Statistics

Frequency Distribution and Measures of Central Tendency

Copyright
This publication © The Northern Alberta Institute of Technology 2002. All Rights Reserved.

LAST REVISED June, 2007
Statistics: Frequency Distribution and Measures of Central Tendency

Statement of Prerequisite Skills
Complete all previous TLM modules before beginning this module.

Required Supporting Materials
Internet Explorer 5.5 or greater.
Macromedia Flash Player.

Rationale
Why is it important for you to learn this material?
Statistical applications occur in nearly every technology. Curve fitting, research and development, analyzing engineering problems, chemical levels in emissions, analysis of electrical circuits and production of goods are a few examples of statistical applications.

Learning Outcome
When you complete this module you will be able to…
Determine the frequency distribution and measures of central tendency of a set of data.

Learning Objectives
1. Understand the requirements for Collecting Data
2. Calculate measures of Central Tendency, Mean, Median, Mode
3. Calculate the Range and the Standard Deviation

Connection Activity
Recall the last time you received a grade on an exam. It is natural to ask how the rest of the class did on the exam so that you can put your grade in some context. Knowing the mean or median tells you the "center" or "middle" of the grades, but it would also be helpful to know some measure of the spread or variation in the grades.
OBJECTIVE ONE

When you complete this objective you will be able to…

Understand the requirements for Collecting Data

Exploration Activity

Statistics is a branch of mathematics that is useful for collecting, organizing, analyzing and interpreting numerical information called data.

Statistics helps us make decisions and predictions about the whole of a set of items using only a small group of items. The whole set of items is called the population and the small group of items is called the sample. A sample represents the population.

To determine the popularity of a political party in the country, it would not be very practical to ask the opinion of everyone in the country. Hence a sample of the population is surveyed for their opinion.

The most important requirement of a sample is that it must be random. This means that each item of the population has an equal chance of being selected for the sample.

The results of the political party survey using a sample of the people in the country would then be organized and analyzed using graphs, tables, and measures like the mean, median, and mode.

Based on the analysis of the data, decisions would be made regarding the popularity of the political party. In what part of the country is the party the strongest and where is it the weakest? If an election were called would the party gain or lose seats? What makes the party popular? What doesn’t?

In any statistical process it is very important how the sample is chosen. A badly chosen sample will only result in poor predictions. A representative sample of a population must contain the relative characteristics of the population in the same proportion. For example the people in Alberta do not represent the whole country. Hence to draw conclusions for the whole country from only the opinions of Albertans would be wrong. In order to make a fair assessment of the political party in the whole country, the sample for each province should be proportional to the people in the province.

To obtain information about a population (not necessarily people) the following methods can be used:

- questionnaires
- telephone surveys
- personal interviews
- measurements
- tests
EXAMPLE 1
Out of 120 drivers surveyed in Edmonton, 72 said they wore seatbelts when driving. If there are 245 000 drivers in the city, approximately how many are wearing seatbelts when driving?

Solution: This is a ratio and proportion problem. We set up the following ratio:

\[
\frac{72}{120} = \frac{X}{245000}
\]

Solving for \(X\), we get

\[
X = \frac{72 \times 245000}{120} = 147000
\]

Therefore, about 147 000 drivers wear seatbelts.

EXAMPLE 2
One way to estimate the fish population of a pond is to use the "capture - recapture" technique. 100 fish were captured, tagged, and then returned to the pond. A week later, 60 fish were caught, 4 of which wore tags. How many fish, approximately, are in the pond?

Solution: Let \(X\) be the number of fish in the pond. Then since 100 fish were initially captured and tagged, the ratio 4/60 must be the same as 100/\(X\). Hence solving for \(X\) we get

\[
X = \frac{60 \times 100}{4} = 1500
\]

Hence the pond contains approximately 1500 fish.
Experiential Activity One

1. List three examples of sampling where taking of the sample destroys the product.

2. The table below shows the results of an election survey in Edmonton, which has 380,000 potential voters. Six hundred people were asked to name their favorite candidate, with the following results.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>62</td>
</tr>
<tr>
<td>Rutger</td>
<td>121</td>
</tr>
<tr>
<td>Johnson</td>
<td>183</td>
</tr>
<tr>
<td>Mulder</td>
<td>71</td>
</tr>
<tr>
<td>Parsons</td>
<td>163</td>
</tr>
</tbody>
</table>

If an election were held after the survey was taken, approximately how many votes would each candidate receive? Show Me.

