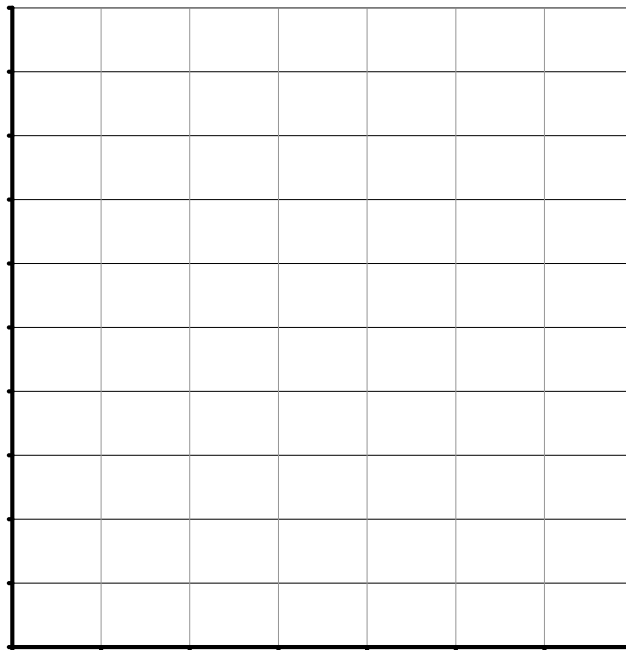
1. Calculate the mean , median, mode, and range for the values: {45, 96, 8, 54, 48, 30, 49, 52, 38, 44}. Do it manually and try in DESMOS too!

2. Kevin had four tests and a final exam. All the tests had the same 'weight'. The final exam was worth twice as much as a single test. If his marks were as follows:
Test 1: 67%, **Test 2:** 87%; **Test 3:** 90%; **Test 4:** 35%; and his final exam was a 92%; what was his final course average?
3. Scott presently has a course average of **42%**. If the final exam is worth 25% of the course average, what exam score (%) does Scott need to get an overall course average of at least 50%.
4. Colleen has been jogging every day! She wants to jog *at least* an average of 4 km per day. The last six days she has jogged 22.2 km; how far will she need to jog today to get an average of at least 4 km / day for the week?

5. A Quality Assurance supervisor from the Canadian Nuts and Bolts company measured the masses of 300 bolts to see how consistently they were being manufactured. Here are the results of his quality testing sample.

Mass [g]	7.40	7.50	7.60	7.70	7.80	7.90	8.00
Frequency	3	5	38	86	91	60	17

- Determine the mean mass of the bolts. (You may want to do your own frequency data table)
- Calculate the median mass.
- do a properly labelled histogram and clearly mark the mean, median, mode, the P_{50} (Q_2), the Q_3 , and the approximate P_{80} . Try it on a Spreadsheet too if you know how to graph with spreadsheets!



- d. What is the **range** of the masses of this sample?
- e. what is the range of the 'inner-quartile' sizes, from Q_1 (P_{25}) to Q_3 (P_{75}). Does this suggest that the masses tend to be clumped more in the centre. What the heck is a quartile? LOL, a median; breaking the data into four equal chunks.

8. Problem Solving. Mr. F normally has an 80% success rate with his curling shots. Today he only had three out of his first six shots that were good. If he has ten more shots to go in the game (there are only 16 shots total for each player in a game), how many good shots must he make in his last 10 shots to at least maintain his 80% success average.

A frequency data table to record and calculate large samples

Frequency Data Table (to calculate statistics of large samples)					
x Value of variable being measured	Tally ticks (if doing a survey)	f frequency each value happens [count]	Cumulative Frequency (running total)	f*x <i>f times x</i>	
					Mode; most frequent x: _____
					Mean, μ or $\bar{x} =$ $\frac{\Sigma(f * x)}{n} =$
					Median, \tilde{x} Halfway into the data; in between two values if n is EVEN. = _____
		sum: n = _____		sum: <i>Σall the f * x's</i> _____	

*A quick way to find the middle place of a string of numbers in order is to take $(n + 1)/2$. That will tell you where the middle place would fall. If the result is a half value then you then you are in between the two places. So in a string of 83 numbers the middle number would be in the 42nd place. In a string of an even number of numbers however, say 180, the middle place would be in the $181 \div 2$ place or the 'ninety and a halfth' place; so you would need find the mean of the two numbers either side; so the mean of the two numbers in the 90th and the 91st place.

Use the Cumulative Frequency column to find the half way value of the data.

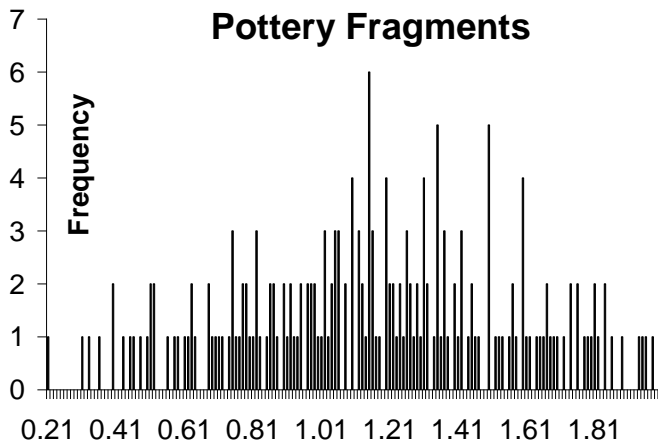
GROUPING DATA

35. Often, there is a very large amount of data which is categorized as many discrete data values over a large range of values. It is often desirable to just group the data together into larger groups of values for more ready analysis.

36. The following table represents the data for the size, n , of a sample of **200** pottery fragments measured in centimeters.

