Measures of Variation: The Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Where $\overline{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Suppose we draw n independent observations from a population with mean μ and variance σ^2.

Usually unknown

Usually unknown

The sample mean $\bar{x}$ estimates the population mean μ.

The sample variance s^2 estimates the population variance σ^2.

Ideally we would estimate σ^2 with:

$$\frac{\sum(x_i - \mu)^2}{n}$$

This is the average squared distance from the true mean

Problem: μ is unknown!

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

On average, this estimator equals the population variance σ^2.

We could try:

$$\frac{\sum(x_i - \bar{x})^2}{n}$$

Tends to underestimate σ^2

The Degrees of Freedom

Suppose we draw 3 independent observations from a population where $\mu = 8$.

i	x_i	$x_i - \mu$
1	9	$9 - 8 = 1$
2	4	$4 - 8 = -4$
3	?	?

Suppose the population mean is known to be 8

We lost one degree of freedom when we estimated the population mean with the sample mean

Suppose we have the same situation, but μ is unknown. We find $\bar{x} = 5$.

i	x_i	$x_i - \bar{x}$
1	9	$9 - 5 = 4$
2	4	$4 - 5 = -1$
3	2	$2 - 5 = -3$

The deviations from the sample mean always sum to 0

$$s^2 = \frac{(9-5)^2 + (4-5)^2 + (2-5)^2}{3-1}$$

Once we know two of these values, we know what the third value must be.

When estimating the population variance, we typically divide by the degrees of freedom as opposed to the sample size.

Measures of Variation: The Sample Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the same units as the original data

- Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Example

Sample Data (X_i) : 10 12 14 15 17 18 18 24

n = 8 Mean = $\overline{X}$ = 16

$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \Lambda + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \Lambda + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

Example

Observation (x) (1)	Mean ($\bar{x}$) (2)	$x - \bar{x}$ (1) − (2)	$(x - \bar{x})^2$ [(1) − (2)]^2	x^2 (1)^2
863	1,351	−488	238,144	744,769
903	1,351	−448	200,704	815,409
957	1,351	−394	155,236	915,849
1,041	1,351	−310	96,100	1,083,681
1,138	1,351	−213	45,369	1,295,044
1,204	1,351	−147	21,609	1,449,616
1,354	1,351	3	9	1,833,316
1,624	1,351	273	74,529	2,637,376
1,698	1,351	347	120,409	2,883,204
1,745	1,351	394	155,236	3,045,025
1,802	1,351	451	203,401	3,247,204
1,883	1,351	532	283,024	3,545,689
			$\Sigma(x - \bar{x})^2 \rightarrow$ 1,593,770	23,496,182 $\leftarrow \Sigma x^2$

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

$$= \frac{1{,}593{,}770}{11}$$

$= 144{,}888$ (or 144,888 [thousands of dollars]^2) $\leftarrow$ Sample variance

$$s = \sqrt{s^2}$$

$$= \sqrt{144{,}888}$$

$= 380.64$ (that is, \$380,640) $\leftarrow$ Sample standard deviation

SD for Grouped Data

Frequency Distribution of Return on Investment of Mutual Funds

Return on Investment	Number of Mutual Funds
5-10	10
10-15	12
15-20	16
20-25	14
25-30	8
Total	**60**

Solution

$\bar{X} = \frac{\sum fx}{\sum f}$

A	B	C	D	E	F	G	H
1	Return on	Investment		No of			
2			MidPoint	Funds			
3	Lower limit	Upper Limit	X	f	fx	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
4	5	10	7.50	10	75	96.69	966.94
5	10	15	12.50	12	150	23.36	280.33
6	15	20	17.50	16	280	0.03	0.44
7	20	25	22.50	14	315	26.69	373.72
8	25	30	27.50	8	220	103.36	826.89
9				**60**	**1040**		**2448.33**
10				Mean=	17.333		
11				Sample Variance =			41.50
12				Sample Standard Deviation =			6.44

Example

Class	Midpoint x (1)	Frequency f (2)	$f \times x$ (3) = (2) × (1)	Mean μ (4)	$x - \mu$ (1) − (4)	$(x - \mu)^2$ $[(1) - (4)]^2$	$f(x - \mu)^2$ (2) × $[(1) - (4)]^2$
700– 799	750	4	3,000	1,250	−500	250,000	1,000,000
800– 899	850	7	5,950	1,250	−400	160,000	1,120,000
900– 999	950	8	7,600	1,250	−300	90,000	720,000
1,000–1,099	1,050	10	10,500	1,250	−200	40,000	400,000
1,100–1,199	1,150	12	13,800	1,250	−100	10,000	120,000
1,200–1,299	1,250	17	21,250	1,250	0	0	0
1,300–1,399	1,350	13	17,550	1,250	100	10,000	130,000
1,400–1,499	1,450	10	14,500	1,250	200	40,000	400,000
1,500–1,599	1,550	9	13,950	1,250	300	90,000	810,000
1,600–1,699	1,650	7	11,550	1,250	400	160,000	1,120,000
1,700–1,799	1,750	2	3,500	1,250	500	250,000	500,000
1,800–1,899	1,850	1	1,850	1,250	600	360,000	360,000
		100	125,000				6,680,000