3. A stratified sample resembles a scaled-down version of the population in which categories have proportional representation. A school has the following numbers of students:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>245</td>
</tr>
<tr>
<td>11</td>
<td>234</td>
</tr>
<tr>
<td>12</td>
<td>213</td>
</tr>
</tbody>
</table>

If a survey of 100 students is to be stratified, approximately how many students from each grade should be included in the survey?

Experiential Activity One Answers

1. Testing battery lifetime.
   Tensile test of steel sample.
   Cleaning strength of soap.
   Can you think of more?

2. Smith 39 267
   Rutger 76 633
   Johnson 115 900
   Mulder 44 967
   Parsons 103 233.

3. Grade 10: 35
   Grade 11: 34
   Grade 12: 31
OBJECTIVE TWO

When you complete this objective you will be able to…

Calculate measures of Central Tendency, Mean, Median, Mode

Exploration Activity

The Mean, the Median and the Mode are called measures of central tendency

The Mean: the average

The mean, or arithmetic average, of a set of numbers is found by adding the numbers, and then dividing the resulting sum by the number of numbers added.

For the set of values \( x_1 + x_2 + x_3 + \ldots + x_n \), the mean is:

\[
\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}
\]

EXAMPLE 1:

Cathy received the following marks for lab assignments: 64, 78, 73, 82, 84, 78, 68

The mean mark is found as follows:

\[
\bar{x} = \frac{64 + 78 + 73 + 82 + 84 + 78 + 68}{7} = \frac{527}{7} = 75.3
\]

The Median: the half-way

When a set of numbers is arranged in order from largest to smallest, or from smallest to largest, the median is the middle number.

Cathy's marks 64, 78, 73, 82, 84, 78, 68 arranged in ascending order are 64, 68, 73, 78, 78, 82, 84

The median is 78.

If there is an even number of numbers, then the median is the average of the two middle numbers.

For the values 24, 28, 36, 39, 40, 42, the median is the average of 36 and 39.

\[
\text{median} = \frac{36 + 39}{2} = 37.5
\]

Frequency Distributions and Measures of Central Tendency
The MODE: the most common

The mode of a set of numbers is the number that occurs most often. For Cathy's marks, 64, 68, 73, 78, 82, 84, the mode is 78.

Some sets of numbers do not have a mode. For example the set of numbers 45, 46, 56, 58, 67, 78, 54, 72 does not have a mode.
Some sets of numbers have more than one mode.

For example the set of numbers 24, 26, 26, 34, 36, 37, 38, 38, 42, 45 has two modes.

When is it more meaningful to use the mean, the median or the mode?

MEAN: If a newspaper wanted to know how much paper is used, the average number of pages per paper per day is the most meaningful value.

MEDIAN: The median income of families in Alberta is usually more meaningful than the mean because very high and very low incomes tend to distort the mean.

MODE: A shoe store is more interested in the mode of shoe sizes bought than the mean. The owner must know which size sells the most and order accordingly.

CALCULATOR MATH

To find the mean on the calculator, using the SHARP EL–531, proceed as follows:
1. Put the calculator in STAT MODE 1.
2. Key in the data by keying in the number followed by the $M+$ key. After each entry you will see $n=$ displayed indicating the number of entries (sample size).
3. When all the data are keyed in, pressing $RCL$ and then $\hat{\mu}$ will display the mean.

Note: Make sure to clear the calculator before entering a new set of data. Press the $CA$ key to clear.
EXAMPLE 2

A manufacturer of light bulbs tested the lifespan of light bulbs and obtained the following results:
(The numbers represent months)
34, 36, 42, 39, 46, 40, 45, 39, 47, 48, 44, 45, 51, 38, 38, 39, 44, 43, 39, 39, 33, 45, 49, 46, 43, 25, 42, 45

Determine
a) the mean
b) the median
c) the mode

SOLUTION:

a) The mean is given by

\[ \bar{x} = \frac{\text{Total}}{31} = \frac{1265}{31} = 40.81 \]

b) The median is the middle number. Since there are 31 numbers, the median is the sixteenth number. The median is 40.

c) The mode is the number which occurs most often. The mode is 39.

