Section 1.9 – Measures of Dispersion

Basic Concepts of Variation

Range

Definition

The ***range*** of a set of data values is the difference between the maximum data value and the minimum data value.

Range = (maximum data value) – (minimum data value)

It is very sensitive to extreme values; therefore not as useful as other measures of variation.

Example

India has 1 satellite used for military and intelligence purposes, Japan has 3, and Russia has 14. Find the range of the sample values of 1, 3, and 14.

Solution

Range = (maximum data value) – (minimum data value)

$= 14 - 1$

$= 13.0$

Definition

The ***standard deviation*** of a set of sample values, denoted by s, is a measure of variation of values about the mean. It is a type of average deviation of values from the mean that is calculated.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$ Sample standard deviation

$$s = \sqrt{\frac{n\left(\sum x^2\right) - \left(\sum x\right)^2}{n(n-1)}}$$ Sample standard deviation

Standard Deviation - Important Properties

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation s is usually positive.
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation s are the same as the units of the original data values.

Example

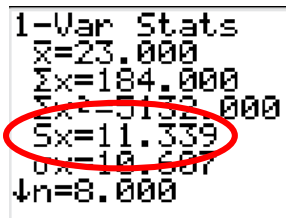
Find the standard deviation of the numbers: 7, 9, 18, 22, 27, 29, 32, 40.

Solution

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$$

$$= \sqrt{\frac{5132 - 8(23)^2}{8-1}}$$

$$\approx 11.34$$



Standard Deviation of a Population

The standard deviation σ (lowercase sigma) of a population is given by the formula

Population standard deviation $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{n-1}}$

Variance of a Sample and a Population

Definition

The variance of a set of values is a measure of variation equal to the square of the standard deviation

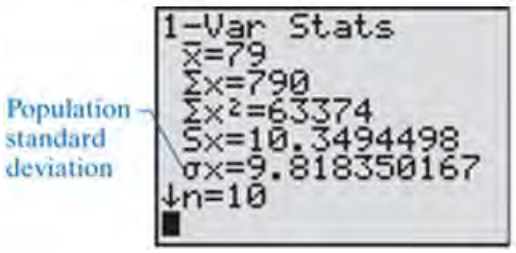
Sample variance: s^2 square of the standard deviation s.

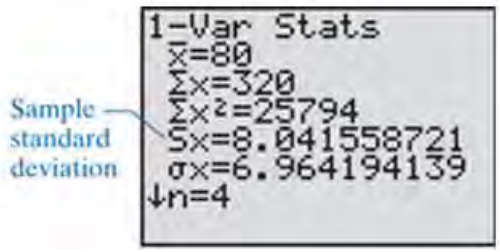
Population variance: σ^2 square of the population standard deviation σ.

Notation

s = *sample* standard deviation σ = *population* standard deviation

s^2 = *sample* variance σ^2 = *population* variance





Unbiased Estimator

The sample variance s^2 is an ***unbiased estimator*** of the population variance σ^2, which means values of s^2 tend to target the value of σ^2 instead of systematically tending to overestimate or underestimate σ^2.

Using and Understanding Standard Deviation

Range Rule of Thumb is based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean

Interpreting a Known Value of the Standard Deviation

Informally define ***usual*** values in a data set to be those that are typical and not too extreme. Find rough estimates of the minimum and maximum "usual" sample values as follows:

Minimum "usual" value = (mean) – 2 × (standard deviation)

Maximum "usual" value = (mean) + 2 × (standard deviation)

Estimating a Value of the Standard Deviation *s*

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{range}{4}$$

Where range = (maximum value) – (minimum value)

Example

The Wechsler Adult intelligence Scale involves an IQ test designed so that the mean score is 100 and the standard deviation is 15. Use the range rule thumb to find the minimum and maximum "usual" IQ scores. Then determine whether an IQ score of 135 would be considered "unusual"

Solution

Mean = 100
Standard deviation = 15

Minimum "usual" value = (mean) – 2 × (standard deviation)
= 100 – 2(15)
= 70

Maximum "usual" value = (mean) + 2 × (standard deviation)
= 100 + 2(15)
= 130

Example

Use the range of thumb to estimate the standard deviation of the sample of 100 FICO credit rating scores listed in the table below.

