Measures of Variability

Range

Variance

Standard Deviation

Coefficient of Variation

Measures of Variability

Table 2.11: Annual Payouts for Two Different Investment Funds

Year	Fund A ($)	Fund B ($)
1	1,100	700
2	1,100	2,500
3	1,100	1,200
4	1,100	1,550
5	1,100	1,300
6	1,100	800
7	1,100	300
8	1,100	1,600
9	1,100	1,500
10	1,100	350
11	1,100	460

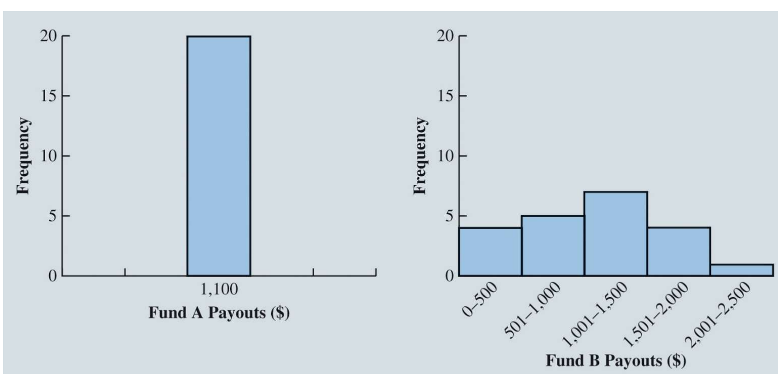
Measures of Variability

Table 2.11: Annual Payouts for Two Different Investment Funds (cont.)

Year	Fund A ($)	Fund B ($)
12	1,100	890
13	1,100	1,050
14	1,100	800
15	1,100	1,150
16	1,100	1,200
17	1,100	1,800
18	1,100	100
19	1,100	1,750
20	1,100	1,000
Mean	1,100	1,100

Measures of Variability

Figure 2.18: Histograms for Payouts of Past 20 Years from Fund A and Fund B



Measures of Variability

Range:

- The **range** can be found by subtracting the smallest value from the largest value in a data set.

- Illustration: Consider the data on home sales in a Cincinnati, Ohio, suburb.
  - Largest home sales price: $456,250.
  - Smallest home sales price: $108,000.

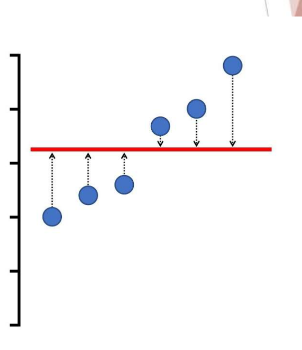
Range = Largest value – *S*mallest value
= $456,250 – $108,000
= $348,250

- Drawback: Range is based on only two of the observations and thus is highly influenced by extreme values.

Measures of Variability

- Variance:
  - is a measure of variability that utilizes all the data.
  - It is based on the deviation about the mean, which is the difference between the value of each observation

Population Variance	Sample Variance
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size



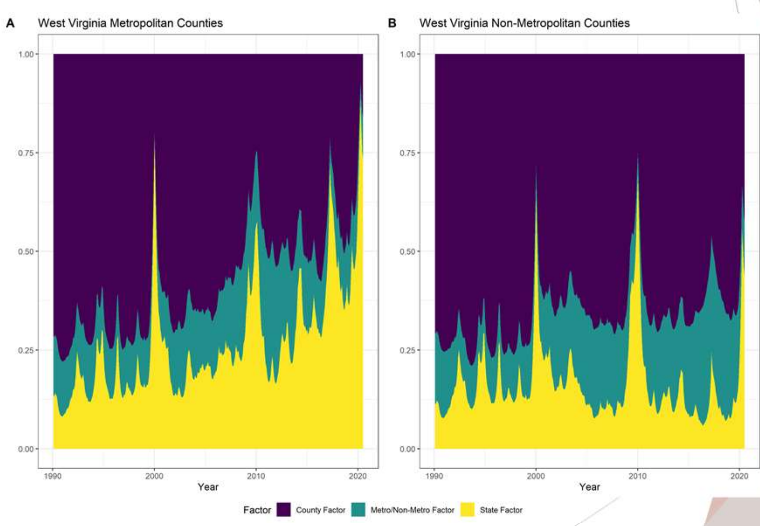
Measures of Variability

Table 2.12: Computation of Deviations and Squared Deviations About the Mean for the Class Size Data

