6

Descriptive Statistics

CHAPTER OUTLINE

Learning Objective for Chapter 6

After careful study of this chapter, you should be able to do the following:

1. Compute and interpret the sample mean, sample variance, sample standard deviation, sample median, and sample range.
2. Explain the concepts of sample mean, sample variance, population mean, and population variance.
3. Construct and interpret visual data displays, including the stem-and-leaf display, the histogram, and the box plot.
4. Explain the concept of random sampling.
5. Construct and interpret normal probability plots.
6. Explain how to use box plots, and other data displays, to visually compare two or more samples of data.
7. Know how to use simple time series plots to visually display the important features of time-oriented data.

Numerical Summaries of Data

- Data are the numeric observations of a phenomenon of interest. The totality of all observations is a population. A portion used for analysis is a random sample.
- We gain an understanding of this collection, possibly massive, by describing it numerically and graphically, usually with the sample data.
- We describe the collection in terms of shape, outliers, center, and spread (SOCS).
- The center is measured by the mean.
- The spread is measured by the variance.

Populations & Samples

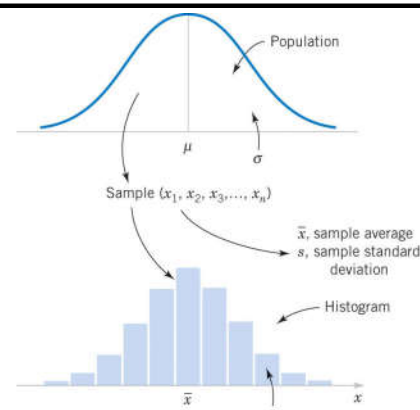


Figure 6-3 (out of order) A population is described, in part, by its parameters, i.e., mean (μ) and standard deviation (σ). A random sample of size *n* is drawn from a population and is described, in part, by its statistics, i.e., mean (*x-bar*) and standard deviation (*s*). The statistics are used to estimate the parameters.

Mean

If the n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6\text{-}1)$$

For the N observations in a population denoted by $x_1, x_2, \ldots, x_N$, the population mean is analogous to a probability distribution as

$$\mu = \sum_{i=1}^{N} x_i \cdot f(x) = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (6\text{-}2)$$

Exercise 6-1: Sample Mean

Consider 8 observations (x_i) of pull-off force from engine connectors from Chapter 1 as shown in the table.

$$\bar{x} = \text{average} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \ldots + 13.1}{8}$$

$$= \frac{104}{8} = 13.0 \text{ pounds}$$

i	x_i
1	12.6
2	12.9
3	13.4
4	12.2
5	13.6
6	13.5
7	12.6
8	13.1
	12.99
= AVERAGE($B2:$B9)	

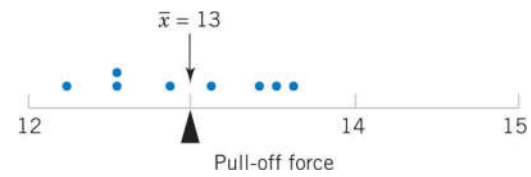


Figure 6-1 The sample mean is the balance point.

Variance Defined

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} \qquad (6\text{-}3)$$

For the N observations in a population denoted by $x_1, x_2, \ldots, x_N$, the population variance, analogous to the variance of a probability distribution, is

$$\sigma^2 = \sum_{i=1}^{N} \left(x_i - \mu\right)^2 \cdot f(x) = \frac{\sum_{i=1}^{N} \left(x_i - \mu\right)^2}{N} \qquad (6\text{-}5)$$

Standard Deviation Defined

- The standard deviation is the square root of the variance.
- σ is the population standard deviation symbol.
- *s* is the sample standard deviation symbol.
- The units of the standard deviation are the same as:
  - The data.
  - The mean.

