Chapter 1

Descriptive statistics

Statistics is a science of data. To study statistics, we have to describe data from some proper perspectives. Generally speaking, there are two ways to describe data, graphical description and numerical description. Those are what we are going to learn in this section.

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.

- 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. Is variable *CLASS* a quantitative or categorical variable?

1.2 Some basic graphs

- **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.

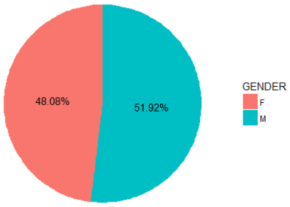


Figure 1.1: Pie chart of the distribution of the *GENDER*

- **Bar graph**

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

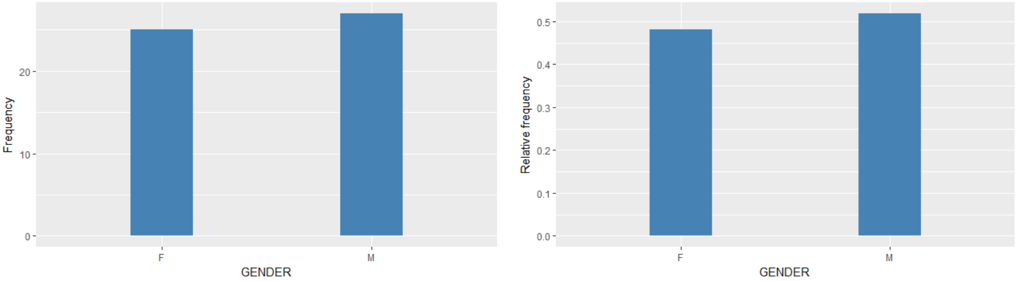


Figure 1.2: Bar graphs with percentage and frequency as vertical axis

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!

- **Dotplot**

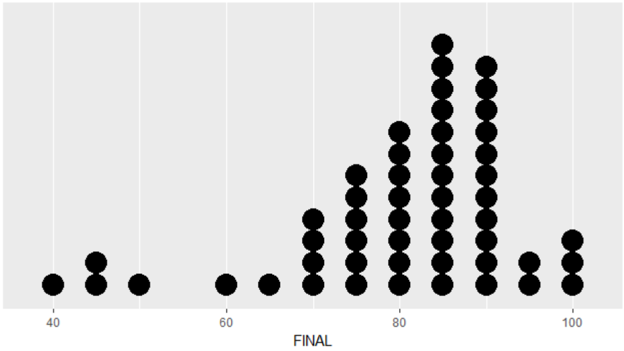


Figure 1.3: Dotplot of the distribution of the *FINAL*

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 $(77.5, 82.5]$.

- **Histogram**

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

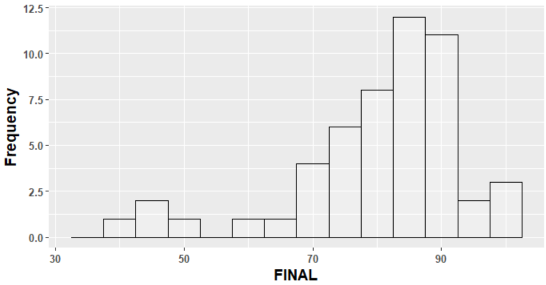


Figure 1.4: Histogram of the distribution of the *FINAL*

The vertical axis can be relative frequency as well.

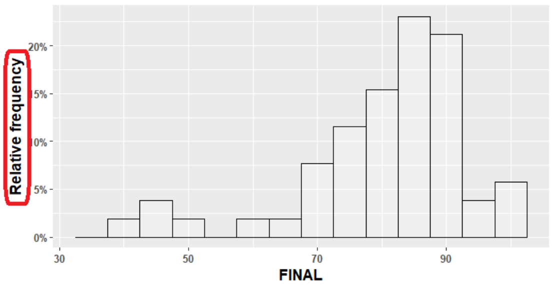


Figure 1.5: Histogram of with vertical axis **relative frequency**

What is the difference between histogram and bar graph?

- **Density curve**

  In figure 1.5, 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 $FINAL \leq 50$. They are 42, 47, 47, 49.)

  Now, if we draw a histogram with bin width 1(figure 1.6), then the percentage a bar can be calculated by

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

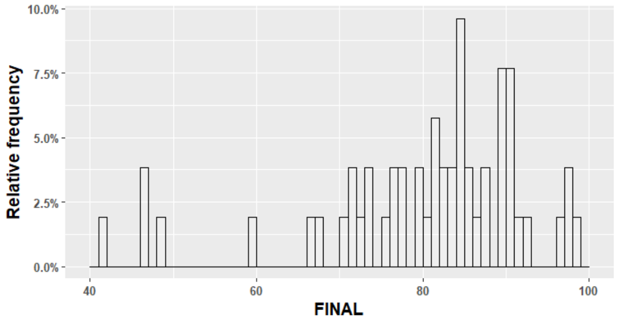


Figure 1.6: Histogram with bin width 1

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

What is the total area of all bars in the histogram?

