The purpose of these notes is to provide a concise overview of the material covered in MATH 1001 at Georgia State University. Note that the chapters are presented in order that they are presented in the course due to exam scheduling.

Chapter 12

12.1

Statistics is the methods of collecting, organizing, analyzing, and interpreting data, and draw conclusions based on the data.

Definition 1. *A **population** is a set of all people or objects whose properties are to be described and analyzed by the data collector.*

*A **sample** is a subset or subgroup of the population. While a **representative sample** is a sample that exhibits characteristics typical of those possessed by target population.*

Example 2. *We are conducting a survey among the citizens of the town of Titus to find their opinion on gambling.*

1. *Describe the population.*

   The population is the citizens of the city.

2. *Would a sample survey of people from the six largest nightclubs be a good idea?*

   No, people that frequent nightclubs may be more favorable to gambling and giving a skewed result.

Definition 3. *A **random sample** is a sample obtained in a way that every element in the population has the equal chance of being selected.*

Example 4. *Continuing the gambling example. Are the following good random samples?*

1. *You interview members of oceanfront condos.*

   No, we are picking people from one area, and they are more likely to have a similar opinion on gambling.

2. *You interview the first 200 people in the phone book.*

   No, we are not randomly picking people, since the phone book is alphabetical order, we are only picking people with A last names.

3. *You interview people by picking random neighborhoods and surveying random people from each.*

   Yes, everyone has an equal chance at being picked.

For this section, we are interested in different ways data can be presented to us. The first is called a **frequency distribution** which achieves the following:

1. takes a set of data items where some may be duplicates.
2. counts how many times a data item occurs in the data set.

For example consider the following data set that contains the year a child reached its maximum growth as a frequency distribution chart. Note that this chart can be achieved by counting the number of times each age is repeated in the data set.

Age	Frequency
10	2
11	3
12	6
13	8
14	10
15	7
16	4
17	2
18	2

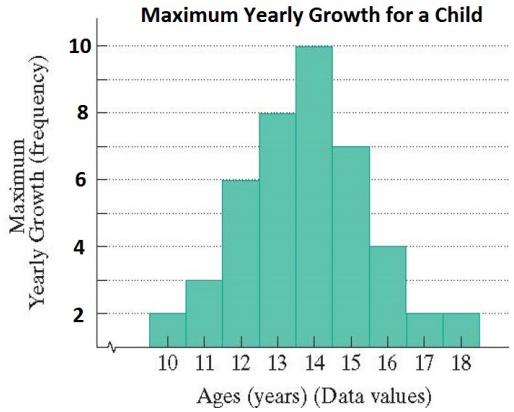
Meaning that 10 children reached their maximum growth at age 14.

Another type of frequency distribution is a grouped distribution which occurs when we group our data into ranges. For each group, the lower class limit is the left most value and the upper class limit is right most value. For example consider the following example of grades on an exam.

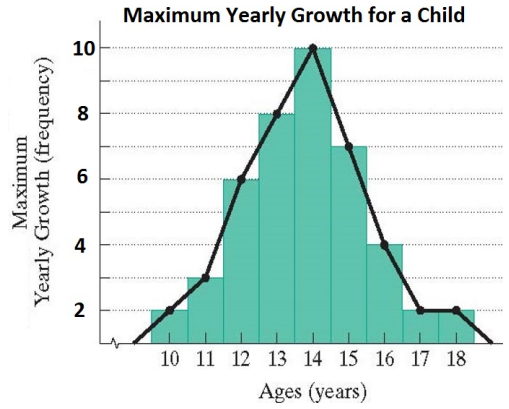
Range	Frequency
90-99	13
80-89	14
70-79	20
60-69	14
0-59	10

In this example, 0, 60, 70,80, and 90 are the lower class limits and 59,69,79,89, and 99 are the upper class limit. Another thing we like to look at is called the **class width** which is the difference between the upper and lower class limits. Note, in our example all the widths a 10 with the exception of the group 0-59 which is 60.

Another type of way that information can be displayed is in a **histogram** which is a bar graph where each bar is touching. The following is an example for the data on maximum growth heights for children.



The last way information can be displayed is as a **frequency polygon** which connects the midpoints of each bar of the histogram.



12.2

In this section we look at central tendencies of data which include mean, median, mode, and midrange.

Definition 5. *The **mean** of a data set is found by adding all of the data items and dividing by the total number of items.*

$$\textit{Mean} = \overline{x} = \frac{\sum x}{n}$$

where $\sum x$ is the sum of all data items and n is the number of items in the data set.