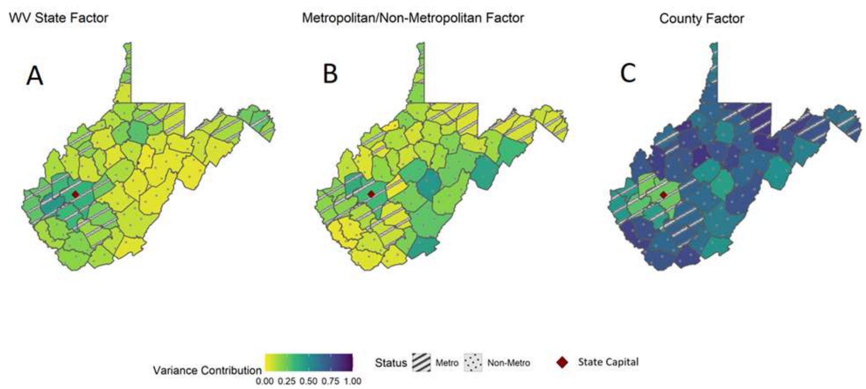
Number of Students in Class (x_i)	Mean Class Size ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

- Computation of Sample Variance: $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64 \text{ (students)}^2$

Variation in the Change in Labor Force Participation Rates



Variation in the Change in Labor Force Participation Rates



Measures of Variability

- **Standard deviation:**
  - is the positive square root of the variance.
- Measured in the same units as the original data.

SAMPLE STANDARD DEVIATION

$$s = \sqrt{s^2} \tag{2.5}$$

- For population, $\sigma = \sqrt{\sigma^2}$.

Measures of Variability

- The **coefficient of variation**
  - is a descriptive statistic that indicates how large the standard deviation is relative to the mean.
- Expressed as a percentage.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\% \qquad \textbf{(2.6)}$$

Measures of Variability

Illustration:

- Consider the class size data:

  46 54 42 46 32

- Mean,

  $\bar{x} = 44.$

- Standard deviation, s = 8 (students).

- Coefficient of variation =

$$\left(\frac{8}{44} \times 100\right)\% = 18.2\%.$$

The coefficient of variation tells that the sample standard deviation is 18.2% of the value of the sample mean.

Analyzing Distributions

Percentiles
Quartiles
z-Scores
Empirical Rule
Identifying Outliers
Boxplots

Analyzing Distributions

- A **percentile**
  - is the value of a variable at which a specified (approximate) percentage of observations are below that value.
- The *p*th percentile tells us the point in the data where:
  - Approximately *p* percent of the observations have values less than the *p*th percentile.
  - Approximately:

$(100-p)$ percent of the observations have values greater than the *p*th percentile.

Location of the *p*th Percentile

$$L_p = \frac{p}{100}(n+1) \quad \textbf{(2.7)}$$

Analyzing Distributions

Illustration:

- To determine the 85th percentile for the home sales data:
- Arrange the data in ascending order:

108,000	138,000	138,000	142,000	186,000	199,500
208,000	254,000	254,000	257,500	298,000	456,250

2. Compute

$$L_{85} = \frac{p}{100}(n+1) = \left(\frac{85}{100}\right)(12+1) = 11.05.$$

3. The interpretation of $L_{85} = 11.05$ is that the 85th percentile is 5% of the way between the value in position 11 and value in position 12.

Analyzing Distributions

Illustration (cont.):

- To determine the 85th percentile for the home sales data:
  - The value in the 11th position is 298,000.
  - The value in the 12th position is 456,250.
  - \$305,912.50 represents the 85th percentile of the home sales data:

$$\begin{aligned}\text{85th percentile} &= 298{,}000 + 0.05(456{,}250 - 298{,}000)\\ &= 298{,}000 + 0.05(158{,}250)\\ &= 305{,}912.50\end{aligned}$$

Analyzing Distributions

- **Quartiles:** When the data is divided into four equal parts:
  - Each part contains approximately 25% of the observations.
  - Division points are referred to as quartiles.

