Chapter 1

One-variable data analysis

In this chapter, we start with the simplest case, where there is only one variable. We will learn some basic concepts and how to describe the data both numerically and graphically. Some more advanced concepts, such as density curve, normal distribution are also introduced.

Table 1.1: Example Data[1]

NAME	CLASS	GENDER	MID	FINAL	BASKETBALL
James	23	M	74	86	N
Andrew	23	M	74	86	Y
Jim	23	M	23	47	Y
Kim	23	M	61	78	Y
Mark	23	M	97	98	N
Owen	23	M	73	85	Y
Cook	23	M	98	99	Y
Albert	23	M	81	90	Y
Donald	23	M	70	84	N
Peter	23	M	53	72	Y
Vince	23	M	68	82	Y
Davis	23	M	83	91	Y
Alan	23	M	82	90	Y
Nick	23	M	64	80	N
Elina	23	F	72	85	N
Daisy	23	F	68	82	Y
Crystal	23	F	53	72	N
Karida	23	F	66	81	N
Linda	23	F	83	91	N
Dale	23	F	70	83	Y
Sandy	23	F	56	74	N
Emma	23	F	65	80	N
Angela	23	F	72	85	N
Katie	23	F	84	91	N
Eileen	23	F	73	85	N
Meggie	23	F	68	82	N
Jack	24	M	45	67	Y
Stan	24	M	23	47	Y
Ryan	24	M	60	77	Y
Murphy	24	M	36	60	N
Mike	24	M	82	90	Y
Antony	24	M	18	42	Y
Clare	24	M	86	93	Y
David	24	M	83	91	Y
Taylor	24	M	69	83	Y
Park	24	M	78	88	N
Gary	24	M	51	71	Y
Carson	24	M	85	92	Y
Elvis	24	M	25	49	Y
Kelly	24	F	59	76	N
Sara	24	F	77	88	N
Cherry	24	F	61	78	N
Lucy	24	F	54	73	Y
Hellen	24	F	46	68	N
Chloe	24	F	95	97	Y
Dorothy	24	F	82	90	N
Natalie	24	F	73	85	N
Vivien	24	F	76	87	N
Cathy	24	F	70	84	N
Carol	24	F	55	74	N
Bella	24	F	96	98	Y
Veronica	24	F	60	77	N

[1]MID is the scores in the midterm exam. FINAL is the scores in the final exam. BASKETBALL indicates whether a student plays basketball or not.

1.1 Basic concepts

- In table 1.1, each student is an **individual**.
- All students are described through perspectives of *NAME, CLASS, GENDER, MID, FINAL,* and *BASKETBALL*. Those different perspectives are called **variables**, for they may take different values for different students.
- The values of *MID* and *FINAL* can be operated on like normal numbers, such as taking average, subtraction. Those variables are called **quantitative variables**.
- The values of *NAME, CLASS, GENDER* and *BASKETBALL* only play the role of sorting individuals into different categories. Those variables are called **categorical variables**
- The way a variable takes different values is called the **distribution** of this variable.
- All the individuals we want to study is called the **population**.
- A subset of the population is called a **sample**.

  Samples and populations are relative. If you take all the Chinese people as the population, people in Shanghai is a sample. If you take all the people of the whole world as the population, then Chinese people is a sample.

  Is *CLASS* a quantitative or categorical variable?
- The number of individuals in the sample is called the **sample size**.

A variable takes values of numbers doesn't mean it is quantitative variable.

1.2 Graphs to describe categorical variables

- **Pie chart**

  There are 25 girls and 27 boys in table 1.1. The pie chart of the distribution of *GENDER* is given by figure 1.1.

