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.

Sample Mean

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

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

$$\mu = \sum_{i=1}^{N} x_i \cdot f(x) = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example 6-1: Sample Mean

Consider 8 observations (x_i) of pull-off force from engine connectors 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	12.3
5	13.6
6	13.5
7	12.6
8	13.1
	13.00
=AVERAGE($B2:$B9)	

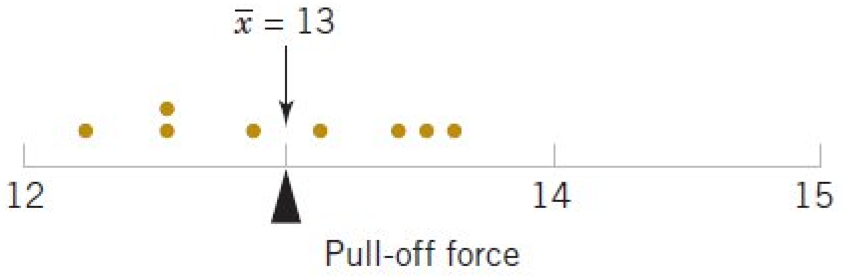


Figure 6-1 The sample mean is the balance point.

Variance Defined

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

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

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

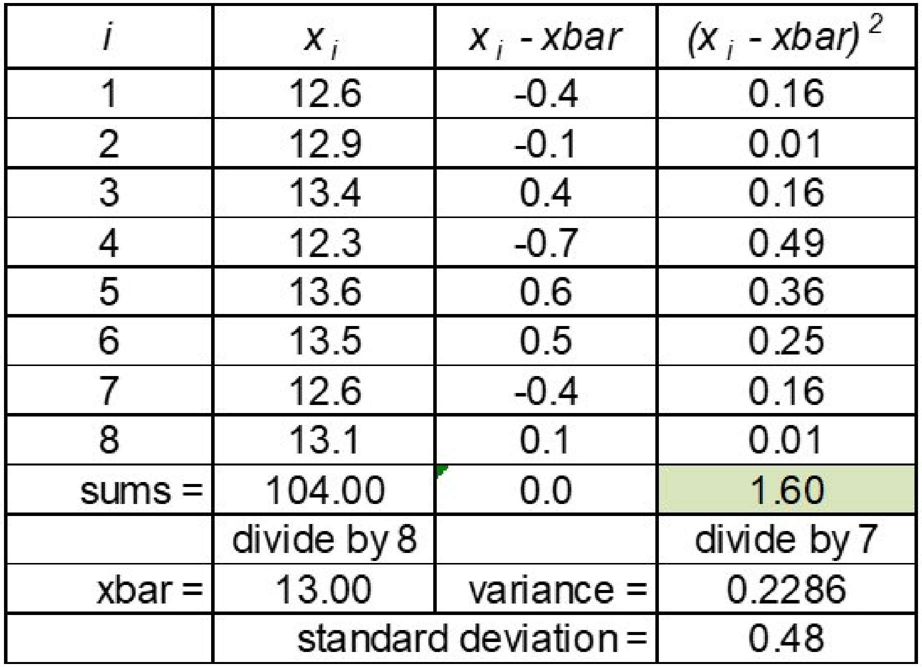
$$\sigma^2 = \sum_{i=1}^{N}\left(x_i - \mu\right)^2 \cdot f\left(x\right) = \frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}$$

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.

Example 6-2: Sample Variance

Table 6-1 displays the quantities needed to calculate the sample variance and sample standard deviation.



i	x_i	x_i - xbar	$(x_i - \text{xbar})^2$
1	12.6	-0.4	0.16
2	12.9	-0.1	0.01
3	13.4	0.4	0.16
4	12.3	-0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	-0.4	0.16
8	13.1	0.1	0.01
sums =	104.00	0.0	1.60
	divide by 8		divide by 7
xbar =	13.00	variance =	0.2286
	standard deviation =		0.48

Table 6-1

<u>Dimension</u> 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.

