Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 1 Spring 2012

Exploratory data analysis

In Probability, we start with a probability distribution for some random variable X, and deduce things like the expectation $\mathbb{E}(X)$ or $\mathbb{P}(X \geqslant 80)$.

In Statistics, we start with some data, and try to work out properties of the underlying probability distribution.

Before we do anything formal, we *picture* the data and *summarize* it: this is called *exploratory data analysis*.

Types of variable

There are two major types of variable.

1. *Quantitative* variables are measurements or counts. They may be (a) continuous or (b) discrete.

   (a) *Continuous* variables are measurements like height in cm, weight in kg, or temperature in °C.
   Their values are real numbers (rounded to the accuracy of measurement).
   There are few repeated values.
   The scale is called a *ratio* scale if 0 is meaningful but changing units by $x \mapsto cx$ makes no essential difference (for example, height in cm or in inches).
   The scale is called an *interval* scale if changing the units by $x \mapsto ax + b$ makes no essential difference (for example, temperature in °C or °F).
   For proportions, both 0 and 1 are meaningful.

   (b) *Discrete* quantitative variables are usually counts, such as number of seeds germinating, or number of faulty parts.
   Their values are usually non-negative integers.
   There may be many repeated values.

2. *Qualitative* variables (also called *factors*) (also sometimes called *categorical* variables, which is confusing—see below) classify objects into categories. They may be (a) categorical or (b) ordinal.

   (a) *Categorical* variables (also called *nominal* variables) have no sense of order in the categories (for example, favourite sport, or colour of eyes).

   (b) *Ordinal* qualitative variables have a natural order on the categories (for example, the catgories might be
   'inactive', 'slightly active' and 'very active',
   or
   'strongly disagree', 'slightly disagree', 'slightly agree' and 'strongly agree').

Notation

It is very common to use n to denote the number of observations. If you use n in this way, it is better to define it explicitly.

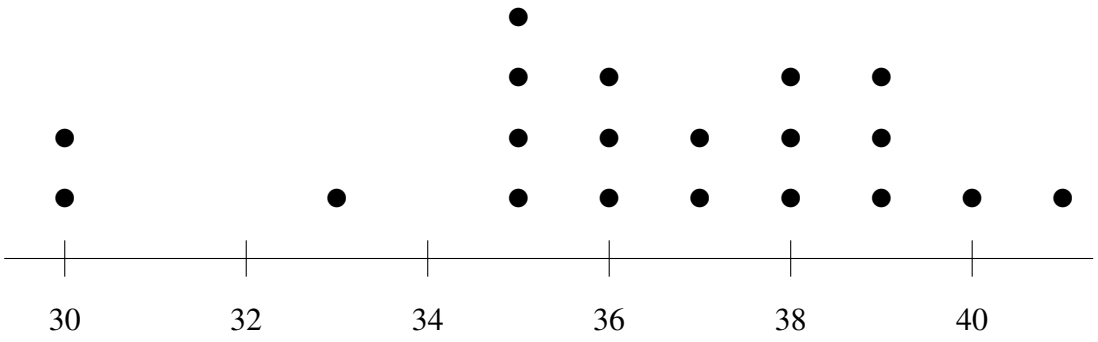
Dot plots

The simplest plot of quantitative data is the *dot plot*.

Draw a horizontal axis for the scale. For each observation, put a dot above the corresponding value. Stack dots for the same value up vertically, using a constant vertical spacing.

Example A driver noted the petrol consumption (in miles per gallon) of his VW car between successive fill-ups.

41, 35, 39, 40, 39, 39, 38, 36, 35, 38, 38, 37, 36, 30, 33, 35, 35, 30, 36, 37.



Dot plots are good for showing clusters, gaps and outliers, for up to about 25 data.

Dot plots can help us to read off the data in order. If the original data are called x_1, x_2, …, x_n, then we write

$$\begin{aligned} x_{(1)} &= \text{smallest value} \\ x_{(2)} &= \text{second smallest value} \\ &\quad \text{and so on } \ldots \\ x_{(n)} &= \text{largest value.} \end{aligned}$$

For the VW data, $x_{(1)} = 30$, $x_{(2)} = 30$ and $x_{(3)} = 33$.