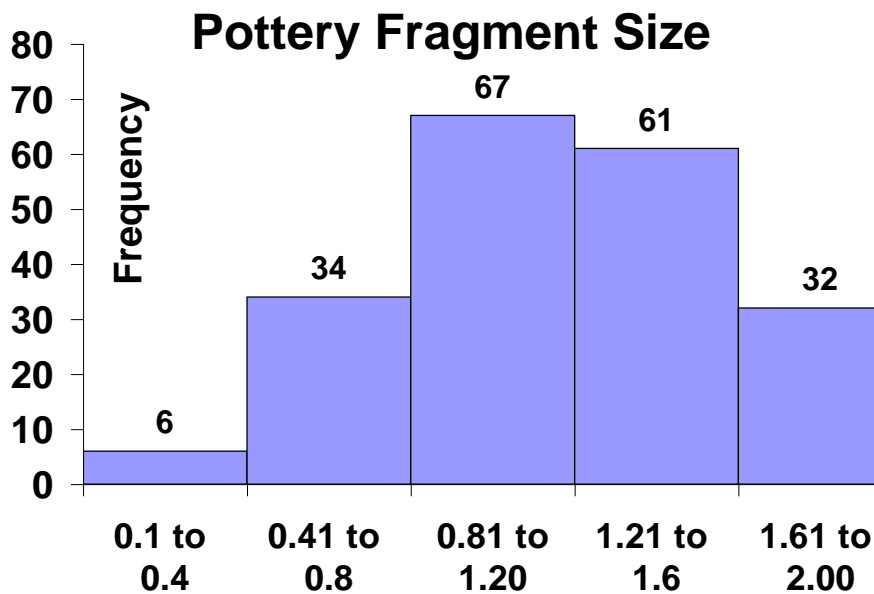
0.20	0.36	0.50	0.52	0.62	0.72	0.87	0.91	1.02	1.16
0.33	0.43	0.51	0.56	0.75	0.79	0.85	1.05	1.12	1.20
0.31	0.40	0.52	0.63	0.78	0.82	0.97	1.04	1.13	1.21
0.45	0.46	0.64	0.74	0.78	0.88	1.04	1.15	1.21	1.35
0.40	0.58	0.68	0.77	0.82	0.98	1.02	1.10	1.28	1.31
0.48	0.59	0.75	0.81	0.86	1.05	1.05	1.17	1.29	1.36
0.51	0.63	0.70	0.90	0.99	1.06	1.10	1.20	1.31	1.45
0.61	0.75	0.80	0.92	1.03	1.06	1.24	1.27	1.38	1.52
0.69	0.71	0.90	0.98	1.08	1.15	1.23	1.35	1.42	1.60
0.68	0.83	0.92	1.02	1.10	1.16	1.26	1.37	1.50	1.61
0.76	0.82	1.00	1.08	1.15	1.25	1.40	1.45	1.53	1.65
0.79	0.87	0.95	1.15	1.24	1.30	1.41	1.50	1.64	1.72
0.86	0.99	1.06	1.20	1.26	1.37	1.44	1.54	1.67	1.76
0.93	0.97	1.12	1.15	1.31	1.35	1.47	1.57	1.69	1.76
0.94	1.10	1.12	1.29	1.35	1.40	1.57	1.68	1.79	1.80
0.95	1.14	1.15	1.32	1.35	1.50	1.58	1.67	1.81	1.89
1.01	1.20	1.22	1.31	1.50	1.56	1.66	1.74	1.84	2.00
1.13	1.16	1.27	1.42	1.46	1.62	1.74	1.81	1.86	1.96
1.18	1.26	1.34	1.42	1.60	1.60	1.78	1.84	1.98	2.00
1.22	1.32	1.37	1.50	1.60	1.70	1.82	1.94	1.95	2.00

37. The mean of this all data is **1.17** cm. A histogram of this data could have as many as 180 different categories and bars! (from 0.20 to 2.00 in intervals of 0.01 cm).



38. We can group the data into **class intervals** of equal width and get a more useful and manageable histogram of the data.

Range	Class Mark	Freq
0.01 to 0.40	0.20	6
0.41 to 0.80.	0.60	34
0.81 to 1.20	1.00	67
1.20 to 1.60	1.40	61
1.61 to 2.00	1.80	32



39. By grouping the data in a table and in the histogram we break the data values into ranges, or classes, of data, in this case ranges of width 0.4 cm. The centre of each range of data value is called the '**class mark**'.

40. It is important to note that by grouping data there is an error introduced into any sample statistics. However; normally this is a very slight error except under special circumstances or in the case of intentionally misleading and manipulated statistics.

41. Using the data above for example, we find the grouped data has a grouped mean of

$$\bar{x} = \frac{6*0.20 + 34*0.60 + 67*1.00 + 61*14.00 + 32*18.00}{200} = 1.16 \quad \text{grouped mean}$$

instead of a 1.17 mean when the data is *not* grouped.

VARIABILITY AND DISPERSION

42. Variability and Dispersion are two words that mean the same thing. It represents how *spread out* the data is, or conversely how closely it is grouped to a central tendency

QUARTILES AND DECILES

43. Quartiles and Deciles are statistics that represent what the data looks like when broken into quarters (quartile sections) or into tenths (decile sections).

a. Quartiles. The three quartiles of a sample of data is found by arranging the data in order of ascending (increasing) value and then finding the datum that is *one quarter* of the way through, *one half* of the way through, and *three quarters* of the way through. To be more precise in explaining:

(1) arrange the data in numerical increasing order. Locate and calculate the Median.

(2) Q_1 or Q_{lower} or lower quartile is found as the **median** of the half of the data to the left of the median.

