

IIT KHARAGPUR



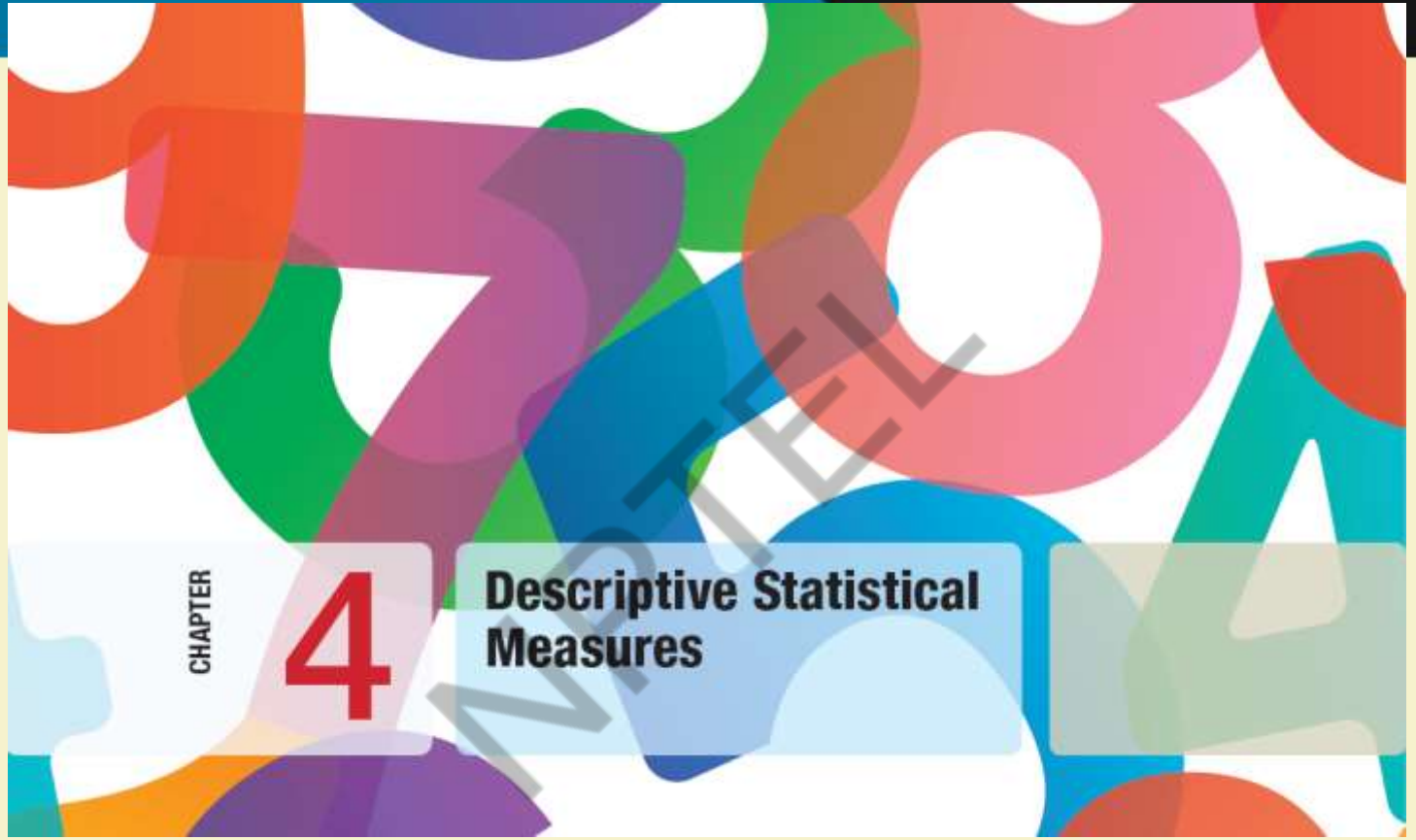
NPTEL ONLINE CERTIFICATION COURSES

Business Analytics for Management Decision

Chapter 4 Descriptive Statistical Measures

Rudra P. Pradhan
Vinod Gupta School of Management
IIT KHARAGPUR

NPTEL



NPTEL

Chapter 4 Topics

- Populations and Samples
- Measures of Location
- Measures of Dispersion
- Measures of Shape
- Excel *Descriptive Statistics* Tool
- Descriptive Statistics for Grouped Data
- Descriptive Statistics for Categorical Data: The Proportion
- Statistics in PivotTables
- Measures of Association
- Outliers
- Statistical Thinking in Business Decisions

NPTEL

Populations and Samples

- Population - all items of interest for a particular decision or investigation
  - *all* married drivers over 25 years old
  - *all* subscribers to Netflix
- Sample - a subset of the population
  - a list of individuals who rented a comedy from Netflix in the past year
- The purpose of sampling is to obtain sufficient information to draw a valid inference about a population.

NPTEL

Measures of Location

Arithmetic Mean

- For a population of size *N:*

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- For a sample of *n* observations*:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Excel function: =AVERAGE(*data range*)
- Property of the mean: $\sum_i (x_i - \bar{x}) = 0$
- Outliers can affect the value of the mean.

NPTEL

Measures of Location

Example 4.1 Computing Mean Cost per Order

(*Purchase Orders* data)

- Using formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean = \$2,471,760/94

= \$26,295.32

	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months)**	**Order Date**	**Arrival Date**
4	Spacetime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	\$ 1.10	17,500	\$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	\$ 3.75	4,250	\$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	\$ 1.10	16,500	\$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	\$195.00	120	\$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	\$ 3.75	4,200	\$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	\$255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	\$255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	\$ 4.05	1,500	\$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	\$ 3.00	1,100	\$ 3,300.00	25	10/05/11	10/10/11

Figure 4.1

NPTEL

Measures of Location

Example 4.1 (continued)

Computing Mean Cost per Order

Applying Formula

=Sum(B2:B95)/Count(B2:B95)

Using Average Function

=Average(B2:B95)

	A	B
1	Observation	Cost per order
2	x1	$2,700.00
3	x2	$19,250.00
4	x3	$15,937.50
5	x4	$18,150.00
6	x5	$23,400.00
92	x91	$16,625.00
93	x92	$74,375.00
94	x93	$72,250.00
95	x94	$6,562.50
96	Sum of cost/order	$2,471,760.00
97	Number of observations	94
98		
99	Mean cost/order	$26,295.32
100		
101	Excel AVERAGE function	$26,295.32

Figure 4.2

NPTEL

Measures of Location

Example 4.2 Finding the Median Cost per Order
(*Purchase Orders* data)

Median - middle value of the data when arranged from least to greatest

Sort the data in column B.

Since $n = 94$,

Median = \$15,656.25

= average of 47th and 48th observations.