Rationale for the Variance

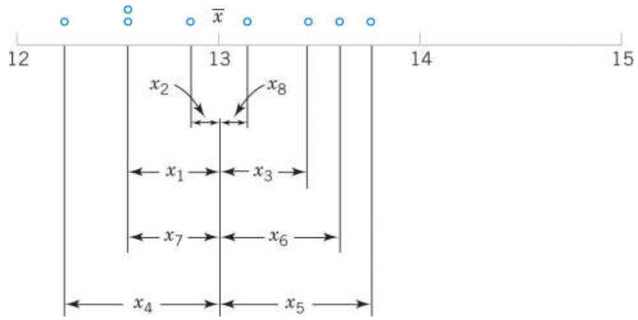


Figure 6-2 The x_i values above are the deviations from the mean. Since the mean is the balance point, the sum of the left deviations (negative) equals the sum of the right deviations (positive). If the deviations are squared, they become a measure of the data spread. The variance is the average data spread.

Example 6-2: Sample Variance

Table 6-1 displays the quantities needed to calculate the summed squared deviations, the numerator of the variance.

Dimension of:
x_i is pounds
Mean is pounds.
Variance is pounds^2.
Standard deviation is pounds.

Desired accuracy is generally accepted to be one more place than the data.

i	x_i	x_i - xbar	$(x_i - xbar)^2$
1	12.6	-0.40	0.1600
2	12.9	-0.10	0.0100
3	13.4	0.40	0.1600
4	12.3	-0.70	0.4900
5	13.6	0.60	0.3600
6	13.5	0.50	0.2500
7	12.6	-0.40	0.1600
8	13.1	0.10	0.0100
sums =	104.00	0.00	1.6000
	divide by 8		divide by 7
mean =	13.00	variance =	0.2286
	standard deviation =		0.48

Computation of s^2

The prior calculation is definitional and tedious. A shortcut is derived here and involves just 2 sums.

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}\sum_{i=1}^{n} x_i}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}\cdot n\overline{x}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1} \qquad (6\text{-}4)$$

Example 6-3: Variance by Shortcut

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1}$$

$$= \frac{1{,}353.60 - (104.0)^2 / 8}{7}$$

$$= \frac{1.60}{7} = 0.2286 \text{ pounds}^2$$

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

i	x_i	x_i^2
1	12.6	158.76
2	12.9	166.41
3	13.4	179.56
4	12.3	151.29
5	13.6	184.96
6	13.5	182.25
7	12.6	158.76
8	13.1	171.61
sums =	104.0	1,353.60

What is this “n–1”?

- The population variance is calculated with *N*, the population size. Why isn’t the sample variance calculated with *n*, the sample size?
- The true variance is based on data deviations from the true mean, μ.
- The sample calculation is based on the data deviations from *x-bar*, not μ. *X-bar* is an estimator of μ; close but not the same. So the n-1 divisor is used to compensate for the error in the mean estimation.

Degrees of Freedom

- The sample variance is calculated with the quantity *n*-1.
- This quantity is called the “degrees of freedom”.
- Origin of the term:
  - There are *n* deviations from *x-bar* in the sample.
  - The sum of the deviations is zero. (Balance point)
  - *n*-1 of the observations can be freely determined, but the n^{th} observation is fixed to maintain the zero sum.

Sample Range

If the *n* observations in a sample are denoted by x_1, x_2, …, x_n, the sample range is:

$$r = max(x_i) - min(x_i)$$

It is the largest observation in the sample less the smallest observation.

From Example 6-3:

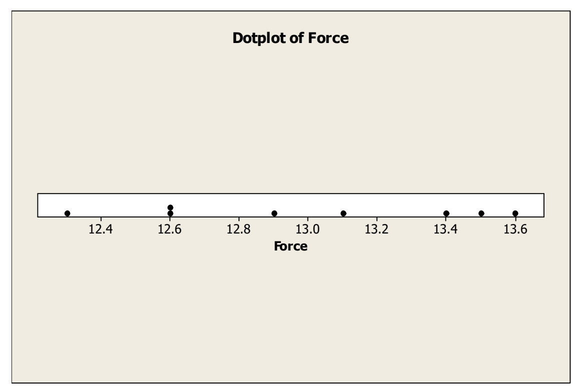
$$r = 13.6 - 12.3 = 1.30$$

Note that: population range ≥ sample range

Intro to Stem & Leaf Diagrams

First, let’s discuss dot diagrams – dots representing data on the number line.