Some measures of location and spread

One measure of location is the *median*. This is

—the middle value if n is odd;
—the average of the two middle values otherwise.

For the VW data, $n = 20$, so

$$\text{median} = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right) = \frac{1}{2}(36+37) = 36.5.$$

One measure of spread is the *range*. This is the highest value (maximum) minus the lowest value (minimum). This is very sensitive to extreme values or outliers.

For the VW data, range $= 41 - 30 = 11$.

The *quartiles* divide the data into four equal parts, just as the median divides them into two equal parts. The lower quartile, which is also called the first quartile and written Q_1, is defined by

$$\text{at least } 1/4 \text{ of values} \leqslant Q_1 \leqslant \text{at least } 3/4 \text{ of values.}$$

If n is not divisible by 4, then

$$Q_1 = \left\lceil \frac{n}{4} \right\rceil \text{-th value,}$$

where $\lceil x \rceil$ denotes the smallest integer z with $z \geqslant x$, which is called the *ceiling* of x. For example, if $n = 11$ then $\left\lceil \frac{n}{4} \right\rceil = \lceil 2.75 \rceil = 3$, so $Q_1 = x_{(3)}$.

$$\begin{array}{ccc} x_{(1)},\ x_{(2)},\ x_{(3)} & \leqslant Q_1 \leqslant & x_{(3)},\ x_{(4)}, \ldots, x_{(11)} \\ \text{3 values} & & \text{9 values} \\ 3 \geqslant \dfrac{11}{4} & & 9 \geqslant 11 \times \dfrac{3}{4} \end{array}$$

If n is divisible by 4, say $n = 4r$, then both $x_{(r)}$ and $x_{(r+1)}$ satsify our condition, so Q_1 could be anywhere between $x_{(r)}$ and $x_{(r+1)}$. The convention we shall use is that

$$Q_1 = \text{average of } x_{(r)} \text{ and } x_{(r+1)}.$$

The second quartile is just the median. It may be written as Q_2.

The upper quartile, which is also called the third quartile and written as Q_3, is defined by

$$\text{at least } 3/4 \text{ of values} \leqslant Q_3 \leqslant \text{at least } 1/4 \text{ of values.}$$

If n is not divisible by 4 then

$$Q_3 = \left\lceil \frac{3n}{4} \right\rceil \text{-th value.}$$

If $n = 4r$, then by convention we put

$$Q_3 = \text{average of } x_{(3r)} \text{ and } x_{(3r+1)}.$$

For the VW data,

$$Q_1 = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(35+35) = 35$$

and

$$Q_3 = \frac{1}{2}\left(x_{(15)} + x_{(16)}\right) = \frac{1}{2}(38+39) = 38.5.$$

The *interquartile range* is $Q_3 - Q_1$: this is another measure of spread. It is not sensitive to extreme values. Half of the observations lie in the interval $[Q_1, Q_3]$.

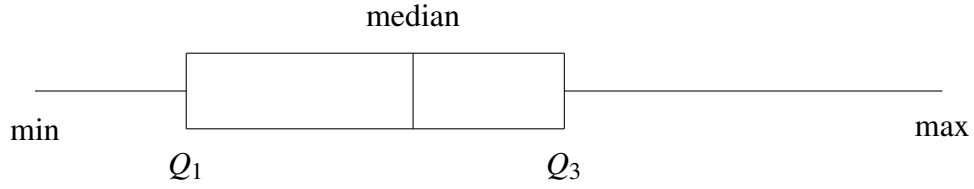
For the VW data, IQR $= 38.5 - 35 = 3.5$.

These five numbers—minimum, lower quartile, median, upper quartile, maximum—are called the *five-number summary*.

