STAT 509: Statistics for Engineers

Chapter 6: Descriptive Statistics

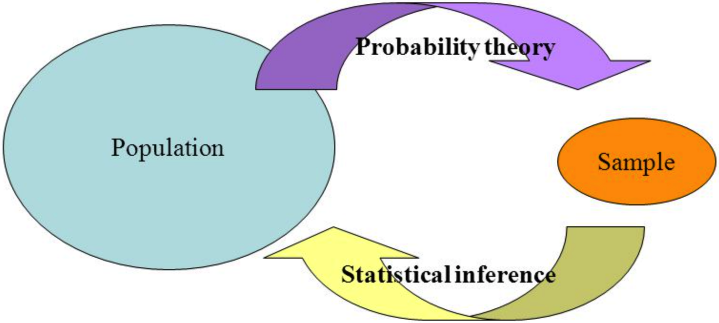
Dr. Dewei Wang
Associate Professor
Department of Statistics
University of South Carolina
deweiwang@stat.sc.edu

Chapter 6: Descriptive Statistics

Learning Objectives:

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

Statistical Inference



Population: Our goal; i.e., time to next failure, lifetime of a product, salary after graduation, COVID-19 prevalence...

Sample: (Randomly) selected from the population. We analyze sample to learn about the population.

Numerical Summaries of Data

Well-constructed data summaries and displays are essential to good statistical thinking because they can focus the engineer on important features of the data or provide insight about the type of model that should be used in solving the problem.

We often find it useful to describe data features

numerically and/or visually.

Now we learn some commonly-used descriptive statistics.

Population mean/variance/standard deviation

In previous chapters, we have introduced the mean/variance/standard deviation of a probability distribution, denoted by $\mu/\sigma^2/\sigma$.

- μ characterizes the **center** of the distribution
- σ^2 and σ characterize the **variability** of the distribution.

If we think of a probability distribution as a model for the population. Then the mean/standard deviation are the center/variability of the population. We call

- μ: *population* mean
- σ^2: *population* variance
- σ: *population* standard deviation

Sample mean/variance/standard deviation

Suppose we observe a sample of size n from the population. We denoted the observed data by $x_1, \dots, x_n$ (lower case x). From these data, we can calculate

- the sample mean:

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

  characterizing the **center** of the observed sample;
- the sample variance:

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

- and the sample standard deviation:

  $$s_n = \sqrt{s_n^2},$$

  characterizing the **variability** of the observed sample.

Sample range & sensitive to outlier

- Another useful measure of variability (**spread**) is the sample range:

$$r_n = \max_i x_i - \min_i x_i.$$

Remark: all these statistics $\bar{x}_n$, $s_n^2(s_n)$, and r_n are **sensitive** to outliers.

Example

Say we have observed $x_i = i$ for $i = 1, \ldots, 20$. Then

$$\bar{x}_n = 10.5, s_n^2 = 35, s_n = 5.9161, r_n = 19.$$

Suppose **One** careless input brought $x_{20} = 20$ to $x_{20} = 2000$. Then

$$\bar{x}_n = 109.5, s_n^2 = 198035, s_n = 4455.0112, r_n = 1999.$$

$\frac{19}{20} = 95\%$ data are unchanged, but these statistics are changed dramatically.

Robust statistics

- Sample percentiles:
  - 1st sample quartile (Q1): The 25th percentile of the sample, denoted by q_1.
  - Sample median (Q2): The 50th percentile of the sample, denoted by q_2.
  - 3rd sample quartile (Q3): The 75th percentile of the sample, denoted by q_3.
- Interquartile range (IQR): $IQR = q_3 - q_1$.