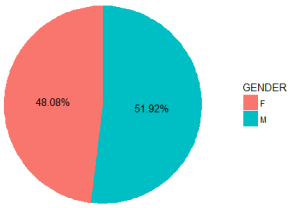


Figure 1.1: Pie chart of the distribution of the *GENDER*

- **Bar graph**

  Similarly, we can draw bar graph to show the information about *GENDER*

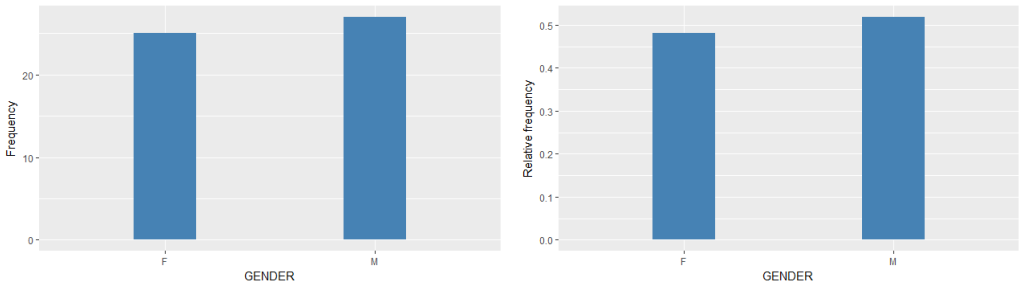


Figure 1.2: Bar graphs with percentage and frequency as vertical axis

In figure 1.2, the vertical axes are **frequency** and **percentage** or (**relative frequency**) respectively. When the sample size is too big, it is better to use relative frequency as the vertical axis.

Be sure to label the axes whenever a graph is drawn!

1.3 Graphs to describe Quantitative Variables

- **Dotplot**

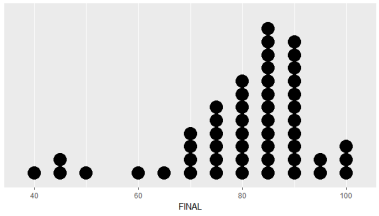


Figure 1.3: Dotplot of the distribution of the *FINAL*

In figure 1.3, the **bin width** is 5. For example, there is only one score lies in the interval $(37.5, 42.5]$, which is "42" from the student whose name is "Antony", and the width of this interval is $42.5 - 37.5 = 5$, which is the bin width. Similarly, there are eight scores lies in the interval $(77.5, 82.5]$.

- **Stemplot**

The **stemplot** is also called **stem-leaf plot**, as shown in figure 1.4. The numbers to the left side of the vertical line are the "stems", and the numbers to the right side of the vertical line are the leaves.

```
1 | 2: represents 12
  4  | 2779
  5  |
  6  | 078
  7  | 12234467788
  8  | 00122233445555566788
  9  | 00001111237889
```

Can we erase the empty stem?

Figure 1.4: The stemplot for the *FINAL*

Sometimes the leaves are too long, we have to split them, and this type of stemplot is the steamplot with **splitting stems**. As shown in figure 1.5, each stem splits into two stems. The first stem holds digits from 0 to 4, the second from 5 to 9. They hold the same number of digits.

```
1 | 2: represents 12
  4* | 2
  4. | 779
  5* |
  5. |
  6* | 0
  6. | 78
  7* | 122344
  7. | 67788
  8* | 0012223344
  8. | 5555566788
  9* | 0000111123
  9. | 7889
```

Why should each stem hold the same number of digits?

Figure 1.5: The stemplot with splitting stems for the *FINAL*

- **Histogram**

  If the dots in dotplot is replaced by bars, the graph will be histogram, as shown in figure

What is the difference between histogram and bar graph?

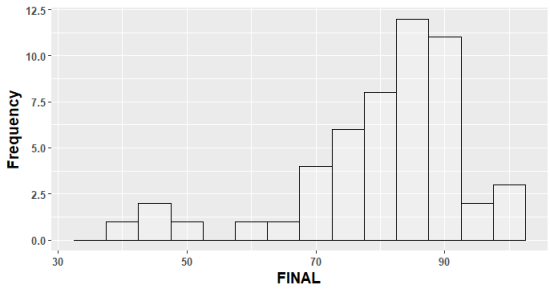


Figure 1.6: Histogram of the distribution of the *FINAL*

The vertical axis can be relative frequency as well.

Compare histogram with stemplot. What are the advantages and disadvantages?