Computation of s^2

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

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

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

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

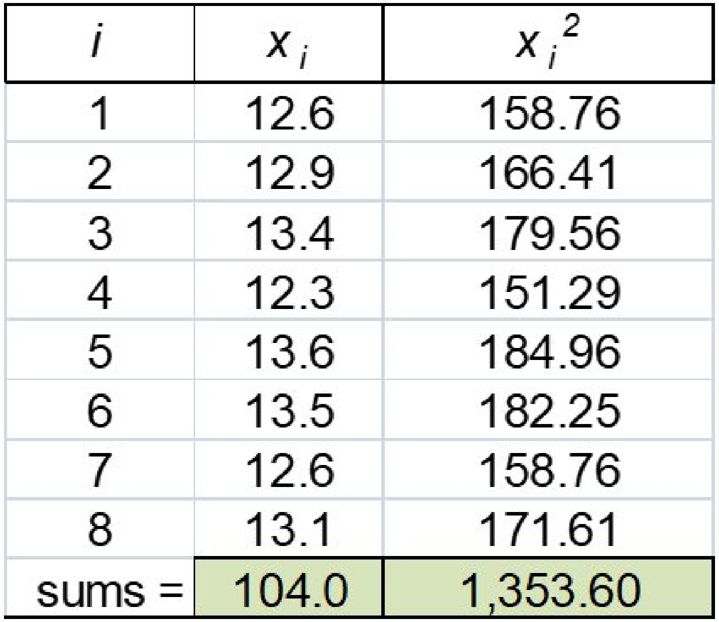
Example 6-3: Variance by Shortcut

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

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

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

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



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.
  - *n*-1 of the observations can be freely determined, but the n^{th} observation is fixed to maintain the zero sum.

Stem-and-Leaf Diagrams

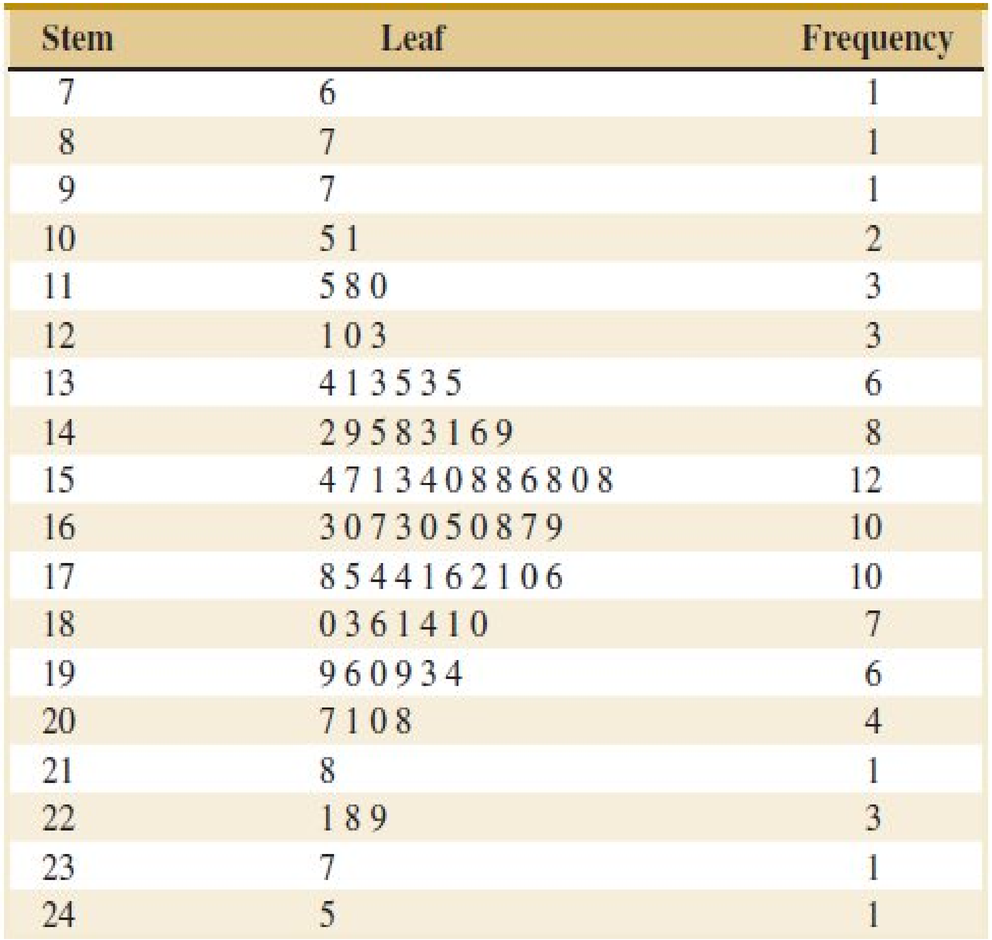
- Dot diagrams (dotplots) are useful for small data sets. Stem & leaf diagrams are better for large sets.
- Steps to construct a stem-and-leaf diagram:
  1) Divide each number (x_i) into two parts: a stem, consisting of the leading digits, and a leaf, consisting of the remaining digit.
  2) List the stem values in a vertical column.
  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

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2.

