3 Descriptive statistics

Engineering data are always variable. Given precise enough measurement, even constant process conditions produce different responses. Thus, it is not the individual data values that are important, but their **distribution**. We will discuss simple methods that describe important distributional characteristics of data.

Definition 3.1. *Descriptive statistics* is the use of plots and numerical summaries to describe data without drawing any formal conclusions.

Through the use of *descriptive statistics*, we seek to find the following features of data sets:

1. Center

2. Spread

3. Shape

4. Outliers

3.1 Graphical and tabular displays of quantitative data

Almost always, the place to start a data analysis is with appropriate graphical and tabular displays. When only a few samples are involved, a good plot can tell most of the story about data and drive an analysis.

3.1.1 Dot diagrams and stem-and-leaf plots

When a study produces a small or moderate amount of **univariate quantittative data**, a *dot diagram* can be useful.

Definition 3.2. A *dot diagram* shows each observation as a dot placed at the position corresponding to its numerical value along a number line.

Example 3.1 (Heat treating gears, cont'd)**.** Recall the example from Chapter 1. A process engineer is faced with the question, "How should gears be loaded into a continuous carburizing furnace in order to minimize distortion during heat treating?" The engineer conducts a well-thought-out study and obtains the runout values for 38 gears laid and 39 gears hung.

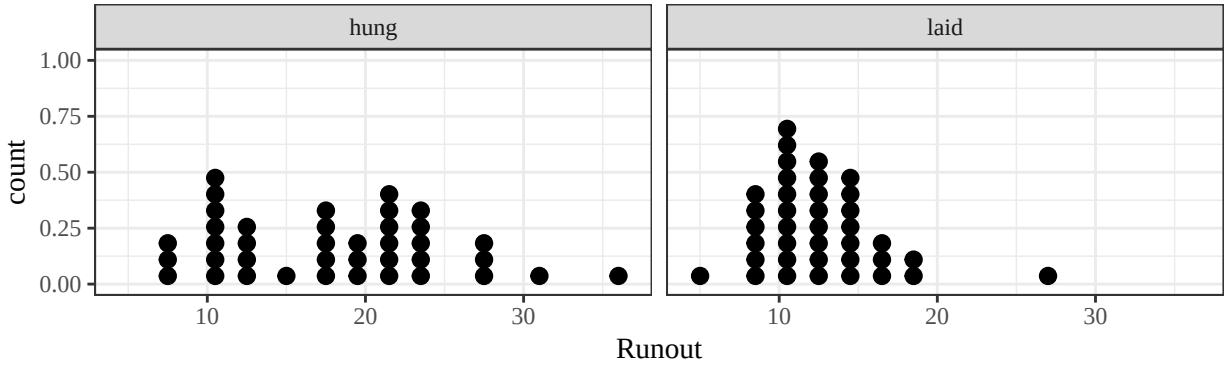


Figure 1: Dot diagrams of runouts.

Example 3.2 (Bullet penetration depth, pg. 67)**.** Sale and Thom compared penetration depths for several types of .45 caliber bullets fired into oak wood from a distance of 15 feet. They recorded the penetration depths (in mm from the target surface to the back of the bullets) for two bullet types.

200 grain jacketed bullets	230 grain jacketed bullets
63.8, 64.65, 59.5, 60.7, 61.3, 61.5, 59.8, 59.1, 62.95, 63.55, 58.65, 71.7, 63.3, 62.65, 67.75, 62.3, 70.4, 64.05, 65, 58	40.5, 38.35, 56, 42.55, 38.35, 27.75, 49.85, 43.6, 38.75, 51.25, 47.9, 48.15, 42.9, 43.85, 37.35, 47.3, 41.15, 51.6, 39.75, 41

Table 1: Bullet penetration depths (mm)