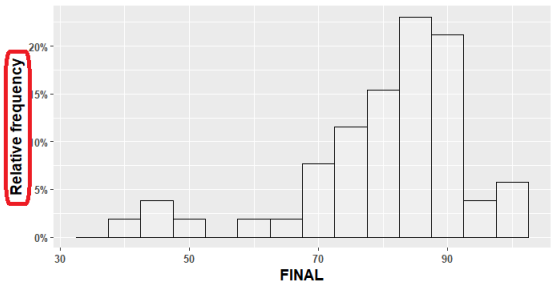


Figure 1.7: Histogram of with vertical axis **relative frequency**

- **Density curve**

  In figure 1.7, we can tell the percentage of $FINAL \leq 50$ is approximately 8% by adding up the percentages of the first three columns. Here, the percentages are indicated by the height of the bars. (4 out of 52 students with $FINAL \leq 50$. Those values are 42, 47, 47, 49.). If the histogram is with bin width 1(figure 1.8), the percentage a bar takes can be calculated by

  $$\textbf{percentage} = \textbf{bar height} = \textbf{bin width} * \textbf{bar height} = \textbf{area of the bar}.$$

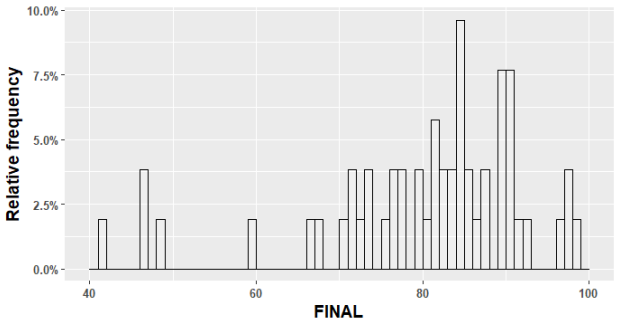
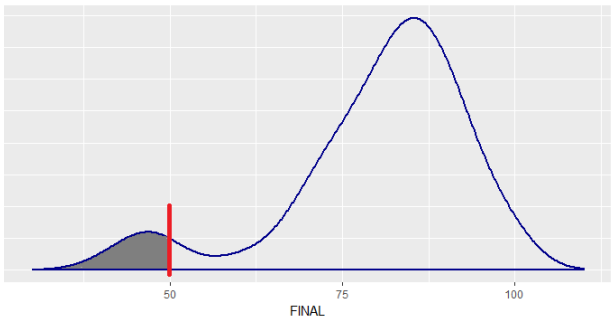


Figure 1.8: Histogram with bin width 1

To calculate the percentage of the students with $FINAL \leq 50$, just add up the areas of the three bars to the left side of 50.

Take a step further. A smooth curve can be drawn such that the area to the left side of **x** gives the percentage of the number of individuals $\leq$ **x**.This graph is called **density curve**. The function of the density curve is called the **probability density function(pdf)**.



What is the total area under the density curve?

Figure 1.9: Smoothed density curve of the *FINAL*

In figure 1.9 , the shaded area gives the percentage of $FINAL \leq 50$, which is approximately 8%.

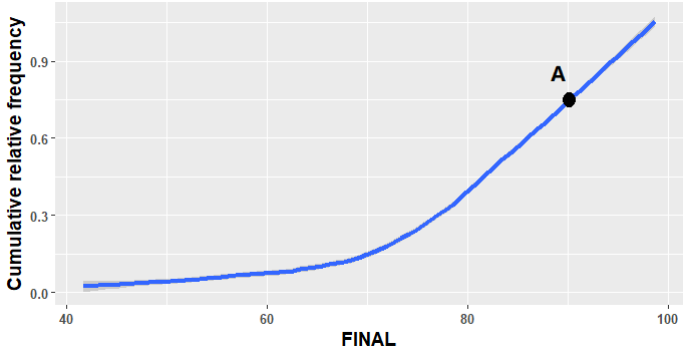
Sometimes the vertical axis is suppressed, because it doesn't mean too much in this book.

- **Cumulative relative frequency curve**

In figure 1.9, for each value of **x** there is an area to the left side of this value. Therefore we can get a function F,such that

$$F(x) = \textbf{Area to the left of x}.$$

If we draw a smooth graph of $F(x)$, it will be like figure 1.10. This curve is called the **cumulative relative frequency curve.** Function $F(x)$ is called **cumulative distribution function(cdf)**.



How to interpret point **A** in figure 1.10?

What is the theoretical relation between a density curve and its cumulative relative frequency curve?

Figure 1.10: Cumulative relative frequency curve of *FINAL*

- **Shape, center and spread of the graphs**

  Shape describes the general outlook of the distribution, by describing **Skewness, clusters, gaps** and **outliers**.