Same example

Say we have observed $x_i = i$ for $i = 1, \dots, 20$. Then

$$q_1 = 5.75, q_2 = 10.5, q_3 = 15.25, IQR = 9.5.$$

Suppose **One** careless input brought $x_{20} = 20$ to $x_{20} = 2000$. Then

$$q_1 = 5.75, q_2 = 10.5, q_3 = 15.25, IQR = 9.5.$$

The outlier does not affect these statistics (robustness).

data.summary in StatEngine

Name your data: e.g,

```
IornMan=1:20
Mulan=seq(2,50,by=3)
Tenet=seq(1,40,by=20)
x=c(12.6,12.9,13.4,12.3,13.6,13.5,12.6,13.1)
```

We use the textbook Table 6.2. These data are the compressive strengths in pounds per square inch (psi) of 80 specimens of a new aluminum-lithium alloy undergoing evaluation as a possible material for aircraft structural elements. The data can be read by

```
x=scan("https://raw.githubusercontent.com/Harrindy
        /StatEngine/master/Data/CompressiveStrength.csv")
```

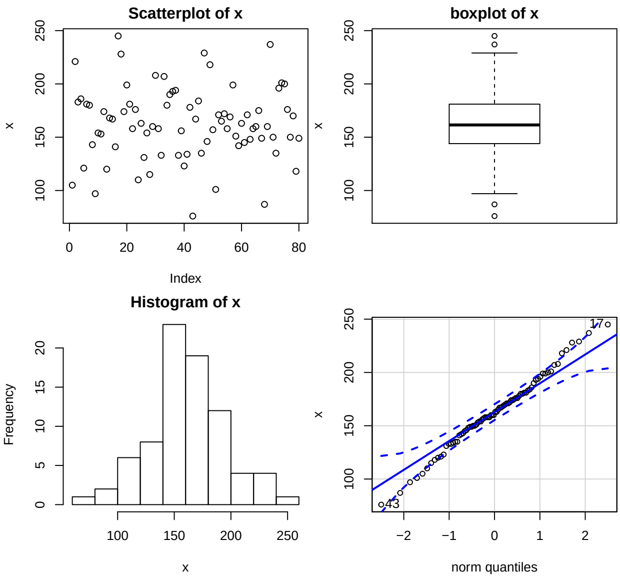
Then run

```
data.summary(x,plot=TRUE)
```

data.summary in StatEngine

```
A stem and leaf diagram is

  The decimal point is 1 digit(s) to the right of the |

   6 | 6
   8 | 77
  10 | 15058
  12 | 013133455
  14 | 12356899001344678888
  16 | 00033577890112445668
  18 | 0011346034699
  20 | 01788
  22 | 1897
  24 | 5

There is no missing value.

Summary:
           [,1] [,2]     [,3]      [,4]    [,5] [,6]  [,7]  [,8]   [,9] [,10]
statistics min  mean     variance  std     max  range Q1    Median Q3   IRQ
result     76   162.6625 1140.6315 33.7732 245  169   144.5 161.5  181  36.5
```

data.summary in StatEngine


Scatterplot of x

boxplot of x

Histogram of x

Stem and Leaf Diagram

Steps to Construct a Stem-and-Leaf Diagram:

(1) Divide each number x_i into two parts: a stem, consisting of one or more 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 stems and leaves on the display.

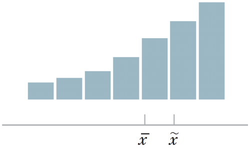
Histogram

Constructing a Histogram (Equal Bin Widths)

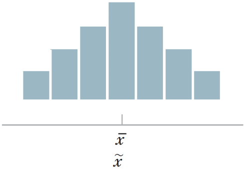
(1) Label the bin (class interval) boundaries on a horizontal scale.

(2) Mark and label the vertical scale with the frequencies or the relative frequencies.

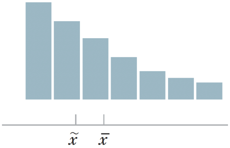
(3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.



Negative or left skew



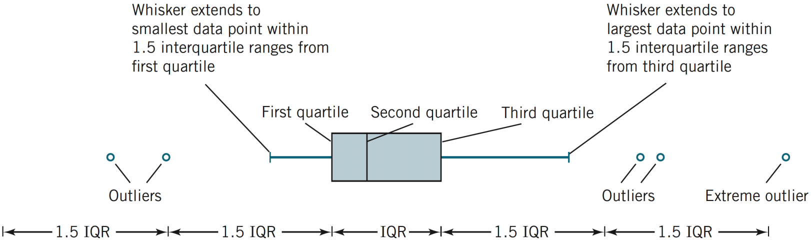
Symmetric



Positive or right skew

where $\bar{x}$ is the sample mean and $\tilde{x}$ is the sample median.