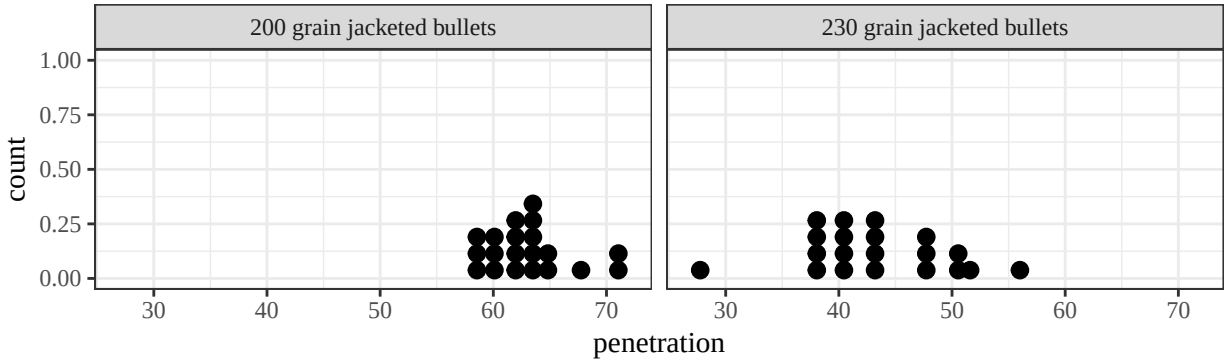


Figure 2: Dot diagrams of penetration depths.

Dot diagrams are good for getting a general feel for the data (and can be done with pencil and paper), but do not allow the recovery of the exact values used to make them.

Definition 3.3. A *stem-and-leaf plot* is made by using the last few digits of each data point to indicated where it falls.

Example 3.3 (Heat treating gears, cont'd)**.**

hung	laid
7, 8, 8, 10, 10, 10, 10, 11, 11,	5, 8, 8, 9, 9, 9, 9, 10, 10, 10,
11, 12, 13, 13, 13, 15, 17, 17,	11, 11, 11, 11, 11, 11, 11, 12,
17, 17, 18, 19, 19, 20, 21, 21,	12, 12, 12, 13, 13, 13, 13, 14,
21, 22, 22, 22, 23, 23, 23, 23,	14, 14, 15, 15, 15, 15, 16, 17,
24, 27, 27, 28, 31, 36	17, 18, 19, 27

Table 2: Thrust face runouts (.0001 in.)

3.1.2 Frequency tables and histograms

Dot diagrams and stem-and-leaf plots are useful for getting to know a data set, but they are not commonly used in papers and presentations.

Definition 3.4. A *frequency table* is made by first breaking an interval containing all the data into an appropriate number of smaller intervals of equal length. Then tally marks can be recorded to indicate the number of data points falling into each interval. Finally, frequencies, relative frequencies, and cumulative relative frequencies can be added.

Example 3.4 (Heat treating gears, cont'd)**.**

Runout (.0001 in)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5-8	|||	3	.079	.079
9-12	𝍸 𝍸 𝍸 |||	18	.474	.553
13-16	𝍸 𝍸 ||	12	.316	.868
17-20	||||	4	.105	.974
21-24		0	0	.974
25-28	|	1	.026	1.000
		38	1.000	

Table 3: Frequency table for laid gear thrust face runouts.

Example 3.5 (Bullet penetration depth, cont'd)**.**

Runout (.0001 in)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
58-59.99	𝍸	5	.25	.25
60.00-61.99	|||	3	.15	.40
62.00-63.99	𝍸 |	6	.30	.70
64.00-65.99	|||	3	.15	.85
66.00-67.99	|	1	.05	.90
68.00-69.99		0	0	.90
70.00-71.99	||	2	.10	1.000
		20	1.000	

Table 4: Frequency table for 200 grain penetration depths.

After making a frequency table, it is common to use the organization provided by the table to create a histogram.

Definition 3.5. A *(frequency or relative frequency) histogram* is a kind of bar chart used to portray the shape of a distribution of data points.