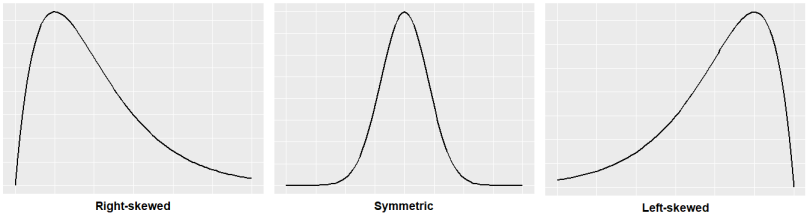


Figure 1.11: Shapes of distributions

As shown in figure 1.11, the skewness of the distributions are **right-skewed, symmetric** and **left-skewed** respectively.

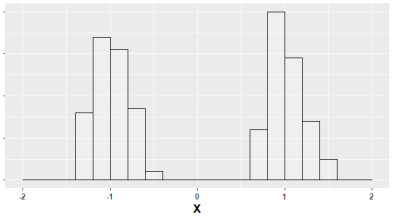


Figure 1.12: Two clusters and a gap

As show in figure 1.12, we say the distribution has two **clusters(modes)** with a **gap**.

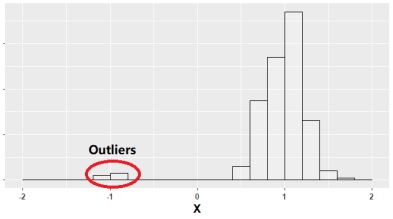


Figure 1.13: Outliers

If some values have striking departures from the pattern of the majority, those values are called **outliers**(as shown in figure 1.13).

Outliers need special attention, for they may be generated by mistakes or some other unconsidered mechanisms.

Center describe the value around which the data is distributed. It will be introduced later

Spread describes how widely the data are scattered. It will be learned latter.

1.4 Summarizing distributions

- **Center**

  There are different ways to describe the center of a distribution. Here we only consider two primary ways of denoting the center:**median** and **mean**.

Median

Arrange the data in increasing or decreasing order, median is the middle one or the average of the middle two.

Mean

For data set $\{x_1, x_2, \cdots, x_n\}$, the mean $\bar{x}$ is given by

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

Notation μ is used for population mean and $\bar{x}$ is for sample mean, though the formulas for them are the same.

For example, we draw a sample of 100 students from a high school, and the mean weight of the whole high school is 65kg, and the mean weight of the 100 students is 60kg. We say

$$\mu = 65\text{kg}, \quad \overline{x} = 60\text{kg}.$$

Median is resistant

For data $\{1, 2, 3, 4, 5\}$, both the mean and the median are 3. If 5 is changed into 500, the mean is 102, while the median is still the same. Median is not easily influenced by extreme values, it is **resistant**

Mean, median and skewness

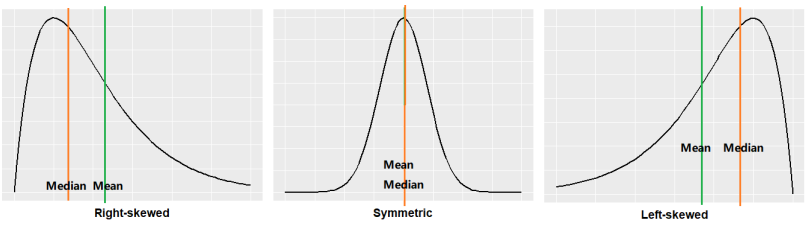


Figure 1.14: The relationship between mean and median

Is mean or median a better description of the center of the distribution of personal incomes, which is strongly right-skewed?

Generally speaking

$$\textbf{symmetric distribution} \implies \textbf{mean} = \textbf{median}$$

$$\textbf{right-skewed distribution} \implies \textbf{mean} > \textbf{median}$$

$$\textbf{left-skewed distribution} \implies \textbf{mean} < \textbf{median}$$

If a distribution is strongly skewed, it is better to use the **resistant** measurement to describe the distribution.

- **Spread**

  Spread is to measure the variability or the dispersion of the data. It can be **range, IQR, variance, standard deviation**.

  Range

  $$\textbf{range} = \textbf{maximun} - \textbf{minimum}$$

  Interquartile range(IQR)

  First quartile(Q_1) is the value with one quarter of the data less than(or equal) to it. For data $\{1, 2, 3, 4, 5, 6, 7, 8\}$, 2 is the first quartile.