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.

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

Frequency Distribution for the data in Table 6-2

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 construct a histogram 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 corresponding to that bin.

Histogram of the Table 6-2 Data

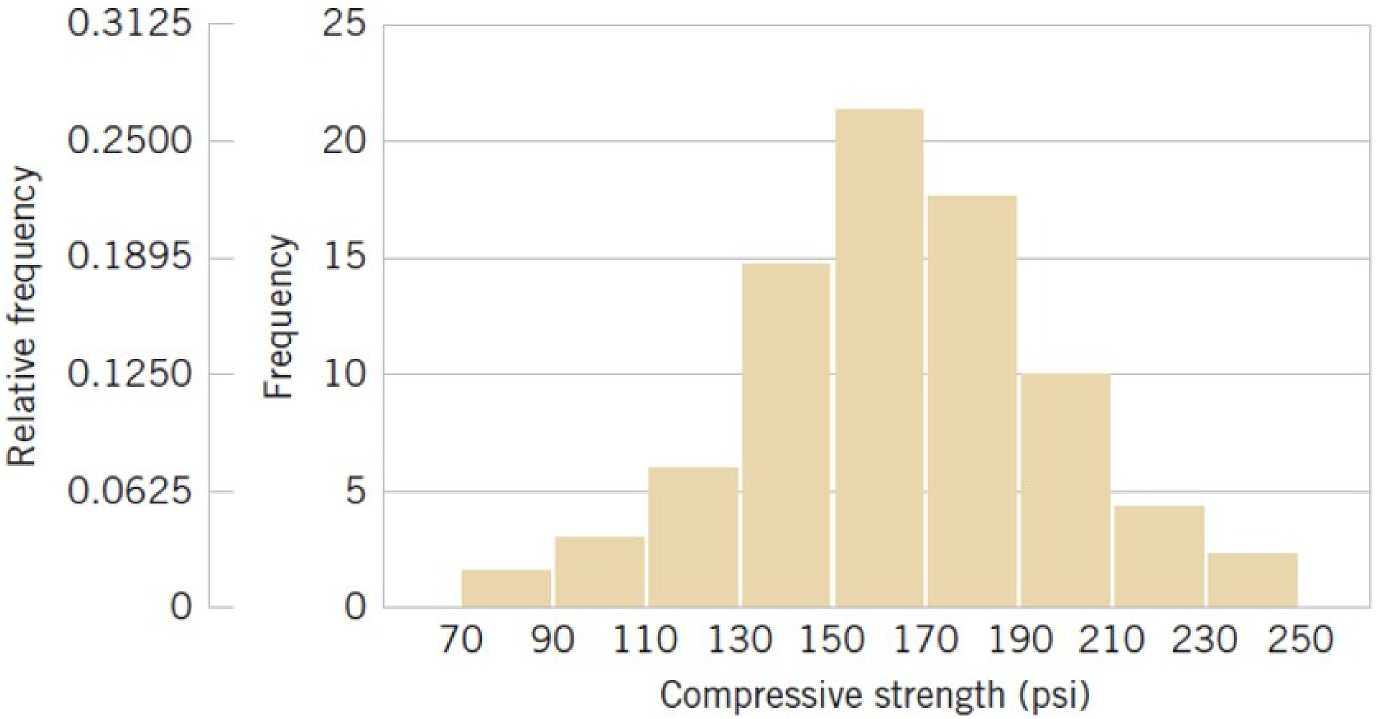


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

Cumulative Frequency Plot

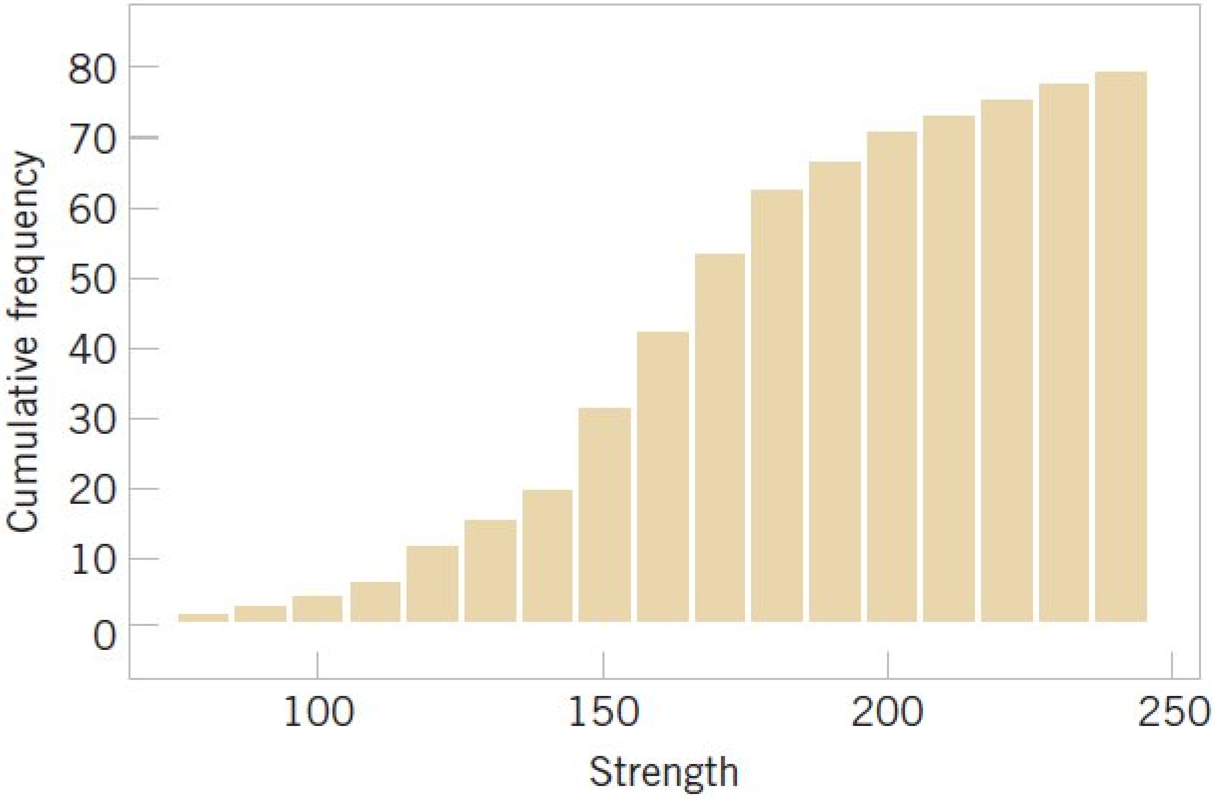


Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

Shape of a Frequency Distribution