Guidelines for making histograms:

Example 3.6 (Bullet penetration depth, cont'd)**.**

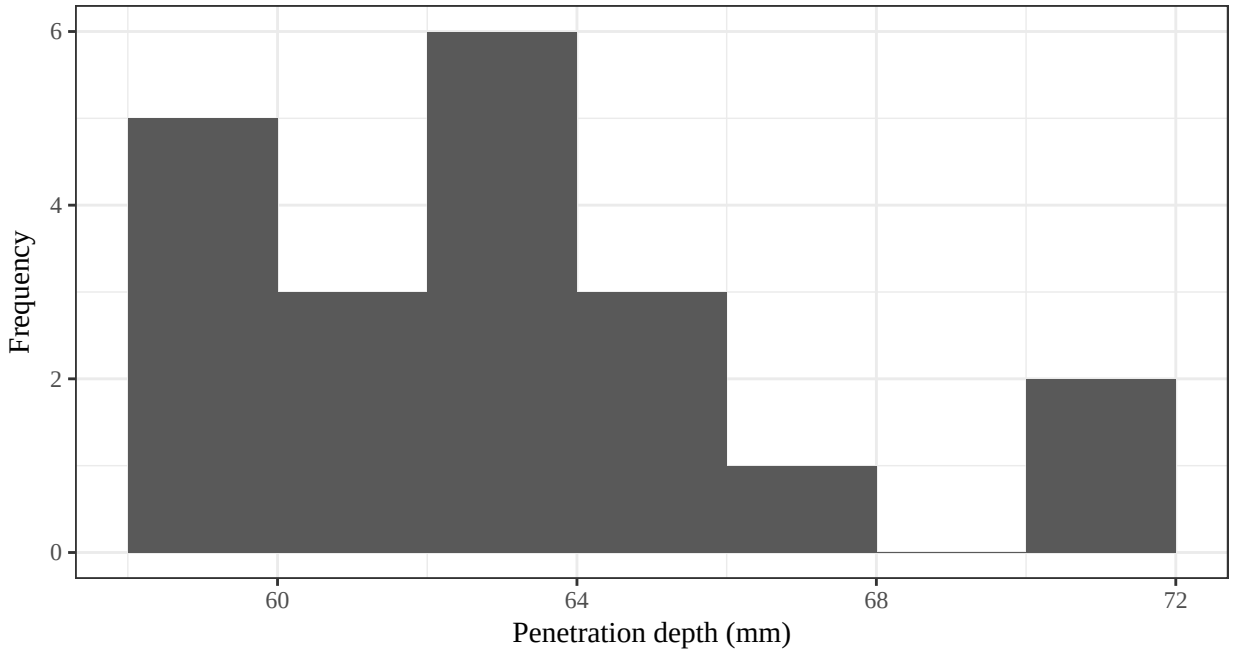


Figure 3: Histogram of the 200 grain penetration depths.

Example 3.7. Suppose you have the following data:

$$74, 79, 77, 81, 68, 79, 81, 76, 81, 80, 80, 78, 88, 83, 79, 91, 79, 75, 74, 73$$

. Create the corresponding *frequency table* and *freqency histogram*.

Why do we plot data? Information on location, spread, and shape is portrayed clearly in a histogram and can give hints as to the functioning of the physical process that is generating the data.

