2.4

Measures of Variation

Measures of Variation

Definition: A measure of variation is a way to indicate the spread (i.e. how far apart the entries are from each other) of a data set.

The easiest measure of variation to compute is the range.

Recall: The range of a data set is the difference between the maximum and minimum data value.

$$\text{Range} = \text{Max Value} - \text{Min Value}$$

Example

The starting salaries (in thousands of dollars) of all 10 new employees at a company are:

41, 38, 39, 45, 47, 41, 44, 41, 37, 42

$$\text{Range} = 47 - 37 = 10$$

i.e. $10,000

Advantages/Disadvantages of the Range

The biggest advantage of the range is that it is very easy to compute as it only uses two numbers.

The disadvantage, however, is that it doesn't tell us much about any of the numbers in between. Other measures of variation that use all data values in a given data set can tell us more.

Definition:

The deviation of a data entry *x* in a population data set is given by

$$\text{deviation of } x = x - \mu$$

where μ is the population mean

Example: Using the company salary data set

To compute the deviation, you first need to compute the population mean

$$\mu = \frac{\sum x}{N} = \frac{415}{10} = 41.5$$

Note: The sum of the deviations should always be 0. This is one way you can check to make sure you've done your work correctly

Salary (x)	Deviation ($x - \mu$)
41	-0.5
38	-3.5
39	-2.5
45	3.5
47	5.5
41	-0.5
44	2.5
41	-0.5
37	-4.5
42	0.5

More Key Terms

Population Variance (σ^2)	$$\sigma^2 = \frac{SS_x}{N} = \frac{\sum (x-\mu)^2}{N}$$ **N is the population size and SS_x is called the "sum of squares"**
Population Standard Deviation (σ)	**Standard deviation is the square root of variance** $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Note: Standard deviation and variance are 0 if all of the entries have the same value. The further apart data entries are from each other, the larger these values

Salary (x)	Deviation ($x-\mu$)	$(x-\mu)^2$
41	-0.5	0.25
38	-3.5	12.25
39	-2.5	6.25
45	3.5	12.25
47	5.5	30.25
41	-0.5	0.25
44	2.5	6.25
41	-0.5	0.25
37	-4.5	20/25
42	0.5	0.25

Example: Using the same salary data as before

$$\sigma^2 = \frac{SS_x}{N} = \frac{88.5}{10} = \frac{\sum (x-\mu)^2}{10} \approx 8.9$$

$$\sigma = \sqrt{\frac{88.5}{10}} \approx 3.0$$

If you have a sample:

Sample Variance (s^2)	$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ ***n* is the sample size, is the sample mean**
Population Standard Deviation (s)	**Standard deviation is still the square root of variance** $$s = \sqrt{s} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

For a frequency distribution with classes, use the class midpoint as your *x* value.

The reason we divide by *n-1* instead of *n* is because calculations have proven that dividing by *n-1* gives a better measurement in practice than dividing by *n*

Time (x)	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$
4	-3.5	12.25
7	-0.5	0.25
6	-1.5	2.25
7	-0.5	0.25
9	1.5	2.25
5	-2.5	6.25
8	0.5	0.25
10	2.5	6.25
9	1.5	2.25
8	0.5	0.25
7	-0.5	0.25
10	2.5	6.25

Example: The data to the left gives the recovery time (in days) for a sample of football players with concussions

$$\bar{x} = \frac{\sum x}{n} = \frac{90}{12} = 7.5$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{39}{11} \approx 3.5$$

$$s = \sqrt{s^2} = \sqrt{\frac{39}{11}} \approx 1.9$$

Finding standard deviation with a TI-84

Input Data

1) Press STAT and then EDIT.

2) Type all data entries in one of the lists

If you are using a frequency table, put data values in one list and frequencies in another list

Remember which lists you used

Calculate

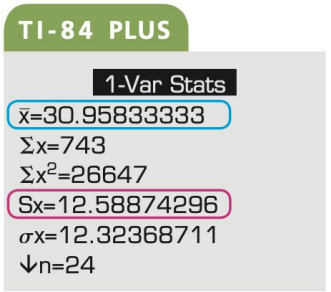
1) Press STAT, move right to CALC, choose option 1 – 1 Var Stats, and then hit ENTER

2) If you are using only one list of data then type the list number in the List row and hit CALCULATE.