=MEDIAN(B2:B94)

=Average(B48,B49)

	A	B	C	D
1	Rank	Cost per order		
2	1	\$68.75		
3	2	\$82.50		
4	3	\$375.00		
5	4	\$467.50		
6	5	\$525.00		
44	43	\$13,650.00		
45	44	\$14,910.00		
46	45	\$14,910.00		
47	46	\$15,087.50		
48	47	\$15,562.50		\$15,562.50
49	48	\$15,750.00		\$15,750.00
50	49	\$15,937.50	Average	\$15,656.25
51	50	\$16,276.75		

Figure 4.3

NPTEL

Measures of Location

Example 4.3 Finding the Mode of A/P terms (*Purchase Orders* data)

- Mode - observation that occurs most often or, for grouped data, the group with the greatest frequency.
- Mode of A/P terms:
  = 30 months
  =MODE.SNGL(H4:H97)

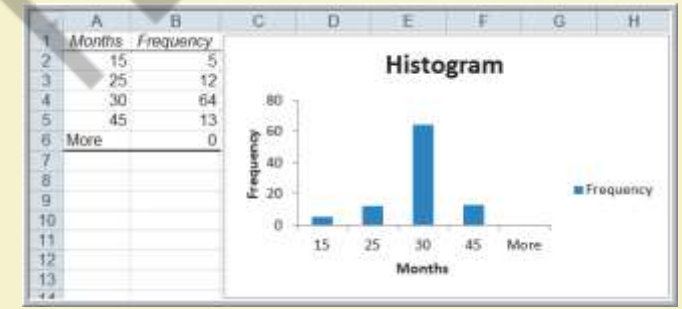


Figure 3.29

- For multiple modes: =MODE.MULT(*data range*)

NPTEL

Measures of Location

Example 4.4 Computing the Midrange

(*Purchase Orders* data)

- Midrange = Average of greatest and least values
- Use the Excel MIN and MAX functions or Sort the data and find them easily.
- Cost per order midrange:

  = ($68.78 + $127,500)/2

  = $63,784.89

  =AVERAGE(MIN(B2:B95), MAX(B2:B95))

NPTEL

Measures of Location

Example 4.5 Quoting Computer Repair Times

Data set includes 250 repair times for customers.

- What repair time would be reasonable to quote to a new customer?
- Median repair time is 2 weeks; Mean and Mode are about 15 days.
- Let's look at a histogram to get a better idea.

	A	B
1	Computer Repair Times	
2		
3	Sample	Repair Time (Days)
4	1	18
5	2	15
6	3	17
7	4	9
250	247	31
251	248	6
252	249	17
253	250	13
254		
255	Mean	14.9
256	Median	14
257	Mode	15

Figure 4.4

NPTEL

Measures of Location

Example 4.5 (continued)

Quoting Computer Repair Times

	A	B	C	D
1	Computer Repair Times			
2				
3			Relative	Cumulative
4	Days	Frequency	Frequency	Percentage
5	0	0	0.000	0.0%
6	1	0	0.000	0.0%
7	2	0	0.000	0.0%
8	3	0	0.000	0.0%
9	4	0	0.000	0.0%
10	5	1	0.004	0.4%
11	6	2	0.008	1.2%
12	7	5	0.020	3.2%
13	8	12	0.048	8.0%
14	9	14	0.056	13.6%
15	10	19	0.076	21.2%
16	11	19	0.076	28.8%
17	12	23	0.092	38.0%
18	13	22	0.088	46.8%
19	14	20	0.080	54.8%
20	15	23	0.092	64.0%
21	16	18	0.072	71.2%
22	17	16	0.064	77.6%
23	18	9	0.036	81.2%
24	19	9	0.036	84.8%
25	20	8	0.032	88.0%
26	21	5	0.020	90.0%



90% are completed within 3 weeks

Figure 4.5

NPTEL

Measures of Dispersion

- Dispersion refers to the degree of variation in the data.
- Range is the difference between the maximum and minimum data values.
- Interquartile Range (IQR) is difference between the third and first quartiles.
- Variance is an average of the squared deviations form the mean (uses all data values).
- Standard Deviation is the square root of the variance.

NPTEL

Measures of Dispersion

Example 4.6 Computing the Range (*Purchase Orders* data)

For the cost per order data:

- Maximum = $127,500
- Minimum = $68.78
- Range = $127,431.22

=MAX(B2:B95)−MIN(B2:B95)

NPTEL

Measures of Dispersion

Example 4.7 Computing the Interquartile Range (*Purchase Orders* data)

For the cost per order data:

- Third Quartile = Q_3 = $27,593.75

  =QUARTILE.INC(B2:B95,3)

- First Quartile = Q_1 = $6,757.81

  =QUARTILE.INC(B2:B95,1)

- Interquartile Range = $20,835.94

The middle 50% of the data is concentrated in a small range of $20,836.
The range of the full data set is affected by extreme values.

NPTEL

Measures of Dispersion

Computing the Variance

- For a population:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- In Excel: =VAR.P(*data range*)

- For a sample:

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

- In Excel: =VAR.S(*data range*)

NPTEL

Measures of Dispersion

Example 4.8 Computing the Variance
(*Purchase Orders* data)

	A	B	C	D
1	Observation	Cost per order	(xi - mean)	(xi - mean)^2
2	x1	$2,700.00	-$23,595.32	$556,739,085.74
3	x2	$19,250.00	-$7,045.32	$49,636,521.91
4	x3	$15,937.50	-$10,357.82	$107,284,417.52
5	x4	$18,150.00	-$8,145.32	$66,346,224.04
6	x5	$23,400.00	-$2,895.32	$8,382,872.97
92	x91	$16,625.00	-$9,670.32	$93,515,072.44
93	x92	$74,375.00	$48,079.68	$2,311,655,710.74
94	x93	$72,250.00	$45,954.68	$2,111,832,692.12
95	x94	$6,562.50	-$19,732.82	$389,384,151.56
96	Sum of cost/order	$2,471,760.00	Sum of squared deviations	$82,825,295,365.68
97	Number of observations	94		
98				
99	Mean cost/order	$26,295.32	Variance	890,594,573.82
100				
101			Excel VAR.S function	890,594,573.82

Figure 4.6

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

= VAR.S(B2:B95)

NPTEL

Measures of Dispersion

Computing the Standard Deviation

- For a population:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

=STDEV.P(*data range*)

- For a sample:

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

=STDEV.S(*data range*)

NPTEL

Measures of Dispersion

Example 4.9 Computing the Standard Deviation (*Purchase Orders* data)

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

$$= \sqrt{\frac{82825295365.68}{93}}$$

$$= \$29{,}842.83$$

=STDEV.S(B2:B95)

	A	B	C	D
1	Observation	Cost per order	(xi - mean)	(xi - mean)^2
2	x1	$2,700.00	-$23,595.32	$556,739,085.74
3	x2	$19,250.00	-$7,045.32	$49,636,521.91
4	x3	$15,937.50	-$10,357.82	$107,284,417.52
5	x4	$18,150.00	-$8,145.32	$66,346,224.04
6	x5	$23,400.00	-$2,895.32	$8,382,872.97
92	x91	$16,625.00	-$9,670.32	$93,515,072.44
93	x92	$74,375.00	$48,079.68	$2,311,655,710.74
94	x93	$72,250.00	$45,954.68	$2,111,832,692.12
95	x94	$6,562.50	-$19,732.82	$389,384,151.56
96	Sum of cost/order	$2,471,760.00	Sum of squared deviations	$82,825,295,365.68
97	Number of observations	94		
98				
99	Mean cost/order	$26,295.32	Variance	890,594,573.82
100				
101			Excel VAR.S function	890,594,573.82

Figure 4.6

NPTEL

Measures of Dispersion

Mean & Standard Deviation of Closing Stock Prices

Intel (INTC):
Mean = $18.81
Stdev. = $0.50

General Electric (GE):
Mean = $16.19
Stdev. = $0.35

INTC is a higher risk investment than GE.