(3) Q_3 or Q_{upper} or upper quartile is the **median** of the half of the data to the right of the median.

b. Median is the Second Quartile. Notice by the definition that the second quartile is really just identical to the *median*. You could call it the Q_2 if you wanted.

c. Deciles. The concept of deciles is exactly the same as quartiles, but the data is broken into tenths. Deciles will not be covered further in these notes.

Example - Finding Quartiles.

44. Find the Lower and Upper quartile of the distinct sample data:

12, 15, **15**, 15, 18, **18**, 19, 20, **21**, 23, 24

45. The Lower Quartile, Q_L is 15, it is a quarter of the way through the data, or in another sense, it is the median of the lower half of the data.

46. The Upper Quartile, Q_U , is 21. It is the value that is three-quarters of the way through the data, or in another sense, the median of the upper half of the data.

Notice that the median, (really the second quartile) is 18, the very centre number of the data.

47. The methods to actually calculate quartiles vary depending on software programs and whether the data is *grouped* or *ungrouped*, and whether it is *discrete* or *continuous*. Regardless, this introduction above to calculating quartiles is sufficient for now. And since we will be using the TI-83 for calculating statistics, we will simply use its method. Check out the Appendix 1 in these notes for how to use the TI-83 for Statistics.

RANGE

48. Range is a very basic measure of Variability. It is simply the difference between the lowest data value and the highest data value. Range is $X_{\max} - X_{\min}$.

Example: Sample shoe sizes in a Math class are: **5, 6, 8, 9, 9, 10, 11, 11, 12**. The range is simply **(12 - 5) = 7**. And you know that somewhere in that '*range*' the mean, and median, and mode is there. So it isn't really a very good statistic.

INTER-QUARTILE RANGE

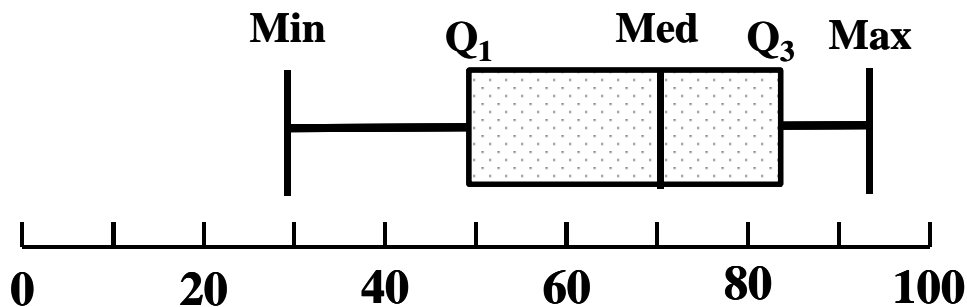
49. The **Inter-Quartile Range** (sometimes called the Semi-quartile Range) is a measure of the *range* of the *central half* of the data. It is always calculated as: $Q_U - Q_L$ ($Q_{\text{Upper}} - Q_{\text{Lower}}$; or $Q_3 - Q_1$). So the inter-quartile range of the shoe sizes above is: $Q_U - Q_L = 11 - 7 = 4$.

50. Summary. The range of the above shoe sizes tells us that all of the shoes cover a full spread **range** of **7** sizes, the **inter-quartile range** tells us that the middle half of the data is spread out over **4** sizes. So if we were concerned about shoe sizes for only the half our class that are near the centre average, we only had to look in a range of **4** different shoe sizes. We would exclude all those 'abnormal' people with tiny or big feet.

WHISKER AND BOX PLOTS

14. Most students will be familiar with these from report cards. It is a convenient way to show the statistics of scores and how you or your child is doing.

15. Whisker and Box plots are hand way to show the five key statistic numbers in a set of data: The Minimum, 1st Quartile (Q_1), Median, 3rd Quartile (Q_3), and Maximum.



The whisker and box plot here represents the Mathematics marks in a class. The lowest mark was 29, the First Quartile Mark was 49, The Median was 71, the Third Quartile mark was 83, the Maximum mark was 93.

STANDARD DEVIATION (s) [or sometimes σ]

51. The standard deviation is the most common measure of variability. It provides a single number statistic that represents how spread out the data is from the average. It is really a measure of the average difference of each individual datum from the average. The symbol we use for a Standard Deviation is the lower case Greek letter 'sigma': σ

52. The Standard Deviation of a sample of observations is calculated as follows:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$$

which is the mathematical way of saying: 'take the difference of every datum from the mean, square each difference, add up all those squares of differences from the mean, divide that sum of all those

squares of differences from the mean by the number of data, N, and then take the square root of all of that'. Easy!

$$\frac{\Sigma(x_i - \bar{x})^2}{N-1}$$

The value: before it is rooted is called the variance or s^2 .

You will also find that we often inter-changeably use the symbol ' σ ' instead of to represent the deviation, there being a subtle difference. It is pronounced '**Sig Mah**' (*It is the lower case version of Σ*)

53. Example of calculating a standard deviation. We will normally use technology to calculate standard deviation, s . But it is important to understand how it is calculated by computers and calculators. Follow the table below for $n=10$ data points.

A	B	C
Data (x)	$x - \bar{x}$	$(x - \bar{x})^2$
3	-3.9	15.21
3	-3.9	15.21
4	-2.9	8.41
6	-0.9	0.81
7	0.1	0.1
8	1.1	1.21
9	2.1	4.41
10	3.1	9.61
10	3.1	9.61
9	2.1	4.41
$\Sigma x = 69$		$\Sigma(x - \bar{x})^2 = 68.99$
$\bar{x} = 69/10$		$\Sigma(x - \bar{x})^2 / (n - 1) = 7.66$
$\bar{x} = 6.9$		$s = \sqrt{7.66} = 2.77$

The symbol Σ is the capital Greek letter **SIGMA**. It stands for **summation**. It just means **add 'em all up!** So Σx means add up all the x s.