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

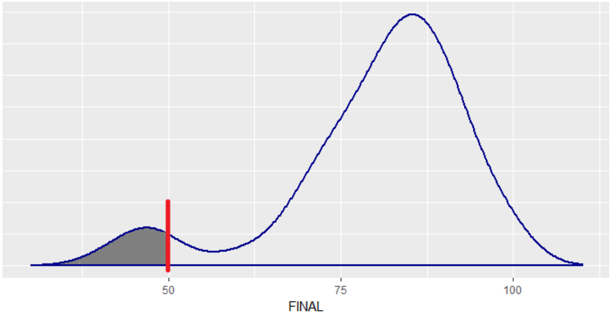


Figure 1.7: Density curve of the distribution of the *FINAL*

Figure 1.7 is a density curve of the districution of the *FINAL*, the shaded area gives the percentage of $FINAL \leq 50$, which is approximately 8%.

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

What is the total area under the density curve?

- **Cumulative relative frequency curve**

  In figure 1.7, 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 1.8. This curve is called the **cumulative relative frequency curve.** Function $F(x)$ is called **cumulative density function(cdf)**.

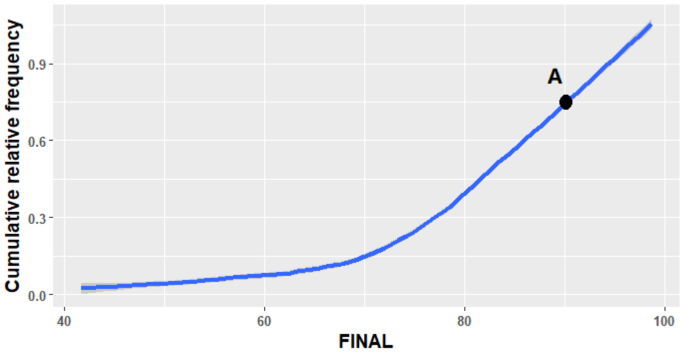


Figure 1.8: Cumulative relative frequency curve of *FINAL*

How to interpret point **A** in figure 1.8?

What is the theoretical relation between figure 1.8 and figure 1.7?

1.3 Some terms to describe graphs

- **Shape**

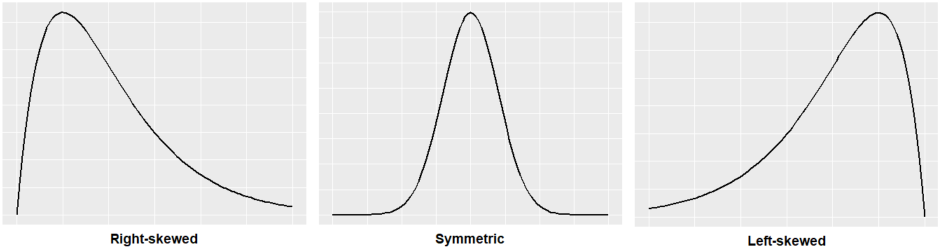


Figure 1.9: Shapes of distributions

As shown in figure 1.9, the shapes of the distributions are **right-skewed, symmetric** and **left-skewed** respectively.

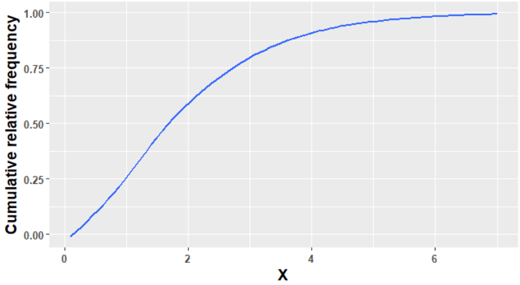


Figure 1.10: A cumulative relative frequency curve

Can you tell whether the distribution in figure 1.10 is right-skewed, left-skewed or symmetric?

- **Clusters and gaps**

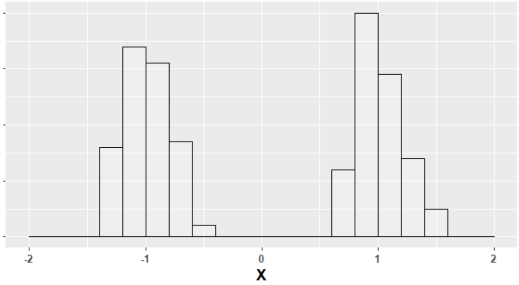


Figure 1.11: Two clusters and a gap

As show in figure 1.11, we say the distribution has two **clusters(modes)** with a **gap**.

- **Outliers**

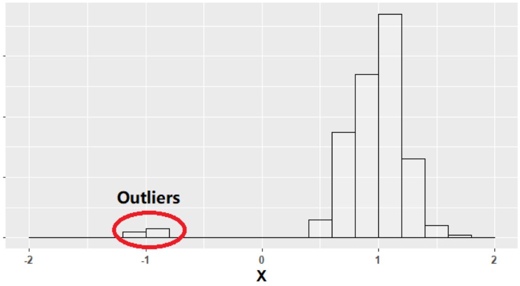


Figure 1.12: Outliers

If some values have striking departures from the pattern of the majority, those values are called **outliers**.

Outliers need special attention, for they may be generated by mistakes or some other mechanisms that are not considered.

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

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

Sometime the notation of the mean is μ. μ is used for population mean, $\bar{x}$ is for sample mean. For example, we draw a sample of 100 students from a high school, and the mean weight of those 100 students is 60kg. We use the notation $\bar{x}$, because the mean weight 60kg comes from the sample of 100 students. If the mean weight of all the students in this high school is 60kg, we use the notation μ. The formulas for μ and $\bar{x}$ are the same.