Minitab produces this graphic using the Example 6-1 data.



Stem-and-Leaf Diagrams

- Dot diagrams (dotplots) are useful for small data sets. Stem & leaf diagrams are better for large sets.
- Steps to construct a stem-and-leaf diagram:
  1) Divide each number (x_i) into two parts: a stem, consisting of the leading digits, and a leaf, consisting of the remaining digit.
  2) List the stem values in a vertical column (no skips).
  3) Record the leaf for each observation beside its stem.
  4) Write the units for the stems and leaves on the display.

Example 6-4: Alloy Strength

Table 6-2 Compressive Strength (psi) of Aluminum-Lithium Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Figure 6-4 Stem-and-leaf diagram for Table 6-2 data. Center is about 155 and most data is between 110 and 200. Leaves are unordered.

Split Stems

- The purpose of the stem-and-leaf is to describe the data distribution graphically.
- If the data are too clustered, we can split and have multiple stems, thereby increasing the number of stems.
  - Split 2 for 1:
    - Lower stem for leaves 0, 1, 2, 3, 4
    - Upper stem for leaves 5, 6, 7, 8, 9
  - Split 5 for 1:
    - 1st stem for leaves 0, 1
    - 2nd stem for leaves 2, 3
    - 3rd stem for leaves 4, 5
    - 4th stem for leaves 6, 7
    - 5th stem for leaves 8, 9

Example 6-5: Chemical Yield Displays

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(a)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

(b)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

(c)

Figure 6-5 (a) Stems not split; too compact
(b) Stems split 2-for-1; nice shape
(c) Stems split 5-for-1; too spread out

Stem-and-Leaf by Minitab

- Table 6-2 data: Leaves are ordered, hence the data is sorted.
- Median is the middle of the sorted observations.
  - If n is odd, the middle value.
  - If n is even, the average or midpoint of the two middle values. Median is 161.5.
- Mode is 158, the most frequent value.

Figure 6-6

Stem-and-leaf of Strength

Count	Stem	Leaves
1	7	6
2	8	7
3	9	7
5	10	15
8	11	058
11	12	013
17	13	133455
25	14	12356899
37	15	001344678888
(10)	16	0003357789
33	17	0112445668
23	18	0011346
16	19	034699
10	20	0178
6	21	8
5	22	189
2	23	7
1	24	5

Quartiles

- The three quartiles partition the data into four equally sized counts or segments.
  - 25% of the data is less than q_1.
  - 50% of the data is less than q_2, the median.
  - 75% of the data is less than q_3.
- Calculated as *Index* = $f(n+1)$ where:
  - *Index* (I) is the I^{th} item (interpolated) of the sorted data list.
  - f is the fraction associated with the quartile.
  - n is the sample size.
- For the Table 6-2 data:

		Value of indexed item		
f	*Index*	I^{th}	$(I+1)^{th}$	quartile
0.25	20.25	143	144	143.25
0.50	40.50	160	163	161.50
0.75	60.75	181	181	181.00

Percentiles

- Percentiles are a special case of the quartiles.
- Percentiles partition the data into 100 segments.
- The *Index* = $f(n+1)$ methodology is the same.
- The 37%ile is calculated as follows:
  - Refer to the Table 6-2 stem-and-leaf diagram.
  - *Index* = 0.37(81) = 29.97
  - 37%ile = 153 + 0.97(154 – 153) = 153.97

Interquartile Range

- The interquartile range (IQR) is defined as:

  $$IQR = q_1 - q_3.$$

- From Table 6-2:

  IQR = 181.00 – 143.25 = 37.75 = 37.8

- Impact of outlier data:
  - IQR is not affected
  - Range is directly affected.

Minitab Descriptives

- The Minitab selection menu:

  Stat > Basic Statistics > Display Descriptive Statistics

calculates the descriptive statistics for a data set.