Boxplots

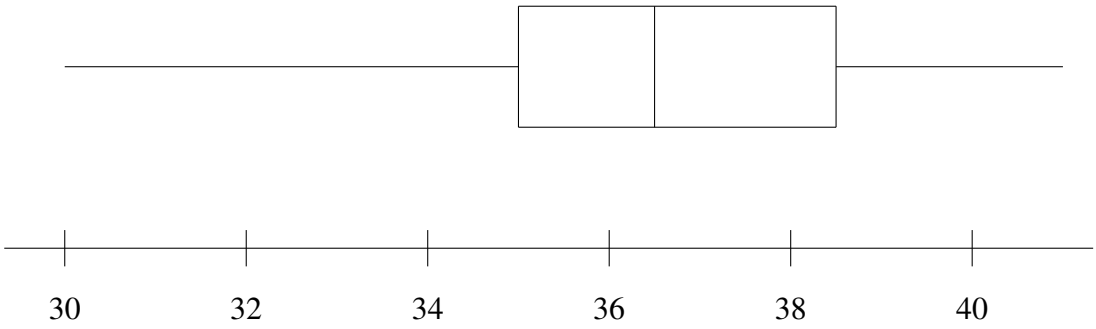
These five numbers are shown in a *boxplot*, also called a *box-and-whisker plot*. Here we shall show boxplots with the axis horizontal, but Minitab draws them with the axis vertical. The *box* is a rectangle positioned above the interval $[Q_1, Q_3]$. The median is shown by a vertical line for the full height of the box, positioned at Q_2.

The *whiskers* are horizontal lines outside the box, positioned half-way up it. The right-hand whisker extends from Q_3 to the maximum, while the left-hand whisker extends from Q_1 to the minimum.



Note that a boxplot always needs a scale; it must be drawn accurately to scale, and it must be drawn with a ruler.

Here is the box-and-whisker plot for the VW data.



In a more elaborate version of the boxplot, a data value is deemed extreme if it is more than $1.5 \times$ IQR outside the box. The whiskers exclude the extreme values, which are shown by stars or dots.

More measures of location and spread

Here are some summary measures that do not depend on putting the data in order.

The *sample mean* $\bar{x}$ is defined by

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

This is measure of location, because it indicates where the centre of the data is.

The *sample variance* s^2 is defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right].$$

Note the divisor $n-1$ carefully: this will be explained later.

The *sample standard deviation* s is defined to be the positive square root of s^2. The sample variance and the sample standard deviation are both measures of spread.

For the VW data, $\sum x_i = 727$ and so $\bar{x} = 36.35$. Moreover, $\sum x_i^2 = 26591$ and so $s^2 = 8.6605$ and $s = 2.94$.

What happens to the summary measures if the data are transformed?

Sometimes data may be transformed by a monotonic function before being analysed. For example, height in inches may be transformed into height in centimetres, temperature in degrees Fahrenheit may be converted into temperature in degrees Celsius, or petrol consumption may be converted from miles per gallon into litres per 100 kilometres. Positive data with a long right tail may be transformed by taking logarithms.

What effect do these transformations have on measures of location and spread?

Theorem 1 *Let $x_1, \ldots, x_n$ be data with sample mean $\bar{x}$, sample standard deviation s_x, minimum $x_{(1)}$, lower quartile Q_{1x}, median m_x, upper quartile Q_{3x} and maximum $x_{(n)}$. For $i = 1, \ldots, n$, put $y_i = ax_i + b$, where a and b are constants and $a \neq 0$. Let $y_1, \ldots, y_n$ have sample mean $\bar{y}$, sample standard deviation s_y, minimum $y_{(1)}$, lower quartile Q_{1y}, median m_y, upper quartile Q_{3y} and maximum $y_{(n)}$. Then*

(a) $\bar{y} = a\bar{x} + b$;

(b) $s_y = |a|\, s_x$;

(c) $m_y = am_x + b$;

(d) *if $a > 0$ then $y_{(1)} = ax_{(1)} + b$ and $y_{(n)} = ax_{(n)} + b$,*
but if $a < 0$ then $y_{(1)} = ax_{(n)} + b$ and $y_{(n)} = ax_{(1)} + b$;

(e) *if $a > 0$ then $Q_{1y} = aQ_{1x} + b$ and $Q_{3y} = aQ_{3x} + b$,*
but if $a < 0$ then $Q_{1y} = aQ_{3x} + b$ and $Q_{3y} = aQ_{1x} + b$.

