Measures of Association Between Two Variables

Scatter Charts

Covariance

Correlation Coefficient

96

Measures of Association Between Two Variables

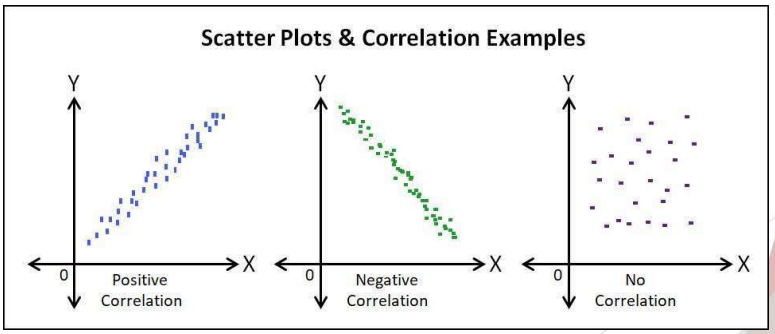
Table 2.14: Data for Bottled Water Sales at Queensland Amusement Park for a Sample of 14 Summer Days

High Temperature (°F)	Bottled Water Sales (cases)
78	23
79	22
80	24
80	22
82	24
83	26
85	27
86	25
87	28
87	26
88	29
88	30
90	31
92	31

97

Measures of Association Between Two Variables

- A **scatter chart:**
  - is a useful graph for analyzing the relationship between two variables.

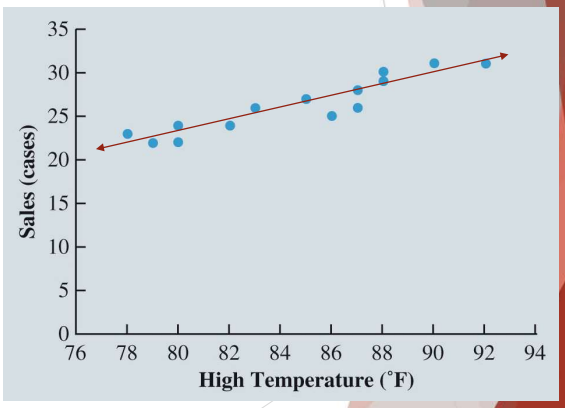


98

Measures of Association Between Two Variables

- Positive relationship:
  - As high temperature increases, sales of bottled water generally also increases.
- Best Fit :
  - a straight line could be used as an approximation for the relationship between high temperature and sales of bottled water.

Sales and High Temperatures



99

Measures of Association Between Two Variables

- **Covariance:**
  - is a descriptive measure of the linear association between two variables:

Sample Covariance

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad (2.9)$$

Population covariance, $\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)\Sigma(y_i - \mu_y)}{N}$.

100

Measures of Association Between Two Variables

Daily High Temperature and Bottled Water Sales at Queensland Amusement Park

- **If Covariance:**
  - Large and > 0 => Strong Positive Relationship
  - Near 0 => Not Linearly Related
  - Large and <0 => Strong Negative Relationship

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	78	23	−6.6	−3.3	21.78
	79	22	−5.6	−4.3	24.08
	80	24	−4.6	−2.3	10.58
	80	22	−4.6	−4.3	19.78
	82	24	−2.6	−2.3	5.98
	83	26	−1.6	−0.3	0.48
	85	27	0.4	0.7	0.28
	86	25	1.4	−1.3	−1.82
	87	28	2.4	1.7	4.08
	87	26	2.4	−0.3	−0.72
	88	29	3.4	2.7	9.18
	88	30	3.4	3.7	12.58
	90	31	5.4	4.7	25.38
	92	31	7.4	4.7	34.78
Totals	1,185	368	0.6	−0.2	166.42

$\bar{x} = 84.6$

$\bar{y} = 26.3$

$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{166.42}{14 - 1} = 12.8$

101

Measures of Association Between Two Variables

- The **correlation coefficient:**
  - Measures the relationship between two variables.
  - Not affected by the units of measurement for *x* and *y*.

Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

r_{xy} = sample correlation coefficient
s_{xy} = sample covariance
s_x = sample standard deviation of x
s_y = sample standard deviation of y

$$r_{xy} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}}$$

102

Measures of Association Between Two Variables

Interpretation of Correlation Coefficient:

$-1 \le r \le +1$

***r* value**	**Relationship between the *x* and *y* variables**
< 0	Negative linear
Near 0	No linear relationship
> 0	Positive linear

103

Measures of Association Between Two Variables

Illustration:

- To determine the sample correlation coefficient for bottled water sales at Queensland Amusement Park:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{12.8}{(4.36)(3.15)} = 0.93$$

- There is a very strong linear relationship between high temperature and sales.