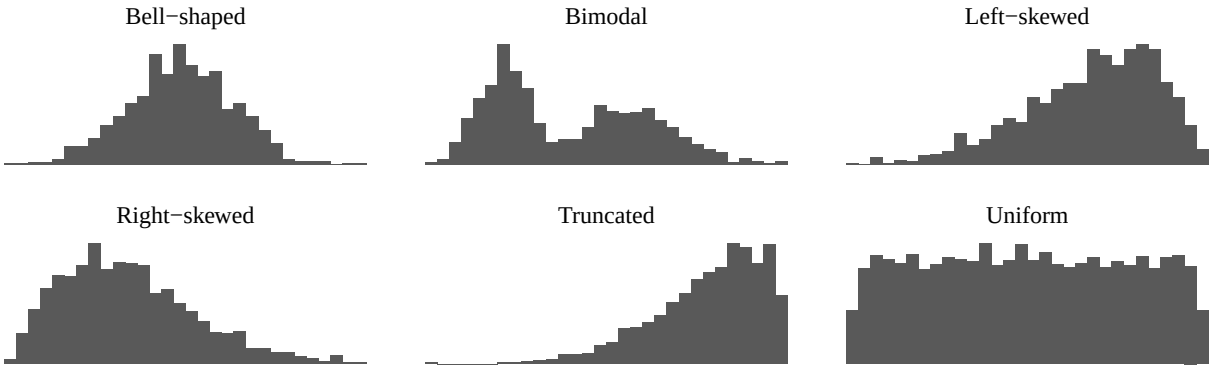


Figure 4: Common distributional shapes.

If data on the diameters of machined metal cylinders purchased from a vendor produce a histogram that is decidedly **bimodal**, this suggests

If the histogram is **truncated**, this might suggest

3.1.3 Scatter plots

Dot-diagrams, stem-and-leaf plots, frequency tables, and histograms are univariate tools. But engineering questions often concern multivariate data and *relationships between the variables*.

Definition 3.6. A *scatterplot* is a simple and effective way of displaying potential relationships between two quantitative variable by assigning each variable to either the x or y axis and plotting the resulting coordinate points.

Example 3.8. Jim and Jane want to know the relationship between an orange tree's age (in days since 1968-12-31) and its circumference (in mm). They recorded the data for 35 orange trees.