When to use mean and when to use median? Let's take a look at a simple example. Say, we have a set of data $\{1, 2, 3, 4, 5\}$. Both the mean and the median are 3. If 5 is recorded as 500 by accident. Now, the mean is 102, while the median is still the same. That is to say, the median is not easily influenced by the extreme values. If a value is not easily influenced by extreme values, it is **resistant**.

The relationship of the mean and median can be shown by figure 1.13.

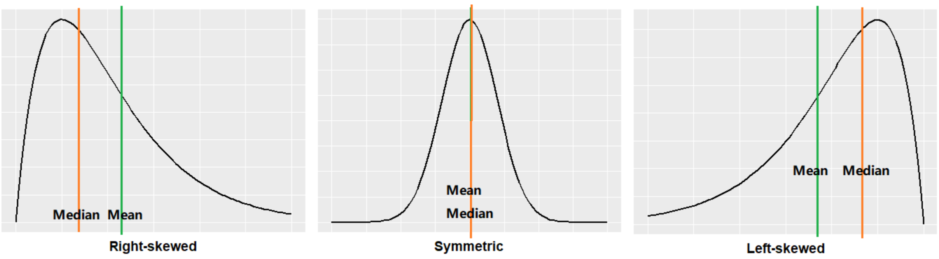


Figure 1.13: The relationship between mean and median

Generally speaking, if the distribution is symmetric, **mean** $=$ **median**; if the distribution is right-skewed, **mean** $>$ **median**; if the distribution is left-skewed, **mean** $<$ **median**. A simple explanation comes as: if the distribution is right-skewed, there are more extreme values to the right side. While the mean is not so resistant as median, it can be more easily dragged to the right than the median. Thus **mean** $>$ **median**.

Therefore, if the distribution is strongly skewed, it is better to use the **median** to describe the center of the distribution.

Is mean or median a better description of the center of the distribution of personal incomes?

- **Spread**

  Spread is to measure the variability or the dispersion of the data.

  - **Range**

    The difference between the maximun and the minimum,

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

  - **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.

    Interquartile range is given by

    $$\mathbf{IQR} = \mathbf{Q_3} - \mathbf{Q_1}.$$

    For data $\{1, 2, 3, 4, 5, 6, 7, 8\}$, the interquartile range **IQR** $= 6 - 2 = 4$

  - **Variance(Var)**

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

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

Variance gives the average square of the difference between the mean and data.

- **Standard deviation**

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

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

Standard deviation gives the average distance of the data from the mean.

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

* Calculate the mean and standard deviation of the *FINAL* of students in class 23 by calculator.

- **Location**

- **Percentile**

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. What percentiles are Q_1, median and Q_3?

* **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*.

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

* **1.5 IQR rule** gives a simple rule to tell whether a value is an outlier or not.

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

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

- **Boxplot**

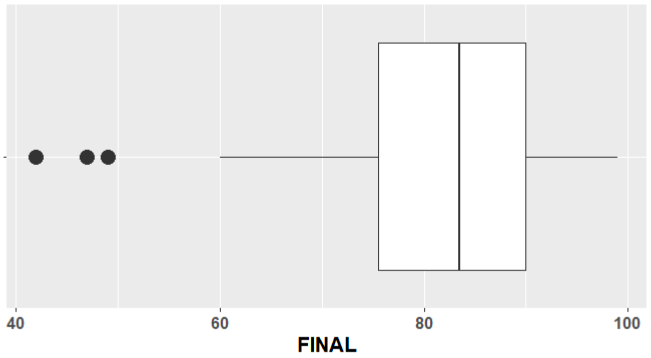


Figure 1.14: The boxplot of the *FINAL*

In figure 1.14, 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**

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

$$z = \frac{x-\mu}{\sigma}.$$

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

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

Calculate the percentile of the *FINAL* of Vince.

Of the measurements of the spread, which are more resistant and which are not?

- **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

$$\mathbf{X} + c = \{x_1 + c, x_2 + c, \cdots, x_n + c\}.$$

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

The mean is added by the same constant c.

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

The standard deviation doesn't change.

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

- Multiply the data by constant a

$$a\mathbf{X} = \{ax_1, ax_2, \cdots, ax_n\}.$$

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

The mean is multiplied by the same constant a.

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

The standard deivation is multiplied by the absolute value of constant a.

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

1.5 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**.

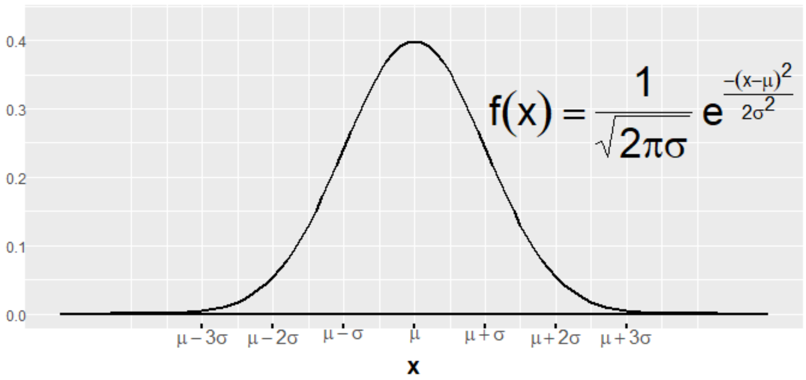


Figure 1.15: Normal Distribution with mean μ and standard deviation σ

- **Notation**

  Figure 1.15, 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 1.16. It is called the **The 68-95-99.7 rule**.