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index
4	9/3/2010	$127.58	$18.43	$21.04	$15.39	10447.93
5	9/7/2010	$125.95	$18.12	$20.58	$15.44	10340.69
6	9/8/2010	$126.08	$17.90	$20.64	$15.70	10387.01
7	9/9/2010	$126.36	$18.00	$20.61	$15.91	10415.24
8	9/10/2010	$127.99	$17.97	$20.62	$15.98	10462.77
9	9/13/2010	$129.61	$18.56	$21.26	$16.25	10544.13
10	9/14/2010	$128.85	$18.74	$21.45	$16.16	10526.49
11	9/15/2010	$129.43	$18.72	$21.59	$16.34	10572.73
12	9/16/2010	$129.67	$18.97	$21.93	$16.23	10594.83
13	9/17/2010	$130.19	$18.81	$21.86	$16.29	10607.85
14	9/20/2010	$131.79	$18.93	$21.75	$16.55	10753.62
15	9/21/2010	$131.98	$19.14	$21.64	$16.52	10761.03
16	9/22/2010	$132.57	$19.01	$21.67	$16.50	10739.31
17	9/23/2010	$131.67	$18.98	$21.53	$16.14	10662.42
18	9/24/2010	$134.11	$19.42	$22.09	$16.66	10860.26
19	9/27/2010	$134.65	$19.24	$22.11	$16.43	10812.04
20	9/28/2010	$134.89	$19.51	$21.86	$16.44	10858.14
21	9/29/2010	$135.48	$19.24	$21.87	$16.36	10835.28
22	9/30/2010	$134.14	$19.20	$21.90	$16.25	10788.05
23	10/1/2010	$135.64	$19.32	$21.91	$16.36	10829.68

Figure 4.7

NPTEL

Measures of Dispersion

Chebyshev's Theorem

For any data set, the proportion of values that lie within k ($k > 1$) standard deviations of the mean is

- *at least* $1 - 1/k^2$

Empirical Rules

For many data sets encountered in practice:

- Approximately 68% of the observations fall within one standard deviation of the mean; that is, between $\bar{x} - s$ and $\bar{x} + s$
- Approximately 95% fall within 2 → $\bar{x} \pm 2s$
- Approximately 99.7% fall within 3 → $\bar{x} \pm 3s$

NPTEL

Measures of Dispersion

Example 4.10 Applying Chebyshev's Theorem
(*Purchase Orders* data)
For the cost-per-order data

- When $k = 2$, $1-1/k^2 = 75\%$
  Mean $\pm$ 2(Stdev.) = [-$33,390.34, $85,980.98]
  89 of the 94 data values (94.68%)
- When $k = 3$, $1-1/k^2 = 89\%$
  Mean $\pm$ 3(Stdev.) = [-$63,233.17, $115,823.81]
  92 of the 94 data values (97.9%)

NPTEL

Measures of Dispersion

Process Capability Index (C_p)

- To measure how well a manufacturing process can achieve specifications, take a sample of output, measure dimensions, compute the total variation using the third empirical rule.
- Compare results to specifications using:

$$Cp = \frac{\text{upper specification} - \text{lower specification}}{\text{total variation}}$$

NPTEL

Measures of Dispersion

Example 4.11 Using the Empirical Rule to Measure the Capability of a Manufacturing Process

$C_p = 0.57$

$C_p < 1$ indicates variation is wider than specified.

Want $C_p \geq 1$ or $C_p \geq 1.5$

	A	B	C	D	E	F	G	H	I	J
1	Manufacturing Measurements									
2										
3	5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11	Mean	4.99
4	5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11	Standard deviation	0.117
5	4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08		
6	5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03	Mean - 3*Stdev	4.640
7	5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94	Mean + 3*Stdev	5.340
8	5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88	Total variation	0.700
9	4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91		
10	5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86	Lower Specification	4.8
11	4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89	Upper Specification	5.2
12	5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91	Specification range	0.4
13	5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87		
14	5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93	Cp	0.57
15	4.85	5.04	5.12	4.97	5.02	4.97	5.02	5.14		
16	4.90	5.09	4.89	5.07	4.99	5.04	5.03	4.87		
17	4.97	5.07	4.91	5.03	5.02	4.94	5.18	4.98		
18	5.09	4.99	4.97	4.81	5.03	4.98	5.08	4.88		
19	4.89	5.01	4.98	4.95	5.02	5.03	5.14	4.88		
20	4.87	4.88	5.01	4.89	5.07	5.05	4.92	5.01		
21	5.01	4.93	5.01	5.08	4.95	4.91	4.97	4.93		
22	4.97	5.10	5.09	4.93	4.95	5.09	4.92	4.93		
23	4.76	4.94	4.93	4.99	4.94	5.21	5.14	4.99		
24	4.94	4.88	5.04	4.94	5.12	4.87	4.92	4.91		
25	4.92	4.89	5.11	5.13	5.08	5.02	5.03	4.96		
26	4.91	4.89	5.07	5.02	4.91	4.81	4.98	4.78		
27	4.96	5.02	5.13	5.13	4.92	4.98	4.89	4.88		

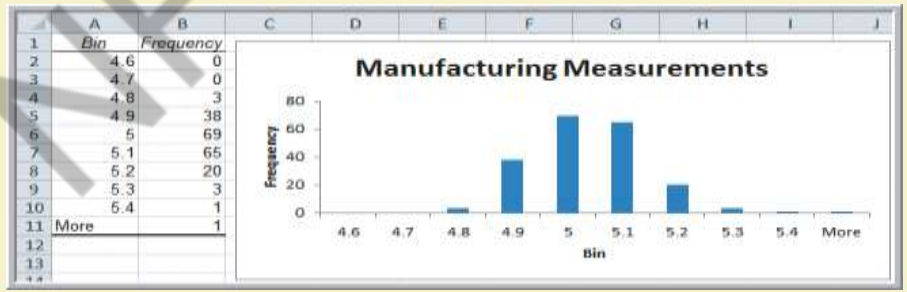
Figure 4.8

NPTEL

Measures of Dispersion

Example 4.11 (continued)

- 3+3+1+1 = 8 of 200 (4%) fall outside the specification limits of between 4.8 and 5.2.
- 3rd Empirical Rule: approximately 0.3% of the data falls outside 3 standard deviations of the mean.
- Chebyshev's Theorem: less than 11% fall outside.



Bin	Frequency
4.6	0
4.7	0
4.8	3
4.9	38
5	69
5.1	65
5.2	20
5.3	3
5.4	1
More	1

Figure 4.9

NPTEL

Measures of Dispersion

Example 4.12 Computing *z*-scores
(*Purchase Orders* data)

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

=STANDARDIZE(*x, mean, standard deviation*)

	A	B	C
1	Observation	Cost per order	z-score
2	x1	$2,700.00	-0.79
3	x2	$19,250.00	-0.24
4	x3	$15,937.50	-0.35
5	x4	$18,150.00	-0.27
6	x5	$23,400.00	-0.10
92	x91	$16,625.00	-0.32
93	x92	$74,375.00	1.61
94	x93	$72,250.00	1.54
95	x94	$6,562.50	-0.66
96			
97	Mean	$26,295.32	
98	Standard Deviation	$29,842.83	

Figure 4.10

NPTEL

Measures of Dispersion

Coefficient of Variation (CV)

- Provides a relative measure of dispersion

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

- Sometimes measured as a percentage.
- Provides a relative measure of risk to return.
- Return to risk = 1/CV
- Sharpe ratio is a related measure in finance.
- Sharpe ratio = excess returns/standard deviation

NPTEL

Measures of Dispersion

Example 4.13 Applying the Coefficient of Variation

- Intel (INTC) is slightly riskier than the other stocks.
- The Index fund has the least risk (lowest CV).