If you are using a frequency distribution type one list under list number and the other for Frequency

Find S.D.

Find the appropriate row for the value you are looking for

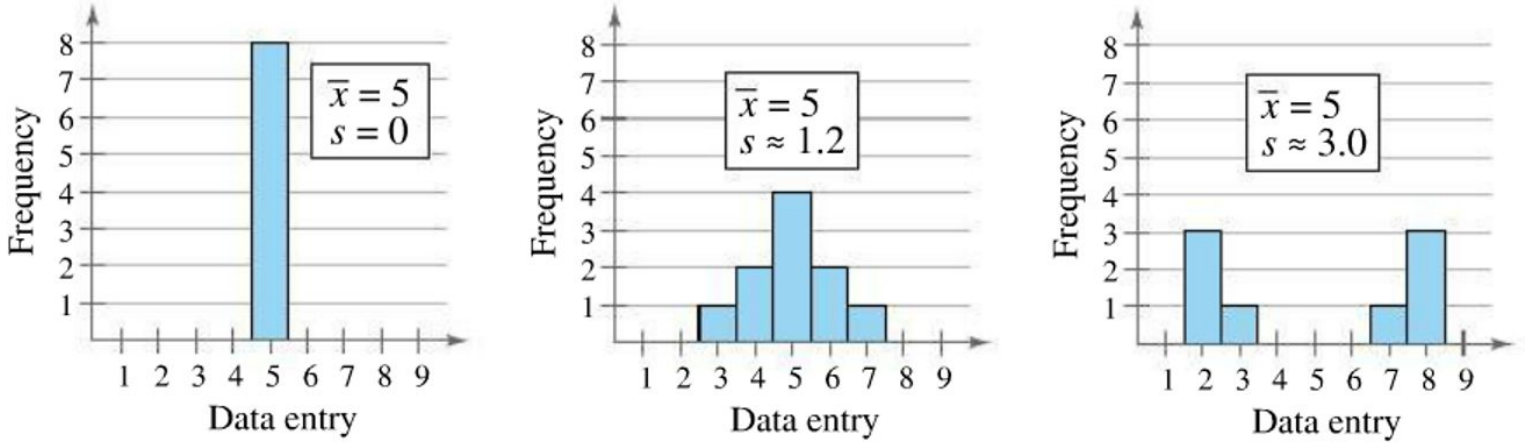


Sample Mean

Sample Standard Deviation

Interpreting Standard Deviation

Standard deviation tells you how far a typical entry is from the mean. The further apart data is, the larger the standard deviation

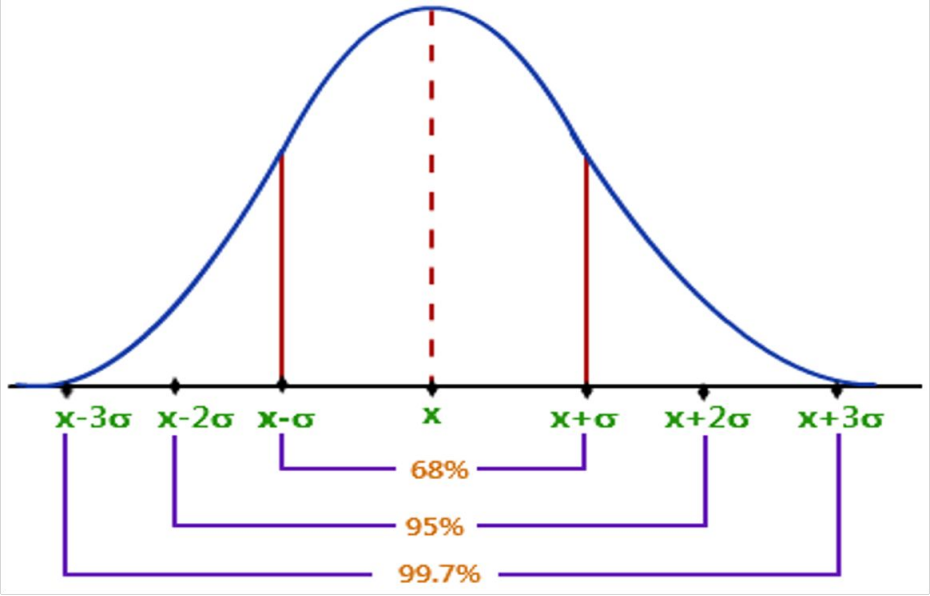


A data entry that is more than two standard deviations from the mean is considered unusual. If it is more than three standard deviations away from the mean is considered very unusual.