  Third quartile(Q_1) is the value with 3/4 of the data less than(or equal) to it. For data $\{1, 2, 3, 4, 5, 6, 7, 8\}$, 6 is the third quartile.

Find the *interquartile range* of $\{1, 2, 3, 4, 5, 6, 7, 8\}$

$$\textbf{IQR} = Q_3 - Q_1$$

  Variance(Var)

Calculate the mean, the variance and the standard deviation of sample data $\{1, 2, 3\}$ by hand.

$$\textbf{Population variance} \quad \textbf{Var} = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.$$

$$\textbf{Sample variance} \quad \textbf{Var} = s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

  Standard deviation

What is the unit of the standard deviation?

$$\textbf{Population standard deviation} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

$$\textbf{Sample standard deviation} \quad \bar{x} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Standard deviation is interpreted as the average distance of the data from the mean. It has the same unit as the original data.

Calculate the standard deviation of the *FINAL* in table 1.1 by calculator and interpret it.

What percentiles are Q_1, median and Q_3?

- **Location**

Percentile

What percentile is the *FINAL* of Vince? Interpret.

n^{th} percentile is the value with n percent of the data smaller or equal to it. For data $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, 1 is the 10^{th} percentile, 6 is the 60^{th} percentile.

Five number summary

There are five important locations for a given set of data, they are **min**, $\mathbf{Q_1}$, **median**, $\mathbf{Q_3}$ and **max**. They are called a **five number summary**. The following is a five number summary of the *FINAL* in table 1.1.

```
##    Min.     Q1.   Median     Q3.    Max.
##   42.00   75.50   83.50    90.00   99.00
```

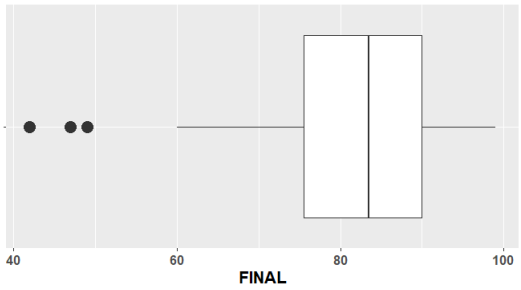
1.5 IQR rule

Find out the outliers of the *FINAL* by the 1.5 IQR rule.

This is rule of thumb to tell whether a value is an outlier or not.

$$x \notin [Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR] \implies \text{x is an outlier.}$$

Boxplot(box-whisker plot)



Can you tell the IQR of the *FINAL* according to figure 1.15?

Figure 1.15: The boxplot of the *FINAL*

In figure 1.15, the three dots are outliers, the left vertical line of the box indicates the value of the Q_1, the middle vertical line indicates the median and the right vertical line indicates the Q_3.

z-score

Calculate and interpret the z-score of the *FINAL* of Vince in table 1.1.

For a value x, its z-score is given by

$$z = \frac{x - \mu}{\sigma} \quad \text{or} \quad z = \frac{x - \bar{x}}{s_x}.$$

The z-score gives the distance from the mean in terms of standard deviation.

Exercise:

(1) Find the IQR, median and the skewness of the distribution in figure 1.16.

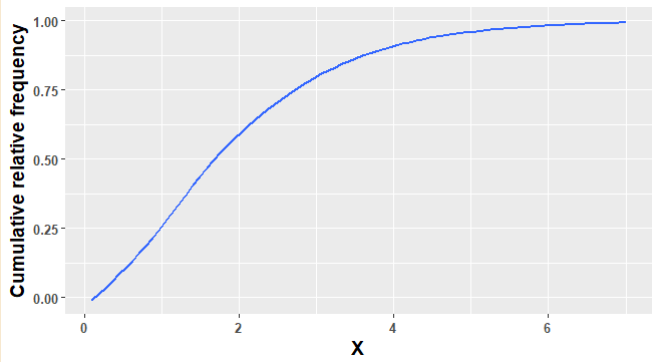


Figure 1.16: A cumulative relative frequency curve

(2) In order to select candidates to bid for a government contract on basis of the prices the submitted and rule out those who just mess around, you want to find a range in the form of $[\text{center} - \text{margin}, \text{center} + \text{margin}]$, where the margin is a constant. Any price falls within the range is qualified to be a candidate. Is this center better to be the mean or the median? Why?

(3) Which of the measurements of the spread are resistant?