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index
4	9/3/2010	$127.58	$18.43	$21.04	$15.39	10447.93
5	9/7/2010	$125.95	$18.12	$20.58	$15.44	10340.69
6	9/8/2010	$126.08	$17.90	$20.64	$15.70	10387.01
7	9/9/2010	$126.36	$18.00	$20.61	$15.91	10415.24
21	9/29/2010	$135.48	$19.24	$21.87	$16.36	10835.28
22	9/30/2010	$134.14	$19.20	$21.90	$16.25	10788.05
23	10/1/2010	$135.64	$19.32	$21.91	$16.36	10829.68
24	Mean	$130.93	$18.81	$21.50	$16.20	$10,639.98
25	Standard Deviation	$3.22	$0.50	$0.52	$0.35	$171.94
26	Coefficient of Variation	0.025	0.027	0.024	0.022	0.016

Figure 4.11

NPTEL

Measures of Shape

- Skewness describes lack of symmetry.
- Coefficient of Skewness = $CS = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$

  =SKEW(*data, range*)
- CS is negative for left-skewed data.
- CS is positive for right-skewed data.
- |CS| > 1 suggests high degree of skewness.
- $0.5 \leq |CS| \leq 1$ suggests moderate skewness.
- |CS| < 0.5 suggests relative symmetry.

NPTEL

Measures of Shape

Example 4.14 Measuring Skewness

- For cost per order: CS = 1.66 (high)
- For A/P terms: CS = 0.60 (moderate)

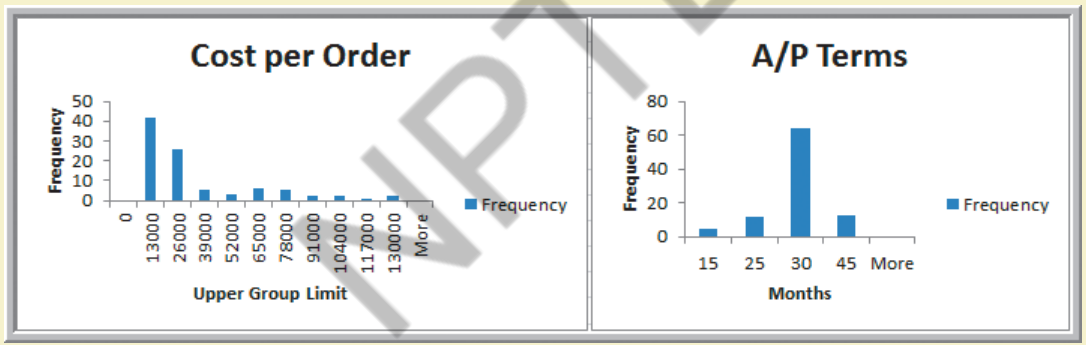


Figure 4.12

NPTEL

Measures of Shape

- Kurtosis refers to peakedness or flatness.
- Coefficient of Kurtosis = $CK = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$

  =KURT(*data, range*)
- CK < 3 indicates the data is somewhat flat with a wide degree of dispersion.
- CK > 3 indicates the data is somewhat peaked with less dispersion.

NPTEL

Measures of Shape

Negatively skewed

Mean < Median < Mode

Positively skewed

Mode < Median < Mean

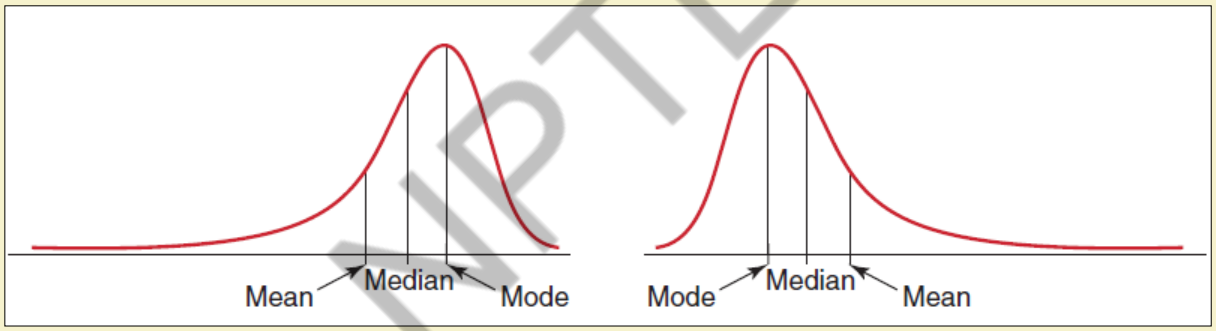


Figure 4.13

NPTEL

Excel Descriptive Statistics Tool

This tool provides a summary of numerical statistical measures for sample data.

Data
Data Analysis
Descriptive Statistics

- Enter *Input Range*
- *Labels* (optional)
- *Summary Statistics*

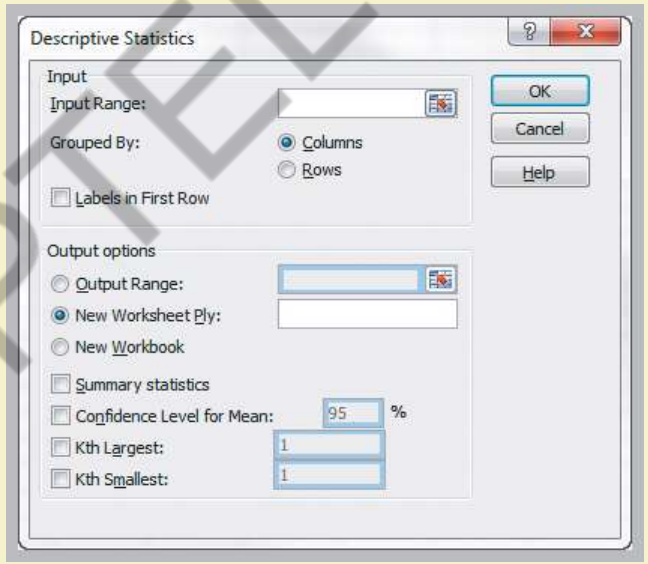


Figure 4.14

NPTEL

Excel Descriptive Statistics Tool

Example 4.15 Using the Descriptive Statistics Tool

Results of the *Analysis Toolpak* do not change when changes are made to the data itself.

Descriptive Statistics for Cost per order and A/P terms

	A	B	C	D
1	*Cost per order*		*A/P Terms (Months)*	
2				
3	Mean	26295.31915	Mean	30.63829787
4	Standard Error	3078.053014	Standard Error	0.702294026
5	Median	15656.25	Median	30
6	Mode	14910	Mode	30
7	Standard Deviation	29842.8312	Standard Deviation	6.808993205
8	Sample Variance	890594573.8	Sample Variance	46.36238847
9	Kurtosis	2.079637302	Kurtosis	1.512188562
10	Skewness	1.664271519	Skewness	0.599265003
11	Range	127431.25	Range	30
12	Minimum	68.75	Minimum	15
13	Maximum	127500	Maximum	45
14	Sum	2471760	Sum	2880
15	Count	94	Count	94

Figure 4.15

NPTEL

Descriptive Statistics for Grouped Data

Computing Statistical Measures from Frequency Distributions

- Mean formulas:

$$\mu = \frac{\sum_{i=1}^{N} f_i x_i}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

- Variance formulas:

$$\sigma^2 = \frac{\sum_{i=1}^{N} f_i (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n - 1}$$

NPTEL

Descriptive Statistics for Grouped Data

Example 4.16 Computing Statistical Measures from Frequency Distributions *(Computer Repair Times)*

	A	B	C	D	E	F
1	Computer Repair Times					
2						
3	Days (x)	Frequency (f)	Frequency*Days	Days - Mean	(Days - mean)^2	Frequency*(Days - Mean)^2
4	0	0	0	-14.912	222.368	0.000
5	1	0	0	-13.912	193.544	0.000
6	2	0	0	-12.912	166.720	0.000
7	3	0	0	-11.912	141.896	0.000
43	39	1	39	24.088	580.232	580.232
44	40	1	40	25.088	629.408	629.408
45	41	0	0	26.088	680.584	0.000
46	42	0	0	27.088	733.760	0.000
47	Sum	250	3728			8840.064
48						
49		Mean	14.912		Variance	35.50226506

Figure 4.16

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

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

NPTEL

Descriptive Statistics for Grouped Data

Example 4.17 Computing Descriptive Statistics for a Grouped Frequency Distribution

We can use group midpoints as approximate percentages of household income spent on rent (except in rows 13, 14).