Proof (a) $\sum_{i=1}^n y_i = \sum_{i=1}^n (ax_i + b) = \sum_{i=1}^n ax_i + \sum_{i=1}^n b = a\sum_{i=1}^n x_i + nb = an\bar{x} + nb$, and so

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{an\bar{x} + nb}{n} = a\bar{x} + b.$$

(b) For $i = 1, \ldots, n$, we have $y_i - \bar{y} = (ax_i + b) - (a\bar{x} + b) = a(x_i - \bar{x})$, using the result (a). Hence

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^n a^2(x_i - \bar{x})^2 = a^2 s_x^2.$$

Taking the positive square root of both sides gives $s_y = |a|\, s_x$.

(c), (d) and (e) We have $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$. If $a > 0$ then $ax_{(1)} \leqslant ax_{(2)} \leqslant \cdots \leqslant ax_{(n)}$, so $ax_{(1)} + b \leqslant ax_{(2)} + b \leqslant \cdots \leqslant ax_{(n)} + b$, and so $y_{(i)} = ax_{(i)} + b$ for $i = 1, \ldots, n$.

On the other hand, if $a < 0$ then $ax_{(1)} \geqslant ax_{(2)} \geqslant \cdots \geqslant ax_{(n)}$, so $ax_{(1)} + b \geqslant ax_{(2)} + b \geqslant \cdots \geqslant ax_{(n)} + b$, and so the y data have the opposite order from the x data, so that $y_{(i)} = ax_{(n+1-i)} + b$ for $i = 1, \ldots, n$.

This proves (d).

If $n = 2r + 1$ then $m_x = x_{(r+1)}$, $m_y = y_{(r+1)}$ and $n + 1 - (r + 1) = r + 1$. If $a > 0$ then $m_y = y_{(r+1)} = ax_{(r+1)} + b = am_x + b$. If $a < 0$ then $m_y = y_{(r+1)} = ax_{(n+1-r-1)} + b = ax_{(r+1)} + b = am_x + b$.

If $n = 2r$ then m_x is the average of $x_{(r)}$ and $x_{(r+1)}$, while m_y is the average of $y_{(r)}$ and $y_{(r+1)}$. If $a > 0$ then $y_{(r)} = ax_{(r)} + b$ and $y_{(r+1)} = ax_{(r+1)} + b$, while if $a < 0$ then $y_{(r)} = ax_{(r+1)} + b$ and $y_{(r+1)} = ax_{(r)} + b$. In both cases,

$$m_y = \frac{y_{(r)} + y_{(r+1)}}{2} = \frac{ax_{(r)} + b + ax_{(r+1)} + b}{2} = am_x + b.$$

This proves (c).

The proof of (e) is similar. ■

For other monotonic transformations, the sample mean and sample standard deviation do not transform in this predictable way. For example, let $n = 9$, $x_i = 1$ for $i = 1, \ldots, 8$ and $x_9 = 2$. Put $z_i = 2/x_i$ for $i = 1, \ldots, 9$. Then $\bar{x} = 10/9$ and $\bar{z} = 17/9 \neq 2/\bar{x}$.

However, the five-number summary behaves in a more predictable way. Suppose that $y_i = f(x_i)$ for $i = 1, \ldots, n$. If f is monotonic increasing then the y_i have the same order as the x_i; if f is monotonic decreasing then they have the opposite order. Thus the minimum and maximum of the x_i are transformed into the minimum and maximum of the y_i, possibly in the opposite order. If n is odd then m_x and m_y are both the middle value, so $m_y = f(m_x)$. If n is even then each median is the average of the two middle values, so m_y will be slightly different from $f(m_x)$ in general. Similar remarks apply to the upper and lower quartiles.