- For the Table 6-2 data, Minitab produces:

Variable	N	Mean	StDev		
Strength	80	162.66	33.77		
	Min	Q1	Median	Q3	Max
	76.00	143.50	161.50	181.00	245.00
	5-number summary				

Frequency Distributions

- A frequency distribution is a compact summary of data, expressed as a table, graph, or function.
- The data is gathered into bins or cells, defined by class intervals.
- The number of classes, multiplied by the class interval, should exceed the range of the data. The square root of the sample size is a guide.
- The boundaries of the class intervals should be convenient values, as should the class width.

Frequency Distribution Table

Considerations:
Range = 245 – 76 = 169

Sqrt(80) = 8.9

Trial class width = 18.9

Decisions:
Number of classes = 9

Class width = 20

Range of classes = 20 * 9 = 180

Starting point = 70

Table 6-4 Frequency Distribution of Table 6-2 Data

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
$70 \le x < 90$	2	0.0250	0.0250
$90 \le x < 110$	3	0.0375	0.0625
$110 \le x < 130$	6	0.0750	0.1375
$130 \le x < 150$	14	0.1750	0.3125
$150 \le x < 170$	22	0.2750	0.5875
$170 \le x < 190$	17	0.2125	0.8000
$190 \le x < 210$	10	0.1250	0.9250
$210 \le x < 230$	4	0.0500	0.9750
$230 \le x < 250$	2	0.0250	1.0000
	80	1.0000	

Histograms

- A histogram is a visual display of a frequency distribution, similar to a bar chart or a stem-and-leaf diagram.
- Steps to build one with equal bin widths:
  1) Label the bin boundaries on the horizontal scale.
  2) Mark & label the vertical scale with the frequencies or relative frequencies.
  3) Above each bin, draw a rectangle whose height is equal to the frequency or relative frequency.

Histogram of the Table 6-2 Data

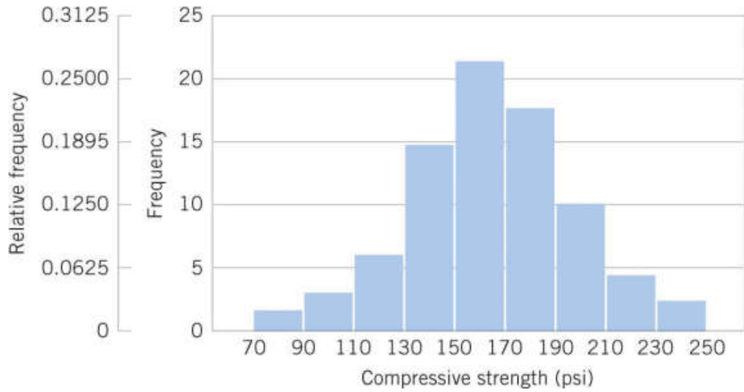


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

Histograms with Unequal Bin Widths

- If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow class widths in the clustered region and wide class widths in the scattered areas.
- In this approach, the rectangle area, not the height, must be proportional to the class frequency.

$$\text{Rectangle height} = \frac{\text{bin frequency}}{\text{bin width}}$$

Poor Choices in Drawing Histograms-1

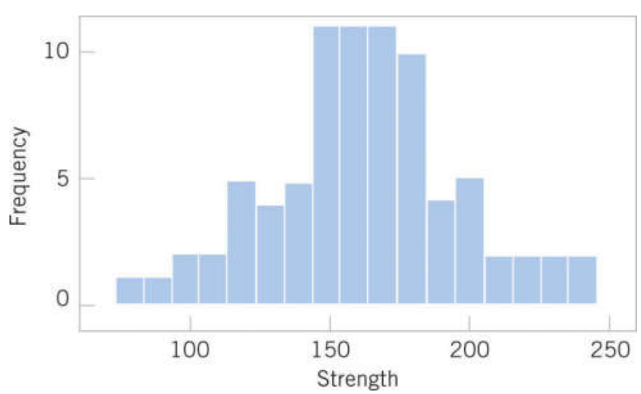


Figure 6-8 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Errors: too many bins (17) create jagged shape, horizontal scale not at class boundaries, horizontal axis label does not include units.