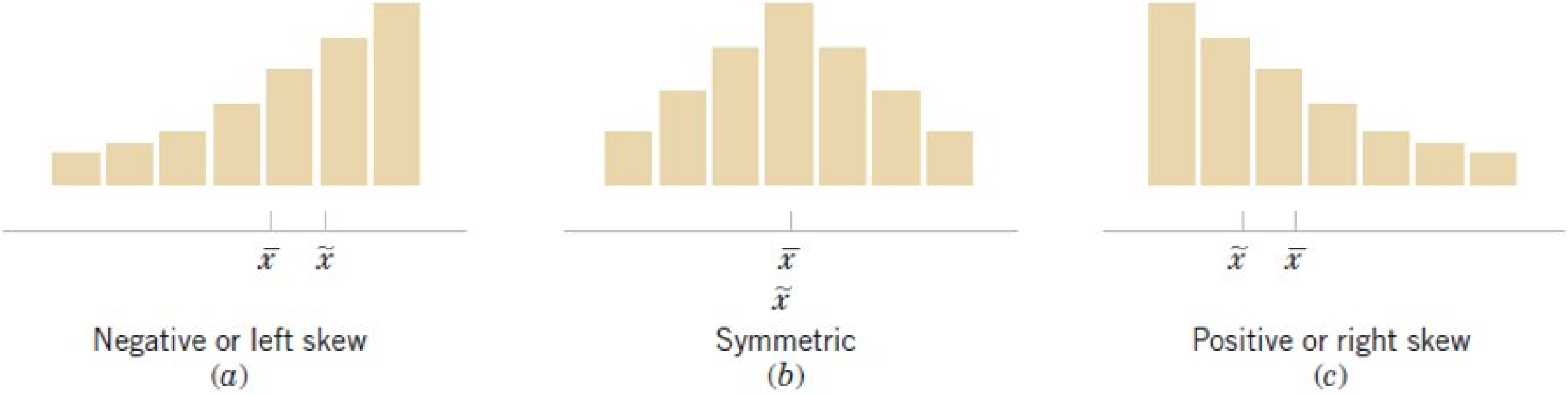


Figure 6-11 Histograms of symmetric and skewed distributions.

(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.☺

Constructing a Probability Plot

- To construct a probability plot:
  - Sort the data observations in ascending order: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
  - The observed value $x_{(j)}$ is plotted against the observed cumulative frequency $(j - 0.5)/n$.
  - The paired numbers are plotted on the probability paper of the proposed distribution.
- If the paired numbers form a straight line, then the hypothesized distribution adequately describes the data.