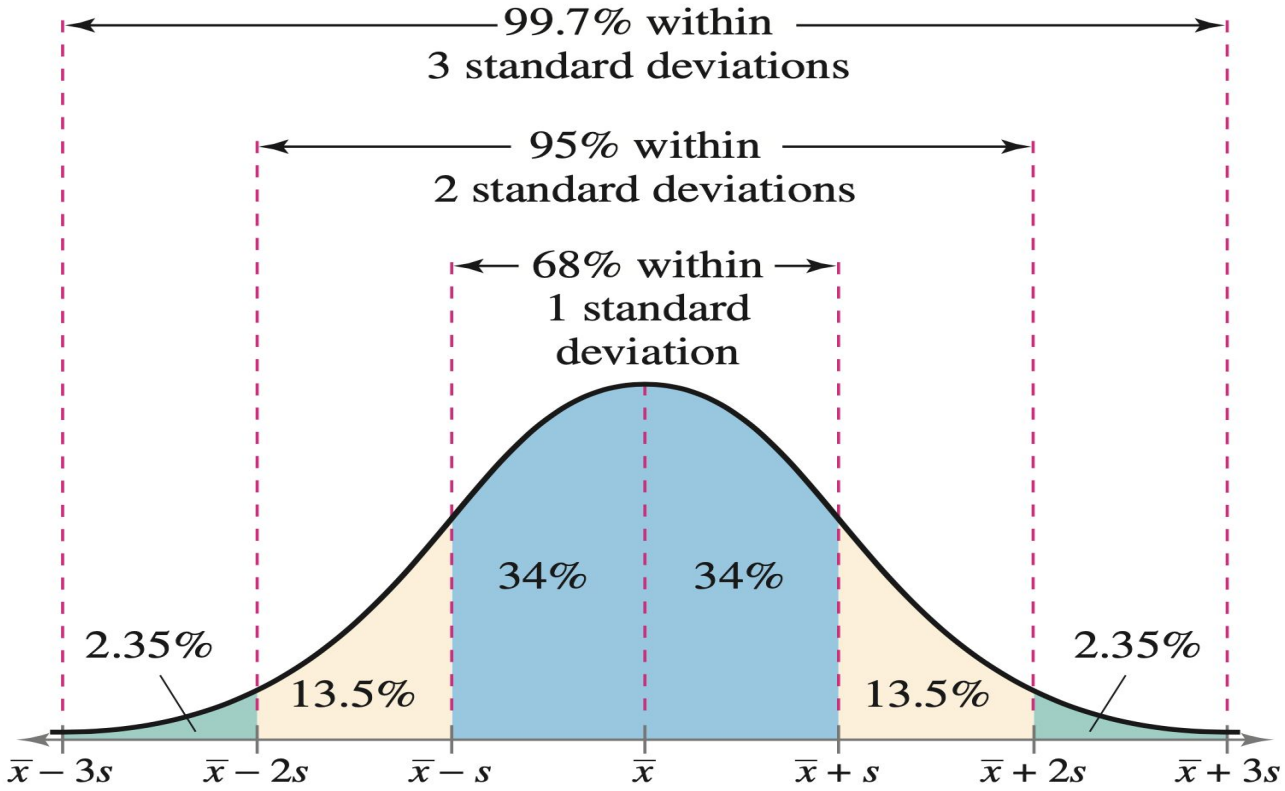
The Empirical Rule (68, 95, 99.7 Rule)

For data with a (symmetric) bell-shaped distribution, the standard deviation has the following characteristics:

