**Math 11 – Statistics**

**Chapter 3 – Measures of Central Tendency for ungrouped data**

3 different measures

Mean (what we normally think of as average)

: population  sample 

Median – the middle term  (This gives the position of middle term)

Mode: The term that occurs the most often.

Unimodal – a single value occurs most times

Bimodal – two terms occur the maximum number of times

Multimodal – many values occur the maximum number of times

Shapes of distributions and the affect on position of mean, median and mode

Median

Mean

Mode

Mean

Median

Mode

Median

Mode

Mean

Measures of Dispersion for ungrouped data

**The Range** 

**Standard Deviation** – the positive square root of the variance

**Variance -** population: 

Sample: 

 or  is called the deviation of the  value from the mean.

If we add all the deviations from the mean we get zero.

 

Therefore, we standardize by squaring and then taking the square root.

 

There are shortcut formulas for the variance – you can use whichever is most convenient for the problem.



The variance and the standard deviation are both positive.

 because they are squared

 because they are the positive square roots



The numerical measure ( mean, median, mode, variance, or standard deviation ) of a population is called a **population parameter** or just a **parameter.** 

The numerical measure of a sample is called a **sample statistic** or just a **statistic** 

**3.3 Measures of Central Tendency for Grouped Data**

Mean, Variance and Standard Deviation of Grouped Data

When data is already grouped, we don’t know the individual values of the variables.

ex The following table gives the grouped data on the weights of all 100 babies born at a hospital in 2005. (Notice, this is the total population)

|  |  |
| --- | --- |
| **Weight (pounds)** | **# of babies** |
| 3 to less than 5 | 5  Notice that this is continuous data, therefore the class limits are the same as the class boundaries. |
| 5 to less than 7 | 30 |
| 7 to less than 9 | 40 |
| 9 to less than 11 | 20 |
| 11 to less than 13 | 5 |

To calculate the mean we will have to assume that observations in each class are evenly distributed throughout the class. Then the average value of the class would be the midpoint.

So, find the midpoint of each class and multiply it by the frequency of the class (i.e. the number of observations in that class) This gives us the approximate total of that class. Next – add the totals together to get the sum of all values in all classes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Weight (pounds)** | **Frequency (# of babies)**  ***f*** | **Midpoint of class**  ***m*** | **Approx total of values in class ( *mf* )** |
| 3 to less than 5 | 5 | 4 | 20 |
| 5 to less than 7 | 30 | 6 | 180 |
| 7 to less than 9 | 40 | 8 | 320 |
| 9 to less than 11 | 20 | 10 | 200 |
| 11 to less than 13 | 5 | 12 | 60 |
|  | . |  | . |

The mean now becomes  for a population and 



The mean weight of babies born in 2005 was about 7.8 pounds.

Variance and Standard Deviation for Grouped Data

 or shortcut 

or shortcut 

Back to babies

We need to find  and 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***weight*** | ***f*** | ***m*** | ***mf*** | ***m2f*** |
| 3 to less than 5 | 5 | 4 | 20 | 80 |
| 5 to less than 7 | 30 | 6 | 180 | 1080 |
| 7 to less than 9 | 40 | 8 | 320 | 2560 |
| 9 to less than 11 | 20 | 10 | 200 | 2000 |
| 11 to less than 13 | 5 | 12 | 60 | 720 |
|  |  |  |  |  |

Substitute into the formula

 This is the variance.

 pounds

When using single valued classes, the value of the class will be used as ***m.***

ex This table gives the number of errors committed by a college baseball team in all 45 games played in the 2005-2006 season. (Note – this is a population)

Find the mean, variance and standard deviation

|  |  |  |  |
| --- | --- | --- | --- |
| ***# of errors*** | ***# of games (f)*** | ***mf*** | ***m2f*** |
| 0 | 11 | 0 | 0 |
| 1 | 14 | 14 | 14 |
| 2 | 9 | 18 | 36 |
| 3 | 7 | 21 | 63 |
| 4 | 3 | 12 | 48 |
| 5 | 1 | 5 | 25 |
|  |  |  |  |

Mean:  errors per game.

Variance: 

Standard Deviation: 

3.4 How do we use Standard Deviation?

When we have a symmetric bell-shaped curve (i.e. unimodal) then we can use ***The Empirical Rule.***

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.7% of the observations lie within three standard deviations of the mean

**95%**

**68%**

img18

      

**99.7%**

* 1. **Measures of Position**

Position means where in the data set a specific value falls.

**Quartiles** are summary measures that divide the data set into 4 equal parts.

(The data must be ranked first.)

25%

25%

25%

25%

2nd Quartile

1st Quartile

3rd Quartile

Q3

Q2

Q1

Q2 = the median

IQR – Interquartile Range 

Let’s look at the kitten data again. The data must be ranked first

37 38 40 41 44 45 49 51 52 52 54 56 57 57 58 60 63 63 65 69 72 76 77 84 91

Q2 = the median which is 

So the 13th entry is 57

To find Q1 – find the median of the lower half:  So halfway between the 6th and 7th items



The interquartile range is 

5 number summary: min, , ,, max

We can see that 50% of the kittens weighed between 47 and 67 grams

25% weighed less than 47 grams, and 25% weighed more than 67 grams.

Where is 60 grams located? Between 50% and 75% or in the 3rd quartile.

**Box and Whisker Plots**

Definition: A plot that shows the center, the spread, and the skewness of a data set.

The following 5 steps are used to construct a box-and-whisker plot.

Back to kitten weights

1. Find the median = 57

  IQR 

1. Find the points that are  below , and  above .

These are called the **lower inner fence** and the **upper inner fence** respectively.

For the kittens 

Lower Inner Fence () 

Upper Inner Fence () 

1. Determine: the smallest value within the two inner fences kittens – 37

The largest value within the two inner fences kittens – 91

1. Draw a horizontal line and mark observation values on it so that all data in the set are covered.

| | | | | | | | | | | | |

35 40 45 50 55 60 65 70 75 80 85 90 95

1. Draw a box that has Q1 as one side and Q3 as the other. Draw a vertical line at Q2.

whisker

whisker

This shows the kitten data is skewed to the right.

(the right whisker is longer)

Make a mark at the lowest value inside the lower inner fence. Make a mark at the largest value inside the upper inner fence. Connect these to the box. These are the whiskers

Any values outside the two inner fences would be indicated by asterisks. These would be outliers.

For the kittens there are none outside the fences. So no outliers for the kittens.

Another example:

Problem 30 page 114 This is the number of runners left on bases by each of 30 Major league teams.

15.5

8th

3 4 4 5 5 5 5 5 6 6 6 6 6 6 6 7 7 8 8 8 8

8th

8 8 9 9 10 10 11 13 18

Place of Q2 Remember this is the place of the value.



Place of Q1 Therefore 

Place of Q1 Therefore, 

IQR 

Lower inner fence: 

Upper inner fence: 

Notice, there are two values not within the fences: 13 and 18. These are outliers and we represent them with asterisks

\*

\*

We can see from the box and whisker plot that the data is skewed right with two outliers.

| | | | | | | | | | |

0 2 4 6 8 10 12 14 16 18 20