Example 6-7: Battery Life

The effective service life (X_j in minutes) of batteries used in a laptop are given in the table. We hypothesize that battery life is adequately modeled by a normal distribution. To this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies and plot them.

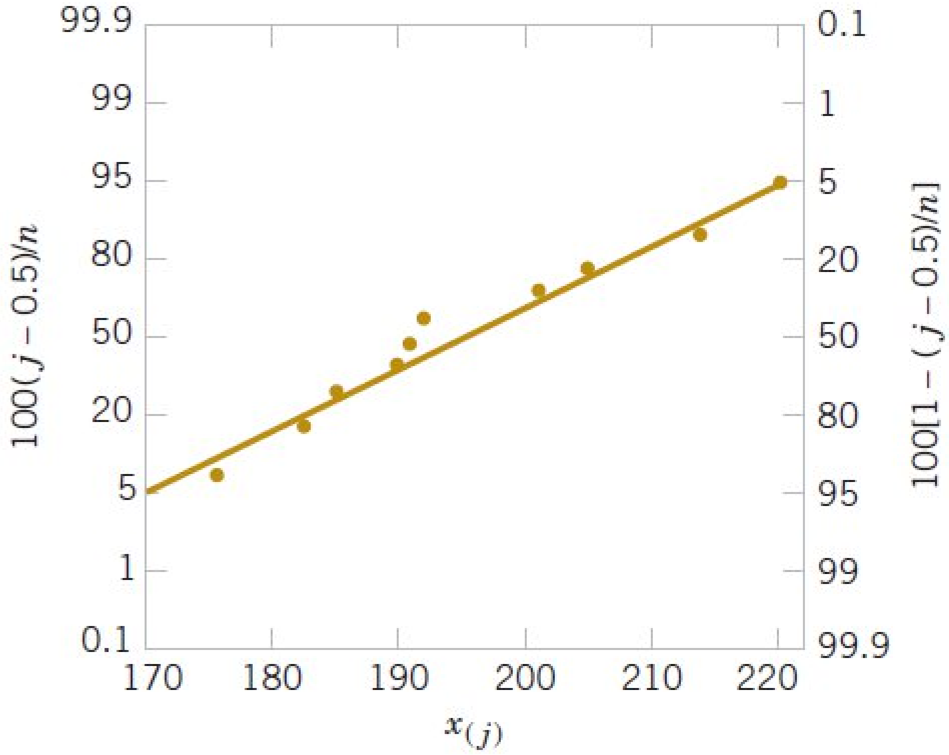


Figure 6-22 Normal probability plot for battery life.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j\text{-}0.5)/10$	$100(j\text{-}0.5)/10$
1	176	0.05	5
2	183	0.15	15
3	185	0.25	25
4	190	0.35	35
5	191	0.45	45
6	192	0.55	55
7	201	0.65	65
8	205	0.75	75
9	214	0.85	85
10	220	0.95	95