When we take the **summation** of 'square of the difference of every data point from the mean' we get what is called variance. In this case the variance is 6.899. By taking the square root of that we get **Standard Deviation**. So the standard deviation, s , of this population is **2.77**

(You might see on websites, or other text books, reference to the 'population deviation', σ , and the 'sample deviation', s . They are normally rather close for sufficiently large samples, we will generally only use the population deviation, σ , on this course)

54. Calculating Standard Deviation with the TI-83. Calculating statistics with the TI-83 is easy, consult Appendix 1 in these notes on how to find Standard Deviation.

*** there are subtle reasons why we divide by ' $n - 1$ ' and not by ' n '. But we shall not investigate that further.

55. We can generally get a sense for what samples have a smaller 'deviation' or spread or variance just by looking at the data. Check out the data for two different pallets with bags of potatoes on them. Just by visual inspection which pallet seems to have bags of potatoes that are more closely grouped to a central value?

Weight in pounds of several bags of potatoes on two different pallets								
Pallet A	19	19	20	20	20	21	20	19
Pallet B	14	17	19	22	17	21	21	24

56. Which pallet seems to have bags of potatoes that are more closely grouped to a central value? And what are the means and standard deviations of the weights of bags of potatoes on each pallet? Sketch some histograms of this data to see how the spread looks.

VARIABILITY CONCLUSION

57. To summarize all of our studies this far, take the data regarding pottery fragment sizes on page 9 and do the following:

a. manually graph (on separate paper) a grouped histogram in class intervals of 2 cm. (you will need to make a frequency table first)

b. find the mean, median, and mode of all the data (use TI-83 if possible or figure out the statistics functions on your own personal calculator)

c. find the range, inter-quartile range, and standard deviation of all the pottery fragment data.

d. graph on the graphing calculator the histogram of the pottery fragment data in class widths of 4 cm on the graphing calculator

COMBINING STATISTICS

58. There may be times when you want to combine two statistics.
Be very careful doing this!

Example: A survey of **100** students was done, they were **sample A**. Their mean income was **\$20,000**. Another survey was done of **20** professors, their mean income was **\$40,000**, and they were **sample B**. The researcher therefore says that the average income of everyone is **\$30,000** because that is the average of the two averages. But he is wrong. The mean (or average) income is really only **\$23,333.33**.

59. The reason you can't necessarily just add or otherwise easily combine statistics is because you can't give equal '*weight*' to different sized samples. In this case, there are five times as many students in sample group A, so more weight has to be given to their mean. The point is: *you cannot just average averages to get a new average!* But politicians do it all the time!

60. The proper way to calculate the above statistics is:

$$\bar{x}_{combined} = \frac{\sum incomes}{N_{combined}} = \frac{100 * \$20,000 + 20 * \$40,000}{120} = \$23,333.33$$

61. Notice how in the calculation that Group A is '*weighted*' five times more heavily than Group B.

62. It is like how teachers do marks also. For example: A *test* may be worth four times as much '*weight*' as an *assignment*. So just because you get 100 in the assignment and 50 in the test, your average is *not* 75! It is really 60!

We won't discuss this any further, you will get more in college! But this is another example of how if you can master statistics, you can manipulate information to your advantage and BS those who do not understand statistics.

DISTRIBUTIONS

63. A 'distribution' relates to the 'shape' of your sample or population data when plotted in a histogram. In other words, what sort of pattern the data takes on and the frequency of different values of the data.

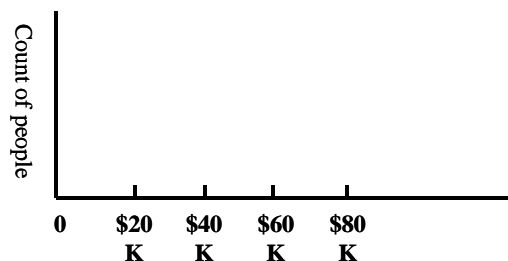
64. Distributions are normally given in percentages of data frequency vs data value.

64. For each of the following populations conjecture (ie: intelligently guess) and sketch a shape for the distribution.

State whether they **seem** 'symmetrical' or in which direction they are 'skewed'. Skewed distributions have a long thin tail where the data variable is skewed.

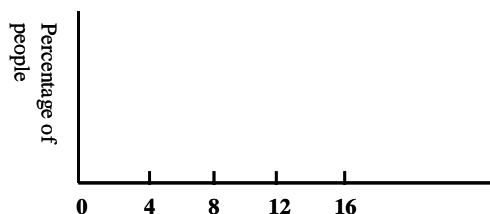
a. The annual income of people in Point Douglas.

Shape? _____
 Symmetrical? _____
 Skewed? _____



b. The number of hours of sleep for each adult Canadian over the past 24 hour period.

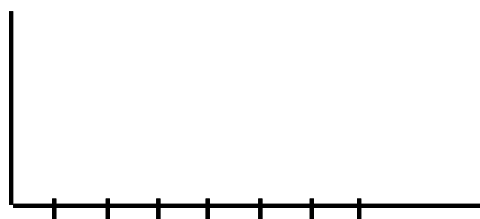
Shape? _____
 Symmetrical? _____
 Skewed? _____



(Normal Distribution}

c. The age of students in a regular high school. Compare it to Adult Ed school too!

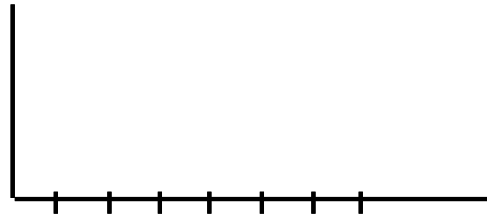
Shape? _____
 Symmetrical? _____
 Skewed? _____



d. The number of girls in a family with three children.