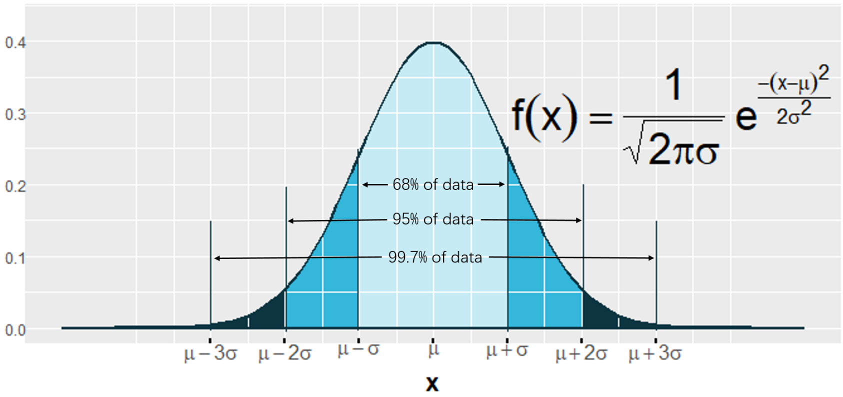


Figure 1.16: The 68-95-99.7 rule

By this rule, we can give some estimations about the parameters μ and σ of the normal distributions.

exercise .

- **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.

exercise .

1.6 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**. If there are **outliers, more than one clusters** and **gaps**, you are suppose to enunciate them.

  For example, by referring to figure 1.14, 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**

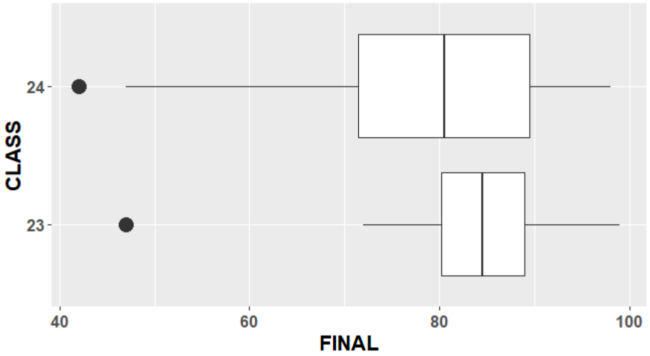
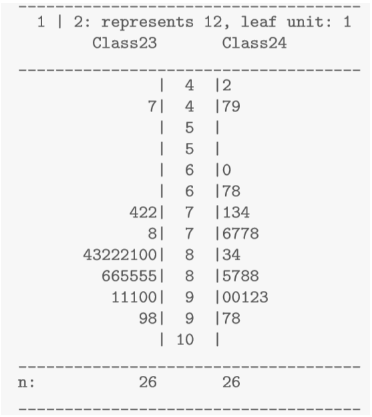


