Mathematical Statistics

MAS 713

Tutorial about Chapter 1
SOLUTIONS

Exercise 1

Answer by yes or no, and **explain**.

1. Will the sample mean always correspond to one of the observations of the sample?
2. Will exactly half of the observations in a sample always fall below the mean?
3. Will the sample mean always be the most frequently occurring data value in the sample?
4. Can the sample standard deviation be equal to zero?
5. Can the sample median be equal to the sample mean?

Will the sample mean always correspond to one of the observations of the sample?

Will the sample mean always correspond to one of the observations of the sample?

No, the mean need not be an observed value. Consider the sample $\{0, 1\}$, its mean is 0.5.

Will exactly half of the observations in a sample always fall below the mean?

Will exactly half of the observations in a sample always fall below the mean?

No, that is the definition of the median.

Will the sample mean always be the most frequently occurring data value in the sample?

Will the sample mean always be the most frequently occurring data value in the sample?

No, as the mean need not even be an observed value.

Definition: The most frequently occurring data value in the sample is called the sample **mode**.

Can the sample standard deviation be equal to zero?

Can the sample standard deviation be equal to zero?

Yes, when all the observations are equal, there is no dispersion in the sample.

Can the sample median be equal to the sample mean?

Can the sample median be equal to the sample mean?

Yes,
but this only happens when
the distribution of the observed values is **exactly symmetric**.

Exercise 2

Answer by yes or no, and **explain**.

1. Suppose that you add +10 to all of the observations in a sample.
   How does this change the sample mean?
   How does it change the sample standard deviation?
2. Suppose that you multiply all of the observations in a sample by 2.
   How does this change the sample mean?
   How does it change the sample standard deviation?
3. A sample of temperature measurements in a furnace yielded a sample average of 446°Celsius and a sample standard deviation of 5.8°Celsius. You would like to communicate this information to an American colleague.
   What are the sample average and the sample standard deviation expressed in °Fahrenheit?
   (*Hint :* temperature in °C = (temperature in °Fahrenheit - 32)$\times 5/9$)

Suppose that you add +10 to all of the observations in a sample.
How does this change the sample mean?

Suppose that you add +10 to all of the observations in a sample.
How does this change the sample mean?

- Adding + 10 to all observations is like **shifting the whole sample by a distance of +10**.
- The new **mean** is also **shifted by +10**.

- However the dispersion in the sample is **not affected by this shift** so
- the **standard deviation is unchanged**.

Suppose that you multiply all of the observations in a sample by 2.
How does this change the sample mean?
How does it change the sample standard deviation?

Suppose that you multiply all of the observations in a sample by 2.
How does this change the sample mean?
How does it change the sample standard deviation?

Multiplying all the observations by 2 is like
shifting and stretching the sample.

new mean = initial mean multiplied by 2

new standard deviation = initial standard deviation multiplied by 2.

A sample of temperature measurements in a furnace yielded a sample average of 446°Celsius and a sample standard deviation of 5.8°Celsius. You would like to communicate this information to an American colleague.
What are the sample average and the sample standard deviation expressed in °Fahrenheit?
(*Hint :* temperature in °C = (temperature in °Fahrenheit - 32)$\times 5/9$)

A sample of temperature measurements in a furnace yielded a sample average of 446°Celsius and a sample standard deviation of 5.8°Celsius. You would like to communicate this information to an American colleague.
What are the sample average and the sample standard deviation expressed in °Fahrenheit?
(*Hint :* temperature in °C = (temperature in °Fahrenheit - 32)$\times 5/9$)

We have $\bar{x} = 446$ and $s_x = 5.8$ in °C.

Denote y the temperature in °F.
Similarly to above, we have

$$\bar{y} = \frac{9}{5} \times \bar{x} + 32 = 834.8$$

and

$$s_y = \frac{9}{5} s_x = 10.44,$$

both expressed in °F.

Exercise 3

An experiment to investigate the survival time (in hours) of an electronic component consists of placing the parts in a test cell and running them for 100 hours under elevated temperature conditions (this is called an 'accelerated life test'). Eight components were tested with the following resulting failure times :

$$75, 63, 100^{+}, 36, 51, 45, 80, 90$$

The observation 100^{+} indicates that the unit still functioned at 100 hours.