	A	B	C
1	Gross Rent as a Percentage of Household Income in 1999		
2	Source: US Census Bureau		
3			
4	Group	Number of Households	
5	Less than 10 percent	2,239,346	
6	10 to 14 percent	4,130,917	
7	15 to 19 percent	5,037,981	
8	20 to 24 percent	4,498,604	
9	25 to 29 percent	3,666,233	
10	30 to 34 percent	2,585,327	
11	35 to 39 percent	1,809,948	
12	40 to 49 percent	2,364,443	
13	50 percent or more	6,209,568	
14	Not computed	2,657,135	

Figure 4.17

NPTEL

Descriptive Statistics for Grouped Data

Example 4.17 (continued)

Our calculations indicate that the typical renter spends about 30% of household income on rent.

	A	B	C	D	E	F	G
1							
2	**Group**	**Percent (x)**	**Number (f)**	**f*x**	**x - mean**	**(x - mean)^2**	**f*(x - mean)^2**
3	Less than 10 percent	5%	2,239,346	111967.30	-24.8645%	0.0618	138446.0126
4	10 to 14 percent	12%	4,130,917	495710.04	-17.8645%	0.0319	131834.1452
5	15 to 19 percent	17%	5,037,981	856456.77	-12.8645%	0.0165	83376.1701
6	20 to 24 percent	22%	4,498,604	989692.88	-7.8645%	0.0062	27823.9852
7	25 to 29 percent	27%	3,666,233	989882.91	-2.8645%	0.0008	3008.2636
8	30 to 34 percent	32%	2,585,327	827304.64	2.1355%	0.0005	1179.0089
9	35 to 39 percent	37%	1,809,948	669680.76	7.1355%	0.0051	9215.4310
10	40 to 49 percent	44.50%	2,364,443	1052177.14	14.6355%	0.0214	50645.9048
11	50 percent or more	60%	6,209,568	3725740.80	30.1355%	0.0908	563921.1249
12		**Sum**	32,542,367	9718613.24			1009450.0462
13							
14			**Mean**	29.86%		**Variance**	0.031019565
15						**Standard Dev.**	17.61%

Figure 4.18

NPTEL

Descriptive Statistics for Categorical Data: The Proportion

Example 4.18 Computing a Proportion

- Proportion of orders placed by Spacetime Technologies

  =COUNTIF(A4:A97, “Spacetime Technologies”)/94

  = 12/94 = 0.128

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months	Order Date	Arrival Date
4	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc	A0115	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc	A0204	5319	Shielded Cable/ft.	$ 1.10	16,500	$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc	A0205	5677	Side Panel	$195.00	120	$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	$255.00	500	$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	$255.00	406	$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	$ 4.05	1,500	$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	$ 3.00	1,100	$ 3,300.00	25	10/05/11	10/10/11

Figure 4.1

NPTEL

Statistics in PivotTables

Statistical Measure Choices in PivotTables

Under *Value Field Settings:*

- Average
- Max and Min
- Product
- Standard deviation
- Variance

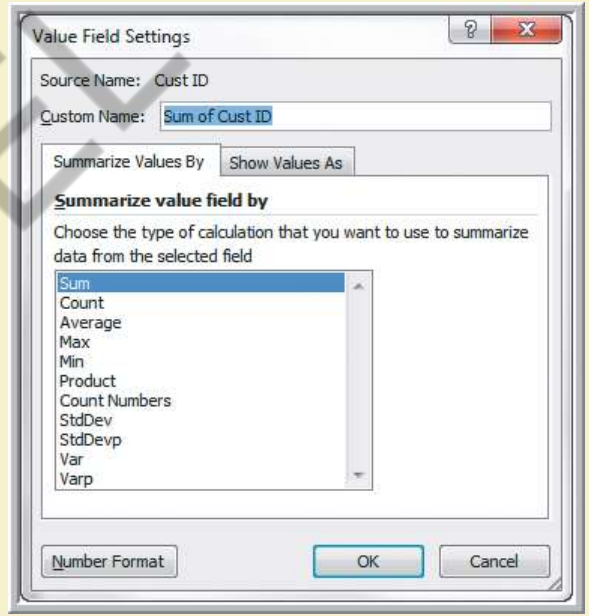


Figure 4.19

NPTEL

Statistics in PivotTables

Example 4.19 Statistical Measures in PivotTables

(Credit Risk Data)

Fields: Checking
Savings
Job Classif.
Row Labels: Job
Σ Values:
Average Checking
Average Savings

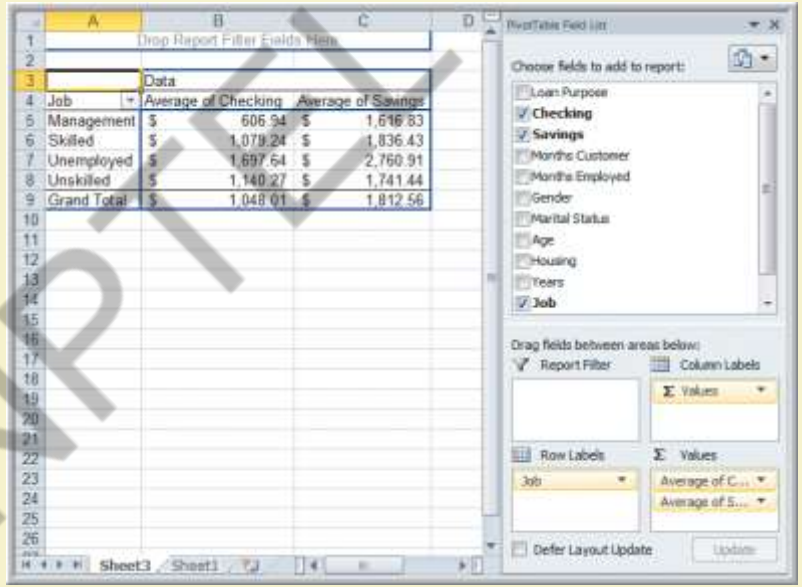


Figure 4.20

NPTEL

Measures of Association

- Data from 49 top liberal arts and research universities can be used to answer questions:
- Is *Top 10% HS* related to *Graduation %*?
- Is *Accept. Rate* related to *Expenditures/Student?*
- Is *Median SAT* related to *Acceptance Rate?*

	A	B	C	D	E	F	G
1	**Colleges and Universities**						
2							
3	**School**	**Type**	**Median SAT**	**Acceptance Rate**	**Expenditures/Student**	**Top 10% HS**	**Graduation %**
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	$ 18,847	70	84