Figure 1.17: Distributions of the *FINAL* of *CLASS* 23 and *CLASS* 24



```
 -------------------------------------
   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 1.18: Back-to-back stem plot with splitting stems

There are many graphs to compare distributions. Figure 1.17 and figure 1.18 are about the distributions of the *FINAL* of *CLASS* 23 and *CLASS* 24.

- **Compare two distributions**

  The distributions are compared through the same perspectives as when we describe the distributions. They are **shape, center** and **spread**, and **outliers, clusters** and **gaps** if necessary.

  The terms to describe the shape are **right-skewed, symmetric** and **left-skewed**.

The terms to describe center are **mean** and **median**. Choose a appropriate one.

The terms to describe the spread are **range, IQR** and **standard deviation**. Choose a appropriate one.

For example, if figure 1.17 is given. We can say:

Both of the distributions are approximately symmetric.

The median of the *FINAL* of *CLASS* 24 is approximately 80, less than the median of the *FINAL* of *CLASS* 23, which is approximately 85.

Then IQR of *CLASS* 24 is approximately 20, larger than the IQR of *CLASS* 23, which is about 10. Thus the *FINAL* of *CLASS* 24 is more widely spread than that of *CLASS* 23 in the angle of IQR.

Both of the distributions have an outlier to the lower end.

Distribution comparison or distribution description problem always show up in AP exam.

1.7 The relation between two variables

- **The relation between two categorical variables**

By reading table 1.1, we may suspect that the is a relation between categorical variable *BASKETALL* and *GENDER*. Maybe boys are more likely to play basketball than girls. How can we describe the relation between *BASKETALL* and *GENDER*?

- **Two-way table**

		BASKETBALL		Total
		Yes	No	
GENDER	Male	21	6	27
	Female	5	20	25
Total		26	26	52

Table 1.2: Two-way table of Gender × Basketball

Table 1.2 gives a simple description about the relation between *BASKETALL* and *GENDER*.

∗ **Conditional distribution**

The **conditional distribution** of a categorical variable is defined as the distribution of this variable while the value of the other variable is fixed.

BASKETBALL	Percentage of Male	Percentage of Female
Yes	$\frac{21}{26} \approx 81\%$	$\frac{5}{26} \approx 19\%$
No	$\frac{6}{26} \approx 23\%$	$\frac{20}{26} \approx 77\%$

Table 1.3: Conditional distribution

Table 1.3 gives the conditional distribution of *GENDER* conditioned on different different values of *BASKETBALL*. For example, the conditional distribution of *GENDER* among those who play basketball is that: about 81% of them are boys, 19% are girls.

If a student plays basketball, this students is more likely to be a boy than a girl. Clearly, there is some association between *GENDER* and *BASKETBALL*.

- **Association**

If the conditional distributions of a variable are different while conditioned on different values of the other variable, we say there is an **association** between those variables. Otherwise, they are **independent**.

The conditional distributions of *GENDER* are different conditioned on different values(Yes, No) of *BASKETBALL*. Therefore, there is an association between *GENDER* and *BASKETBALL*.

- **Marginal distribution**