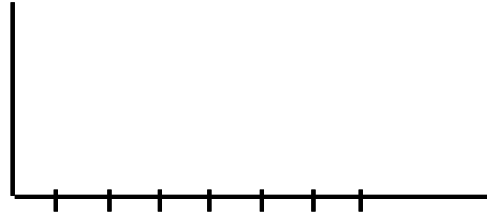
Shape? _____

Symmetrical? _____
 Skewed? _____
 This is a common distribution, it
 is called binomial.



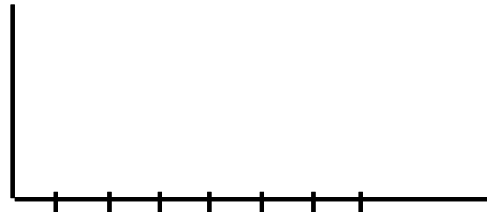
e. The number of litres of
 water consumed by each
 Winnipeg household in the last
 24 hour period

Shape? _____
 Symmetrical? _____
 Skewed? _____



f. The pay of employees and
 owners in the cities'
 MacDonalds' restaurants.

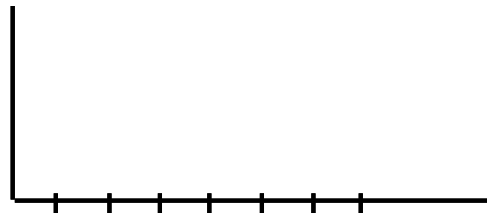
Shape? _____
 Symmetrical? _____
 Skewed? _____



[a bi-modal distribution]

g. The height of the cute little
 prairie dogs at the Assiniboine
 Zoo.

Shape? _____
 Symmetrical? _____
 Skewed? _____



*This is a very important
 distribution, it is called a normal
 distribution*

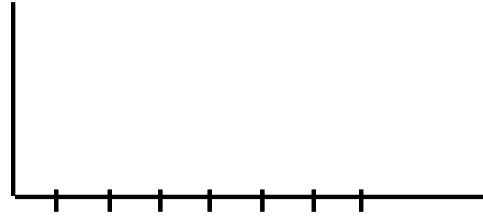
h. The count or frequency or probability of each digit from 0 to 9 on the license plates in the parking lot.

Shape? _____

Symmetrical? _____

Skewed? _____

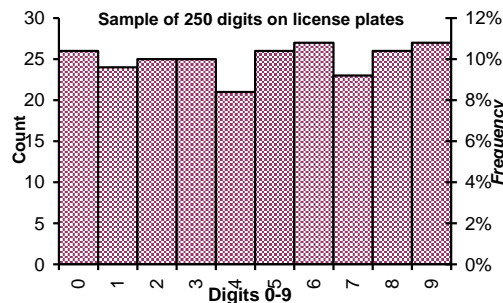
This is a boring distribution, it is called a uniform distribution.



THE UNIFORM DISTRIBUTION

65. The uniform distribution is as basic as there is. Every data value has the same frequency or percentage of occurrence; the probability of any particular value of the variable being selected in a sample are all equal.

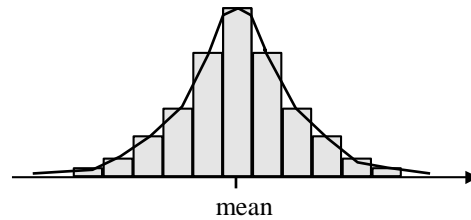
66. Here is a '*uniform distribution*', the value of the 250 different digits on about 75 cars in the parking lot. Each digit has about a 10% chance or frequency of occurring. If you were to take another sample on a different day you would likely get close to the same results. Notice this particular sample is discrete data, no such thing as a '2.3' digit.



THE NORMAL DISTRIBUTION

67. The normal distribution is the most common and most important. It helps find statistics as to how 'normal' something is or if it is a rare event. IQ test results are 'normally distributed', so are measurements in manufactured parts, so are school marks, people's shoe sizes and their heights and their weights, weights of newborns, navigation errors, ...

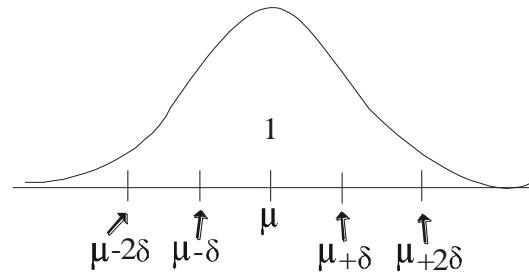
68. You may have heard about the 'bell' curve; that is actually just the normal distribution. Its histogram looks like a bell.



PROPERTIES OF A NORMAL DISTRIBUTION

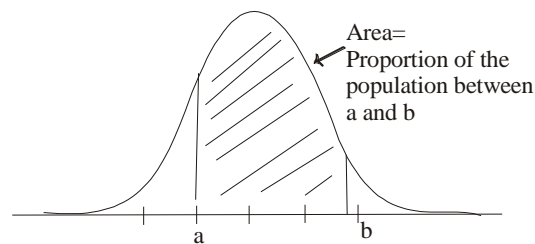
69. Every normal curve has a mean, μ , and a standard deviation, σ .

The graph or histogram is “**bell-shaped**” and **symmetric** about its mean. So the **mean, median, and mode are all exactly the same**.



The total area under every normal curve is 1.00 or 100% just as it is under any distribution measured in percentages.