Is there any meaningful measure of location that can be calculated for these data?
What is its numerical value?

Answer:

Yes, the median is a meaningful location measure, as it is unaffected by extreme values.
So the unobserved value beyond 100 hours could be anything, the median would remain unchanged. Specifically, no matter the exact value of 100^+, the median is

$$m = \frac{1}{2}(63 + 75) = 69 \text{ hours},$$

as the sample size is even and 63 and 75 are the central values of the sample.

Exercise 4

Consider a sample of observations $x_1, x_2, \ldots, x_n$.

- For what value a is the quantity $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - a)^2$ minimised?
- Interpret in terms of location and dispersion parameters you know.

- Differentiate the function $f(a) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - a)^2$ with respect to a :

$$\frac{d}{da}f(a) = \frac{-2}{n-1}\sum_{i=1}^{n}(x_i - a)$$

Set this to 0 :

$$\sum_{i=1}^{n}(x_i - a) = 0$$

it follows

$$\sum_{i=1}^{n} x_i = na \qquad \text{or equivalently} \qquad a = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

(you can check that the second derivative of $f(a)$ is always positive)

- $\rightsquigarrow$ the sample mean is the value
  that minimises the sum of the squared deviations.

Exercise 5

The following data is a sample of shear strength, (MPa) of a joint bonded in a particular manner :

$$22.4, 40.4, 16.4, 73.7, 36.6, 109.9, 30.0, 4.4, 33.1, 66.7, 81.5$$

- Determine the 5-number summary.
- Determine the iqr. Are there any outliers (by the $1.5\times$ iqr rule)?
- Construct a box-plot and comment on its features.
- Determine the mean $\bar{x}$ and the sample standard deviation s.
- By how much could the largest observation be decreased without affecting the iqr?

- First order the observations :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

- The minimal value is 4.4, the maximal value is 109.9.
- There are $n = 11$ observations (odd number of observations), so the median is the $(n+1)/2 = 6$th largest observation, that is,

$$m = 36.6 \text{ (MPa)}$$

- This splits the sample in two equal parts :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$

and

$$36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

(include the median in each half).

- First order the observations :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

- The minimal value is 4.4, the maximal value is 109.9.
- There are $n = 11$ observations (odd number of observations), so the median is the $(n+1)/2 = 6$th largest observation, that is,

$$m = 36.6 \text{ (MPa)}$$

- This splits the sample in two equal parts :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$

and

$$36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

(include the median in each half).

- First order the observations :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$
$$< 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

- The minimal value is 4.4, the maximal value is 109.9.
- There are $n = 11$ observations (odd number of observations), so the median is the $(n+1)/2 = 6$th largest observation, that is,

$$m = 36.6 \text{ (MPa)}$$

- This splits the sample in two equal parts :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$

and

$$36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

(include the median in each half).

- First order the observations :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$
$$< 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

- The minimal value is 4.4, the maximal value is 109.9.
- There are $n = 11$ observations (odd number of observations), so the median is the $(n+1)/2 = 6$th largest observation, that is,

$$m = 36.6 \text{ (MPa)}$$

- This splits the sample in two equal parts :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$

and

$$36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

(include the median in each half).

- The first quartile is the median of the first half and the third quartile is the median of the second half.
- As these subsamples have even numbers of observations, their medians are

$$q_1 = \frac{22.4 + 30}{2} = 26.2 \text{ (MPa)}$$

and

$$q_3 = \frac{66.7 + 73.7}{2} = 70.2 \text{ (MPa)}$$

$\rightsquigarrow$ the 5-number summary of the sample is thus

$$\{4.4, 26.2, 36.6, 70.2, 109.9\}$$

- The first quartile is the median of the first half and the third quartile is the median of the second half.
- As these subsamples have even numbers of observations, their medians are