If we consider the distribution of *GENDER* regardless of the *BASKETBALL*, we just look at the data at the right margin of tabel 1.2.

$$\text{Percentage of girls: } \frac{25}{52} = 48\%, \qquad \text{Percentage of boys: } \frac{27}{52} = 52\%$$

Similarly, we can find out the distribution of *BASKETBALL* regardless of the *GENDER* by looking at the data at the bottom margin of tabel 1.2.

All the distributions are **marginal distribuions**, for we only consider the data at the margin of the two-way table.

Is it true that if two variables have no association(independent), the marginal distributions and the conditional distributions are the same?

– **Side-by-side bar graph**

Side-by-side bar graph is a graph to show the relation between two categorical variables.

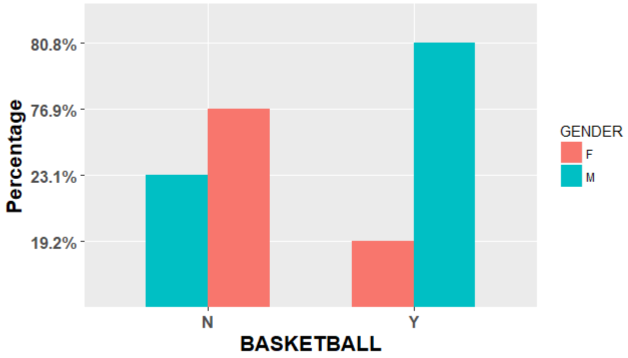


Figure 1.19: Side-by-side bar graph of *GENDER* and *BASKETBALL*

– **Stacked bar graph**

Stacked bar graph can play the same role as side-by-side bar graph.

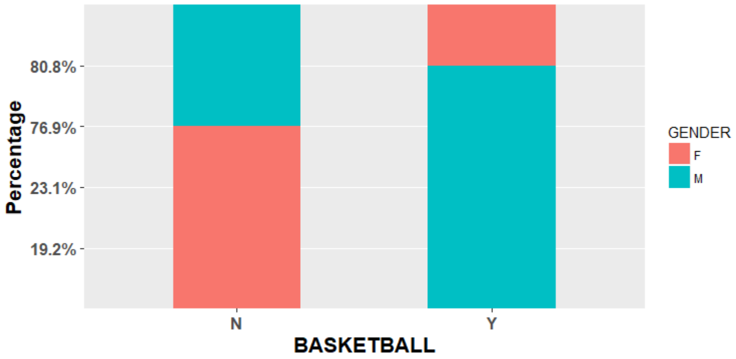


Figure 1.20: Stacked bar graph

- **The relation between two quantitative variables**

For convenience, we will use data in table 1.4.

Table 1.4: Small data set

Name	Calculus	Physics
James	75	74
Sam	87	83
Crystal	88	92
Evelyne	84	82
Phoebe	89	86
Vince	76	85
Mike	71	76
Lucy	81	89
Kitty	86	91
Owen	88	82
Angela	96	93
Christina	87	81
Jamie	84	80
Meggie	80	82
Kevin	97	99
Tom	90	86
John	81	88
Jason	72	80

From the table we may have a feeling about the "positive association" between *Calculus* and *Physics*. But how to describe this relationship?

- **Scatter plot**

  For each student we can get a two dimensional coordinates $(Calculus, Physics)$. For example, the coordinates of 'James' is $(75, 74)$. If we draw all those coordinates out in a coordinates system, the graph is called **scatter plot**, as shown in figure 1.21.

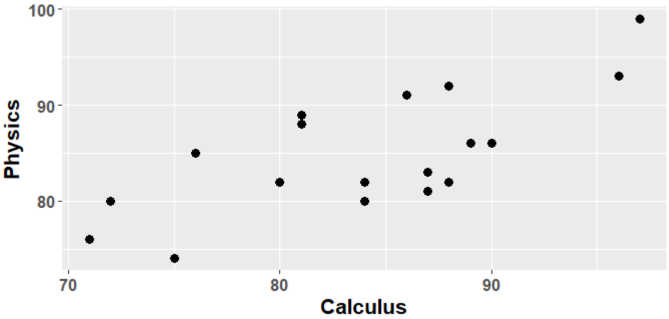


Figure 1.21: Scatter plot of *FINAL* **vs** *MID*

* **Explanatory variable, Response variable**

  If we want to use *MID* to explain *FINAL*, then *MID* is called **explanatory variable** and *FINAL* is called **response variable**. The explanatory variable goes to the x-axis and the response variable goes to the y-axis in the scatter plot.

- ### Direction

  The direction of a scatter plot gives a general trend, and can be described by terms **positive association** or **negative association**. If one variable increases while the other increases, those two variables have a positive association. If one variable increases while the other decreases, they have a negative association. *FINAL* and *MID* have a positive association according to figure 1.21.

- ### Form

  We describe the form of a scatter plot by terms **linear** or **curved**. If the points in the scatter plot forms a linear pattern, the scatter plot has a linear form, otherwise a curved form. The scatter plot in figure 1.21 has a linear form.

- ### Strength

  Strength describes how strong is the form. Terms used to describe the strength are: **strong, moderate, weak**. The scatter plot in figure 1.21 has a strong linear relation.

- ### Outliers

  Outliers are points don't follow the pattern of the majority either in the **x** direction or in the **y** direction. We modify figure 1.21 a little bit as shown by figure 1.22. Point **A** and **B** are outliers in the **x** direction and the **y** direction respectively.