Q_1 = first quartile, or 25th percentile.

Q_2 = second quartile, or 50th percentile (also the median).

Q_3 = third quartile or 75th percentile.

- The difference between the third and first quartiles is often referred to as the **interquartile range**, or IQR.

Analyzing Distributions

- The **z-score:**
  - measures the relative location of a value in the data set.
- Helps to determine how far a particular value is from the mean relative to the data set's standard deviation.
- Often called the standardized value.
- The z -score can be interpreted as the number of standard deviations x_i is from the mean $\bar{x}$.

Analyzing Distributions

- z-Scores (cont.):

If $x_1, x_2, \dots, x_n$ is a sample of n observations:

z-SCORE

$$z_i = \frac{x_i - \bar{x}}{s} \tag{2.8}$$

where

z_i = the z-score for x_i
$\bar{x}$ = the sample mean
s = the sample standard deviation

Analyzing Distributions

Table 2.13: z-Scores for the Class Size Data

No. of Students in Class (x_i)	Deviation About the Mean ($x_i - \bar{x}$)	z-Score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	–2	−2/8 = −0.25
46	2	2/8 = 0.25
32	–12	−12/8 = −1.50

For class size data, $\bar{x} = 44$ and $s = 8$.

For observations with a value > mean, z-score > 0.

For observations with a value < mean, z-score < 0.

Analyzing Distributions

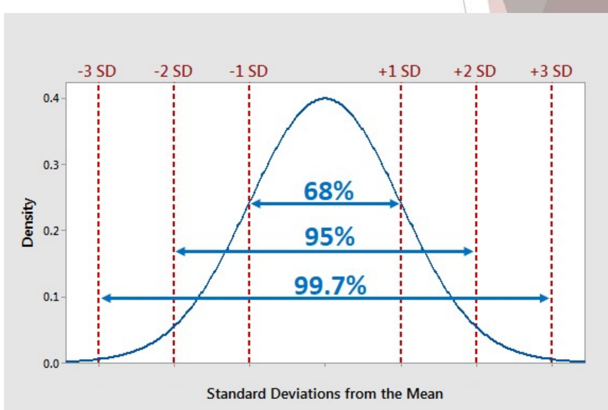
Empirical Rule:

- For data having a bell-shaped distribution:
  - Approximately 68% of the data values will be within 1 standard deviation.
  - Approximately 95% of the data values will be within 2 standard deviations.
  - Almost all the data (99.7%) values will be within 3 standard deviations.

Analyzing Distributions

- The height of adult males in the United States
  - Mean 69.5 inches and
  - Standard deviation of 3 inches.
- 1 SD -> 68% of U.S. males are between 66.5 and 72.5 in tall
- 2 SD -> 95% of U.S. males are between 63.5 and 75.5 in tall
- 3 SD -> 99.7% of U.S. males are between 60.5 and 78.5 in tall

A Symmetric Bell-Shaped Distribution



Analyzing Distributions

- **Outliers:**
  - Extreme values in a data set.

- They can be identified using standardized values (z-scores).
- Any data value with a z-score less than –3 or greater than +3 is an outlier.
- Such data values can then be reviewed to determine their accuracy and whether they belong in the data set.

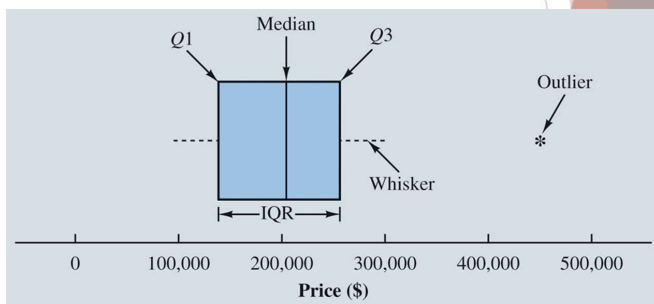
Analyzing Distributions

- **Outliers:**
  - Data value incorrectly recorded;
    - Correct before further analysis.
  - Data value incorrectly included
    - It can be removed.
  - Unusual data value that has been recorded correctly
    - The observation should remain.

