Descriptive Statistics

Population, sample, and statistics. When a data set consists of the n observations $x_1, \ldots, x_n$ recorded or collected, they must be considered as the observed values of *independent* random variables

$$X_1, \ldots, X_n$$

having an *identical* distribution. To judge the quality of data, it is useful to envisage a **population** from which the sample should be drawn. A **random sample** is chosen at random from the population to ensure that the sample is representative of the population. A numerical summary measure of random sample is called *statistic*.

Sample mean and median. The sample mean $\bar{X}$, which is commonly called the average, is defined as

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample median is the value of the "middle" data point. When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the sample median is the mean of the two middle numbers. Thus, the sample median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5.

Sample variance. The sample variance

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{6.1}$$

is the average squared deviation of each number from the sample mean $\bar{X}$, and the **sample standard deviation** S is $\sqrt{S^2}$. The alternative formula of (6.1) becomes

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right),$$

which simplifies the calculation of S^2.

Order statistics. Let $X_1, \ldots, X_n$ be *independent* and *identically distributed* (iid) random variables from the common population. When we sort $X_1, \ldots, X_n$ as

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)},$$

the random variable $X_{(k)}$ is called the **k-th order statistic**. When n is an odd number, using $k_2 = (n+1)/2$ the sample median becomes $X_{(k_2)}$. When n is even, using $k_2 = n/2$ we can define the sample median by

$$\frac{X_{(k_2)} + X_{(k_2+1)}}{2}$$

Lower and upper sample quartiles. The 25-sample percentile is the value indicating that 25% of the observations takes values smaller than the value. Similarly, we can define 50-percentile, 75-percentile, and so on. Note that 50-percentile is the median. We call 25-percentile

the **lower sample quartile** and 75-percentile the **upper sample quartile**. For their actual calculation, let k_1 be the smallest integer exceeding $(n-1)/4$ so that $0 \le f_1 = k_1-(n-1)/4 < 1$, and similarly let k_3 be the smallest integer exceeding $3(n-1)/4$ so that $0 \le f_3 = k_3-3(n-1)/4 < 1$. Then we can obtain the sample lower and upper quartiles respectively by

$$\begin{aligned}(\text{Lower Quartile}) &= f_1 X_{(k_1)} + (1-f_1)X_{(k_1+1)}\\ (\text{Upper Quartile}) &= f_3 X_{(k_3)} + (1-f_3)X_{(k_3+1)}\end{aligned}$$

Exploratory Data Analysis

Stem and leaf plot.

Example 1. The weight data (in grams) of 40 candy bars are collected. Construct a stem and leaf plot.

The first value 20.5 is recorded with 20 treated as the stem and 5 as the leaf. The second value 20.7 with 20 and 7, and so on.

```
20.5 20.7 20.8 21.0 21.0 21.4
21.5 22.0 22.1 22.5 22.6 22.6
22.7 22.7 22.9 22.9 23.1 23.3
23.4 23.5 23.6 23.6 23.6 23.9
24.1 24.3 24.5 24.5 24.8 24.8
24.9 24.9 25.1 25.1 25.2 25.6
25.8 25.9 26.1 26.7
```

Solution. $\Longrightarrow$

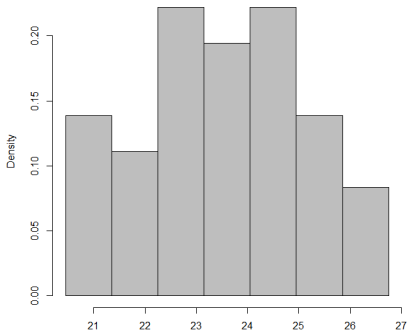
```
20 | 578
21 | 0045
22 | 015667799
23 | 13456669
24 | 13558899
25 | 112689
26 | 17
```

Relative frequency histogram. Once a data set has been collected, it is useful to find an informative way of presenting it. Given the number of observations f_i, called frequency, in the i-th interval, the height h_i of the i-th rectangle above the i-th interval is represented by

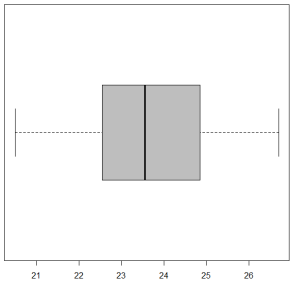
$$h_i = \frac{f_i}{n \times (\text{width of the } i\text{-th interval})}.$$

When the width of each interval is equally chosen, the width w is called **bandwidth** and the height h_i becomes $h_i = \dfrac{f_i}{n \times w}$.