70. The area under a normal curve between $x = a$ and $x = b$ equals the proportion of the population between a and b (or equivalently the probability that a data value in the population is between a and b). The mean, median and mode have the same value in a normal curve.



SAMPLE NORMAL DISTRIBUTION Source: Ministry of Health, B.C.
<http://www.vs.gov.bc.ca/stats/annual/1999>

71. The table and accompanying histograms show the birth weight for all babies born to B.C. mothers in 1999 by percentage.

Birth Weight (kg)	Male Frequency	Female Frequency
< 0.500	19	11
0.500 - 0.749	29	17
0.750 - 0.999	34	30
1.000 - 1.249	46	32
1.250 - 1.499	58	49
1.500 - 1.749	82	71
1.750 - 1.999	116	124
2.000 - 2.249	227	204
2.250 - 2.499	397	435
2.500 - 2.749	830	1 013
2.750 - 2.999	1 713	2 106
3.000 - 3.249	2 953	3 544
3.250 - 3.499	4 120	4 176
3.500 - 3.749	4 166	3 709
3.750 - 3.999	3 109	2 454
4.000 - 4.249	1 993	1 330
4.250 - 4.499	1 028	575
4.500 - 4.749	426	196
4.750 - 4.999	135	71
5.000 - 5.249	36	26
5.250 - 5.499	12	13
> 5.500	3	3
Total	21 532	20 189

Birth Weight (kg) - Males

Birth Weight (kg) - Females

a) The vertical scale axis on the histograms gives the sample percentage.

For each of the male and female populations calculate the sample percentage for the 3.500 - 3.749 kg birth weight interval.

$$\text{Male: } \frac{4166}{21532} = 0.1935 \qquad \text{Female: } \frac{3709}{20189} = 0.1837$$

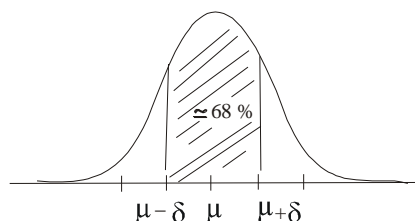
b) Are these histograms symmetric about their means? Comment.

Yes: Close to symmetric

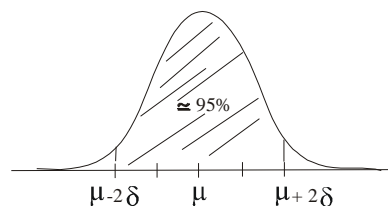
68-95-99 Rule

72. For every normal curve:

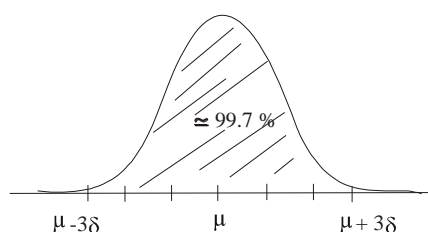
about 68% of the population is within
_____ standard deviation of the mean



about 95% of the population is within
_____ standard deviations of the mean.

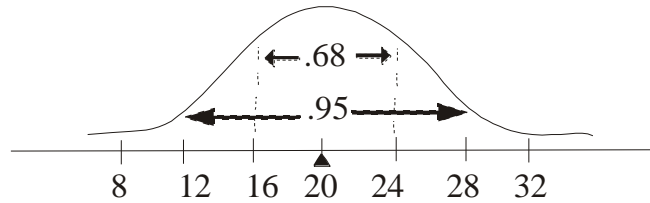


about 99.7% of the population is within
_____ standard deviations of the mean



Example :

73. For the population of Canadian high school students, suppose that the number of hours of TV watched per week is 'normally' distributed with a mean of 20 hours and a standard deviation of 4 hours.



a) Sketch a normal curve to show the distribution of the TV hours watched per week.

Mark the values on the horizontal axis that are 1, 2, and 3 standard deviations from the mean.

b) Approximately, what percentage of high school students watch

i) between 16 and 24 hours per week? = _____

ii) between 12 and 28 hours per week? = _____

iii) between 8 and 32 hours per week? = _____

iv) more than 20 hours per week? = _____

v) more than 28 hours per week? = _____

vi) fewer than 24 hours per week? = _____

The *Standard Normal Distribution*

74. The **normal** distribution occurs so often that we have developed a way to use it as a standard method of handling important statistics for many normally distributed populations with a minimum of effort.

75. The **standard normal distribution** is a special and particular member of the normal distribution family. In many applications it is used as a reference curve for finding 'standardized' values, probabilities, and percentages.

76. Different situations have different deviations. The deviation in the heights of prairie dogs might be $\sigma = 10 \text{ cm}$, whereas the deviation in the heights of ostriches might be $\sigma = 80 \text{ cm}$ because ostriches are bigger.

77. Instead of measuring how many cm a random variable measure, x , is from the mean, we can compare instead to how many standard deviations they are from the mean. So if $x - \mu$ is the distance of a

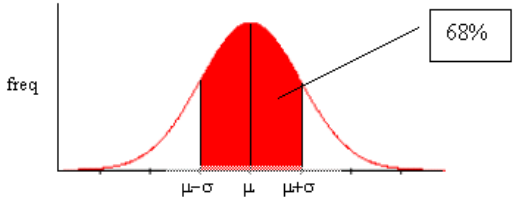
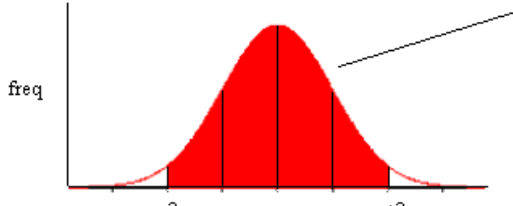
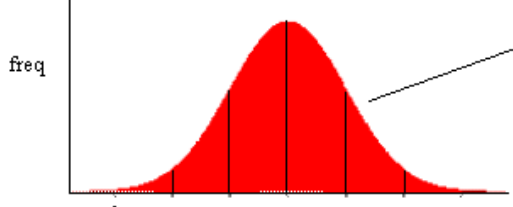
variable from the mean, x , then $\frac{(x - \mu) \text{cm}}{\sigma \text{cm}}$ would be measure of how many standard deviations away the datum was from the mean.