$$q_1 = \frac{22.4 + 30}{2} = 26.2 \text{ (MPa)}$$

and

$$q_3 = \frac{66.7 + 73.7}{2} = 70.2 \text{ (MPa)}$$

$\rightsquigarrow$ the 5-number summary of the sample is thus

$$\{4.4, 26.2, 36.6, 70.2, 109.9\}$$

- The interquartile range is

$$\text{iqr} = q_3 - q_1 = 70.2 - 26.2 = 44 \text{ (MPa)}$$

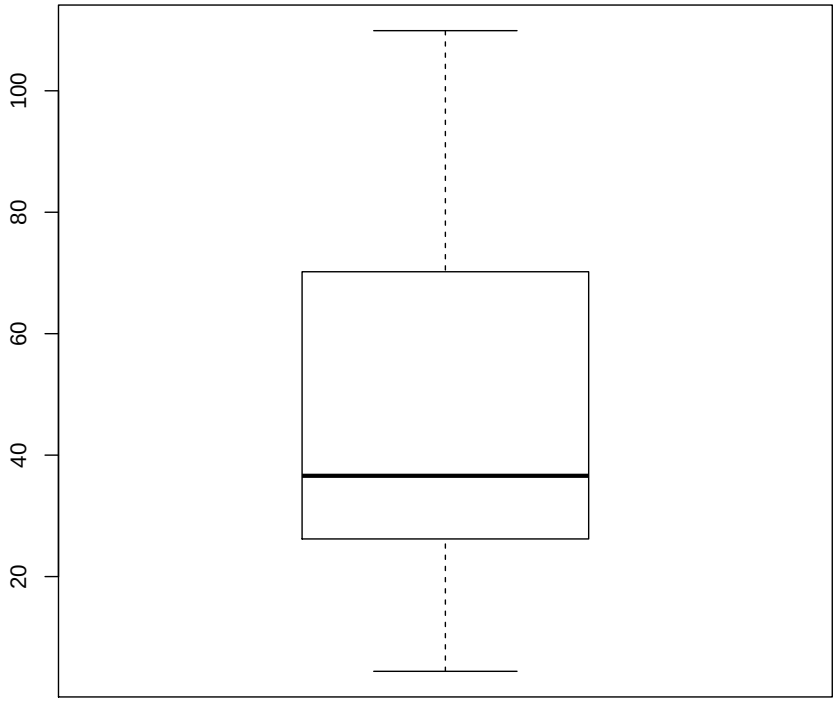
Thus, outliers would be the observations

smaller than $q_1 - 1.5 \times \text{iqr} = 26.2 - 1.5 \times 44 = -39.8$, or

larger than $q_3 + 1.5 \times \text{iqr} = 70.2 + 1.5 \times 44 = 136.2$.

$\rightsquigarrow$ there is **no outlier**.

- Boxplot:



There is a slight positive skew to the data. There is no outlier.

- The mean is

$$\bar{x} = \frac{1}{11}(4.4 + 16.4 + \ldots + 109.9) = 46.83 \text{ (MPa)},$$

the standard deviation is

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

$$= \sqrt{\frac{1}{10}\left((4.4^2 + 16.4^2 + \ldots + 109.9^2) - 11 \times 46.83^2\right)} = 31.996 \text{ (MPa)}$$

- If the data set remained the same except for the largest observation being decreased, the only way that the iqr would change would be for the largest value to become small enough to make q_3 smaller. To do this the value of 109.9 would have to be replaced by one less than 73.7.

Exercise 6

Direct evidence of Newton's universal law of gravitation was provided from a renowned experiment by Henry Cavendish (1731-1810). In the experiment, masses of objects were determined by weighting, and measured force of attraction was used to calculate the density of earth. The values of the earth's density estimated by Cavendish, expressed as a multiple of the density of water (1 g/cm^3), are :

5.50 5.30 5.47 5.10 5.29 5.65 5.55 5.61 5.75 5.63 5.27 5.44 5.57 5.36
4.88 5.86 5.34 5.39 5.34 5.53 5.29 4.07 5.85 5.46 5.42 5.79 5.62
5.58 5.26

(*Source :* Philosophical Transactions, 17 (1798), 469)

1. Find the sample mean, the sample standard deviation and the sample median of these data.
2. Determine the iqr. Are there any outliers (by the 1.5$\times$ iqr rule)?
3. Construct a box-plot and comment on its features.
4. Would you suggest the sample mean or the sample median as single estimate of the density of earth from Cavendish's data?