Probability Plot on Standardized Normal Scores

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

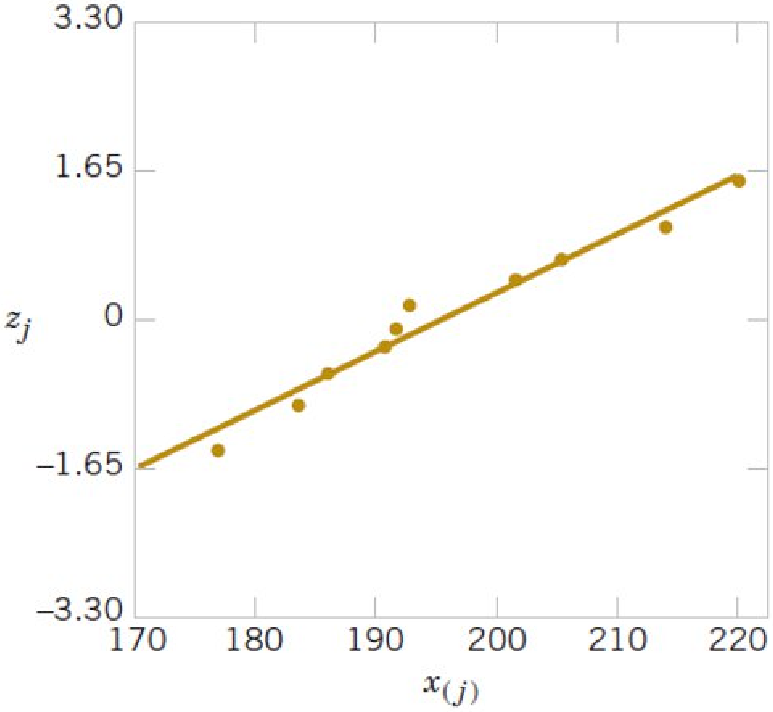


Figure 6-23 Normal Probability plot obtained from standardized normal scores. This is equivalent to Figure 6-19.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j\text{-}0.5)/10$	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

Probability Plot Variations

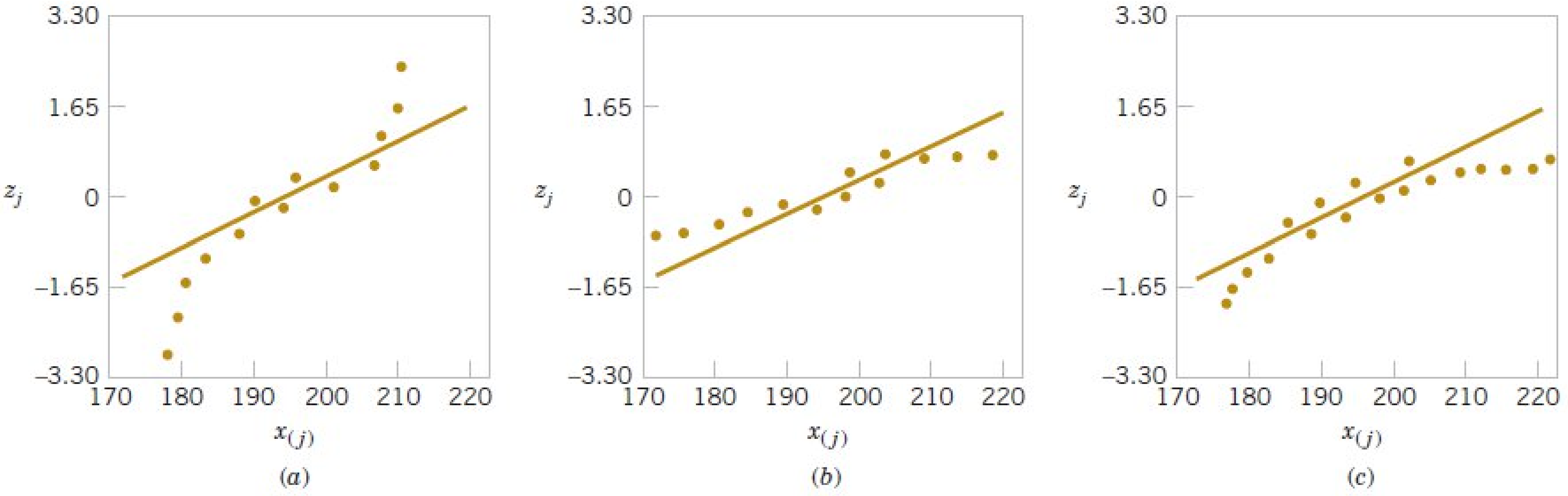


Figure 6-24 Normal probability plots indicating a non-normal distribution.
(a) Light tailed distribution
(b) Heavy tailed distribution
(c) Right skewed distribution