Statistical Resresentaiton of Experimental Data

Measures of Central Tendency: Mode, Median, and Mean

A value that represents a typical, or central, entry of a data set.
Most common measures of central tendency:

- Mean
- Median
- Mode

Mode

The mode of a data set (sample) is the **most frequent** value

Example

DataSet: 5 3 7 2 4 4 2 4 8 3 4 3 4

- The 4 is most freuent

Every data set has not a mode. For example if the grading policy of a course arranged to have equal number of scores 1,2,3, and 4. All values will have equal frequency.

the mode is a useful central value when we want to know the <u>most frequently occurring data value</u>, such as the most frequently requested shoe size.

Median

Another average that is useful is the median, or central value, of an ordered distribution.

When you are given the median, you know there are an equal number of data values in the ordered distribution that are above it and below it

If there is:

- <u>odd number of entries:</u> median is the middle data entry.
- <u>even number of entries:</u> median is the mean of the two middle data entries.

HOW TO FIND THE MEDIAN

The **median** is the central value of an ordered distribution. To find it,

1. Order the data from smallest to largest.
2. For an *odd* number of data values in the distribution,

   Median = Middle data value
3. For an *even* number of data values in the distribution,

$$\text{Median} = \frac{\text{Sum of middle two values}}{2}$$

Measures of Central Tendency: Mode, Median, and Mean

Median Example

Data Set: 19 19 27 28 18 35

(Sorted) data set:

18 19 19 27 18 35

$$\text{Median} = \frac{19 + 27}{2} = 23 \text{ cents}$$

The median uses the position rather than the specific value of each data entry. If the extreme values of a data set change, the median usually does not change.

This is why the median is often used as the average for house prices. If one mansion costing several million dollars sells in a community of much-lower-priced homes, the median selling price for houses in the community would be affected very little, if at all.

for large ordered data sets of size n

$$\text{Postition of the middle value} = \frac{n+1}{2}$$

Mean (average)

The sum of all the data entries divided by the number of entries.

$$« = \frac{\Sigma x}{N} \qquad \bar{x} = \frac{\Sigma x}{n}$$

Population mean Sample mean

Example

Student L needs at least a B in biology. Here are all scores:

58 67 60 84 93 98 100

Compute the mean and determine if Linda's grade will be a B (80 to 89 average) or a C (70 to 79 average).

$$« = \frac{\Sigma x}{N} = \frac{58 + 67 + 60 + 84 + 93 + 98 + 100}{7}$$

$$= \frac{560}{7} = 80$$

Measures of Central Tendency: Mode, Median, and Mean

Example

The unit load of 40 randomly selected students from a college shown below. Find the mean, median and mode:

Sample Mean	**15.0**
Median	**15**
Mode	**12**

17	12	14	17	13	16	18	20	13	12
12	17	16	15	14	12	12	13	17	14
15	12	17	16	12	18	20	19	12	15
18	14	16	17	15	19	12	13	13	15

A disadvantage of the mean, however, is that it can be affected by exceptional values.

The mean is not a resistant measure of center because we can make the mean as large as we want by changing the size of only one data value.

The median, on the other hand, is more resistant. However, a disadvantage of the median is that it is not sensitive to the specific size of a data value.

Measures of Central Tendency: Mode, Median, and Mean

trimmed mean

A measure of center that is more resistant than the mean but still sensitive to specific data values is the trimmed mean.

A trimmed mean is the mean of the data values left after "trimming" a specified percentage of the smallest and largest data values from the data set.

Usually a 5% trimmed mean is used. This implies that we trim the lowest 5% of the data as well as the highest 5% of the data.

HOW TO COMPUTE A 5% TRIMMED MEAN

1. Order the data from smallest to largest.
2. Delete the bottom 5% of the data and the top 5% of the data. *Note:* If the calculation of 5% of the number of data values does not produce a whole number, *round* to the nearest integer.
3. Compute the mean of the remaining 90% of the data.

Example:

The class sizes of 20 randomly chosen Introductory Algebra classes in C are shown.

14 20 20 20 20 23 25 30 30 30
35 35 35 40 40 42 50 50 80 80

Compute the mean
Compute a 5% trimmed mean

Example

Find the mean, median, and mode of the sample ages of a class shown. Which measure of central tendency best describes a typical entry of this data set? Are there any outliers?