Therefore the mean lifetime of the light bulbs is 40.81 months, the median is 40 months and the mode is 39 months.
Experiential Activity Two

1. Calculate the mean of the following sets of numbers to one decimal place.
   a) 4, 6, 6, 8, 3, 5, 9, 4, 7, 2, 4
   b) 4, −3, −5, 4, −7, −2, 6, 0, −3, 5
   c) 8.3, 9.2, 4.8, 5.7, 12.2, 8.9, 7.5, 6.3
   d) 1203, 1803, 4095, 3002, 6281, 4000, 1987, 5672

2. Find the median of each of the following sets of numbers.
   a) 5, 7, 7, 9, 10
   b) 5, 7, 12, 9, 6, 8, 13, 10
   c) −4, −2, 0, 6, 12, 12, 14, 15
   d) 89, 67, 34, 99, 48, 67, 48

3. Find the mode (or modes) of each of the following sets of numbers.
   a) 30, 34, 34, 38, 39, 39, 42, 44, 39, 42
   b) 62, 67, 78, 62, 69, 56, 45, 48
   c) 18, 14, 19, 13, 14, 23, 25, 14, 23, 20, 21, 18, 22
   d) −6, −3, −3, −2, 0, 0, 3, 4, 6, 6, 9, 12

4. The City Parks Department employs 150 students each summer. Fifty students earn $6.00/hour, forty students earn $7.00/hour, 30 students earn $7.50/hour, and 30 students earn $8.50/hour. Find the mean, median, and mode of the given data.

5. Calculate the mean, median, and mode for the following data.
   \[
   \begin{array}{|c|c|}
   \hline
   \text{Value} & \text{Frequency} \\
   \hline
   23 & 2 \\
   25 & 4 \\
   26 & 6 \\
   28 & 3 \\
   30 & 2 \\
   \hline
   \end{array}
   \]

6. The following are the exam marks for a chemistry class.
   72, 65, 78, 45, 56, 90, 68, 77, 34, 67, 56, 78, 71, 95, 48, 63, 60, 33, 82, 92, 32, 32, 67, 73
   Determine the mean, median, and mode of the exam marks.  

Experiential Activity Two Answers

1. a) 5.3, b) −0.1, c) 7.9, d) 3505.4
2. a) 7, b) 8.5, c) 9, d) 67
3. a) 39, b) 62, c) 14, d) -3, 0, 6
4. Mean = 7.07, Median = 7.00, Mode = 6.00
5. Mean = 26.2, Median = 26, Mode = 26
6. Mean = 65.3, Median = 67, Mode = 56, 67, 78

Frequency Distributions and Measures of Central Tendency
OBJECTIVE THREE

*When you complete this objective you will be able to…*

The Range and the Standard Deviation

**Exploration Activity**

In the previous section we have seen that the mean, the median, and the mode describe the central tendencies of the data. However they do not tell as how the data are spread out.

Consider the following sets of data:

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
</tr>
</tbody>
</table>

For both sets of data, the mean = 7, the median = 7, and the mode = 7. Note that set B is more spread out than set A.

The amount to which data are spread out is called the *dispersion*. These are several ways to measure dispersion. The simplest measure of dispersion is called the *range*.

\[
\text{range} = (\text{largest number}) - (\text{smallest number})
\]

For the Set A shove, the range is \(8 - 6 = 2\)
For the Set B above, the range is \(11 - 1 = 10\)

Although the range is easy to calculate, it is not really a good measure of dispersion. It is affected by the extreme values in the set.
**Standard Deviation:**

The most widely used measure of dispersion is called the **standard deviation**. The standard deviation is a measure that tells us how the data is spread out from the mean.

**Standard Deviation (Population)**

The symbol for the standard deviation commonly used is $\sigma$ (pronounced sigma) and is calculated using the formula:

$$
\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}}
$$

where $\bar{x}$ is the mean of the data, and $n$ is the number of values in the data.

**Standard Deviation (Sample)**

The formula for calculating the standard deviation of a sample of a population is the same as the formula for calculating the standard deviation of the whole population except that we divide by $n - 1$

$$
s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}}
$$

where $\bar{x}$ is the mean of the data, and $n$ is the number of values in the data.

**NOTE:** In this module all of the exercises assume you are calculating the standard deviation of a sample ($s$).

To calculate the standard deviation using the formula, follow the steps below:

1. Calculate the mean of the data. (calculate $\bar{x}$)
2. Calculate the difference between each number in the data and the mean. (calculate $x_n - \bar{x}$)
3. Square each difference.
4. Calculate the sum of the squares of the differences.
5. Divide the above sum of the squares by $n - 1$. This is the mean of the squares.
6. Find the square root of the above mean of the squares.
EXAMPLE 1

The masses of several fish taken from a pond on a fish farm are 645 gm, 580 gm, 692 gm, 634 gm, 620 gm, 546 gm, 605 gm, 636 gm, 567 gm, and 625 gm. Find the standard deviation.