Analyzing Distributions

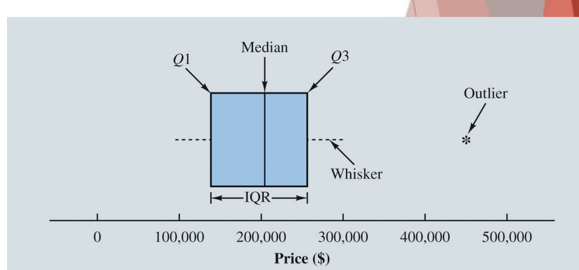
- A **boxplot:**
  - is a graphical summary of the distribution of data.
- Developed from the quartiles for a data set.

Figure 2.22: Boxplot for the Home Sales Data



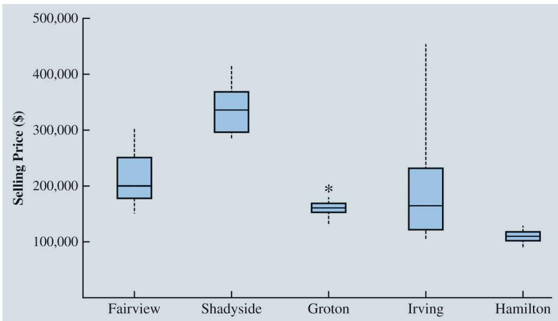
Analyzing Distributions

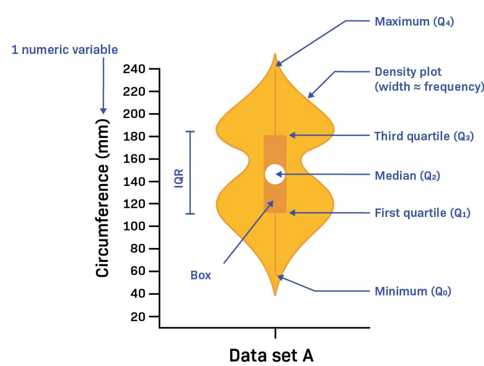
- Steps used to male boxplot:
  - Draw box with ends located at the first and third quartiles.
    - E.g. Q1 = 139,000 and Q3 = 256,625.
  - Draw a line at the median (203,750)
  - Find Outliers
    - IQR = Q3 – Q1
    - Limits = Q1 – 1.5(IQR), Q3 + 1.5(IQR)
      - IQR = 117,625.
      - Limits =
      - 139,000 – 1.5(117,625) = -37,437.5
      - 256,625 + 1.5(117,625) = 433,062.5
- Draw Whiskers
  - Dashed lines from the ends of the box to the smallest and largest values inside the limits computed in Step 3.
    - Values of 108,000 and 298,000.
- Finally, the location of each outlier is shown with an asterisk
  - 456,250.



Analyzing Distributions

Figure 2.23: Boxplots Comparing Home Sale Prices in Different Communities





What do you observe about the differences in Home Prices for different communities?

Measures of Variability

Figure 2.19: Calculating Variability Measures for the Home Sales Data in Excel

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	=AVERAGE(B2:B13)
3	2	254000		Median:	=MEDIAN(B2:B13)
4	3	186000		Mode 1:	=MODE.MULT(B2:B13)
5	4	257500		Mode 2:	=MODE.MULT(B2:B13)
6	5	108000			
7	6	254000		Range:	=MAX(B2:B13)-MIN(B2:B13)
8	7	138000		Variance:	=VAR.S(B2:B13)
9	8	298000		Standard Deviation:	=STDEV.S(B2:B13)
10	9	199500			
11	10	208000		Coefficient of Variation:	=E9/E2
12	11	142000			
13	12	456250		85th Percentile:	=PERCENTILE.EXC(B2:B13,0.85)
14					
15				1st Quartile:	=QUARTILE. EXC(B2:B13,1)
16				2nd Quartile:	=QUARTILE. EXC(B2:B13,2)
17				3rd Quartile:	=QUARTILE. EXC(B2:B13,3)
18					
19				IQR:	=E17-E15