708	713	781	809	797	793	711	681	768	611	698	836	768
532	657	559	741	792	701	753	745	681	598	693	743	444
502	739	755	835	714	517	787	714	497	636	637	797	568
714	618	830	579	818	654	617	849	798	751	731	850	591
802	756	689	789	628	692	779	756	782	760	503	784	591
834	694	795	660	651	696	638	635	795	519	682	824	603
709	777	829	744	752	783	630	753	661	604	729	722	706
594	664	782	579	796	611	709	697	732				

Solution

Those scores have a minimum of 444 and a maximum of 850.

$$s \approx \frac{range}{4}$$

$$= \frac{850 - 444}{4}$$

$$= 101.5$$

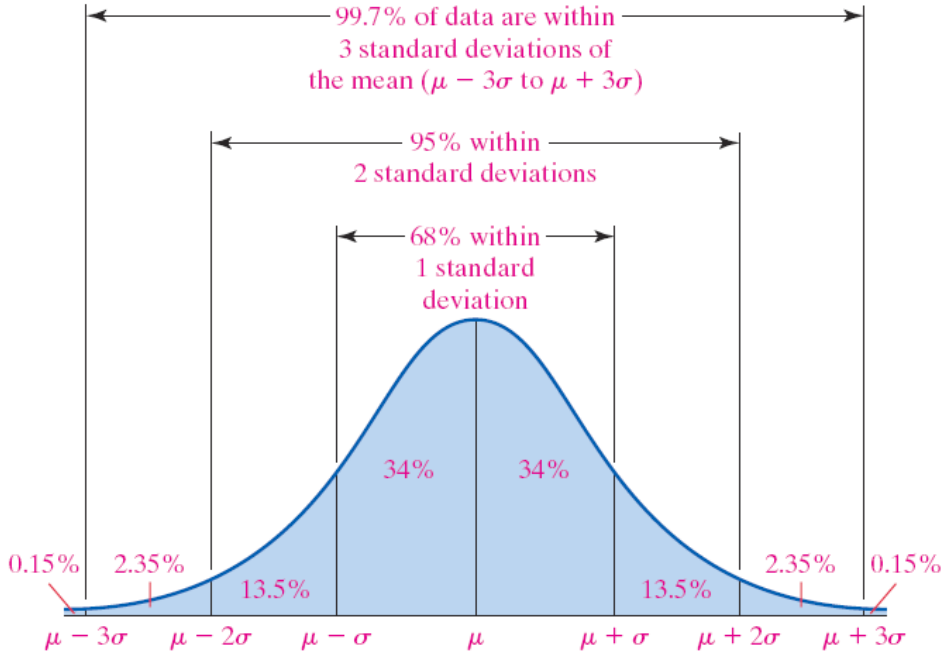
Properties of the Standard Deviation

- ✓ The standard deviation measures the ***variation*** among data values
- ✓ Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- ✓ Has the same units of measurement as the original data
- ✓ For many data sets, a value is *unusual* if it differs from the mean by more than two standard deviations
- ✓ When comparing variation in two different data sets, compare the standard deviation only if they use the same scale and units, and they have means that are approximately the same.

Empirical (or 68-95-99.7) **Rule**

Another concept that is helpful in interpreting the value of a standard deviation is the ***empirical rule***. For data sets having a distribution that is approximately bell shaped, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



Example

Empirical Rule IQ scores have a bell-shaped distribution with mean of 100 and a standard deviation of 15. What percentages of IQ scores are between 70 and 130?

Solution

$130 = 100 + 15 + 15$
$70 = 100 - 15 - 15$

70 and 130 are each exactly 2 standard deviation away from the mean 100.
2 standard deviation = $2s = 2(15) = 30$

Therefore, 2 standard deviation from the mean is

$100 - 30 = 70$
$100 + 30 = 130$

The empirical rule tells us that about 95% of all values are within 2 standard deviation of the mean, so about 95% of all IQ scores are between 70 and 130.

Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within *K* standard deviations of the mean is always at least $\left(1-\frac{1}{K^2}\right)100\%$, where *K* is any positive number greater than 1.

- For $K = 2$, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.
- For $K = 3$, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.

Chebyshev's Inequality

For any data set or distribution, at least $\left(1-\frac{1}{K^2}\right)100\%$ of the observations lie within *k* standard deviations of the mean, where *k* is any number greater than 1. That is, at least $\left(1-\frac{1}{K^2}\right)100\%$ of the data lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

Note: We can also use Chebyshev's Inequality based on sample data.

Example

Chebyshev's Theorem IQ scores have a mean of 100 and a standard deviation of 15. What can we conclude from Chebyshev's Theorem?

Solution

We can conclude that:

At least $\frac{3}{4}$ (or 75%) of IQ scores are within 2 standard deviation of the mean (between 70 and 130).

At least $\frac{8}{9}$ (or 89%) of IQ scores are within 3 standard deviation of the mean (between 55 and 145).

Standard Deviation Defined

For a particular data value of x, the amount of deviation is $x - \bar{x}$. Those deviations could be a negative numbers, and the sum could be zero. To get statistic that measures variation (instead of always zero), we need to avoid canceling out of negative and positive numbers. We can get the ***mean absolute deviation*** (or **MAD**), which is the mean distance of the data from the mean:

$$mean\ absolute\ deviation = \frac{\sum|x-\bar{x}|}{n}$$

Example

The following data represent the serum HDL cholesterol of the 54 female patients of a family doctor.