Example 6. *Find the mean of the data set* $S - \{92, 84, 84, 80, 78, 77, 77, 77, 75, 75\}$.

$$\overline{x} = \frac{\sum x}{n} = \frac{92 + 84 + 84 + 80 + 78 + 77 + 77 + 77 + 75 + 75}{10} = \frac{799}{10} = 79.9$$

therefore the mean of the data set is 79.9.

In the case that we are dealing with a frequency data set, the mean formula changes slightly to

$$\text{Mean} = \overline{x} = \frac{\sum xf}{n}$$

where $\sum xf$ is the sum of all data items times there corresponding frequency and n is the number of items in the data set.

Example 7. *Find the mean of the following data set*

Age	*Frequency*
10	*2*
11	*3*
12	*6*
13	*8*
14	*10*
15	*7*
16	*4*
17	*2*
18	*2*

$$\overline{x} = \frac{10(2) + 11(3) + 12(6) + 13(8) + 14(10) + 15(7) + 16(4) + 17(2) + 18(2)}{44} = \frac{608}{44} \approx 13.8$$

Another central tendency we like to consider is the **median** which splits the data. Medians are found in the following way

1. Arrange the data items from smallest to largest.
2. If the number of items is odd, then the median is the middle item on the list.
3. If the number of items is even, then the median is the average of the middle two items.

Example 8. *Find the median of the data set* $S = \{84, 90, 98, 95, 88\}$
Reordering the set we have $84, 88, 90, 95, 98$ *and since there are five items, then the median is the middle. Here the median is* 90.

Example 9. *Find the median of the data set* $S = \{84, 90, 98, 95, 88, 96, 98, 91\}$
Reordering the set we have $84, 88, 90, 91, 95, 96, 98, 98$ *and since there are eight items, then the median is the average of the middle items. Here the middle two items are 91 and 95, which average to (91+95)/2=93.*

In the case that we are dealing with frequency distributions, you can use the following to find the median. If n data items are in order, then the median is in the $(n + 1)/2$ spot. Note, that if n is odd, then $(n + 1)/2$ is a whole number. However if n is even, then $(n + 1)/2$ is not an integer, it will have a decimal of .5 so you will need to find the average of the middle two items.

For example, if $n = 18$, then $(n + 1)/2 = 9.5$ so you need to average the items in the 9th and 10th spot!

Example 10. *Six people have the salaries $19,700, $20,400, $21,500, $22,600, $23,000, and $95,000. Notice that the mean for these salaries are $33,700 while the median is $22,050. This large discrepancy is that the salary of $95,000 skews the mean towards the larger side.*

The third central tendency is the **mode** of the data set which is found by finding the data item that occurs the most. You can have no mode or more than one mode.

Example 11. *Given the data set 7, 2, 4, 7, 8, 10, we see that the mode is 7 since 7 occurs twice.*

Given the data set 1, 3, 5, 7, 2, we see that there is no mode since every item occurs once.

Given the data set 3, 3, 5, 6, 7, 6, we see that there are two modes, 3 and 6 since they both occur twice. Here we would call the data set bimodal.

The final central tendency we like to look at is **midrange** which is found by

$$\textbf{midrange} = \frac{\text{smallest item + largest item}}{2}.$$

Example 12. *Given the data set $S = \{14, 16, 18, 20, 15, 19, 25, 26, 22, 15\}$, find the midrange. Notice that the smallest item is 14 and the largest item is 26, thus the midrange is found by*

$$\textit{midrange} = \frac{14 + 26}{2} = \frac{40}{2} = 20.$$

12.3

In this section we look at measures of dispersion which describe the spread of the data.

One of the measures of dispersion which we like to look at is **range** which is found by finding the difference between the highest and lowest data item in our data set.

Example 13. *Given the data set $\{3, 4, 7, 9, 2, 1, 10\}$ we can find the range by identifying the smallest value 1 and the largest value 10 first. Next the difference $10 - 1 = 9$ is the range.*

Definition 14. *The second measure of dispersion is called the standard deviation, however it relies on **deviation from the mean**.*
Found by determining how much each data item differs from the mean. Note that this depends on the item! As a formula we could say

deviation for the mean = data item - mean.

Example 15. *Find the deviation from the mean of 70, 69,68,65.*

First we need to find the mean of the five values, here it is $\overline{x} = 68$. Next we find the deviations.

Item	*Deviation from Mean*
70	*70-68=2*
69	*69-68=1*
68	*68-68=0*
65	*65-68=-3*

Notice that if the item is larger than the mean that the deviation is positive, while if it is smaller than the mean it is negative.

NOTE: The sum of the deviations from the mean across all the items is zero!