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Measures of Central Tendency: Mode, Median, and Mean

Weighted Average

Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers.

$$\text{Weighted average} = \frac{\Sigma xw}{\Sigma w}$$

where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.

Example

You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?

Source	Score, *x*	Weight, *w*	*x·w*
Test Mean	86	0.50	86(0.50)= 43.0
Midterm	96	0.15	96(0.15) = 14.4
Final Exam	82	0.20	82(0.20) = 16.4
Computer Lab	98	0.10	98(0.10) = 9.8
Homework	100	0.05	100(0.05) = 5.0
		$\Sigma w = 1$	$\Sigma(x \cdot w) = 88.6$

$$\bar{x} = \frac{\Sigma(x \cdot w)}{\Sigma w} = \frac{88.6}{1} = 88.6$$

Measures of Central Tendency: Mode, Median, and Mean

Mean of Grouped Data

Mean of a Frequency Distribution is Approximated by

$$\overline{x} = \frac{\Sigma(x \cdot f)}{n} \qquad n = \Sigma f$$

where *x* and *f* are the midpoints and frequencies of a class, respectively

1. Find the midpoint of each class. $x = \dfrac{\text{(lower limit)+(upper limit)}}{2}$
2. Find the sum of the products of the midpoints and the frequencies. $\Sigma(x \cdot f)$
3. Find the sum of the frequencies. $n = \Sigma f$
4. Find the mean of the frequency distribution. $\overline{x} = \dfrac{\Sigma(x \cdot f)}{n}$

Example:
Use the frequency distribution to approximate the mean number of minutes that a sample of Internet subscribers spent online during their most recent session.

Class	Frequency, *f*
7 – 18	6
19 – 30	10
31 – 42	13
43 – 54	8
55 – 66	5
67 – 78	6
79 – 90	2

Measures of Variation

Range
The difference between the maximum and minimum data entries in the set.
The data must be quantitative.

Range = (Max. data entry) – (Min. data entry)

Deviation
The difference between the data entry, x, and the mean of the data set.

Population data set:

Deviation of x = x – μ

Sample data set:

Deviation of x = x – x

Variance, and Standard Deviation

The variance can be thought of a kind of average of the squares of the deviations

$$\pm^2 = \frac{\Sigma(x - \text{«})^2}{N}$$

Population Variance

Variance, and Standard Deviation

Standard deviation is a measure of the typical amount an entry deviates from the mean

$$\pm = \sqrt{\pm^2} = \sqrt{\frac{\Sigma(x - \text{«})^2}{N}}$$

Population Standard Deviation

The more the entries are spread out, the greater the standard deviation

$$s^2 = \frac{\Sigma(x - \overline{x})^2}{n-1}$$

Sample Variance

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n-1}}$$

Sample Standard Deviation

Measures of Variation

Population example

A corporation hired 10 graduates. The starting salaries for each graduate are shown. Find the population variance and standard deviation of the starting salaries.

Starting salaries (1000s of dollars)

41 38 39 45 47 41 44 41 37 42

$\mu = 41.5$

Salary, *x*	Deviation: *x* – μ	Squares: (*x* – μ)^2
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
38	38 – 41.5 = –3.5	$(-3.5)^2 = 12.25$
39	39 – 41.5 = –2.5	$(-2.5)^2 = 6.25$
45	45 – 41.5 = 3.5	$(3.5)^2 = 12.25$
47	47 – 41.5 = 5.5	$(5.5)^2 = 30.25$
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
44	44 – 41.5 = 2.5	$(2.5)^2 = 6.25$
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
37	37 – 41.5 = –4.5	$(-4.5)^2 = 20.25$
42	42 – 41.5 = 0.5	$(0.5)^2 = 0.25$

$SS_x = 88.5$

$$\pm^{2} = \frac{\Sigma(x - «)^2}{N} = \frac{88.5}{10} \approx 8.9$$

$$\pm = \sqrt{\pm^{2}} = \sqrt{8.85} \approx 3.0$$

Measures of Variation

Sample Example

The starting salaries are for the Chicago branches of a corporation. The corporation has several other branches, and you plan to use the starting salaries of the Chicago branches to estimate the starting salaries for the larger population. Find the *sample* standard deviation of the starting salaries.