1.5 Data transformation

Suppose we have a sample data set $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ with mean $\bar{x}$ and standard deviation s_x .

- Add a constant c to the data: $\mathbf{X} + c = \{x_1 + c, x_2 + c, \cdots, x_n + c\}$.

$$\begin{aligned}\text{The mean of } \mathbf{X} + c &= \frac{(x_1 + c) + (x_2 + c) + \cdots + (x_n + c)}{n} \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n} + c \\ &= \bar{x} + c.\end{aligned}$$

$$\begin{aligned}\text{The standard deviation of } \mathbf{X} + c &= \sqrt{\frac{\sum_{i=1}^{n}[(x_i + c) - (\bar{x} + c)]^2}{n}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \\ &= s_x.\end{aligned}$$

What about the **IQR**, **range** and **percentiles** if the data is added by a constant c.

- Multiply the data by constant a: $a\mathbf{X} = \{ax_1, ax_2, \cdots, ax_n\}$.

$$\begin{aligned}\text{The mean of } \mathbf{aX} &= \frac{(ax_1) + (ax_2) + \cdots + (ax_n)}{n} \\ &= a\frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= a\bar{x}.\end{aligned}$$

$$\begin{aligned}\text{The standard deviation of } a\mathbf{X} &= \sqrt{\frac{\sum_{i=1}^{n}(ax_i - a\bar{x})^2}{n}} \\ &= |a|\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \\ &= |a|s_x.\end{aligned}$$

Calculate the mean and standard deviation of the z-scores for any give data set.

The **z-score** is also called **standardized score** for the reason of its mean and standard deviation. **It gives us a reasonable way to compare values from different datasets as well.** For example, you get a 4 in **AP statistics** test and 100 in **TOEFL** test. Which score is better? You can compare the z-scores of those two.

Exercise:

The height distribution of a class in inches is given below in table 1.2.

Variable	n	$\bar{x}$	S_x	Min	Q_1	Med	Q_3	Max
Height	25	67	4.29	60	63	66	69	75

Table 1.2: Height distribution in inches

(1) Suppose you convert the class's height form inches to centimeters(1 inch = 2.54 cm). Describe the effect this will have on the shape of the distribution and the values of the variables in table 1.2?

(2) If all the students stand on a 6-inch-high platform and them measure the distance from the top of the heads to the ground, how would the shape compare with the original height distribution and what is the effect on the the values of the variables in table 1.2?

(3) Convert all the height to their z-scores, what would this effect the shape of the distribution and the values of the variables in table 1.2?

1.6 Normal distribution

The **normal distribution** is one of the most important distributions in statistics. A lot of natural phenomena follow the normal distribution, such as the error of repeated measurements. The **central limit theorem** makes it more powerful, which says: the sampling distribution of sample mean approaches normal distribution as the sample size increases. Normal distribution is also called **Gaussian distribution**.

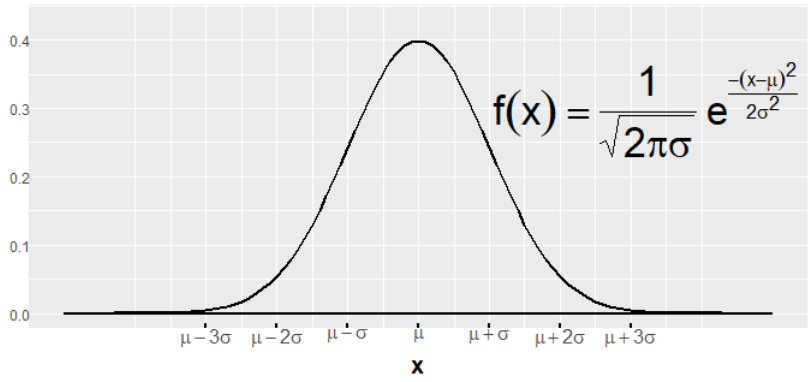


Figure 1.17: Normal Distribution with mean μ and standard deviation σ

- **Notation**

  Figure 1.17, gives the density curve of a normal distribution with mean μ and standard deviation σ, and the normal distribution is completely decided by those two parameters. $\mathcal{X}$ follows a normal distribution with mean μ and standard deviation σ is denoted as $\mathcal{X} \sim \mathcal{N}(\mu, \sigma)$.