Box Plots



Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Box Plots

Description of a box plot.

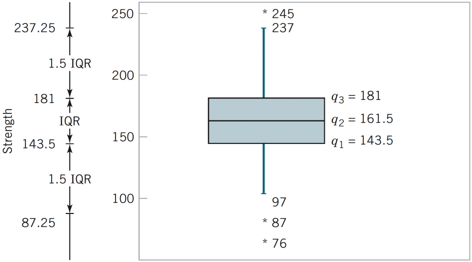


FIGURE 6.14

Box plot for compressive strength data in Table 6.2.

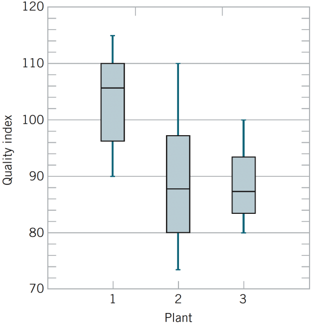


FIGURE 6.15

Comparative box plots of a quality index at three plants.

Box plots facilitate comparison between samples (populations).

Probability Plot (Q-Q plot)

In later chapters, we often assume a normal distribution for the population. To verify this assumption or to check whether the data are from a normal distribution, we often use Q-Q plot:

(i) Rearrange the sample to the order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

(ii) For each j, calculate $p_j = (j - 0.5)/n$.

(iii) Find z_j such that $P(Z \leq z_j) = p_j$; i.e., $p_j = normal.quantile(0, 1, p_j)$

(iv) Plot $(z_j, x_{(j)})$.

Probability Plot (Q-Q plot) checks normality

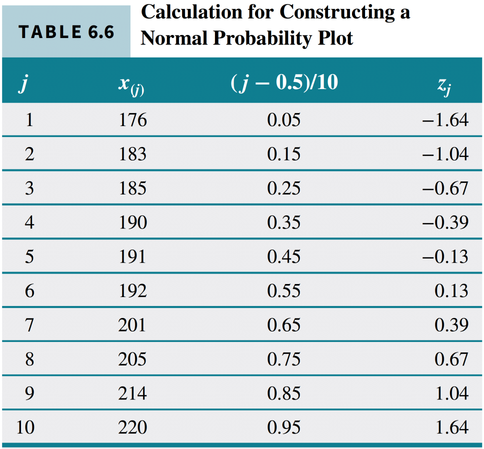
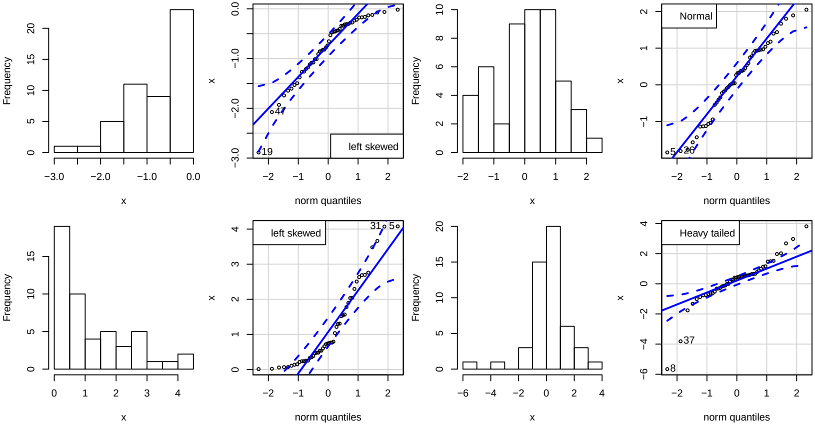


TABLE 6.6 **Calculation for Constructing a Normal Probability Plot**

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

Probability Plot (Q-Q plot)



If the two dashed lines cover all the dots, we could conclude that the samples are from a normal distribution; otherwise, we might not be able to make such a conclusion.