Figure 4.21

NPTEL

Measures of Association

- Covariance is a measure of the linear association between two variables, *X* and *Y*.
- For a population:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

=COVARIANCE.P(*array1, array2*)

- For a sample:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

=COVARIANCE.S(*array1, array2*)

NPTEL

Measures of Association

Example 4.20 Computing the Covariance

- Scatterplot of the *Colleges and Universities* data

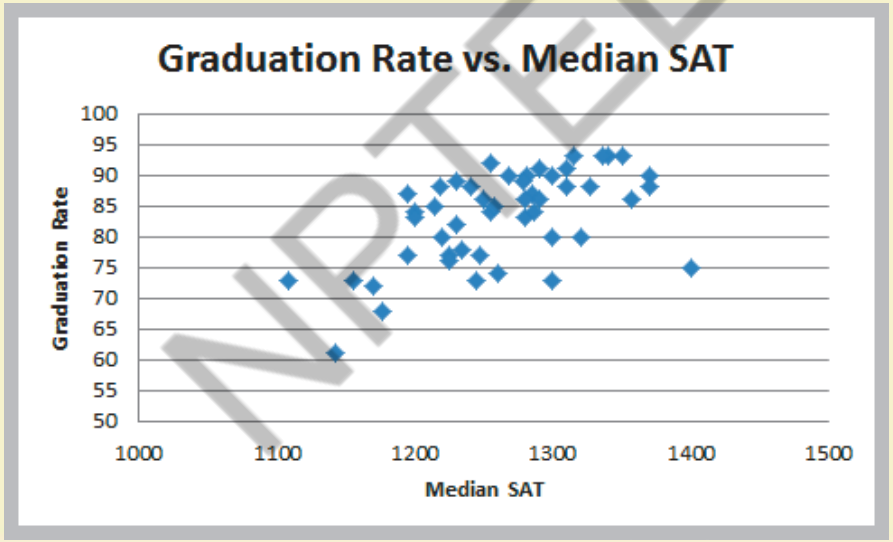


Figure 4.22

NPTEL

Measures of Association

Example 4.20 (continued)

Computing the Covariance

	A	B	C	D	E	F
1		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y-Mean(Y))
2		93	1315	9.755	51.898	506.2698875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
46		78	1234	-5.245	-29.102	152.6372345
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102		Sum	12641.77551
52					Count	49
53					Covariance	263.3703231
54						
55					COVARIANCE.S	263.3703231

Figure 4.23

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

NPTEL

Measures of Association

- Correlation is a measure of the linear association between two variables, *X* and *Y*.
- Correlation Coefficient formulas:

For a population: $\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$

For a sample: $r_{xy} = \frac{\text{cov}(X, Y)}{s_x s_y}$ CORREL=(*array1, array2*)

- The Correlation Coefficient does not depend upon units of measurement (unlike covariance).
- Also known as the:
  Pearson product moment correlation

NPTEL

Measures of Association

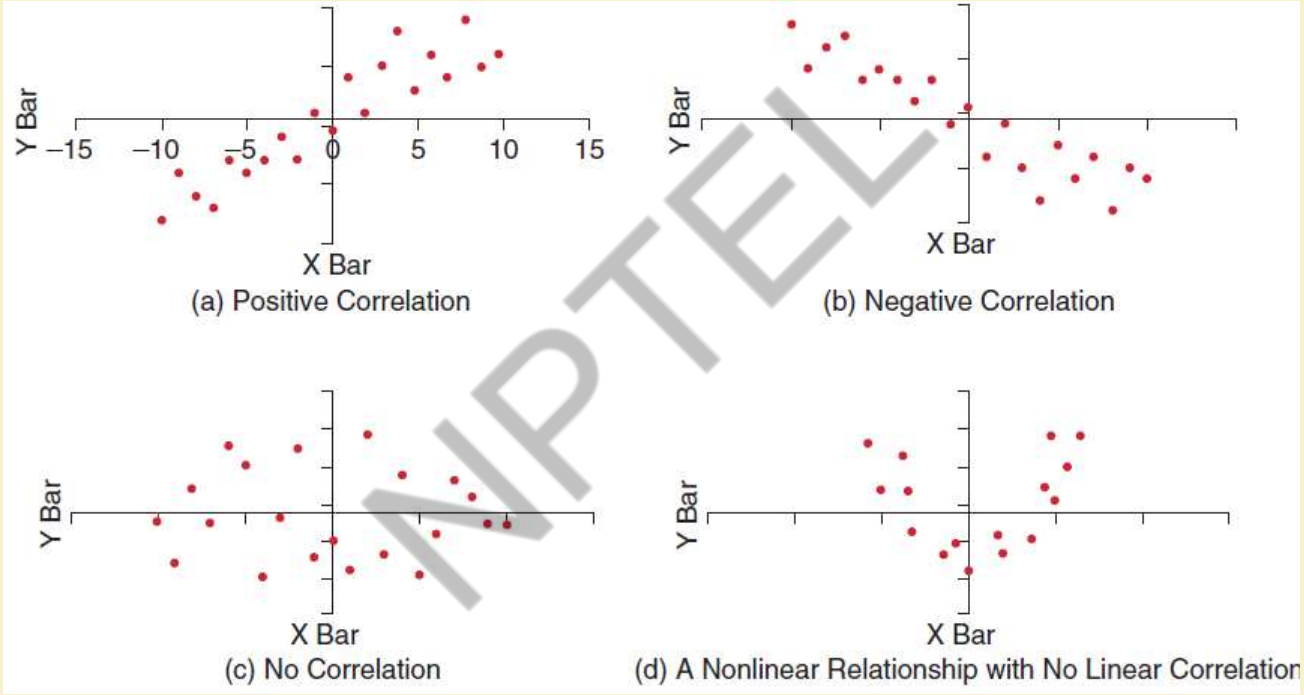


Figure 4.24

NPTEL

Measures of Association

Example 4.21 Computing the Correlation Coefficient *(Colleges and Universities* data)

- *Graduation %* and *Median SAT*

	A	B	C	D	E	F
1		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y-Mean(Y))
2		93	1315	9.755	51.898	506.2698875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102		Sum	12641.77551
52	Standard Deviation	7.449	62.676		Count	49
53					Covariance	263.3703231
54					Correlation	0.564146827
55						
56					CORREL Function	0.564146827

Figure 4.25

$$r_{xy} = \frac{\text{cov}(X,Y)}{s_x s_y}$$

NPTEL

Measures of Association

Excel *Correlation* Tool

- *Data*
- *Data Analysis*
- *Correlation*

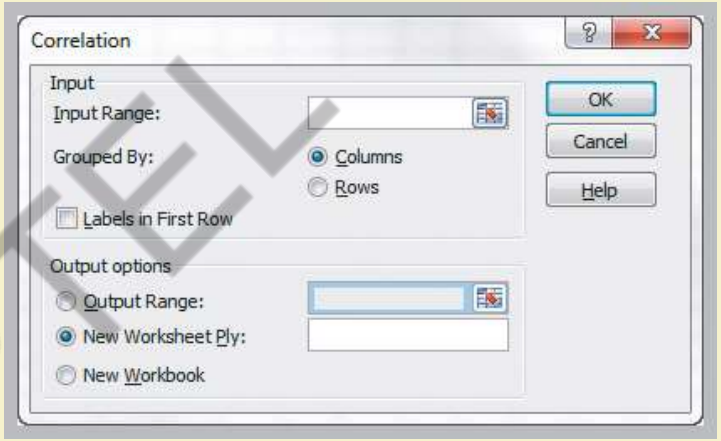


Figure 4.26

Excel computes the correlation coefficient between all pairs of variables in the *Input Range*.