78. The calculation $\frac{(x - \mu)}{\sigma}$ is called the **normalized z score** of a distribution. It is a real number that tells how many standard deviations something is away from 'the normal' mean.

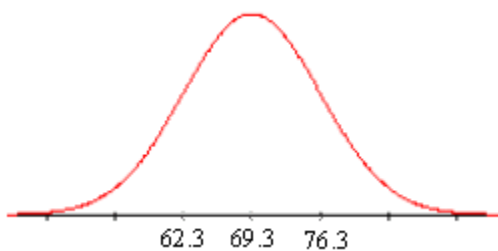
Thus: $z = \frac{(x - \mu)}{\sigma}$. It is a measure of how far from 'normal' or average something is.

79. The z score can be used with a normalized table of standard deviation to calculate how 'normal' and sample of data is. Only one table of information is required, you do not need a separate table for prairie dogs, ostriches etc.

CHARACTERISTICS OF THE NORMAL DISTRIBUTION (summary)

 <p>A normal distribution curve with the area between $\mu - \sigma$ and $\mu + \sigma$ shaded red. A callout box points to this shaded area with the text "68%". The x-axis is labeled with $\mu - \sigma$, μ, and $\mu + \sigma$. The y-axis is labeled "freq".</p>	<p>In a perfect Normal Distribution</p> <p>68% of all data is within $\pm 1\sigma$ of the mean, μ.</p>
 <p>A normal distribution curve with the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ shaded red. A callout box points to this shaded area with the text "95%". The x-axis is labeled with $\mu - 2\sigma$, μ, and $\mu + 2\sigma$. The y-axis is labeled "freq".</p>	<p>95% of all data is within $\pm 2\sigma$ of the mean, μ.</p>
 <p>A normal distribution curve with the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ shaded red. A callout box points to this shaded area with the text "99.7%". The x-axis is labeled with $\mu - 3\sigma$, μ, and $\mu + 3\sigma$. The y-axis is labeled "freq".</p>	<p>99.7% of all data is within $\pm 3\sigma$ of the mean, μ.</p>

80. The graph below represents the marks of a large number of students at the Grand Rapids high school where the mean mark is 69.3 percent and the standard deviation is 7 percent, and the distribution of marks is approximately normal.



81. We know that if the marks are approximately normally distributed, then approximately:

- * 68 percent of the marks are between 62.3 percent and 76.3 percent
- * 34% of the marks are between 69.3 percent and 76.3 percent (i.e., between the mean, $\mu + 1\sigma$)
- * 50% of the marks are below the mean, μ , of 69.3 percent
- * 16% of the marks are above 76.3 percent which is all that marks that are more than $\mu + 1\sigma$

82. But at what mark do the top 25% marks start? And how can you compare these marks at Grand Rapids with another school that may be marks easier?

83. That is the role of the **normalized standard distribution**, and the **Z-score**.

With a z-score we can use tables and computers much more easily. One table and one computer program will fit all data. It is just a matter of applying the following formula:

$$z = \frac{x - \mu}{\sigma}$$

84. This effectively makes the mean of the raw data from every data set zero, and finds a z score for every data point, x . But because now all means are reduced to zero and the deviation has been 'normalized' we can use one table or one computer program for all data.

The normalized standard distribution table is attached at Appendix 5

85. How to use the table. See Appendix 2 for how to read the Standard Normal Distribution Table

(There are many variations of this table in different books, and on the internet, they all achieve the same thing)

The table provides a percentage of the data points that are to the right of the mean for any given z-score.

That is it!

86. Example. Grand Rapids school has a Physics course. The mean mark last year was 75%. The standard deviation of the marks was 10%. The marks are distributed closely to a 'normal' distribution. The top 25% of all the marks were at or above what mark?

87. Solution. 50% of the marks are to the left of the mean. We are looking in the table for where 25% of the marks are at the right end of the tail of the normal curve.. That occurs at a z-score of 0.67 or 0.68. Now

$$z = \frac{x - \mu}{\sigma}$$

So $0.68 = \frac{x - 75}{10}$ so $x = 81.8$. So 25% of the marks are above 81.8.

88. If any student was picked at random as they walked up from the cafeteria, what is the probability that they had less than 60%.

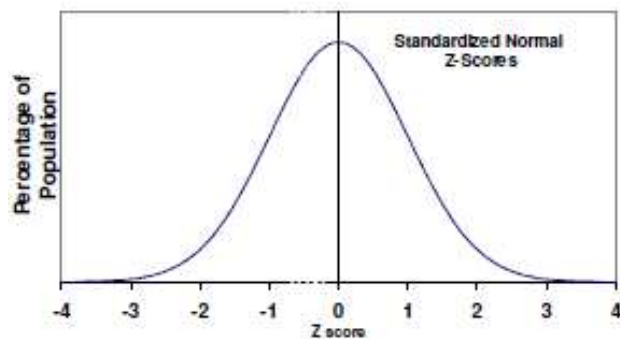
A failure is at 50%. The mean is still 75%. The Standard Deviation is still at 10%. So the z-score that represents all marks less than 50% is

$$z = \frac{60 - 75}{10} = -1.5$$

89. A z-score of -1.5 is not in the table! But we know that the normal curve is symmetrical, so if $+1.5$ means that 43.3% of students were between the mean and 1σ (10%), then 6.7% must be at the far right end of the tail. So, by symmetry, the far left of the tail below a z-score of -1.5 must also have 6.7% of the students. So, there is 6.7% chance that the next person to walk through the door has a score of less than or equal to 60%.