104

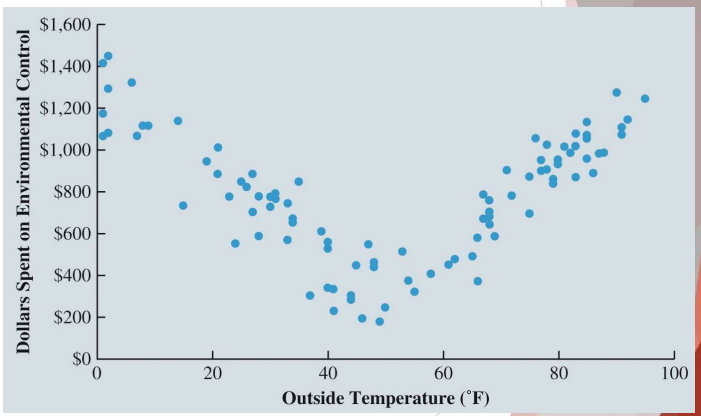
Measures of Association Between Two Variables

Nonlinear Relationship Producing a Correlation Coefficient Near Zero

Correlation Coefficient = $r_{xy} = -0.007$

What does this indicate?

Is there a different type of relationship?



105

Measures of Association Between Two Variables

Figure 2.27: Calculating Covariance and Correlation Coefficient for Bottled Water Sales Using Excel

	A	B
1	High Temperature (°F)	Bottled Water Sales (cases)
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	Covariance:	=COVARIANCE.S(A2:A15,B2:B15)
18	Correlation Coefficient:	=CORREL(A2:A15,B2:B15)

	A	B
1	High Temperature (°F)	Bottled Water Sales (cases)
2	78	23
3	79	22
4	80	24
5	80	22
6	82	24
7	83	26
8	85	27
9	86	25
10	87	28
11	87	26
12	88	29
13	88	30
14	90	31
15	92	31
16		
17	Covariance:	12.80
18	Correlation Coefficient:	0.93

106

Data Cleaning

Missing Data

Blakely Tires

Identification of Erroneous Outliers and other Erroneous Values

Variable Representation

107

Data Cleaning

- Data sets commonly include observations with missing values for one or more variables.
- In some cases, missing data naturally occur; these are called:
- **Legitimately missing data.**
  - Generally, no remedial action is taken for legitimately missing data.

108

Data Cleaning

- **Illegitimately missing data.**
- Addressing such missing data:
  1. Discard observations (rows) with any missing values.
  2. Discard any variable (column) with missing values.
  3. Fill in missing entries with estimated values.
  4. Apply a data-mining algorithm that can handle missing values.

109

Data Cleaning

- **Missing completely at random** (MCAR):
- whether data are missing does not depend on either the value of the missing data or the value of any other variable in the data.

Deletion

Probability of missing data is same for all cases

Rare in the Real World

110

Data Cleaning

- **Missing at random** (MAR):
- The tendency for an observation to be missing a value for some variable is related to the value of some other variable(s) in the data.

Imputation

Probability of missing data is NOT same for all cases

More Realistic!

111

Data Cleaning

- **Imputation:**
  - The systematic replacement of missing values with values that seem reasonable.

Age(yrs)	Ht>5	wt(lbs)	obese
20	no	130	no
24	yes	160	yes
26	no	150	no
25			no

Original data with missing values

Age(yrs)	Ht>5	wt(lbs)	obese
20	no	130	no
24	yes	160	yes
26	no	150	no
25	no	150	no

Filled missing values using strawman imputation

112

Data Cleaning

- **Missing not at random** (MNAR): The tendency for the value of a variable to be missing is related to the value that is missing.

Improve

Heavier weight => Missing Values

Others: Weaker opinionated people are less likely to respond to survey

113

Data Cleaning

Identification of Erroneous Outliers and other Erroneous Values:

- Uncover Data=quality Issues using:
  - summary statistics, frequency distributions, bar charts and histograms, z-scores, scatter plots, correlation coefficients, and other tools
- Many software ignore missing values when calculating various summary statistics.
- Warning:
  - Sometime missing values are indicated with a unique value (such as 9999999)
    - These values may skew summary statistics.
  - Find missing data before running summary statistics OR
  - Look over summary statistics very carefully!

114

Data Cleaning

Variable Representation:

- Data-Mining
  - May get too many variables to work with
- **Dimension reduction:**
  - is the process of removing variables from the analysis without losing crucial information.

- Critical:
  - Determine how to represent measurements of variables and which variables to consider.
  - Consider relationships or combinations of variables for more insights

115

Data Cleaning

Blakely Tires:

- A U.S. producer of automobile tires wants to learn about the conditions of its tires on automobiles in Texas.
- The data obtained includes the position of the tire on the automobile, age of the tire, mileage on the tire, and depth of the remaining tread on the tire.
- Begin assessing the quality of these data by determining which (if any) observations have missing values (see Figure 2.30).