	A	B	C	D	E
1	Home Sale	Selling Price ($)			
2	1	138000		Mean:	$ 219,937.50
3	2	254000		Median:	$ 203,750.00
4	3	186000		Mode 1:	$ 138,000.00
5	4	257500		Mode 2:	$ 254,000.00
6	5	108000			
7	6	254000		Range:	$ 348,250.00
8	7	138000		Variance:	9037501420
9	8	298000		Standard Deviation:	$ 95,065.77
10	9	199500			
11	10	208000		Coefficient of Variation:	43.22%
12	11	142000			
13	12	456250		85th Percentile:	$ 305,912.50
14					
15				1st Quartile:	$ 139,000.00
16				2nd Quartile:	$ 203,750.00
17				3rd Quartile:	$ 256,625.00
18					
19				IQR:	$ 117,625.00

Analyzing Distributions

Figure 2.20: Calculating z-Scores for the Home Sales Data in Excel

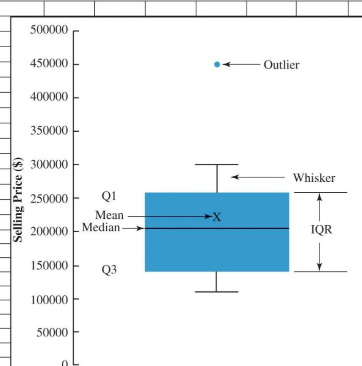
	A	B	C
1	Home Sale	Selling Price ($)	z-Score
2	1	138000	=STANDARDIZE(B2,B15,B16)
3	2	254000	=STANDARDIZE(B3,B15,B16)
4	3	186000	=STANDARDIZE(B4,B15,B16)
5	4	257500	=STANDARDIZE(B5,B15,B16)
6	5	108000	=STANDARDIZE(B6,B15,B16)
7	6	254000	=STANDARDIZE(B7,B15,B16)
8	7	138000	=STANDARDIZE(B8,B15,B16)
9	8	298000	=STANDARDIZE(B9,B15,B16)
10	9	199500	=STANDARDIZE(B10,B15,B16)
11	10	208000	=STANDARDIZE(B11,B15,B16)
12	11	142000	=STANDARDIZE(B12,B15,B16)
13	12	456250	=STANDARDIZE(B13,B15,B16)
14			
15	Mean:	=AVERAGE(B2:B13)	
16	Standard Deviation:	=STDEV.S(B2:B13)	

	A	B	C
1	Home Sale	Selling Price ($)	z-Score
2	1	138,000	–0.862
3	2	254,000	0.358
4	3	186,000	–0.357
5	4	257,500	0.395
6	5	108,000	–1.177
7	6	254,000	0.358
8	7	138,000	–0.862
9	8	298,000	0.821
10	9	199,500	–0.215
11	10	208,000	–0.126
12	11	142,000	–0.820
13	12	456,250	2.486
14			
15	Mean:	$ 219,937.50	
16	Standard Deviation:	$ 95,065.77	

Analyzing Distributions

Figure 2.24: Boxplot Created in Excel for Home Sales Data

	A	B
1	Home Sale	Selling Price ($)
2	1	138000
3	2	254000
4	3	186000
5	4	257500
6	5	108000
7	6	254000
8	7	138000
9	8	298000
10	9	199500
11	10	208000
12	11	142000
13	12	456250
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		



Analyzing Distributions

Figure 2.25: Boxplots for Multiple Variables Created in Excel

	Fairview	Shadyside	Groton	Irving	Hamilton
Selling Prices	302,000	336,000	152,000	201,000	102,000
	265,000	398,000	158,000	365,000	108,000
	280,000	378,000	154,000	115,000	88,000
	220,000	298,000	170,000	105,000	111,000
	149,000	425,000	132,000	225,000	105,000
	155,000	344,000	164,000	115,000	87,000
	198,000	302,000	198,000	108,000	95,000
	187,000	300,000	158,000	218,000	111,000
	208,000	298,000	149,000	454,000	98,000
	174,000	342,000	165,000	103,000	78,000