A Problem for you to try:

90. A camera battery has an average life of 1.5 years with a standard deviation of 3 months. Assume the distribution is normal. What is the probability that a randomly selected battery will last between 18 and 24 months? Sketch and shade a normal curve diagram to support your answer.



CONFIDENCE INTERVAL

Let's say you have been taking their has been a bus route to Dauphin for 50 years. The company advertises that it takes a mean of 225 minutes to get to Dauphin with a Standard Deviation of 15 minutes.

So if x is our variable for time: $\mu_x = 225$ and $\sigma_x = 15$. You realize this means that the bus over its entire history takes between 210 and 240 minutes to get to Dauphin 68% of the time.

Or you could calculate for 95% confidence (2σ) that the bus will take 210 ± 30 minutes to get to Dauphin. Or in other words $\text{Prob}(180 < x < 240)$ is 95%.

So you take the bus **one** weekend (out of the thousands the bus has run!) and find with one trip (sample) that the time is 230 minutes! Can you now write a letter to the paper complaining that the Bus Company is wrong?

It turns out that your sample of size one (one trip) compared to the thousands the bus company has measured is likely not as accurate as the bus company's. A proper statistic needs lots of measurements (samples). Notice if you did 10 runs and calculated a mean time, your mean, \bar{x} should be close to the population mean, $\mu_x = 225$. So how confident can you be in your measurement compared to the bus company?

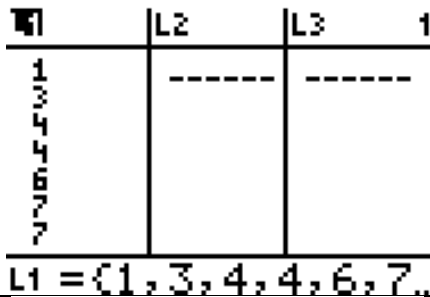
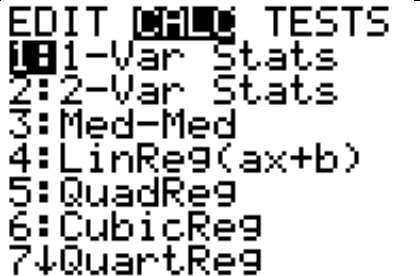
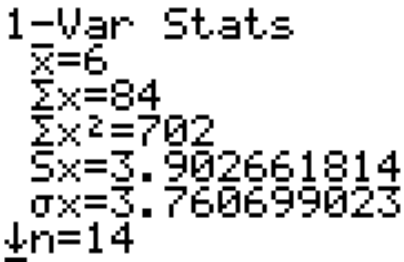
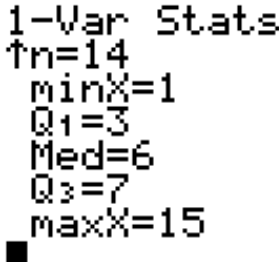
It turns out the formula to calculate how confident you can be in your limited sample is:

$$z = \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}}$$

Normally we calculate this confidence interval to a level of 95% confidence.

Teacher will explain further in class.

USING TECHNOLOGY

<p>There are multiple device and websites that will perform Statistical Calculations.</p> <p>This Appendix will demonstrate the Texas Instruments Device. Most Spreadsheets of course will also do the calculations</p>	
<p>Given the data set: {1, 3, 4, 4, 6, 7, 7, 7, 8, 2, 2, 6, 15, 12}</p> <p>a. determine the mean, median, mode from a single list b. determine the mean, median, mode from a frequency table b. make a histogram of the data from a frequency table c. determine the quartiles</p>	
<p>Enter the 14 data into the TI-83 into L₁ in the familiar STAT EDIT process</p>	
<p>Select</p> <p>STAT CALC 1-VAR STATS</p>	
<p>PRESS ENTER</p>	
	
<p>It is an extended screen in two parts, cursor down for all the values.</p>	

The mean is 6; the sum of all the data is 84, the sum of the each data element squared is 702 (we do not use that directly); the S_x and σ_x are measures of variability (sample deviation and population deviation) which we do not consider here, the **n, sample size**, is the number of data, the **minX** is X_{\min} , **Q1** is the first quartile, **Med** is the median (or Q2), **Q3** is the third quartile, and **maxX** is the X_{\max} of our data set.

ENTERING DATA FROM A FREQUENCY TABLE

Once you get past a couple dozen data values and given that many of them are the same it is very useful to enter data in a frequency table. Often the data has been grouped too into average class values.

Value	Frequency
2	3
3	6
4	9
5	11
6	4
7	5
8	1

Enter data into List 1 and List 2

```

L1      | 2 | L3      |
-----|---|-----|
2.5     | 3 |         |
3.5     | 6 |         |
4.5     | 9 |         |
5.5     |11 |         |
6.5     | 4 |         |
7.5     | 5 |         |
8.5     | 1 |         |
-----|---|-----|
L2 = {3,6,9,11,4,.

```

But this time make sure when you do the Stats Calculation that you tell the Calculator that the two lists you are using for data and for Frequency.

```

1-Var Stats L1,L2

```

```
1-Var Stats
 $\bar{x}$ =4.666666667
 $\Sigma x$ =182
 $\Sigma x^2$ =938
Sx=1.527525232
 $\sigma x$ =1.507814403
↓n=39
```

```
1-Var Stats
↑n=39
minX=2
Q1=4
Med=5
Q3=6
maxX=8
```