1. Direct calculations give $\bar{x} = 5.42$ (g/cm^3) and $s = 0.339$ (g/cm^3).
   - There are 29 observations, so the median is the 15th smallest one (the 'middlemost' observation), which can be found to be

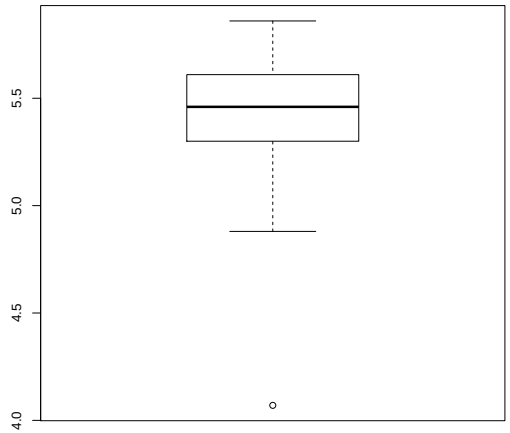
$$m = 5.46 \text{ (g/cm}^3\text{)}$$

2. - This median value splits the sample in two subsamples of 15 observations each (include the median in each half).
   - The 'middlemost' observation in the lower half (the 8th smallest) is the first quartile, found to be $q_1 = 5.30$ g/cm^3, while
   - the 'middlemost' observation in the upper half (the 8th largest) is the third quartile, found to be $q_3 = 5.61$ g/cm^3.
   - The interquartile range is thus

$$\text{iqr} = q_3 - q_1 = 0.31 \text{(g/cm}^3\text{)},$$

   - and by the suggested rule, any observation outside $[q_1 - 1.5 \times \text{iqr}, q_3 + 1.5 \times \text{iqr}] = [4.835, 6.075]$ is an outlier.
   - The value 4.07 is thus an outlier.

3. Boxplot :



The boxplot itself is: - fairly symmetric,
- an outlying value clearly appears.

Conclusion: That observation was most likely corrupted by important measurement and other experimental errors.

4. The sample median should certainly be a more accurate estimate of the true earth density,
as the sample mean would be affected by the outlying value.

Exercise 7

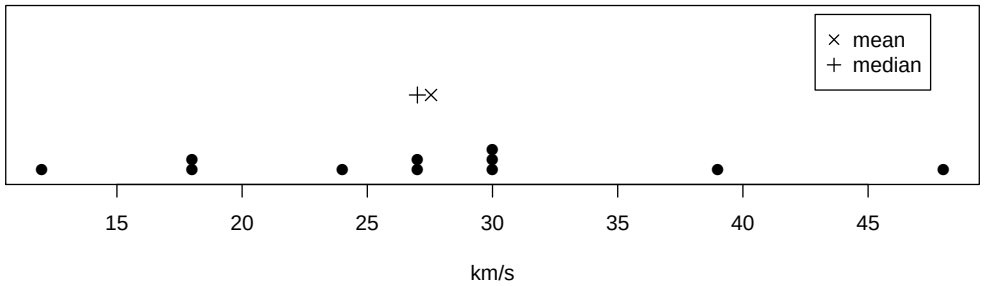
A.A. Michelson (1852-1931) made many series of measurements of the speed of light. Using a revolving mirror techniques, he obtained

12 30 30 27 30 39 18 27 48 24 18

for the differences **[(velocity of light in air) - 299,700 km/s]**.
(*Source :* The Astrophysical Journal, 65 (1927), 11.)

1. Draw a dotplot.
2. Find the median and the mean. Locate both on the dotplot.
3. Find the variance and standard deviation.
4. Find the quartiles.
5. Find the minimum, maximum, range, and interquartile range.

1. Dotplot :



2.
   - Direct calculations give $\bar{x} = 27.55$ km/s.
   - With 11 observations, the median is the 6th smallest, that is $m = 27$ km/s.
3. Direct calculations give
   - $s^2 = 100.47$ km^2/s^2 and
   - $s = 10.02$ km/s.

4. • The median value splits the sample in two subsamples of 6 observations each (include the median in each half).
   • The first quartile is the middle between the 3rd and the 4th smallest observations, that is

$$q_1 = \frac{1}{2}(18 + 24) = 21 \text{ (km/s)}.$$

   • The third quartile is the middle between the 3rd and the 4th largest observations, that is

$$q_3 = \frac{1}{2}(30 + 30) = 30 \text{ (km/s)}.$$

5. • The minimum value is $x_{(1)} = 12$ km/s,
   • the maximum value is $x_{(11)} = 48$ km/s,
   • the range is $x_{(11)} - x_{(1)} = 36$ km/s,
   • the interquartile range (iqr) is $q_3 - q_1 = 9$ km/s.