Starting salaries (1000s of dollars)

41 38 39 45 47 41 44 41 37 42

Salary, *x*	Deviation: *x* – μ	Squares: (*x* – μ)^2
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
38	38 – 41.5 = –3.5	$(-3.5)^2 = 12.25$
39	39 – 41.5 = –2.5	$(-2.5)^2 = 6.25$
45	45 – 41.5 = 3.5	$(3.5)^2 = 12.25$
47	47 – 41.5 = 5.5	$(5.5)^2 = 30.25$
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
44	44 – 41.5 = 2.5	$(2.5)^2 = 6.25$
41	41 – 41.5 = –0.5	$(-0.5)^2 = 0.25$
37	37 – 41.5 = –4.5	$(-4.5)^2 = 20.25$
42	42 – 41.5 = 0.5	$(0.5)^2 = 0.25$

$\mu = 41.5$ **$SS_x = 88.5$**

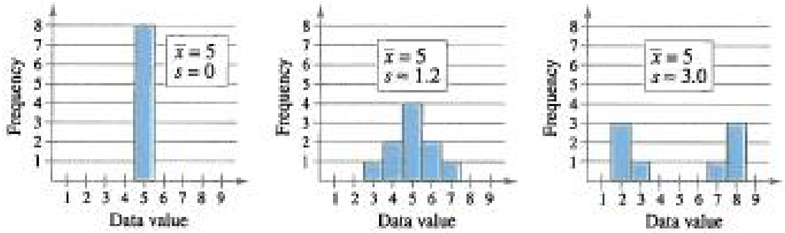
$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1} = \frac{88.5}{10-1} \approx 9.8$$

$$s = \sqrt{s^2} = \sqrt{\frac{88.5}{9}} \approx 3.1$$

Measures of Variation

Interpreting Standard Deviation

Standard deviation is a measure of the typical amount an entry deviates from the mean.
The more the entries are spread out, the greater the standard deviation.



Coefficient of Variation

If $\bar{x}$ and s represent the sample mean and sample standard deviation, respectively, then the sample **coefficient of variation** CV is defined to be

$$CV = \frac{s}{\bar{x}} \cdot 100$$

If μ and σ represent the population mean and population standard deviation, respectively, then the population coefficient of variation CV is defined to be

$$CV = \frac{\sigma}{\mu} \cdot 100$$

This gives us the advantage of being able to directly compare the variability of two different populations using the coefficient of variation.

Measures of Variation

Standard Deviation for Grouped Data

When a frequency distribution has classes, estimate the sample mean and standard deviation by using the midpoint of each class.

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2 f}{n-1}}$$

where $n = \Sigma f$ (the number of entries in the data set)

Sample standard deviation for a frequency distribution

x	*f*	*xf*
0	10	0(10) = 0
1	19	1(19) = 19
2	7	2(7) = 14
3	7	3(7) =21
4	2	4(2) = 8
5	1	5(1) = 5
6	4	6(4) = 24

$\Sigma f = 50$

$\Sigma(xf) = 91$

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{91}{50} \approx 1.8$$

Example:

a random sample of the number of children per household

Number of Children in 50 Households				
1	3	1	1	1
1	2	2	1	0
1	1	0	0	0
1	5	0	3	6
3	0	3	1	1
1	1	6	0	1
3	6	6	1	2
2	3	0	1	1
4	1	1	2	2
0	3	0	2	4

Measures of Variation

- Determine the sum of squares.

x	f			
0	10	0 – 1.8 = –1.8	$(-1.8)^2 = 3.24$	3.24(10) = 32.40
1	19	1 – 1.8 = –0.8	$(-0.8)^2 = 0.64$	0.64(19) = 12.16
2	7	2 – 1.8 = 0.2	$(0.2)^2 = 0.04$	0.04(7) = 0.28
3	7	3 – 1.8 = 1.2	$(1.2)^2 = 1.44$	1.44(7) = 10.08
4	2	4 – 1.8 = 2.2	$(2.2)^2 = 4.84$	4.84(2) = 9.68
5	1	5 – 1.8 = 3.2	$(3.2)^2 = 10.24$	10.24(1) = 10.24
6	4	6 – 1.8 = 4.2	$(4.2)^2 = 17.64$	17.64(4) = 70.56

$$\Sigma(x-\overline{x})^2 f = 145.40$$

- Find the sample standard deviation.

$$s = \sqrt{\frac{\Sigma(x-\overline{x})^2 f}{n-1}} = \sqrt{\frac{145.40}{50-1}} \approx 1.7$$