116

Data Cleaning

Figure 2.30: Portion of Excel Spreadsheet Showing Number of Missing Values for Variables in *TreadWear* Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	13391487	LR	58.4	2.2	2805		# of Missing Values	0	0	1
3	21678308	LR	17.3	8.3	39371					
4	18414311	RR	16.5	8.6	13367					
5	19778103	RR	8.2	9.8	1931					
6	16355454	RR	13.7	8.9	23992					
7	8952817	LR	52.8	3.0	48961					
8	6559652	RR	14.7	8.8	4585					

117

Data Cleaning

Blakely Tires (cont.):

- Sort all of Blakely's data on Miles from smallest to largest value to determine which observation is missing its value of this variable.

Figure 2.31: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted on Miles from Lowest to Highest Value

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	15890813	LF	16.1	8.6	206		# of Missing Values	0	0	1
3	15890813	LR	16.1	8.6	206					
4	15890813	RF	16.1	8.6	206					
455	9306585	RR	45.4	4.1	107237					
456	9306585	LF	45.4	4.1	107237					
457	3354942	LF	17.1	8.5						

118

Data Cleaning

Figure 2.32: Portion of Excel Spreadsheet Showing *TreadWear* Data Sorted from Lowest to Highest by ID Number

54	3121851	LR	17.1	8.4	21378
55	3121851	RR	17.1	8.4	21378
56	3121851	RF	17.1	8.4	21378
57	3121851	LF	17.1	8.5	21378
58	3354942	LF	17.1	8.5	
59	3354942	RF	21.4	7.7	33254
60	3354942	RR	21.4	7.8	33254
61	3354942	LR	21.4	7.7	33254
62	3374739	RR	73.3	0.2	57313
63	3574739	RF	73.3	0.2	57313
64	3574739	LF	73.3	0.2	57313
65	3574739	LR	73.3	0.2	57313

119

Data Cleaning

Figure 2.33: Portion of Excel Spreadsheet Showing the Mean and Standard Deviation for Each Variable in the *TreadWear* Data

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	80441	LR	19.0	8.1	37419		# of Missing Values	0	0	1
3	80441	LF	19.0	8.1	37419		Mean	23.80	7.68	25440.22
4	80441	RR	19.0	8.2	37419		Standard Deviation	31.82	2.62	23600.21
5	80441	RF	19.0	8.1	37419					
6	95990	RR	8.6	9.7	5670					
7	95990	LR	8.6	9.7	5670					
8	95990	LF	8.6	9.7	5670					

120

Data Cleansing

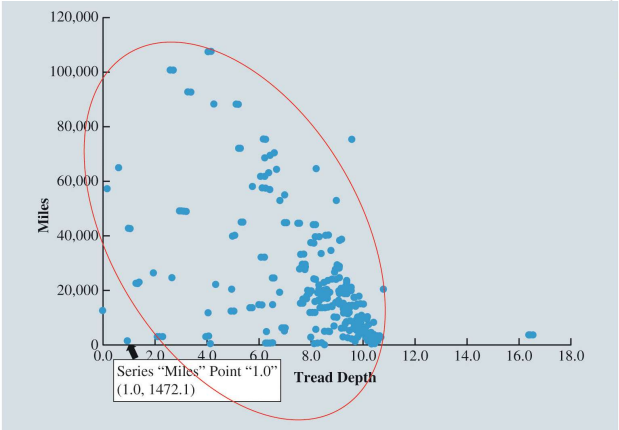
Figure 2.34: Portion of Excel Spreadsheet Showing the *TreadWear* Data Sorted on Life of Tires (Months) from Lowest to Highest Value

	A	B	C	D	E	F	G	H	I	J
1	ID Number	Position on Automobile	Life of Tire (Months)	Tread Depth	Miles			Life of Tire (Months)	Tread Depth	Miles
2	9091771	RF	1.8	10.8	2917		# of Missing Values	0	0	1
3	9091771	RR	1.8	10.7	2917		Mean	23.80	7.68	25440.22
4	9091771	LF	1.8	10.7	2917		Standard Deviation	31.82	2.62	23600.21
5	7712178	LF	2.1	10.7	2186		Minimum	1.8		
6	7712178	RR	2.1	10.7	2186		Maximum	601.0		
452	3574739	RR	73.3	0.2	57313					
453	3574739	LF	73.3	0.2	57313					
454	3574739	LR	73.3	0.2	57313					
455	3574739	LR	73.3	0.2	57313					
456	2122934	LR	111.0	9.3	21000					
457	8696859	LR	601.0	2.0	26129					

121

Data Cleaning

Figure 2.35: Scatter Diagram of Tread Depth and Miles for the *TreadWear* Data



122