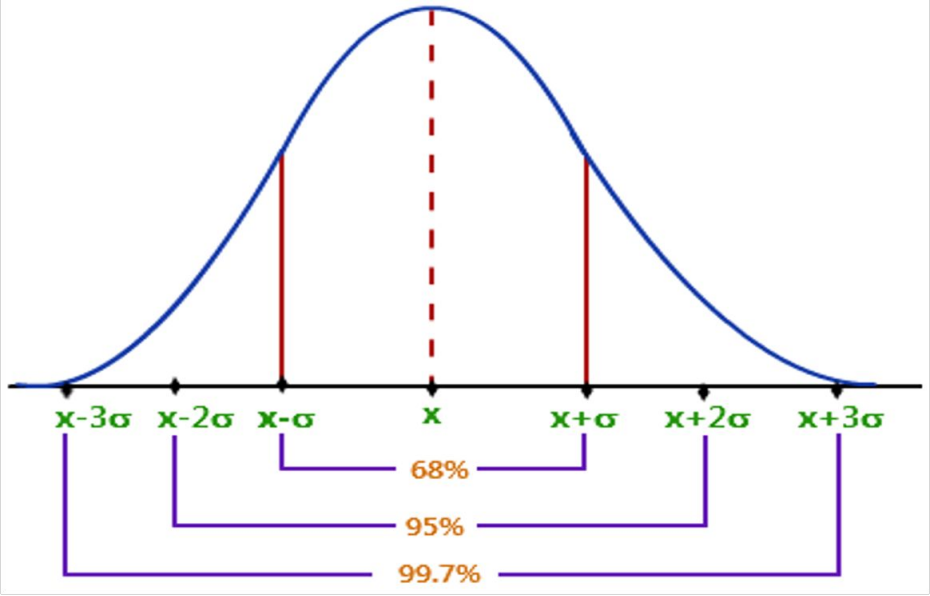
About 68% of the data lie within one standard deviation of the mean.
About 95% of the data lie within two standard deviations of the mean.
About 99.7% of the data lie within three standard deviations of the mean.