SOLUTION:

To simplify the work, we make the following table using the given data.

<table>
<thead>
<tr>
<th>Mass</th>
<th>Difference from mean</th>
<th>Square of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>645</td>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td>580</td>
<td>−35</td>
<td>1225</td>
</tr>
<tr>
<td>692</td>
<td>77</td>
<td>5929</td>
</tr>
<tr>
<td>634</td>
<td>19</td>
<td>361</td>
</tr>
<tr>
<td>620</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>546</td>
<td>−69</td>
<td>4761</td>
</tr>
<tr>
<td>605</td>
<td>−10</td>
<td>100</td>
</tr>
<tr>
<td>636</td>
<td>21</td>
<td>441</td>
</tr>
<tr>
<td>567</td>
<td>−48</td>
<td>2304</td>
</tr>
<tr>
<td>625</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>6150</td>
<td></td>
<td>16146</td>
</tr>
</tbody>
</table>

\[
\text{mean} = \frac{6150}{10} = 615
\]

\[
\text{mean of squares} = \frac{16146}{n-1}
\]

\[
\text{mean of squares} = \frac{16146}{9} = 1794
\]

\[
\text{Square root of mean of squares} = \sqrt{1794} = 42.36
\]

The standard deviation rounded to the nearest tenth is 42.4.
The standard deviation is a useful indicator of the spread of the data. For example, suppose that the standard deviation of the prices of a certain camera sold by several different merchants is $50. If the mean price of the camera is $150, then the standard deviation indicates a wide spread in the selling price. However if the mean price of the camera is $1800, then the standard deviation of $50 indicates little variation in the selling price.

The standard deviation is used by statisticians to make predictions. We will see how that is done in the next section.

CALCULATOR MATH

To find the standard deviation on the calculator, using the SHARP EL−531, proceed as follows:

1. Put the calculator in STAT MODE 1.
2. Key in the data by keying in the number followed by the M+ key. After each entry you will see \[ n = \] displayed indicating the number of entries (sample size).
3. When all the data are keyed in, to find:
   a) Standard Deviation of the sample (s) press [RCL] and then 5
   b) Standard Deviation of the population (σ) press [RCL] and then 6
   c) Mean (\( \bar{x} \)) press [RCL] and then 4

Note: Make sure to clear the calculator before entering a new set of data. Press the CA key to clear.
Experiential Activity Three

1. Find the range of the following sets of numbers:
   a) 3, 5, 8, 2, 5, 6, 9, 3
   b) 7, −1, 9, 6, −4, 8, 0, 7
   c) 45.7, 40.9, 73.8, 65.8, 59.6

2. Find the standard deviation of the following sets of numbers. Round your answer to the nearest tenth.
   a) 5, 8, 9, 12, 5, 3, 7
   b) 209, 233, 195, 256, 239, 202, 187, 267, 245
   c) 17.6, 28.3, 22.1, 27.3, 19.8, 17.9, 23.6, 22.5

3. The percent scores of an entrance exam for a group of students is given below.
   62, 82, 56, 34, 68, 69, 56, 84, 92, 59,
   88, 72, 38, 23, 67, 91, 80, 76, 70, 45,
   58, 56, 87, 59, 78, 45, 92, 66, 78, 20,
   45, 37, 67, 48, 78, 55, 78, 52
   Calculate the range and the standard deviation. Show Me.

4. The gasoline consumption was determined for 10 cars. All cars were the same make and model. The amounts used (in litres per 100 km) were 11.3, 10.8, 10.5, 12.0, 11.5, 12.2, 10.9, 10.5, 11.4, 11.5. Find the mean and the standard deviation.

Experiential Activity Three Answers

1. a) 7, b) 13, c) 32.9
2. a) 2.8, b) 26.9, c) 3.7
3. Range = 72, $\sigma = 18.7$
4. Mean $\bar{x} = 11.3$, $\sigma = 0.55$
OBJECTIVE FOUR

*When you complete this objective you will be able to…*

Use the Standard Normal Distribution Curve and the Standard Deviation

**Exploration Activity**

A frequency distribution tells us how many times (frequencies) certain values occur in a given set of data. One way to analyze data is to see how many times each value occurs. For example, the marks of a recent grade 12 math exam, are tabulated. The test was out of 100% and each student received a whole number mark.

The marks are shown on a bar graph below. Joining the midpoints of the bars gives us a frequency polygon.