Example 2. The weight data (in grams) of 40 candy bars are collected. The bandwidth of 0.9 is chosen, and the first interval $(20.45, 21.35)$, the second interval $(21.35, 22.25)$, and so on, are constructed. Draw a histogram.



Boxplot. A box is drawn stretching from the lower sample quartile (the 25-percentile) to the upper sample quartile (the 75-percentile). The median is shown as a line across the box. Therefore 1/4 of the distribution is between this line and the right of the box and 1/4 of the distribution is between this line and the left of the box. Two separate lines (dotted), called "whiskers," stretch out from the ends of the box to the smallest and the largest value of data.



Interquartile range (IQR) and fence. The difference between the upper and the lower quartile is called the interquartile range (IQR)

$$\text{IQR} = (\text{Upper Quartile}) - (\text{Lower Quartile})$$

In order to investigate extreme values, we need the following quantities.

$$(\text{Lower inner fence}) = (\text{Lower Quartile}) - (1.5) \times \text{IQR}$$
$$(\text{Upper inner fence}) = (\text{Upper Quartile}) + (1.5) \times \text{IQR}$$

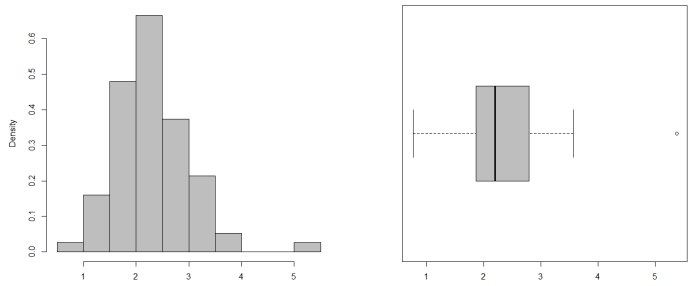
Any data value beyond inner fence on either side is considered as an **outlier**.

Characterizing data: Shape of distribution. Exploratory graphics such as histogram and boxplot can be used to describe the characteristics of the sample distribution. A histogram with one peak is called **unimodal**. When it has two major peak, it is said to be **bimodal**, suggesting a possibility of two distinct populations in the data. A **symmetric** shape should have two symmetric tails on both side, whereas a **skewed** histogram has a longer tail on one side than that on the other. We call it **right-skewed** or **left-skewed** accordingly as we observe the longer right-hand tail or the longer left-hand tail.

Characterizing data: Presence of outliers. Graphical presentations can be also used to identify "odd-looking" value which does not fit in with the rest of the data. Such value is called an **outlier**. In many cases an outlier is discovered to be a misrecorded data value, or represents some special condition that was not in effect when the data were collected. In general data points outside of the fence are considered as outliers.

Characterizing data: Example.

We obtain histogram and boxplot from data set along with the summary statistics.



Minimum	Lower Quartile	Median	Upper Quartile	Maximum
0.760	1.865	2.200	2.795	5.380

Example, continued. The histogram shows that the sample distribution is unimodal and symmetric except for a potential outlier on far right. The maximum value 5.38 appears to be quite separate from the rest of the data as it is shown in the above histogram and boxplot. In fact, IQR is 0.93, the lower and upper inner fence are 0.47 and 4.19. Thus, the maximum value 5.38 is outside of the fence, and it can be considered as an outlier.

Assignment No.6

Supplementary Readings.

TH: Section 3.1–3.3.

Elliot A. Tanis and Robert V. Hogg, *A Brief Course in Mathematical Statistics*. Prentice Hall, NJ.

WM: Section 1.1–1.7.

Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye, *Probability & Statistics for Engineers & Scientists*, 9th ed. Prentice Hall, NJ.

Problem 1. Total of 26 farms are selected in a county, and the sizes of their corn field are measured in acres.

Farm size	sample size	sample mean	sample SD
0-40	15	5.4	8.3
41-160	11	24.3	15.1

(a) Let $X_1, \ldots, X_{15}$ be the numbers of acres of corn sampled from the farms of size at most 40. Find $\sum_{i=1}^{15} X_i$ and $\sum_{i=1}^{15} X_i^2$.

(b) Let $X_1, \ldots, X_{26}$ be all the data collected from 26 farms. Find $\sum_{i=1}^{26} X_i$ and $\sum_{i=1}^{26} X_i^2$.