Poor Choices in Drawing Histograms-2

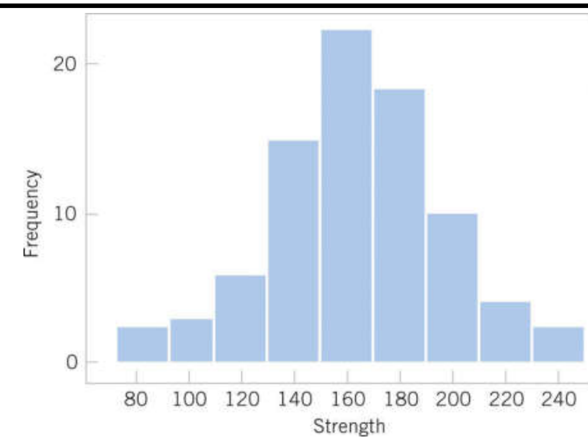


Figure 6-9 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Errors: horizontal scale not at class boundaries (cutpoints), horizontal axis label does not include units.

Cumulative Frequency Plot

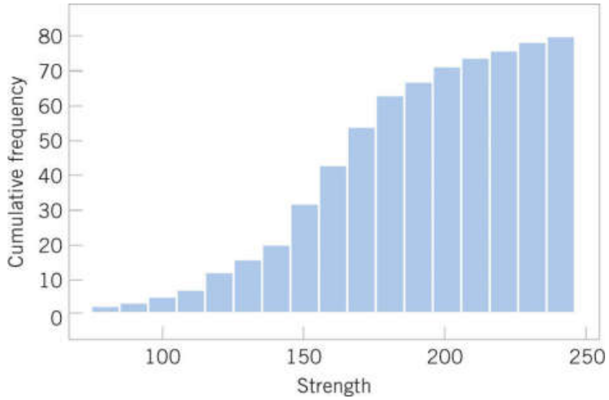


Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

Shape of a Frequency Distribution

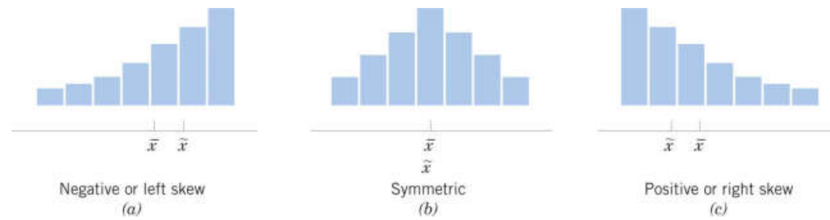


Figure 6-11 Histograms of symmetric and skewed distributions.
(b) Symmetric distribution has identical mean, median and mode measures.
(a & c) Skewed distributions are positive or negative, depending on the direction of the long tail. Their measures occur in alphabetical order as the distribution is approached from the long tail.☺

Histograms for Categorical Data

- Categorical data is of two types:
  - Ordinal: categories have a natural order, e.g., year in college, military rank.
  - Nominal: Categories are simply different, e.g., gender, colors.
- Histogram bars are for each category, are of equal width, and have a height equal to the category's frequency or relative frequency.
- A Pareto chart is a histogram in which the categories are sequenced in decreasing order. This approach emphasizes the most and least important categories.

Example 6-6: Categorical Data Histogram

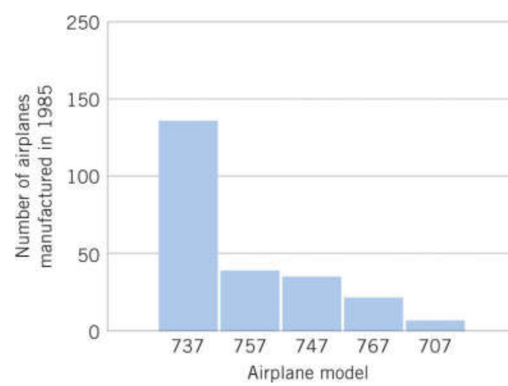


Figure 6-12 Airplane production in 1985. (Source: Boeing Company) Comment: Illustrates nominal data in spite of the numerical names, categories are shown at the bin's midpoint, a Pareto chart since the categories are in decreasing order.