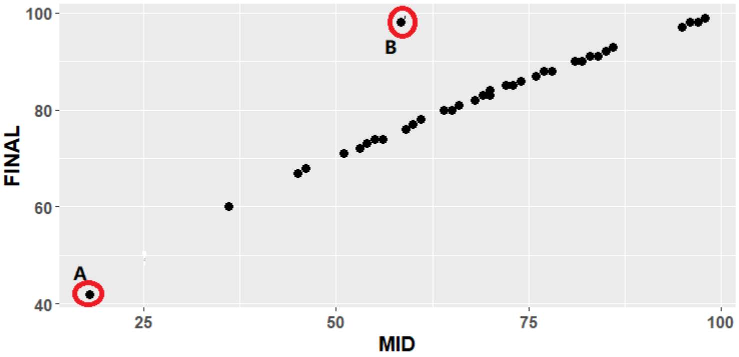


Figure 1.22: Outliers of a scatter plot

– Correlation r

Correlation r measures the strength and the direction of the linear relationship.

Suppose $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n\}$ are the data points, then

$$
\begin{aligned}
r &= \frac{1}{n-1}[(\frac{x_1-\bar{x}}{s_x}\frac{y_1-\bar{y}}{s_y})+(\frac{x_2-\bar{x}}{s_x}\frac{y_2-\bar{y}}{s_y})+\cdots+(\frac{x_n-\bar{x}}{s_x}\frac{y_n-\bar{y}}{s_y})] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}
\end{aligned}
$$

- $\mathbf{r} \in [\mathbf{-1},\mathbf{1}]$ $r > 0$ means there is a positive association. $r < 0$ means there

  is a negative association.

- **Correlation r is the measurement of the strength of the linear relationship.**

  If there is a linear relationship, $|r|$ is close to 1, means the linear relationship is strong. " $|r|$ is close to 1" itself can not guarantee there is a linear relation. There may be a curved relation with $|r|$ close to 1.

  $r = 0$ means there is no linear relation, but there may be a curved relationship. If $r = \pm 1$, the scatter plot is strictly linear.

  Correlation r only works on condition that there is a linear relationship. How can we tell there is a linear relationship? We will learn latter.

- **Correlation r doesn't imply causation**

- **Correlation r only describes the relationship between two quantitative variables and has no unit.**

- **Least-squares regression**

Least-squares regression is a simple model of the relationship between two quantitative variables. It is simple, but it embodies the basic idea of modelling, which plays an important role in machine learning as well.

  - **Regression lines**

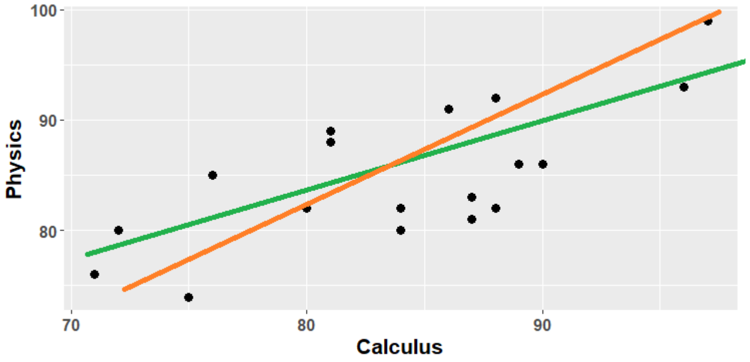


Figure 1.23: Linear regression lines

A regression line is a line drawn to model the scatter plot. There may be more than one regression lines as shown in figure 1.23. We need to set up a criteria and find the best one according to this criteria.

– Criteria

∗ Residuals

Suppose the equation for a regression line is

$$\hat{\mathbf{y}} = \mathbf{a} + \mathbf{b}\mathbf{x}.$$

$\hat{\mathbf{y}}$ represents the predicted value of *Physics* and $\mathbf{x}$ represents *Calculus*. Thus for each given value of *Calculus*, we can find out a predicted value of *Physics* by the equation of the regression line. Our prediction $\hat{\mathbf{y}}$ may be different from the observation $\mathbf{y}$. The difference is called **residual**.

$$\textbf{residual} = \mathbf{y} - \hat{\mathbf{y}}$$

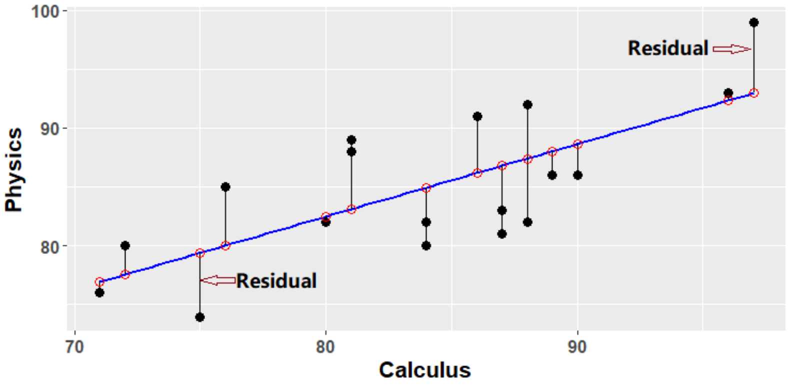


Figure 1.24: Visualizing residuals

A visualization of the residuals is shown in figure 1.24, with black dots observations and red circles predicted value, and the vertical lines connecting the red circles and the black dots represents residuals.

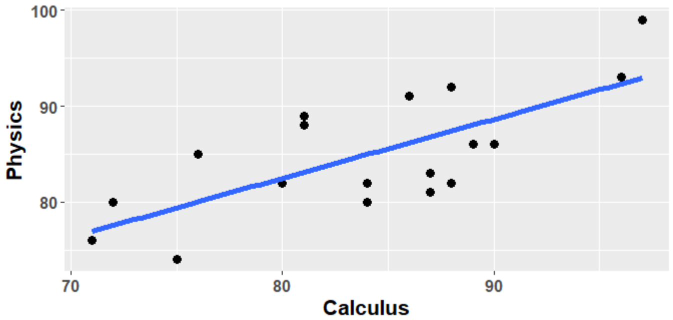
∗ **The criteria**

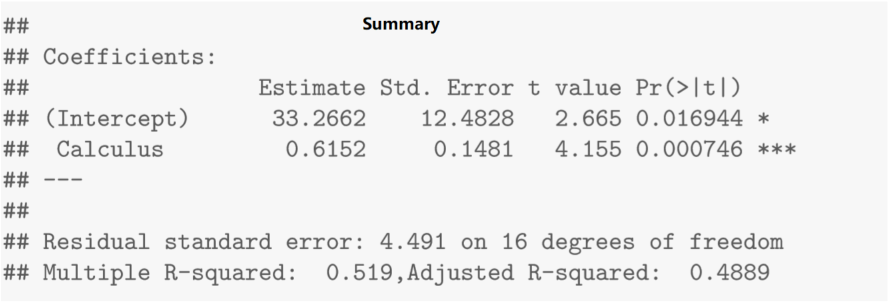
We want to find a line which minimizes the sum of the squares of the residuals:

$$\Sigma(\mathbf{y} - \hat{\mathbf{y}}^2).$$

The line minimizes $\Sigma(\mathbf{y} - \hat{\mathbf{y}}^2)$ is called **the least-squares regression line**

– **Least-squares regression line**