(c) Compute the sample mean $\bar{X} = \frac{1}{26}\sum_{i=1}^{26} X_i$ and the sample variance $S^2 = \frac{1}{25}\sum_{i=1}^{26}(X_i - \bar{X})^2$

Problem 2. The data below shows commuting times (in minutes) from a random sample of 10 employees who work at a company.

```
21, 26, 32, 27, 23, 39, 30, 24, 28, 36
```

(a) Compute the sample mean, the sample median, and the sample variance.

(b) Compute the sample quartiles and the interquartile range.

(c) Find inner fences. Are there any outliers?

Problem 3. A company selected a random sample of 20 households from the city, and found that the average amount of life insurance is $\bar{X} = 132,000$ and the sample standard deviation is $S = 25,500$. Then the company also asked an independent agency to investigate this issue. They collected a random sample of 41 households from the city, and found that the average amount of life insurance is $\bar{X} = 130,400$ and the sample standard deviation is $S = 25,100$. Based on the information you have obtained above, find the mean, the variance, and the standard deviation of the combined data.

Computer project. Characterize data for each of the next two case studies. Then discuss your opinion regarding the question associated with the experiment. Your report must include:

(a) Sample mean, sample variance, summary statistics, and IQR;

(b) histogram, and your comment on the shape of sample distribution;

(c) boxplot and your comment on possible outliers;

(d) discussion on study questions.

Case study 1. In making aluminum castings, an average of 3.5 ounces per casting must be trimmed off and recycled as a raw material. A new manufacturing procedure has been proposed to reduce the amount of aluminum that must be recycled in this way. For a sample of 12 castings made with the new process, the following are the weights of aluminum trimmed and recycled.

4.2	2.0	3.6	3.8	2.6	3.5
3.2	2.6	2.5	1.7	3.3	4.8

Study question: What do you find regarding the evidence that the new process reduces the amount of trimmed aluminum?

Case study 2. A chocolate manufacturer claims that at the time of purchase by a consumer the age of its product is less than 30 days. In an experiment to test this claim a random sample of 40 chocolate are found to have ages at the time of purchase. Download the data set "`chocolate.txt`" from the course website.

Study question: With this data how do you feel about the manufacturer's claim?

Answers

Problem 1.

(a) $\sum_{i=1}^{15} X_i = (15)(5.4) = 81$ and $\sum_{i=1}^{15} X_i^2 = (14)(8.3)^2 + (15)(5.4)^2 = 1401.86$.

(b) $\sum_{i=16}^{26} X_i = (11)(24.3) = 267.3$ and $\sum_{i=16}^{26} X_i^2 = (10)(15.1)^2 + (11)(24.3)^2 = 8775.49$.
Thus, $\sum_{i=1}^{26} X_i = 81 + 267.3 = 348.3$ and $\sum_{i=1}^{26} X_i^2 = 1401.86 + 8775.49 = 10177.35$.

(c) $\bar{X} = 348.3/26 \approx 13.4$ and $S^2 = (10177.35 - (26)(13.4)^2)/25 \approx 220.35$.

Problem 2. We sort the data, and get the order statistics:

21, 23, 24, 26, 27, 28, 30, 32, 36, 39

(a) The sample mean, the sample median, and the sample variance are 28.6, $27.5 = (X_{(5)} + X_{(6)})/2$, and 32.9, respectively.

(b) The lower and the upper quartile are $24.5 = (0.75)X_{(3)} + (0.25)X_{(4)}$, $31.5 = (0.25)X_{(7)} + (0.75)X_{(8)}$, and the IQR is $7 = 31.5 - 24.5$.

(c) The lower and the upper inner fence are $14 = 24.5 - 10.5$ and $42 = 31.5 + 10.5$. There is no outliers.

Problem 3. We can combine the data collected, and calculate the overall sample mean and the sample variance as follows.

$$\begin{aligned}
\bar{X} &= \frac{1}{61}[(20)(132,000) + (41)(130,400)] \approx 130,924.6 \\
\sum X_i^2 &= (19)(25,500)^2 + (20)(132,000)^2 \\
&\quad + (40)(25,100)^2 + (41)(130,400)^2 \approx 1,083,205,710 \\
S^2 &= \frac{1}{60}(\sum X_i^2 - (61)\bar{X}^2) \approx 626,490
\end{aligned}$$

We have the sample standard deviation $S = \sqrt{626,490} \approx 25,029.8$.