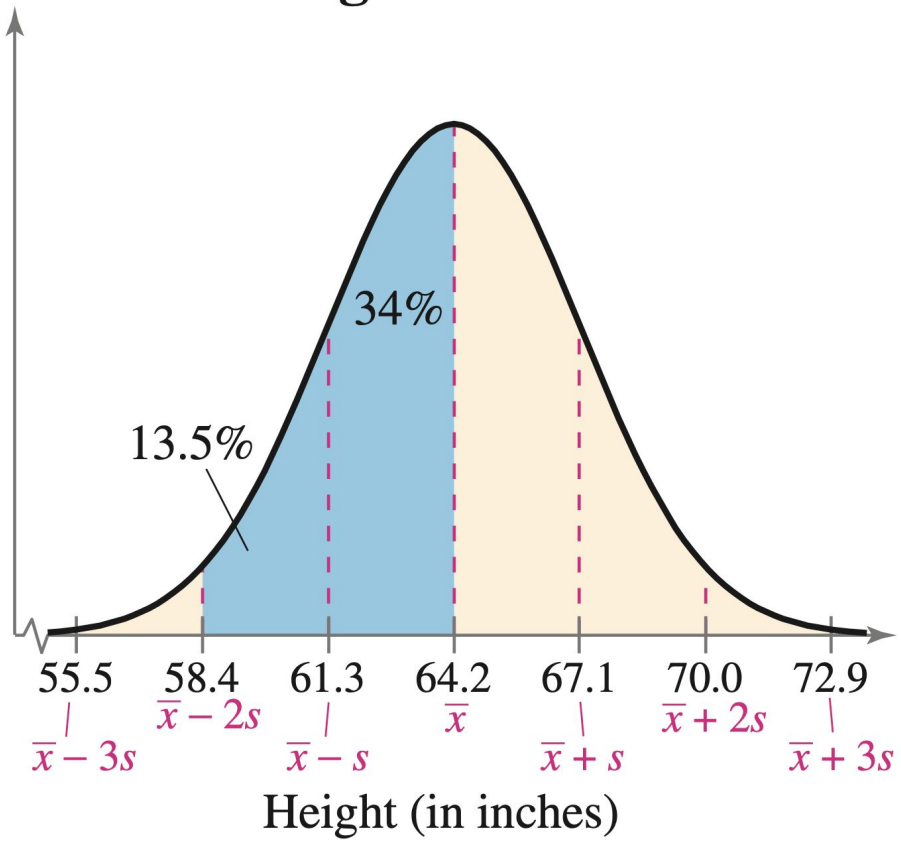


Bell-Shaped Distribution



Example: In a survey conducted by the National Center for Health Statistics, the sample mean height of women in the United States (ages 20–29) was 64.2 inches, with a sample standard deviation of 2.9 inches. Estimate the percent of the women whose heights are between 58.4 inches and 64.2 inches.





$$13.59 + 34.13 = 47.72\%$$

Approximately 47.73% of women have height between 58.4 in and 64.2 in

Additional Examples

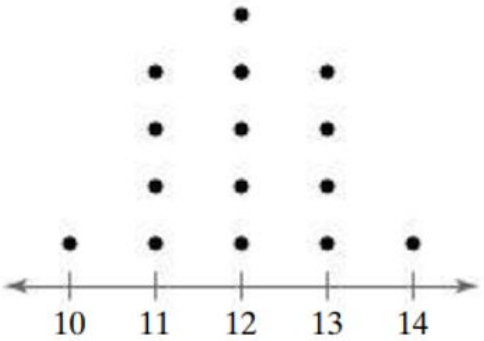
1) You are applying for jobs at two companies, Company C offers starting salaries with μ=$39,000 and σ=$4,000. Company D offers starting salaries with μ=$39,000 and σ=$1500. From which company are you more likely to get an offer of $42,000 or more? Explain your reasoning.

2) The mean monthly utility bill for a sample of households in a city is $70, with a standard deviation of $8.
   a) Between what two values do about 95% of the data lie?
   b) If there are 40 households in the sample, estimate the number of households whose monthly utility bills are between $54 and $86.
   c) The monthly utility bills are for 8 more households are listed. Determine which of the data entries listed below are unusual. Explain your reasoning.

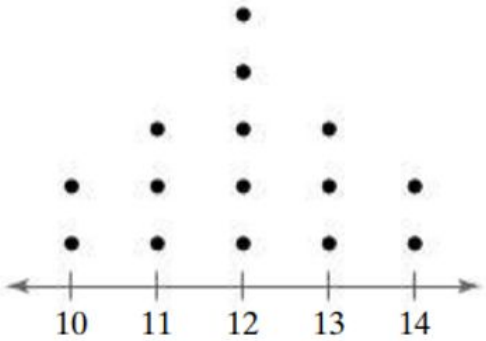
      $65, $52, $63, $83, $77, $98, $84, $70

3) Without doing any calculations, determine which data set (i), (ii), or (iii) is the data set with the greatest standard deviation and which data set is the data set with the least standard deviation. Explain your reasoning.

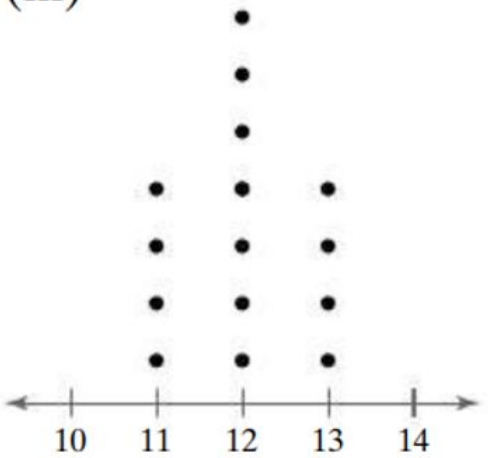
(i)



(ii)



(iii)



4) Approximate the mean and standard deviation of the sample using the data set displayed at the right.

Note: You can solve this problem more than one way. For example:

- **You can use the frequency histogram to expand out all the data entries (e.g. there are three 0s, fifteen 1s, etc.)**
- **You can use the singular entry number as the class midpoint**

Cars per Household The results of a random sample of the number of cars per household in a region are shown in the histogram.