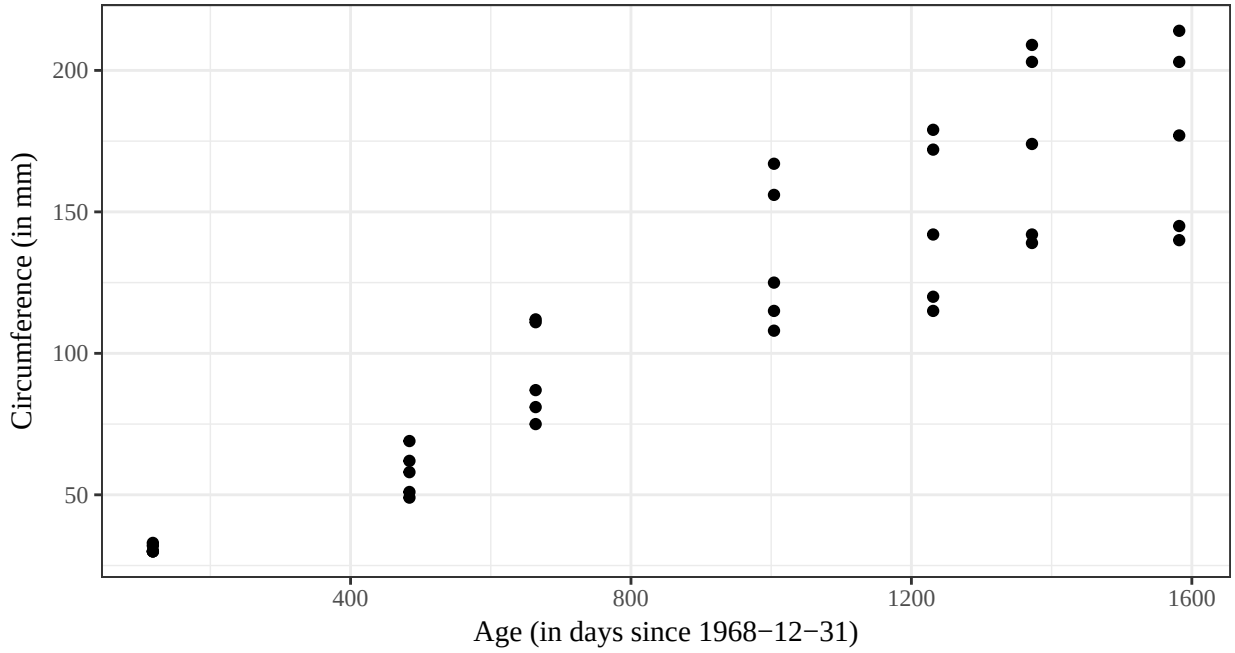
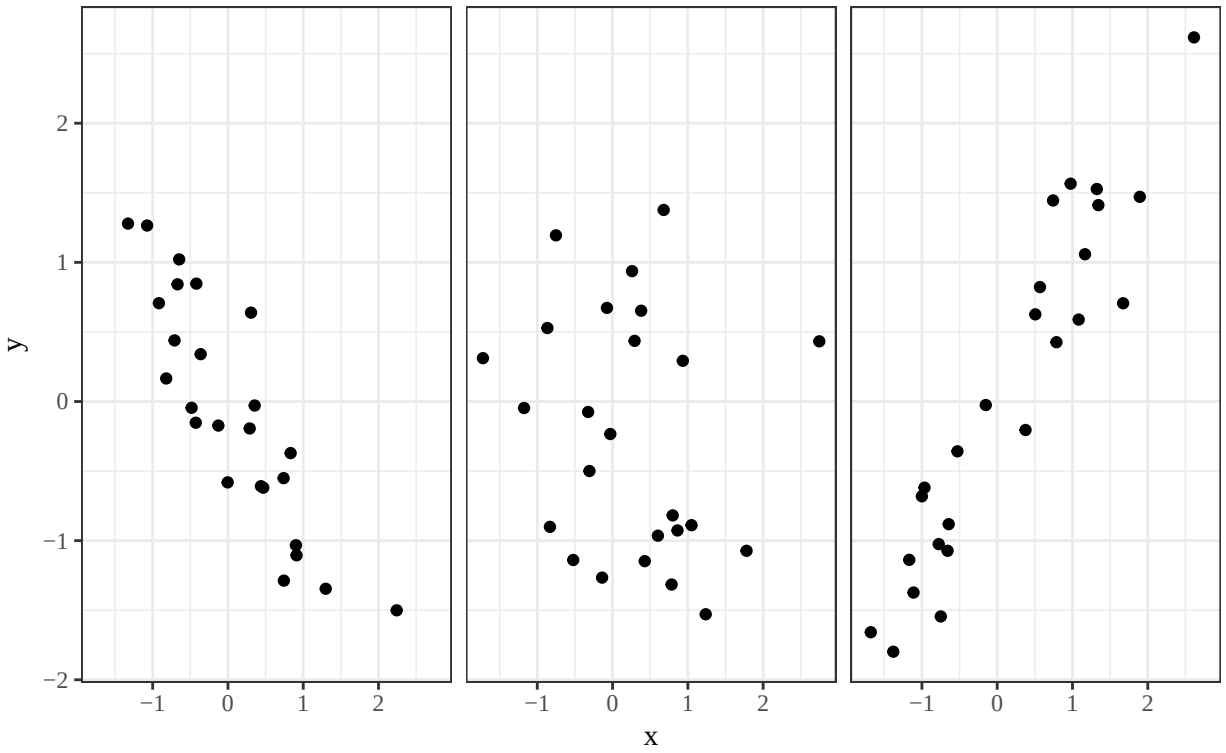


Figure 5: Scatterplot of 35 trees' age and circumference.

There are three typical association/relationship between two variables:



Definition 3.7. A *run chart* is a basic graph that displays data values in a time sequence in the order in which the data were generated.

Example 3.9 (Office hours)**.** A professor collects data on the number of students that come to her office hours per week during the course of the semester.

Week	Attendance
1	0.00
2	1.00
3	4.00
4	5.00
5	40.00
6	2.00
7	5.00
8	10.00
9	7.00
10	30.00
11	0.00
12	4.00
13	3.00
14	19.00
15	60.00

Table 5: Weekly attendance in office hours for a semester.

3.2 Quantiles

3.2.1 Boxplots

3.2.2 Quantile-quantile plots

3.2.3 Theoretical quantile-quantile plots

3.3 Numerical summaries

3.3.1 Location and spread

3.3.2 Statistics and parameters

3.4 Categorical and count data