```
##                          Summary
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.2662    12.4828   2.665 0.016944 *
##  Calculus        0.6152     0.1481   4.155 0.000746 ***
## ---
##
## Residual standard error: 4.491 on 16 degrees of freedom
## Multiple R-squared:  0.519,Adjusted R-squared:  0.4889
```

Figure 1.25: Least-squares regression line

Figure 1.25 gives a graph and a summary of the least-squares regression line. Lets read the summary.

∗ **The equation $\hat{\mathbf{y}} = \mathbf{a} + \mathbf{bx}$**

In the summary, **a** is the estimated coefficient of the `Intercept`: `33.2622`, and **b** is the estimated coefficient of the `Calculus`: `0.6152`. Thus the equation of the least squares regression line is

$$\widehat{\textbf{Physics}} = 33.262 + 0.6152 \times \textbf{Calculus}$$

The **intercept a** is interpreted as: when the scores of *Calculus* is 0, the predicted value of *Physics* is 33.3 points. The **slope b** is interpreted as: when the scores of the *Calculus* increases by 1 point, the scores of the *Physics* will increase by 0.62 points.

Some formulas:

$$\bar{y} = a + b\bar{x}$$

$$b = r\frac{s_y}{s_x}$$

$$\Sigma\textbf{residuals} = \Sigma(y - \hat{y})$$

Mathematical deduction for those formulas.
Find the predicted *Physics* and the residual of James

- **Standard deviation of the residuals: S**

$$\begin{aligned} S &= \Sigma(\textbf{residuals} - \textbf{mean of residuals})^2/(n-2) \\ &= \Sigma(y-\hat{y})^2/(n-2) \end{aligned}$$

S gives the typical error of the predictions.

- **Coefficient of determination: $\mathbf{r^2}$**

Suppose we know nothing about the relationship between *Calculus* and *Physics*. For a given value of *Calculus*, a reasonable prediction of the value of *Physics* is its mean, since it is the center of *Physics*. Now the variance of the predictions of all the students in table 1.4 is given by

$$\frac{\Sigma(y-\bar{y})^2}{n-1}.$$

Suppose we know the relation between *Calculus* and *Physics*, and the relation is given by the least-squares regression line. For a given value **x** of *Calculus*, our predicted value $\hat{y}$ will be given by $\hat{y} = a + bx$. Now the variance of the predictions of all the students in table 1.4 is given by

$$\frac{\Sigma(y-\hat{y})^2}{n-1} = \frac{\Sigma\text{residual}^2}{n-1}.$$

The percentage of the variance of the predictions explained by the least-squares regression line is

$$1 - \frac{\Sigma\text{residual}^2}{n-1} \Bigg/ \frac{\Sigma(y-\bar{y})^2}{n-1}$$

This value is **the coefficient of determination**. Thus

$$\mathbf{r^2} = 1 - \frac{\Sigma\text{residual}^2}{\Sigma(y-\bar{y})^2}.$$

$\mathbf{r^2}$ gives the percentage of the the variance explained by the least-squares regression line.

- **Residual plot**

Residual plot is a scatter plot to show the distributions of the residuals.

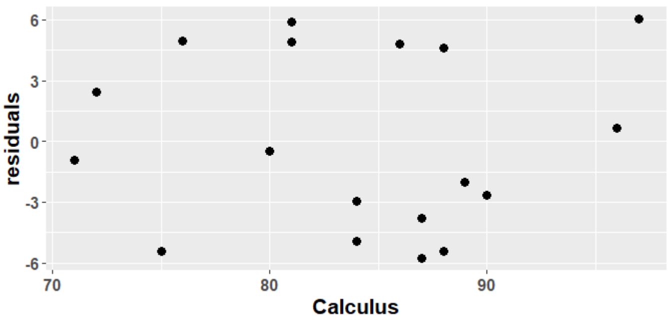


Figure 1.26: Residual plot

Residual plot in figure 1.26 is a scatter plot of residuals against the explanatory variable *Calculus*.

The residual plot gives us information about whether there is any "left-over" pattern after the response variable is explained by our model. If there is no left-over pattern in the residual plot and the residuals are randomly scattered around 0, our model is appropriate. In figure 1.26, there is no left-over pattern and the residuals is randomly scattered around 0. Thus the linear model is appropriate for the relationship between *Physics* and *Calculus*.

Residual plot gives us information about whether a model is appropriate.

- **Extrapolation and influential points**

Extrapolation is the use of the regression line to predict the response values based on the explanatory values far out of the range from which we obtain the regression line. For the linear regression in figure 1.25, if a give *Calculus* score is 0, the predicted value for *Physics* is 33.3.

We have to watch out for those extrapolations. Most of the time they are not accurate.

An **influential point** is the point that can markedly influence the result of the least-squares regression line if it is removed.

Outliers in the x direction are often influential.

- **Transform to achieve linearity**

Linear regression can describe some curved relationships after some transformation.

- **Exponential**

The exponential model is given by

$$\mathbf{y} = \mathbf{a}\mathbf{b}^{\mathbf{x}}, \quad \mathbf{a} \text{ and } \mathbf{b} \text{ are constant.}$$

Now we take natural log of both sides:

$$\ln \mathbf{y} = \ln(\mathbf{a}\mathbf{b}^{\mathbf{x}}) = \ln \mathbf{a} + (\ln \mathbf{b})\ \mathbf{x}.$$

Thus, $\ln y$ and x has a linear relationship, with intercept $\ln a$ and slope $\ln b$.

∗ **Power model**

The power model is given by

$$\mathbf{y} = \mathbf{a}\mathbf{x}^{\mathbf{b}}, \quad \mathbf{a} \text{ and } \mathbf{b} \text{ are constant.}$$

Now we take natural log of both sides:

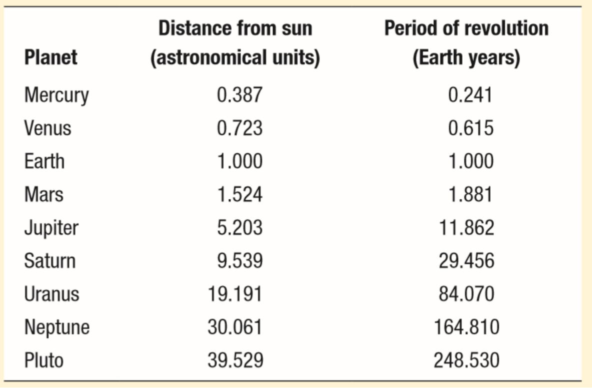
$$\ln \mathbf{y} = \ln(\mathbf{a}\mathbf{x}^{\mathbf{b}}) = \ln \mathbf{a} + \mathbf{b}\ \ \ln \mathbf{x}.$$

Thus $\ln y$ and $\ln x$ has a linear relationship, with intercept $\ln \mathbf{a}$ and slope $\mathbf{b}$.

Sometimes, we don't know whether it is a power model or an exponential model. We can draw two scatter plots to show the relationship of $\ln \mathbf{y} \sim \ln \mathbf{x}$ and $\ln \mathbf{y} \sim \mathbf{x}$, to see which one is more linear.

Example:

Here is the data for the nine planets in figure 1.27.



Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

Figure 1.27: Data for the nine planets (*Adapted from 'The practice of statistics'*)

(1) Is power model or exponential model more appropriate to for the relationship of **Period** $\sim$ **Distance**?

(2) If a new planet called Eris is discovered with distance 102.15 AU, what is the period?