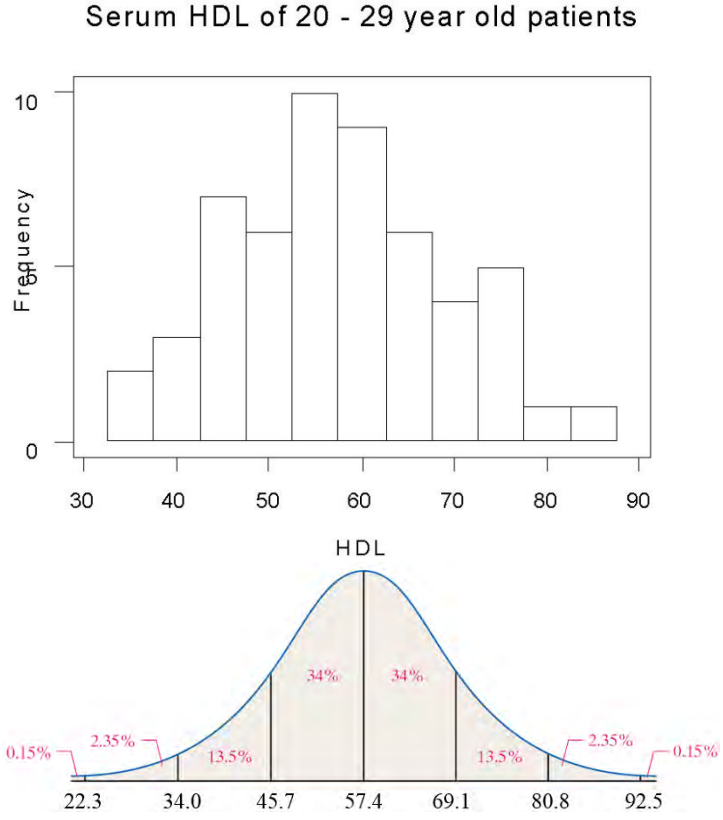
41	48	43	38	35	37	44	44	44
62	75	77	58	82	39	85	55	54
67	69	69	70	65	72	74	74	74
60	60	60	61	62	63	64	64	64
54	54	55	56	56	56	57	58	59
45	47	47	48	48	50	52	52	53

a) Compute the population mean and standard deviation.
b) Draw a histogram to verify the data is bell-shaped.
c) Determine the percentage of all patients that have serum HDL within 3 standard deviations of the mean according to the Empirical Rule.
d) Determine the percentage of all patients that have serum HDL between 34 and 69.1 according to the Empirical Rule.
e) Determine the actual percentage of patients that have serum HDL between 34 and 69.1.
f) Determine the percentage of patients that have serum HDL within 3 standard deviations of the mean.
g) Determine the actual percentage of patients that have serum HDL between 34 and 80.8 (within 3 SD of mean).

Solution

a) Using a TI-83 plus graphing calculator, we find $\mu = 57.4$ and $\sigma = 11.7$

b)



c) According to the Empirical Rule, 99.7% of the all patients that have serum HDL within 3 standard deviations of the mean.

d) 13.5% + 34% + 34% = 81.5% of all patients will have a serum HDL between 34.0 and 69.1 according to the Empirical Rule.

e) 45 out of the 54 or 83.3% of the patients have a serum HDL between 34.0 and 69.1.

f) $\left(1-\frac{1}{3^2}\right)100\% = 88.9\%$

g) $\frac{52}{54} \approx 0.96 = 96\%$

Definition

The ***coefficient of variation*** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

Sample	***Population***
$CV = \frac{s}{\bar{x}} \cdot 100\%$	$CV = \frac{\sigma}{\mu} \cdot 100\%$

Example

Compare the variation in heights of men to the variation in weights of men, using these sample results obtained from data below

Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

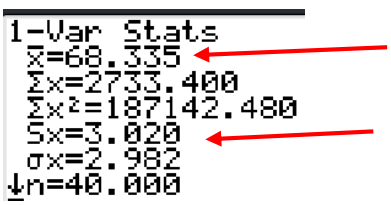
Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

Solution

The heights yield: $\bar{x} = 68.34$ in. and $s = 3.02$ in.

The weights yield: $\bar{x} = 172.55$ lb. and $s = 26.33$ lb.



Heights

Heights $CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02}{68.34} \cdot 100\% = 4.42\%$

Weights $CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33}{172.55} \cdot 100\% = 15.26\%$

We can see that heights (with CV = 4.42%) have considerably less variation than weights (with CV = 1%). This make intuitive sense, because the weights among men vary much more then heights.
It is very rare to see two adult men with one of them being twice as tall as other, but it is much more common to see two men with one of them weighing twice as much as the other.