Exercise 8

An experimental study of the atomisation characteristics of biodiesel fuel was aimed at reducing the pollution produced by diesel engines. Biodiesel fuel is recyclable and has low emission characteristics. One aspect of the study is the droplet size (μm) injected into the engine, at a fixed distance from the nozzle. Consider the following observed droplet size :

2.1 2.2 2.2 2.3 2.3 2.4 2.5 2.5 2.5 2.8 2.9 2.9 2.9 3.0 3.1 3.1 3.2 3.3
3.3 3.3 3.4 3.5 3.6 3.6 3.6 3.7 3.7 4.0 4.2 4.5 4.9 5.1 5.2 5.3 5.7 6.0 6.1
7.1 7.8 7.9 8.9

(*Source :* Kim et al (2008), Energy and Fuels, 22, 2091–2098.)

1. Obtain a frequency table
   using $[2,3)$, $[3,4)$, $[4,5)$, $[5,7)$, and $[7,9)$ as classes,
2. Construct a density histogram.
3. Obtain the sample mean $\bar{x}$ and the sample variance s^2.

1. The frequency distribution is

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Frequency	13	14	4	6	4

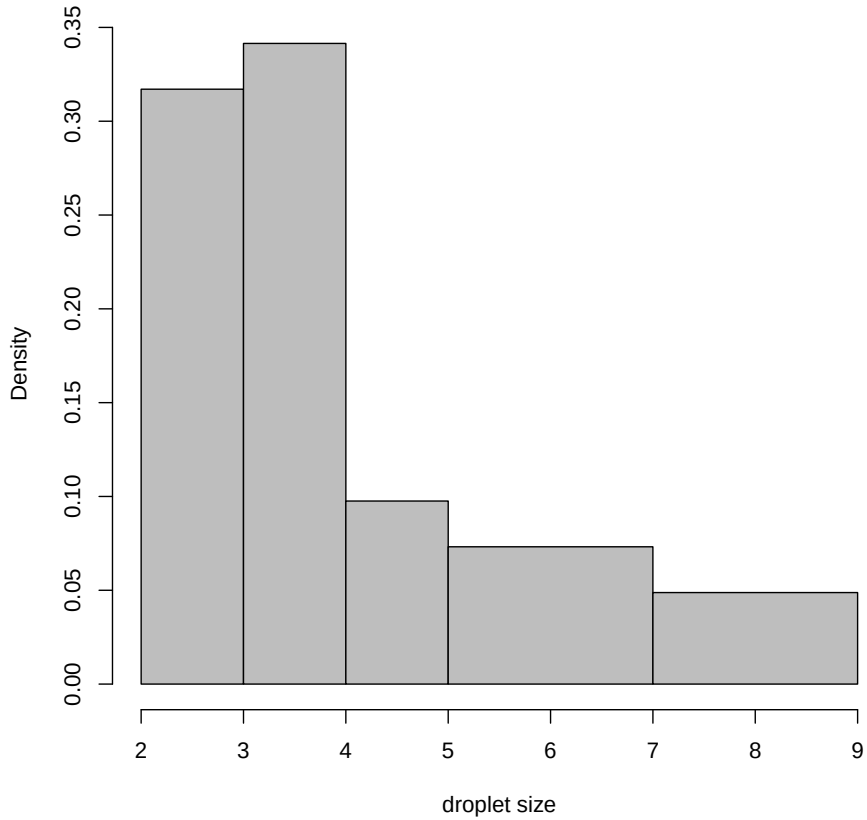
2. - The relative frequency in each class is given by the (absolute) frequency divided by the total number of observations (that is to say, the relative frequency in a class is the proportion of observations in that class).
   - With $n = 41$ observations, we find :

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Relative frequency	0.317	0.341	0.098	0.146	0.098

   Note that these proportions must sum to 1 (as it is the case).
   - The density in each class is given by the relative frequency divided by the class width. Here we find :

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Density	0.317	0.341	0.098	0.073	0.049

- The density histogram is :



3. Direct calculations give $\bar{x} = 3.97\ \mu$m and $s^2 = 2.91\ \mu\text{m}^2$.