Box Plot or Box-and-Whisker Chart

- A box plot is a graphical display showing center, spread, shape, and outliers (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

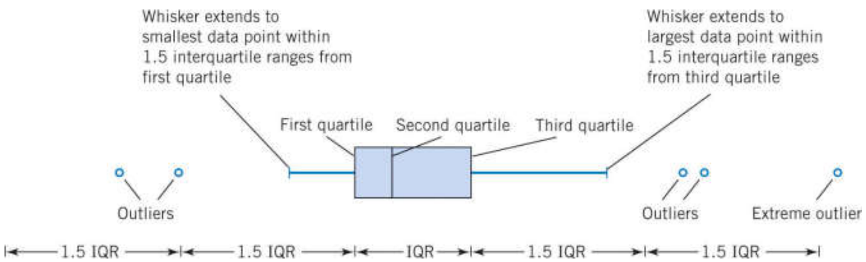


Figure 6-13 Description of a box plot.

Box Plot of Table 6-2 Data

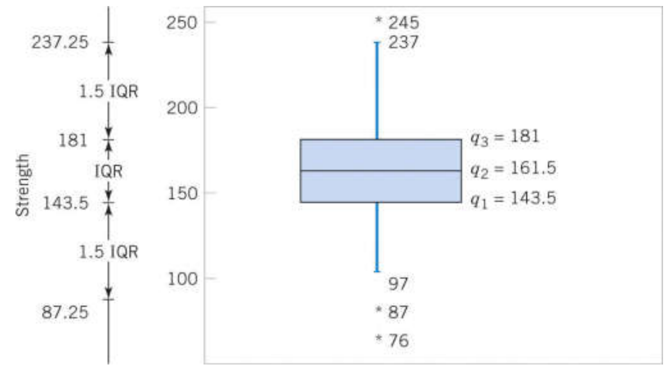


Figure 6-14 Box plot of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Box plot may be shown vertically or horizontally, data reveals three outliers and no extreme outliers. Lower outlier limit is: 143.5 – 1.5*(181.0-143.5) = 87.25.

Comparative Box Plots

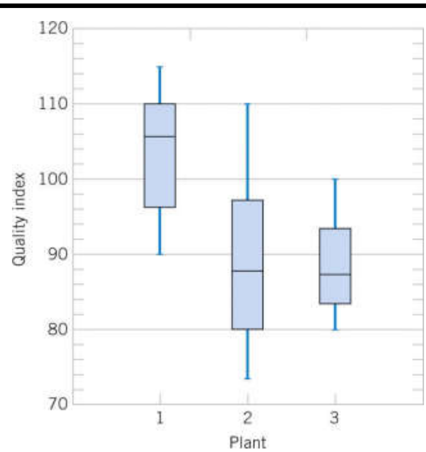


Figure 6-15 Comparative box plots of a quality index at three manufacturing plants. Comment: Plant 2 has too much variability. Plants 2 & 3 need to raise their quality index performance.

Time Sequence Plots

- A time series plot shows the data value, or statistic, on the vertical axis with time on the horizontal axis.
- A time series plot reveals trends, cycles or other time-oriented behavior that could not be otherwise seen in the data.

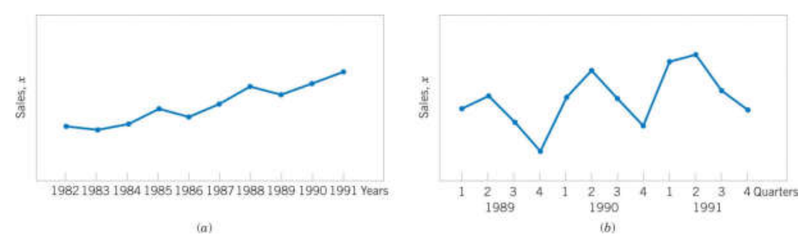


Figure 6-16 Company sales by year (a) & by quarter (b). The annual time interval masks cyclical quarterly variation, but shows consistent progress.