Now, we can think of the standard deviation as an average of the deviations from the mean. To compute the standard deviation, we follow the following steps.

1. Find the mean of the items.
2. Find the deviation from the mean for each item
$$\text{data item } - \overline{x}$$
3. Square each deviation
$$(\text{data item } - \overline{x})^2$$
4. Sum all of the squared deviations from the mean
$$\sum(\text{data item } - \overline{x})^2$$
5. Divide the previous by $n - 1$
$$\frac{\sum(\text{data item } - \overline{x})^2}{n - 1}$$

6. Take the square root, note that most of the time this is denoted by s

$$s = \sqrt{\frac{\sum(\text{data item } - \overline{x})^2}{n-1}}$$

Example 16. *Find the standard deviation for 70, 69,68,65.*

First we need to find the mean of the five values, here it is $\overline{x} = 68$. *Next we find the deviations.*

Item	*Deviation from Mean*	*Deviation*^{2}
70	*70-68=2*	*4*
69	*69-68=1*	*1*
68	*68-68=0*	*0*
65	*65-68=-3*	*9*
Total	*0*	*14*

Next we divide the sum of the deviations squared by $n - 1 = 4 - 1 = 3$.

$$\frac{\sum(\textit{data item } - \overline{x})^2}{n-1} = \frac{14}{3}.$$

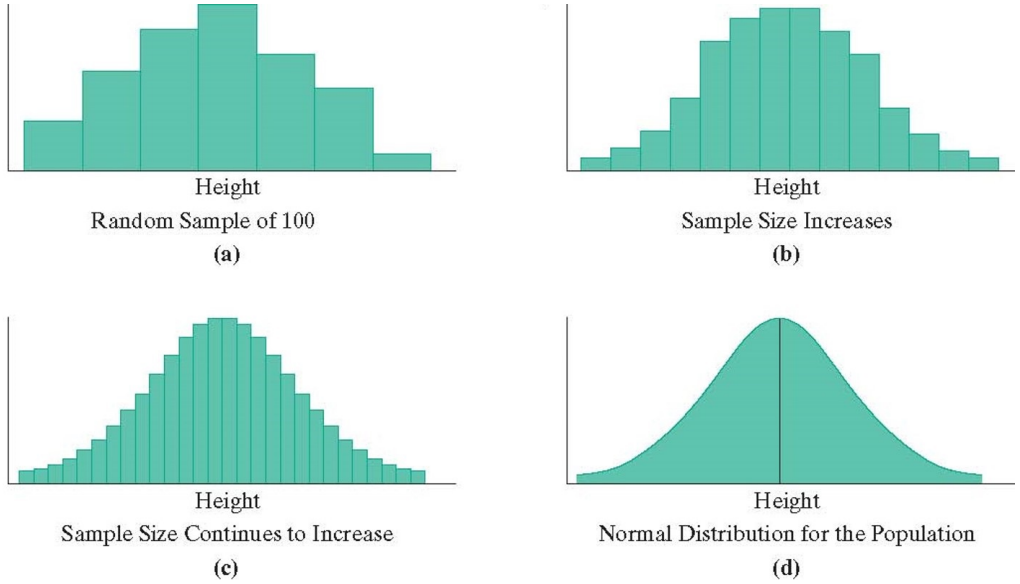
Finally we take the square root

$$s = \sqrt{\frac{\sum(\textit{data item } - \overline{x})^2}{n-1}} = \sqrt{\frac{14}{3}} \approx 2.67$$

The standard deviation measures the variation of the data items. Moreover, the smaller the standard deviation, the closer the data items are too each other while the larger the standard deviation, the more spread out the data is.

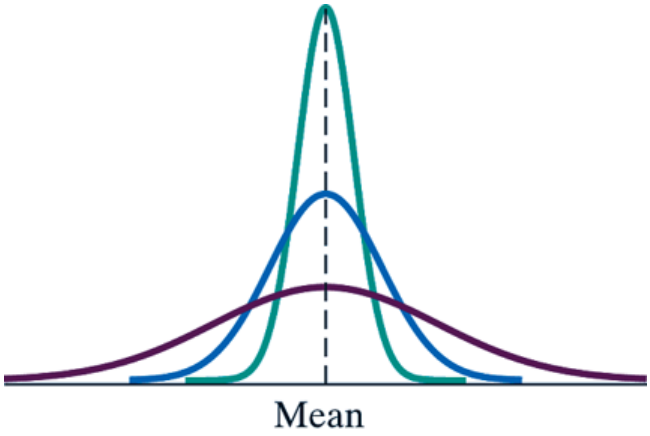
12.4

A type of distribution we like to talk about is the **normal distribution** also called a bell curve due to its shape. It is symmetric with respect to a vertical line through the center and interestingly, this line passes through the mean, median, and mode.