- **The 68-95-99.7 rule**

  The density curve of the normal distribution is symmetric and bell-shaped, with mean equal to median. The total area under the curve and above the horizontal axis is 1. There is an empirical rule for the areas as shown in figure 1.18. It is called the **The 68-95-99.7 rule**.

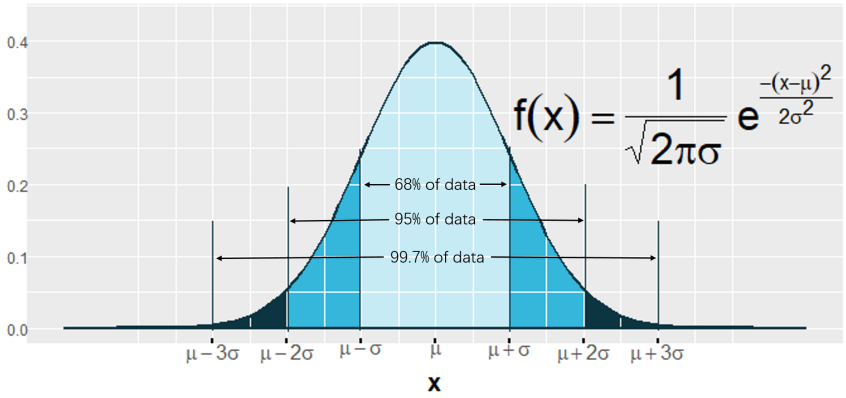


Figure 1.18: The 68-95-99.7 rule

We can estimate the parameters of μ and σ of the normal distributions by the 68-95-99.7 rule.

- **Standard normal distribution**

  Suppose $X \sim \mathcal{N}(\mu, \sigma)$, Let Z be the set of all z-scores of X

  $$Z = \frac{X - \mu}{\sigma}.$$

  Then Z is normally distributed, with mean μ_Z and standard deviation σ_Z. According to the knowledge of data transformation: $\mu_Z = 0$, $\sigma_Z = 1$. Thus,

  $$Z \sim \mathcal{N}(0, 1)$$

  $\mathcal{N}(0, 1)$ is called **standard normal distribution**.

  Suppose the percentage of the data less than x is give by $P(X < x)$, which equals to the area to the left side of x under the density curve. We have the following equations:

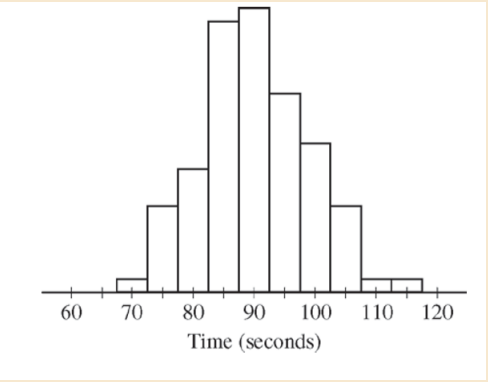
  $$P(X < x) = P(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}) = P(Z < z_x).$$

  If we known the standard normal distribution of Z, we know all the normal distributions.

Normal distribution is very important!!

Exercise

The amount of time required for each of 100 mice to navigate through a maze was recorded. The histogram below shows the distribution of times, in seconds, for the 100 mice.



Which of the following values is closest to the standard deviation of the 100 times?

(a) 2.5 seconds

(b) 10 seconds

(c) 20 seconds

(d) 50 seconds

(e) 90 seconds

Exercise:
At some fast-food restaurant, customers who want a lid for their drinks get them from a large stack left near straws, napkins, and condiments. The lids are made with small amount of flexibility so they can be stretched across the mouth of the cup and then snugly secured. When lids are too small or too large, customers can get very frustrated especially if the end up spilling their drinks. At one particular restaurant, large drink cups require lids with a "diameter" of between 3.95 and 4.05 inches. The restaurant's lid supplier claims that the diameter of their large lids follows a Normal distribution with mean 3.98 inches and standard deviation 0.02 inches. Assume that the supplier's claim is true.

(a) Find the percentage the large lids are too small to fit.

(b) Find the percentage the large lids are too big to fit.

The supplier is considering two changes to reduce the percent of its large-cup lids that are too small to 1%. One strategy is to adjust the mean diameter. The other is to alter the production process, thereby decreasing the standard deviation of the lid diameters.