Digidot Plot of Table 6-2 Data

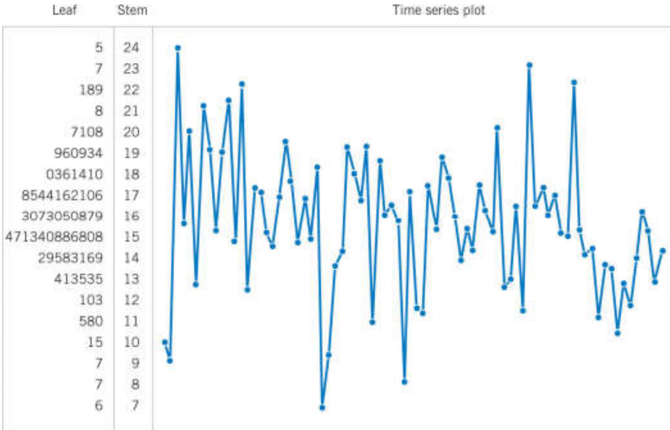


Figure 6-17 A digidot plot of the compressive strength data in Table 6-2. It combines a time series with a stem-and-leaf plot. The variability in the frequency distribution, as shown by the stem-and-leaf plot, is distorted by the apparent trend in the time series data.

Digiplot of Chemical Concentration Data

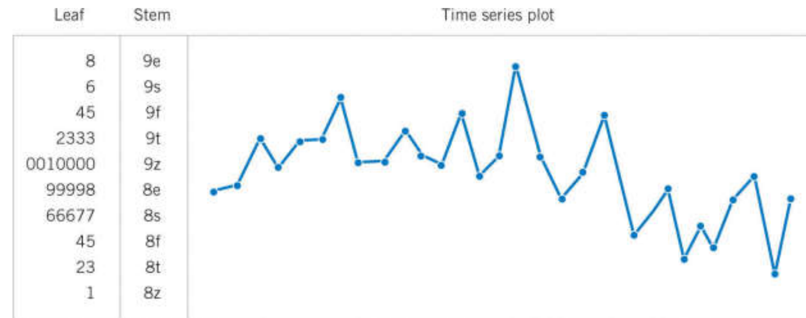


Figure 6-18 A digiplot of chemical concentration readings, observed hourly. Comment: For the first 20 hours, the mean concentration is about 90. For the last 9 hours, the mean concentration has dropped to about 85. This shows that the process has changed and might need adjustment. The stem-and-leaf plot does not highlight this shift.

Probability Plots

- How do we know if a particular probability distribution is a reasonable model for a data set?
- We use a probability plot to verify such an assumption using a subjective visual examination.
- A histogram of a large data set reveals the shape of a distribution. The histogram of a small data set would not provide such a clear picture.
- A probability plot is helpful for all data set sizes.

How To Build a Probability Plot

- To construct a probability plot:
  - Sort the data observations in ascending order: $x_{(1)}$, $x_{(2)}$,…, $x_{(n)}$.
  - The observed value $x_{(j)}$ is plotted against the cumulative distribution $(j - 0.5)/n$.
  - The paired numbers are plotted on the probability paper of the proposed distribution.
  - If the paired numbers form a straight line, it is reasonable to assume that the data follows the proposed distribution.

Example 6-7: Battery Life

The effective service life (minutes) of batteries used in a laptop are given in the table. We hypothesize that battery life is adequately modeled by a normal distribution. The probability plot is shown on normal probability vertical scale.

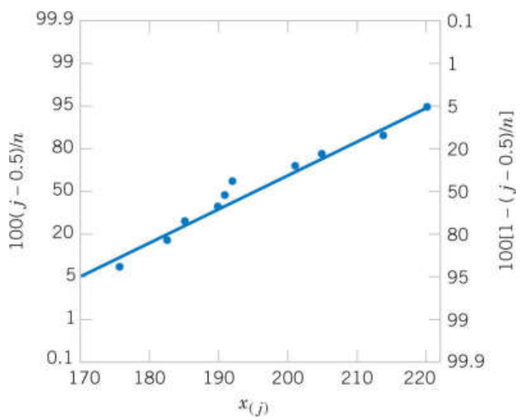


Figure 6-19 Normal probability plot for battery life.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$
1	176	0.05
2	183	0.15
3	185	0.25
4	190	0.35
5	191	0.45
6	192	0.55
7	201	0.65
8	205	0.75
9	214	0.85
10	220	0.95

Probability Plot on Ordinary Axes

A normal probability plot can be plotted on ordinary axes using z-values. The normal probability scale is not used.