In a lot data collection we are not able to sample the entire population and the normal distribution comes from when we are able to sample the entire population.

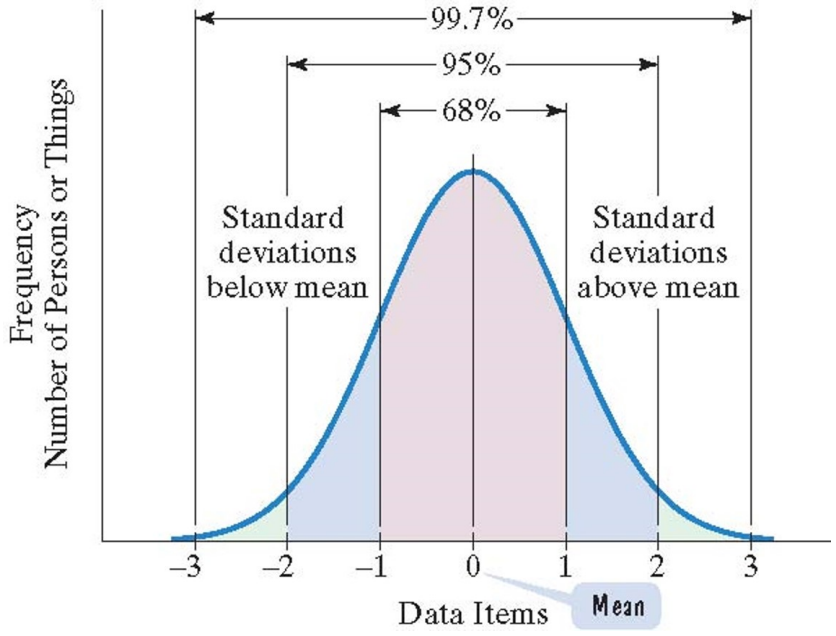
As the standard deviation increases the bell curve starts to flatten out, see the following:



In this example, all three curves have the same mean, the only difference is that green is a small standard deviation (all the data is very close to the mean) and the purple has a larger standard deviation (more data is away from the mean).

The most important rule for the normal distribution is the 68-95-99.7 Rule which states that, see the following figure:

1. Approximately 68% of the data items lie within one standard deviation of the mean in both directions.
2. Approximately 95% of the data items occur within two standard deviations in both directions.
3. Approximately 99.7% of the data items occur within three standard deviation in both directions.
4. Anything happening out side of this is called an outlier.



Some applications:

Example 17. *The male heights in North American are approximately normally distributed with a mean of 70 in and a standard deviation of 4 in. Find the height that is 1.5 standard deviations above and 2 standard deviations below the mean.*

1.5 Standard Deviations above, so add

$$70 + 1.5(4) = 76in$$

2 Standard Deviations below, so subtract

$$70 - 2(4) = 62in$$

A **z-score** describes how many standard deviations a data item in a normal distribution lies above or below the mean.

$$\text{z-score} = \frac{\text{data item-mean}}{\text{standard deviation}}.$$

1. z-score > 0, lies above the mean.
2. z-score $= 0$, lies at the mean.
3. z-score < 0, lies below the mean.

Example 18. *The mean weight of a newborn is 7 lbs with a standard deviation of .8 lbs. If the weights of the babies are distributed normally, find the z-score of a 9 lbs, 7 lbs, and a 6 lbs child.*

$$z_9 = \frac{9-7}{.8} = \frac{2}{.8} = 2.5$$

so a 9 lb baby lies 2.5 standard deviations above the mean.

$$z_7 = \frac{7-7}{.8} = \frac{0}{.8} = 0$$

so a 7 lb baby lies at the mean.

$$z_6 = \frac{6-7}{.8} = \frac{-1}{.8} = -1.25$$

so a 6 lb baby lies 2.5 standard deviations above the mean.

Besides z-scores, percentiles and quartiles measure a data item's position in the distribution.

An item is in the n percentile means that it larger than $n\%$ of the other data items. The quartiles, are just specific percentiles. Meaning the first quartile is the 25th percentile, second quartile is the 50th percentile, and the thrid quartile is the 75th percentile.

Example 19. *If a student scores in the 75th percentile, that it means that they scored better than 75% of the other students who took the test.*

Definition 20. *If n is the sample size, the margin for error is*

$$\pm\frac{1}{\sqrt{2}} \times 100\%.$$

Note, represents two standard deviations above and below the mean.