Input Range Data must be in contiguous columns.

NPTEL

Measures of Association

Example 4.22 Using the *Correlation* Tool
(Colleges and Universities data)

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

Figure 4.27

- Lower acceptance rate, higher median SAT
- Lower acceptance rate, higher % top 10 HS students
- Lower acceptance rate, higher graduation rate
- Higher median SAT, higher graduation rate

NPTEL

Outliers

The Mean and Range are sensitive to outliers.

How do we identify outliers?

Some possible methods to identify outliers are:

- *z*-scores greater than +3 or less than -3
- extreme outliers are more than 3*IQR to the left of Q_1 or right of Q_3
- mild outliers are between 1.5*IQR and 3*IQR to the left of Q_1 or right of Q_3

There is no standard definition of what constitutes an outlier.

NPTEL

Outliers

Example 4.23 Investigating Outliers (*Home Market Value* data)

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00

Figure 4.28

Note that the complete data set has 43 observations.

- Are any homes outliers?

NPTEL

Outliers

Example 4.23 (continued) Investigating Outliers

None of the *z*-scores for Square Feet or Market Value exceed 3.

	A	B	C	D	E
1	Home Market Value				
2					
3	House Age	Square Feet	z-score	Market Value	z-score
4	33	1,812	0.5300	$90,000.00	-0.196
5	32	1,914	0.9931	$104,400.00	1.168
6	32	1,842	0.6662	$93,300.00	0.117
7	33	1,812	0.5300	$91,000.00	-0.101
40	28	1,784	0.4029	$91,300.00	-0.073
41	27	1,484	-0.9592	$81,300.00	-1.020
42	27	1,520	-0.7957	$100,700.00	0.818
43	28	1,520	-0.7957	$87,200.00	-0.461
44	27	1,684	-0.0511	$96,700.00	0.439
45	27	1,581	-0.5188	$120,700.00	2.713
46	Mean	1,695		92,069	
47	Standard Deviation	220.257		10553.083	

Figure 4.29

NPTEL

Outliers

Example 4.23 (continued) Investigating Outliers

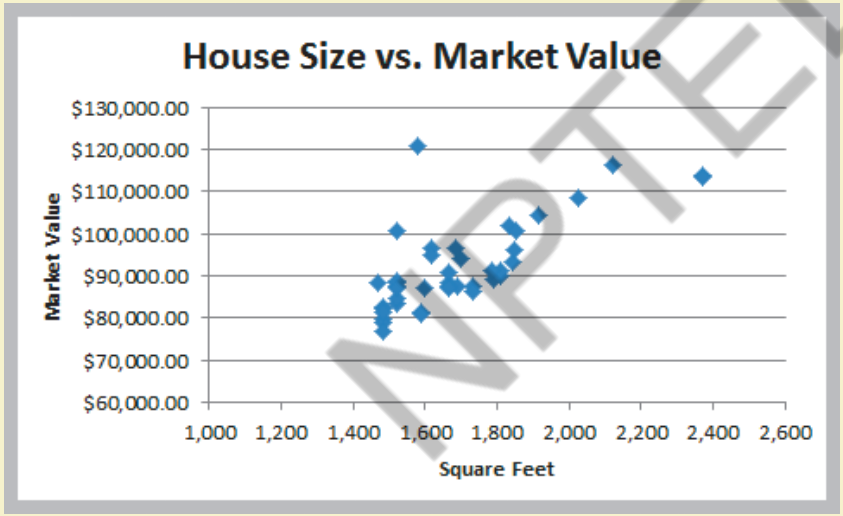


Figure 4.30

The house with a market value near $120,000 and square footage near 1600 does not fall in line with the rest of the data.

NPTEL

Statistical Thinking in Business Decisions

Statistical Thinking is a philosophy of learning and action for improvement, based on principles that:

- all work occurs in a system of interconnected processes
- variation exists in all processes
- better performance results from understanding and reducing variation

Business Analytics provide managers with insights into facts and relationships that enables them to make better decisions.

NPTEL

Statistical Thinking in Business Decisions

Example 4.24 Applying Statistical Thinking

- Average infection rate = 0.0072
- Standard deviation = 0.0053

	A	B	C	D
1	Surgery Infections			
2				
3	Month	Surgeries	Infections	Infection Rate
4	1	208	1	0.0048
5	2	225	3	0.0133
6	3	201	3	0.0149
7	4	236	1	0.0042
8	5	220	3	0.0136
9	6	244	1	0.0041
10	7	247	1	0.0040
11	8	245	1	0.0041
12	9	250	1	0.0040
13	10	227	0	0.0000
14	11	234	2	0.0085
15	12	227	6	0.0264
16	13	213	2	0.0094
17	14	212	1	0.0047
18	15	193	2	0.0104
19	16	182	0	0.0000
35	32	241	1	0.0041
36	33	220	3	0.0136
37	34	278	1	0.0036
38	35	255	3	0.0118
39	36	225	1	0.0044

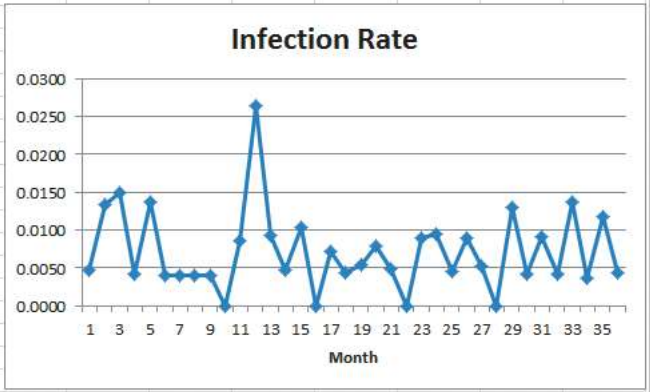


Figure 4.31

NPTEL

Statistical Thinking in Business Decisions

Example 4.24 (continued) Applying Statistical Thinking *(Surgery Infections* data)

- Control limits set at *z*-scores of -3 and +3
- Control limits: -0.009 (set to 0) and 0.0023