(c) If the standard deviation remains at $\sigma = 0.02$ inches, at what value should the supplier set the mean diameter of its large-cup lids so that only 1% are too small to fit?

(d) If the mean diameter stays at $\mu = 3.98$ inches, what value of the standard deviation will result in only 1% of lids that are too small to fit?

(e) Which of the two options in (c) and (d) do you think is preferable?

2011FR1

A professional sports team evaluates potential players for a certain position based on two main characteristics, speed and strength.

(a) Speed is measured by the time required to run a distance of 40 yards, with smaller times indicating more desirable (faster) speeds. From previous speed data for all players in this position, the times to run 40 yards have a mean of 4.60 seconds and a standard deviation of 0.15 seconds, with a minimum time of 4.40 seconds, as shown in the table below.

	Mean	Standard Deviation	Minium
Time to run 40 yards	4.60 seconds	0.15 seconds	4.40 seconds

Based on the relationship between the mean, standard deviation, and minimum time, is it reasonable to believe that the distribution of 40-yard running times is approximately normal? Explain.

(b) Strength is measured by the amount of weight lifted, with more weight indicating more desirable (greater) strength. From previous strength data for all players in this position, the amount of weight lifted has a mean of 310 pounds and a standard deviation of 25 pounds, as shown in the table below.

	Mean	Standard Deviation
Amount of weight lifted	310 pounds	25 pounds

Calculate and interpret the z-score for a player in this position who can lift a weight of 370 pounds.

(c) The characteristics of speed and strength are considered to be of equal importance to the team in selecting a player for the position. Based on the information about the means and standard deviations of the speed and strength data for all players and the measurements listed in the table below for Players A and B, which player should the team select if the team can only select one of the two players? Justify your answer

	Player A	Player B
Time to run 40 yards	4.42 seconds	4.57 seconds
Amount of weight lifted	370 pounds	375 pounds

1.7 Comparing distributions

- **Perspectives to describe a distribution**

  When you are asked to describe the distribution give by a graph, you are supposed to describe the **shape, center** and **spread**.

  For example, by referring to figure 1.15, we say the distribution of the *FINAL* is roughly symmetric, with median around 84, and IQR about 16, and there are 3 outliers.

- **Graphs for comparing distributions**

  Many graphs can be drawn to compare the distributions. Figure 1.19 and figure 1.20 are two of them.

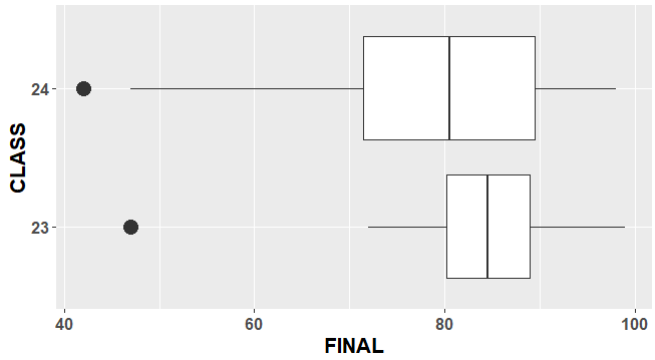


Figure 1.19: Distributions of the *FINAL* of *CLASS* 23 and *CLASS* 24

```
-------------------------------------
  1 | 2: represents 12, leaf unit: 1
       Class23       Class24
-------------------------------------
                |  4  |2
              7 |  4  |79
                |  5  |
                |  5  |
                |  6  |0
                |  6  |78
            422 |  7  |134
              8 |  7  |6778
       43222100 |  8  |34
         665555 |  8  |5788
          11100 |  9  |00123
             98 |  9  |78
                | 10  |
-------------------------------------
n:           26       26
-------------------------------------
```

Figure 1.20: Back-to-back stem plot with splitting stems

- **How to Compare two distributions?**

  The distributions are compared through **shape, center** and **spread**, and use proper terms for each perspective.

  - For shape, we use terms: **right-skewed, symmetric, left-skewed, outliers, gaps clusters**.
  - For center, we use terms: **mean** and **median**.
  - For spread, we use terms: **range, IQR** and **standard deviation**.

Distribution comparison or distribution description problems always show up in AP exam!!

The "read the graph ans say something" problem

Compare the distributions of *FINAL* between different *CLASS* by referring to figure 1.19.