Exercises *Section* 1.9 – Measures of Dispersion

1. In statistics, how do variation and variance differ?

2. Collegiate Dictionary has 1459 pages of defined words. Listed below are the numbers of defined words per page for a simple random sample of those pages. If we use this sample as a basis for estimating the total number of defined words in the dictionary.

51 63 36 43 34 62 73 39 53 79

a) Find the range, variance, and standard deviation.

b) How does the variation of these numbers affect our confidence on the accuracy of the estimate?

3. The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those teats, with the measurements given in hic (standard head injury condition units.

774 249 1210 546 431 612

a) Find the range, variance, and standard deviation

b) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?

4. The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below

$7448 $4911 $9051 $6374 $4277

a) Find the range, variance, and standard deviation

b) Do the different measures of center differ very much?

5. Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA's Space Transport System (space shuttle).

73 95 235 192 165 262 191 376 259 235 381 331 221 244 0

a) Find the range, variance, and standard deviation

b) How might that duration time be explained?

6. Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.

448 242 231 246 246 293 280 227 213 262 239 213 258 255 257

a) Find the range, variance, and standard deviation

b) Is there on time that is very different from the others?

7. Listed below are numbers of satellites in orbit from different countries.

158 17 15 17 7 3 5 1 8 3 4 2 4 1 2 3 1 1 1 1 1 1 1 1

a) Find the range, variance, and standard deviation

b) Does on country have an exceptional number of satellites?

8. Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

Find the coefficient of variation for each of the two sets of data, then compare the variation.

9. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the coefficient of variation for each of the two sets of data, then compare the variation.

10. Find the Standard Deviation from the frequency distribution and find the standard deviation of sample summarized in a frequency distribution table by using the formula

$s = \sqrt{\dfrac{n\left[\sum\left(f \cdot x^2\right)\right] - \left[\sum\left(f \div x\right)\right]^2}{n(n-1)}}$, where x represents the class midpoint, f represents the class frequency, and n represents the total number of sample values.

a)

Tar (mg) in Nonfiltered Cigarettes	***Frequency***
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

Pulse Rates of Females	***Frequency***
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

11. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using the empirical rule, what is the approximate percentage of women between

a) 154 cm and 168 cm?

b) 147 cm and 175 cm?

12. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts, and the voltages have a bell-shaped distribution. Using the empirical rule, what is the approximate percentage of voltage amounts between

a) 124.4 volts and 125.6 volts?

b) 124.1 volts and 125.9 volts?

13. The mean value of land and buildings per acre from a sample of farms is \$1,500, with a standard deviation of \$200. Using the empirical rule, estimate the percent of farms whose land and building values per acre are between \$1,300 and \$1,700. (Assume the data set has a bell-shaped distribution.)

14. The mean value of land and buildings per acre from a sample of farms is \$2,400, with a standard deviation of \$450. Using the empirical rule, between what two values do about 95% of the data lie? (Assume the data set has a bell-shaped distribution.)

15. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using Chebyshev's Theorem, what do we know about the percentage of women with heights that are within 2 standard deviations of the mean? What are the minimum and maximum heights that are within 2 standard deviations of the mean?

16. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts. Using Chebyshev's Theorem, what do we know about the percentage of voltage amounts that are within 3 standard deviations of the mean? What are the minimum and maximum voltage amounts that are within 3 standard deviations of the mean?

17. The mean time in a women's 400-meter dash is 57.07 seconds, with a standard deviation of 1.05 seconds. Apply Chebyshev's Theorem to the data using $k = 2$. Interpret the results.

18. The number of gallons of water consumed per day by a small village are listed. Make a frequency distribution (using five classes) for the data set. Then approximate the population mean and the population standard deviation of the data set.

167	180	192	173	145	151	174	175	178	160
195	224	244	146	162	146	177	163	149	188

19. To get the best deal on a microwave oven, Jeremy called six appliance stores and asked the cost of a specific model. The prices he was quoted are listed below:

\$325 \$384 \$156 \$210 \$219 \$284

Find the variance for the given data.

20. Compare the variation in heights to the variation in weights of thirteen-year old girls. The heights (in inches) and weights (in pounds) of nine randomly selected thirteen-year old girls as listed below

Heights (inches):	59.3	61.2	62.6	64.7	60.1	58.3	64.6	63.7	66.1
Weights (pounds):	87	96	91	119	96	90	123	98	139

Find the coefficient of variation for each of the two sets of data, then compare the variation

21. The amount of Jen's monthly phone bill is normally distributed with a mean of \$56 and a standard deviation of \$9. What percentage of her phone bills are between \$29 and \$83? Use the empirical rule to solve.