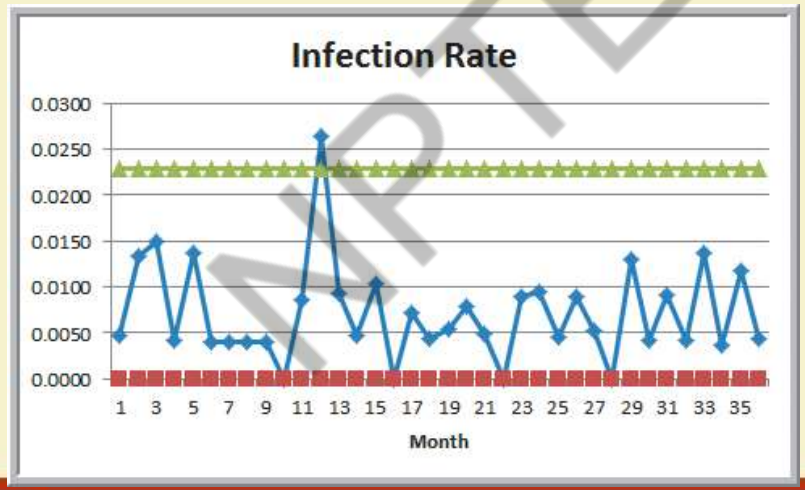


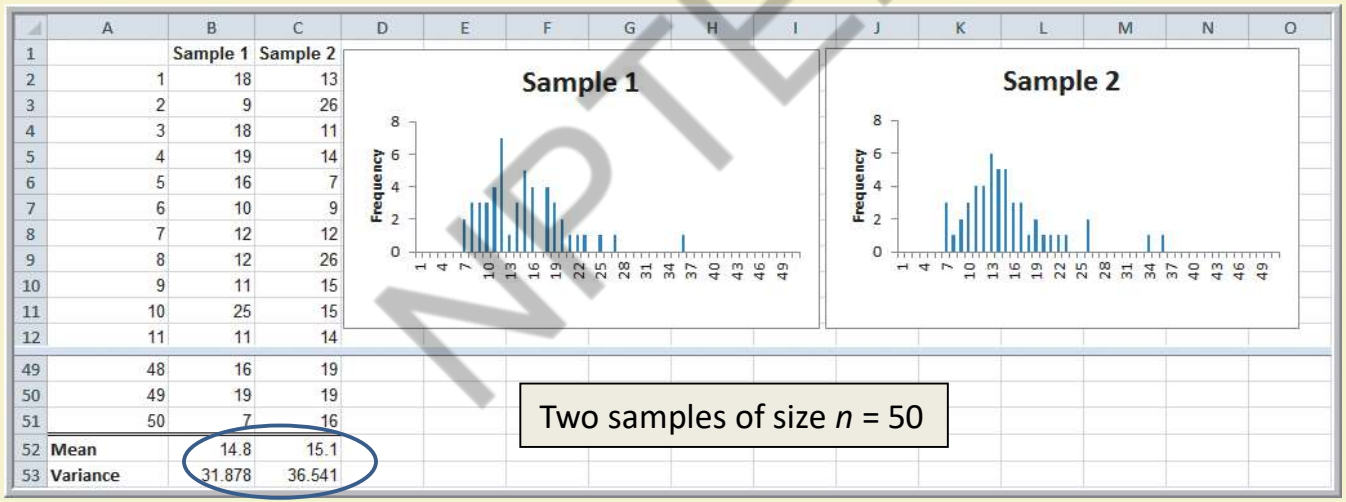
Figure 4.32

NPTEL

Statistical Thinking in Business Decisions

Example 4.25 Variation in Sample Data

- Population: 250 computer repair times
- $\mu = 14.91$ days, $\sigma^2 = 35.5$ days^2



	A	B	C
1		Sample 1	Sample 2
2	1	18	13
3	2	9	26
4	3	18	11
5	4	19	14
6	5	16	7
7	6	10	9
8	7	12	12
9	8	12	26
10	9	11	15
11	10	25	15
12	11	11	14
49	48	16	19
50	49	19	19
51	50	7	16
52	Mean	14.8	15.1
53	Variance	31.878	36.541

Two samples of size $n = 50$

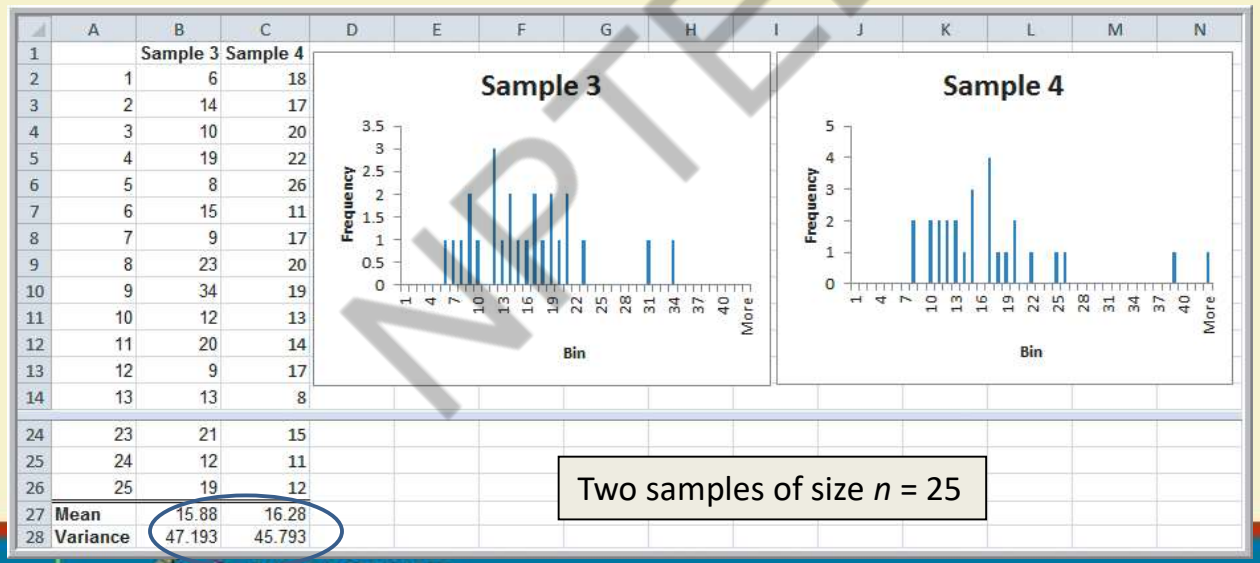
Figure 4.33

NPTEL

Statistical Thinking in Business Decisions

Example 4.25 (continued) Variation in Sample Data

- The two $n = 25$ samples have higher variation than the population and the $n = 50$ samples.



	A	B	C
1		Sample 3	Sample 4
2	1	6	18
3	2	14	17
4	3	10	20
5	4	19	22
6	5	8	26
7	6	15	11
8	7	9	17
9	8	23	20
10	9	34	19
11	10	12	13
12	11	20	14
13	12	9	17
14	13	13	8
24	23	21	15
25	24	12	11
26	25	19	12
27	Mean	15.88	16.28
28	Variance	47.193	45.793

Two samples of size $n = 25$

Figure 4.34

NPTEL

Statistical Thinking in Business Decisions



Analytics in Practice:

Applying Statistical Thinking to Detecting Financial Problems

Sarbanes-Oxley Act (2002)

- helped improve the quality of data that companies disclose to the public but companies can still commit financial fraud.

Anomaly detection scores (a form of *z*-score) are often used by the SEC to detect companies committing financial fraud.

NPTEL

Chapter 4 - Key Terms

- Arithmetic mean (mean)
- Bimodal
- Chebyshev's theorem
- Coefficient of kurtosis
- Coefficient of skewness
- Coefficient of variation
- Correlation
- Correlation coefficient (Pearson product moment correlation coefficient)
- Covariance
- Dispersion
- Empirical rules
- Interquartile range (midspread)
- Kurtosis
- Median
- Midrange
- Mode
- Outlier

NPTEL

Chapter 4 - Key Terms (continued)

- Population
- Process capability index
- Proportion
- Range
- Return to risk
- Sample
- Sample correlation coefficient
- Skewness
- Standard deviation
- Standardized value (*z*-score)
- Statistical thinking
- Unimodal
- Variance

NPTEL

Case Study

Performance Lawn Equipment (





- Recall that PLE produces lawnmowers and a medium size diesel power lawn tractor.
- Compute the mean satisfaction ratings by year and region and provide descriptive statistics for the 2012 customer survey data.
- Determine how responses times change quarterly.
- Examine changes in defects over time.
- Compare sales with industry totals and write a formal report summarizing your results.

NPTEL