$$\bar{x} = \frac{\Sigma(f \times x)}{n}$$

$$= \frac{125,000}{100}$$

$$= 1,250 \text{ (thousands of dollars)} \leftarrow \text{Mean}$$

$$\sigma^2 = \frac{\Sigma f(x - \mu)^2}{N}$$

$$= \frac{6,680,000}{100}$$

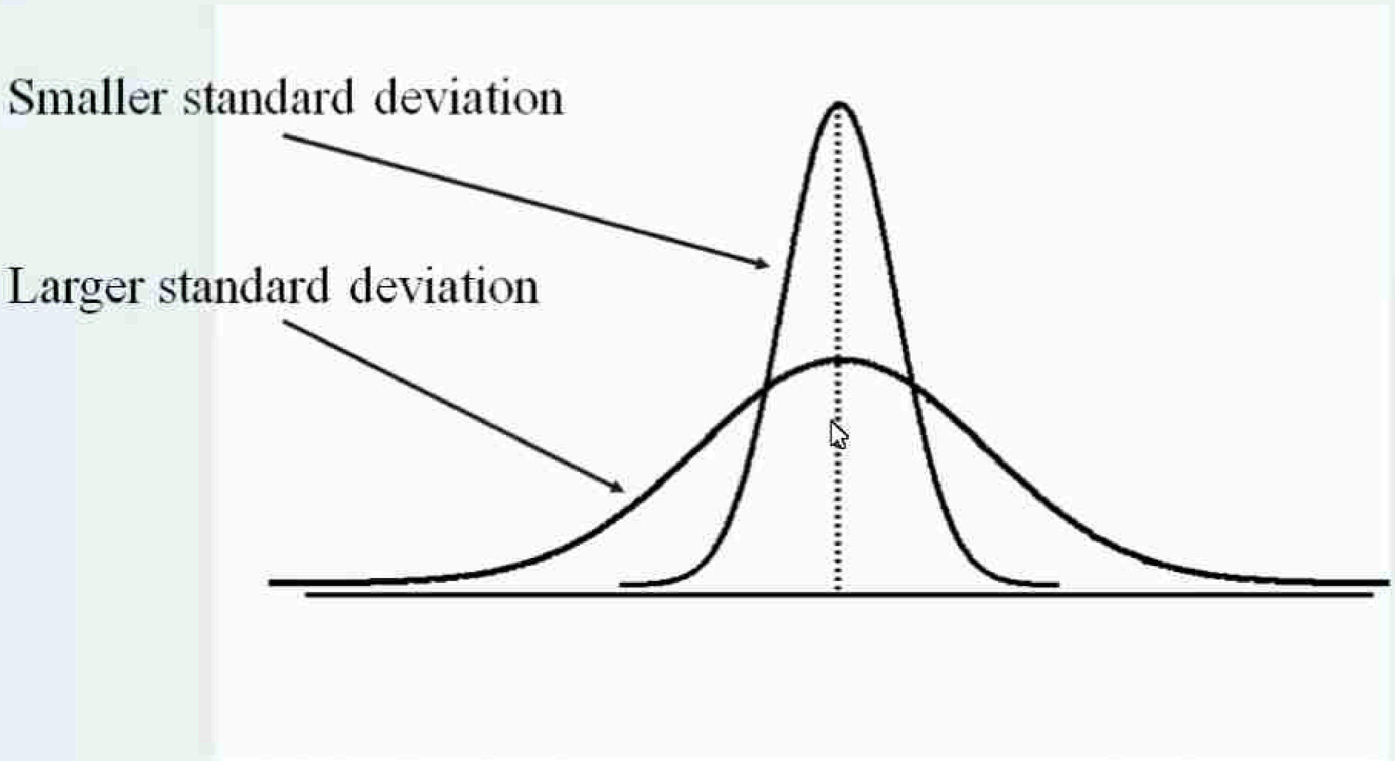
$$= 66,800 \text{ (or } 66,800 \text{ (thousands of dollars)}^2\text{)} \leftarrow \text{Variance}$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{66,800}$$

$$= 258.5 \leftarrow \text{Standard deviation} = \$258,500$$

Measures of Variation: Comparing Standard Deviations



Coefficient of Variation

The **coefficient of variation** (CV) is a measure of relative **variability**.

It is the ratio of the **standard deviation to the mean** (average).

Always in percentage (%)

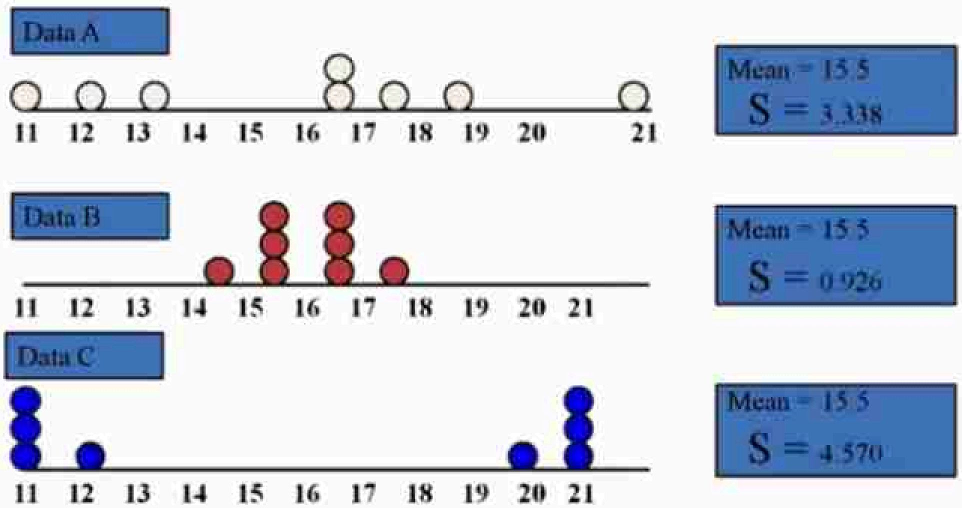
Shows **variation relative to mean**

Can be used to compare the variability of two or more sets of data measured in **different units**

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Measure of Variation

The **coefficient of variation** (CV) is a measure of relative **variability**. It is the ratio of the **standard deviation to the mean** (average).



Cont'd…

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\overline{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\overline{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price