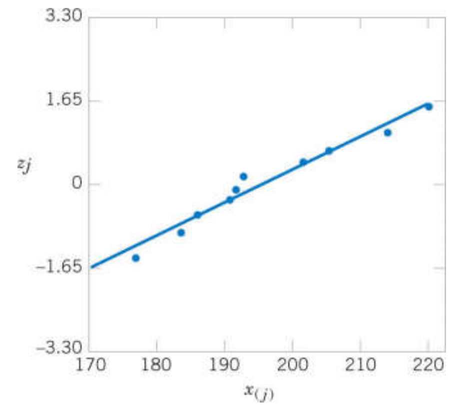


Figure 6-20 Normal Probability plot obtained from standardized normal scores. This is equivalent to Figure 6-19.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

Use of the Probability Plot

- The probability plot can identify variations from a normal distribution shape.
  - Light tails of the distribution – more peaked.
  - Heavy tails of the distribution – less peaked.
  - Skewed distributions.
- Larger samples increase the clarity of the conclusions reached.

Probability Plot Variations

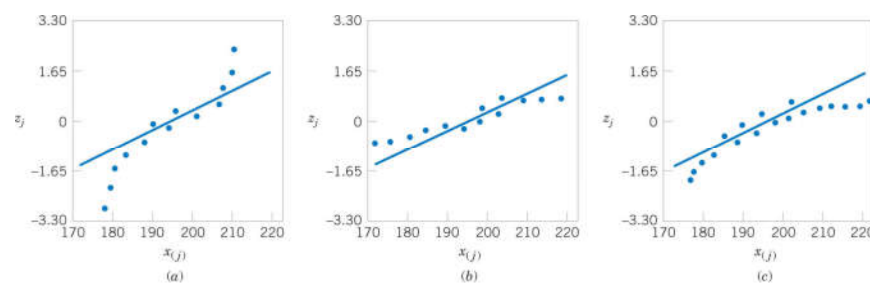
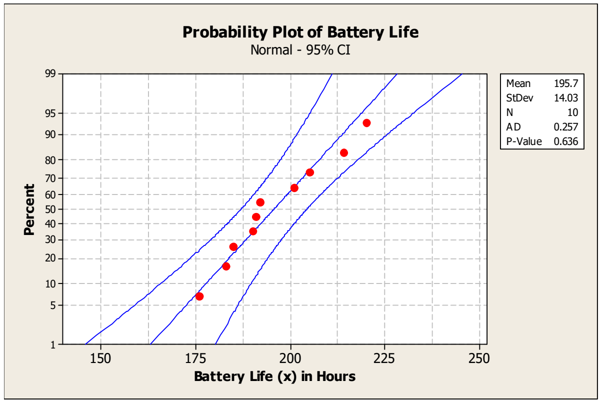


Figure 6-21 Normal probability plots indicating a non-normal distribution.
(a) Light tailed distribution (squeezed together)
(b) Heavy tailed distribution (stretched out)
(c) Right skewed distribution (one end squeezed, other end stretched)

Probability Plots with Minitab

- Obtained using Minitab menu: Graphics > Probability Plot. 14 different distributions can be used.
- The curved bands provide guidance whether the proposed distribution is acceptable – all observations within the bands is good.



Important Terms & Concepts of Chapter 6

- Box plot
- Frequency distribution & histogram
- Median, quartiles & percentiles
- Multivariable data
- Normal probability plot
- Pareto chart
- Population:
  - Mean
  - Standard deviation
  - Variance
- Probability plot
- Relative frequency distribution
- Sample:
  - Mean
  - Standard deviation
  - Variance
- Stem-and-leaf diagram
- Time series plots