<table>
<thead>
<tr>
<th>Mark</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>2</td>
</tr>
<tr>
<td>30-39</td>
<td>4</td>
</tr>
<tr>
<td>40-49</td>
<td>7</td>
</tr>
<tr>
<td>50-59</td>
<td>18</td>
</tr>
<tr>
<td>60-69</td>
<td>30</td>
</tr>
<tr>
<td>70-79</td>
<td>21</td>
</tr>
<tr>
<td>80-89</td>
<td>14</td>
</tr>
<tr>
<td>90-100</td>
<td>3</td>
</tr>
</tbody>
</table>

*Frequency Distributions and Measures of Central Tendency*
If the same test had been given to 500,000 students (all the students in the country), and if we then had made a frequency polygon graph we might obtain a graph as shown. Note that for this much data we would choose the width of each bar to be 1 instead of 10.

The graph is in the shape of a bell and is symmetrical. Distributions that result in such a graph are called normal distributions. The graph is called a normal curve. The normal curve and the distribution for many large sets of numbers are closely matched. A set of numbers is considered to be large if it contains more than 30 values. Furthermore, sets of numbers with a normal distribution have the same value for the mean, median, and mode.

In a normal distribution the area between the x-axis and the bell curve represents the number of values. The total area under the curve represents all the values. As you can see the area is largest near the center of the curve, so most of the data is found here. Towards the sides, the area under the curve is quite small, so not many values of the data are found there.

In a normal distribution, about 68% of the numbers fall within 1 standard deviation on each side of the mean, about 95% are within 2 standard deviations on each side of the mean, and about 99% are within 3 standard deviations on each side of the mean.
Since the normal curve is symmetrical, we can write in the percentages of the values that lie under the curve and are bounded by the vertical lines which separate the number of standard deviations from the mean. This will make the application of the normal distribution easier.
EXAMPLE 1

The lifetimes of 1000 light bulbs are normally distributed with a mean life of 7000 hours and a standard deviation of 600 hours.

a) What percentage of the bulbs should last less than 6400 h?
b) What percentage of the bulbs should last more than 8200 h?
c) What percentage of the bulbs should last less than 7600 h?
d) How many light bulbs will last longer than 6400 h?

SOLUTION:

Sketch a normal curve. The mean is 7000 and the standard deviation is 600 h. The axes are marked accordingly as shown below.

a) To the left of 6400 h, the area under the curve is only 14% + 2% = 16% of the total area. Therefore only 16% of the bulbs should last less than 6400 h.

b) To the right of 8200 h, the area under the curve is only about 2%. This represents the percentage of bulbs that should last more than 8200 h.

c) To the left of 7600 h, the area under the curve is 34% + 34% + 14% + 2% = 84% of the total area. Therefore 84% of the bulbs should last less than 7600 h.

d) To the right of 6400 h, the area under the curve is 34% + 34% + 14% + 2% = 84%. Hence 84% of the 1000 bulbs should last more than 6400 h. Now, 84% of 1000 is 0.84 \times 1000 = 840.

Therefore 840 bulbs should last longer than 6400 h.
Experiential Activity Four

1. The weights of 500 male students at a college were determined. The mean weight was 152 lb and the standard deviation was 141 lb. Assuming that the weight are normally distributed, find how many students weigh
   a) More than 166 lb
   b) Less than 124 lb
   c) Between 138 lb and 152 lb.

2. A traffic survey used radar to determine the speeds of two thousand cars on a highway. The mean speed was found to be 92 km/h and the standard deviation was 10 km/h. Determine:
   a) What percentage of the cars traveled over 82 km/h
   b) How many cars traveled between 72 km/h and 92 km/h
   c) How many cars traveled faster than 102 km/h

3. The mean diameter of 500 ball bearings tested was 0.503 in and the standard deviation was 0.004 in. A ball bearing is considered defective if it is smaller than 0.499 in or larger than 0.507 in. Assuming the diameters are normally distributed, determine:
   a) How many ball bearings are defective.
   b) How many ball bearings are not defective.

4. Five thousand jet pilots took a reaction test. The results were normally distributed with a mean of 0.45 and a standard deviation of 0.03.
   a) How many scored lower than 0.42?
   b) What percentage scored higher than 0.51
   c) If a pass is a score lower than 0.48, how many passed?

Experiential Activity Four Answers

1. a) 80, b) 10, c) 170
2. a) 84%, b) 960, c) 320
3. a) 160, b) 340
4. a) 800, b) 2%, c) 4200

Practical Application Activity

Complete the Frequency Distribution and Measures of Central Tendency module assignment in TLM.

Summary

This module introduced theory on frequency distribution and measures of central tendency.