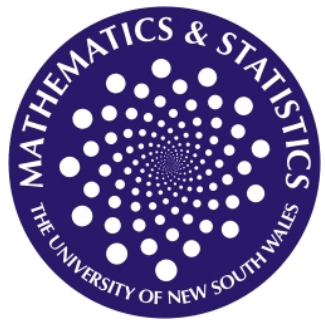
Statistics

MATH2089



Semester 1, 2018 – Lecture 2

This lecture

2. Descriptive Statistics

2.1 Introduction
2.2 Graphical representations

Additional reading:
Sections 1.2, 2.1 (pp. 61-63), 2.2 (pp. 70-73), 2.3 (pp. 79-80, 82-84) in the textbook (2nd edition)
Sections 1.2, 2.1 (pp. 63-65), 2.2 (pp. 72-74), 2.3 (pp. 80-82, 83-86) in the textbook (3rd edition)

2. Descriptive Statistics

Introduction

On Slide 5 (Lecture 1), we defined statistics as **learning from data**.

However, statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are virtually useless unless they are condensed.

$\Rightarrow$**Data should be presented in ways that facilitate their interpretation and subsequent analysis**

The aspect of statistics which deals with organising, describing and summarising data is called **descriptive statistics**.

Essentially, descriptive statistics tools consist of

1. graphical methods and
2. numerical methods

Types of variables

There are essentially two types of variables:

1. **categorical** (or qualitative) variables: take a value that is one of several possible categories (no numerical meaning)

   Eg. gender, hair colour, field of study, status, etc.

2. **numerical** (or quantitative) variables: naturally measured as a number for which meaningful arithmetic operations make sense

   Eg. height, age, temperature, pressure, salary, etc.

Attention: sometimes categorical variables are disguised as quantitative variables.

For example, one might record gender information coded as 0 = Male, 1 = Female. It remains a categorical variable, it is not naturally measured as a number.

Types of variables

- categorical variables
  - ordinal: there is a clear ordering of the categories

    Eg. salary class (low, medium, high), opinion (disagree, neutral, agree), etc.
  - nominal: there is no intrinsic ordering to the categories

    Eg. gender, hair colour, etc.
- numerical variables
  - discrete: the variable can only take a finite (or countable) number of distinct values

    Eg. number of courses you are enrolled in, number of persons in a household, etc.
  - continuous: the variable can take any value in an entire interval on the real line (uncountable)

    Eg. height, weight, temperature, time to complete a task, etc.

Graphical representations

A picture is worth a thousand words

- Graphical representations are often the most effective way to quickly obtain a feel for the essential characteristics of the data.

Fact

Any good statistical analysis of data should always begin with plotting the data.

- Plots often reveal useful information and opens paths of inquiry
- They might also highlight the presence of irregularities or unusual observations ("outliers")

Dotplot

- A **dotplot** is an attractive summary of numerical data when the data set is reasonably small
- Each observation is represented by a dot above the corresponding location on a horizontal measurement scale
- When a value occurs more than one time, there is a dot for each occurrence and these dots are stacked vertically

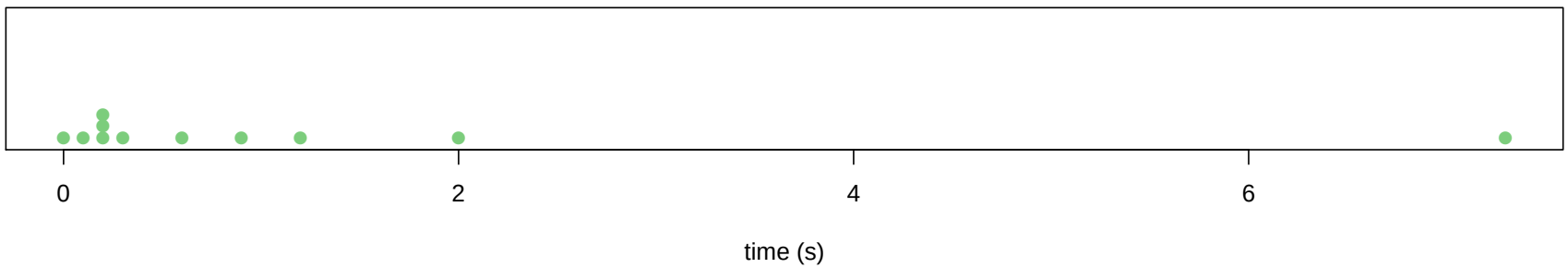
Dotplot: example

Example

In 1987, for the first time, physicists observed neutrinos from a supernova that occurred outside of our solar system. At a site in Kamiokande, Japan, the following times (in seconds) between neutrinos were recorded:

`0.107; 0.196; 0.021; 0.283; 0.179; 0.854; 0.58; 0.19; 7.3; 1.18; 2`

Draw a dotplot:



Note that the largest observation is extremely different to the others. Such an observation is called an outlier.

$\Rightarrow$ Could be: recording error? missed observations? real observation?

Stem-and-leaf plot

- The dotplot is useful for small samples, up to (say) about 20 observations. However, when the number of observations is moderately large, another graphical display may be more useful
- A **stem-and-leaf plot** (or just stemplot) is another effective way to organise numerical data without much effort
- Idea: separate each observation into a **stem** (all but last digit) and a **leaf** (final digit)

$\Rightarrow$ Example: $24 := 2|4$ $139 := 13|9$ $5 := 0|5$

- Write all unique stems in vertical column with the smallest at the top, and draw a vertical line at the right of this column
- Write each leaf in the row to the right of its stem, in increasing order out from the stem

Stem-and-leaf plot: example

Example 1.5 in the textbook

Study of use of alcohol by university students. We are interested in a variable X, the percentage of undergraduate students who are binge drinkers. We observe X on 140 campuses across the US

The sample is:

```
26 57 66 66 41 46 65 35 46 38 44 29 43
14 11 68 37 27 18 46 30 32 35 59 39 32 31 39 21 58
65 50 44 29 53 27 38 52 29 58 45 34 36 56 47 22 59
46 24 51 26 39 23 55 50 42 18 48 64 44 46 66 33 61
38 35 22 57 42 42 26 47 67 37 39 58 26 41 61 51 61
56 48 53 13 28 52 36 62 31 38 42 64 51 54 33 19 25
42 37 36 55 37 56 43 28 56 49 39 57 48 52 60 17 49
61 44 18 67 36 58 47 16 33 27 29 48 45 34 57 56 48
46 49 15 52 04 41 64 37
```

(Source: "Health and Behavioural Consequences of Binge Drinking in College", J. Amer. Med. Assoc., 1994, 1672-1677)

Stem-and-leaf plot: example

$\Rightarrow$ Separate the tens digit ("stem") from the ones digit ("leaf")

0	4
1	1345678889
2	12234566667778899999
3	011223334455566667777788899999
4	111222223344445566666677788888999
5	00111222233455666667777888899
6	0111124445566677 8

This stemplot suggests that

- a typical value is in the stem 4 row (probably in mid-40% range)
- there is a single peak, but the shape is not perfectly symmetric
- there are no observations unusually far from the bulk of the data (no outliers)

Stem-and-leaf plot

The stemplot conveys information about the following aspects of the data:

- identification of a typical value
- extent of spread about the typical values
- presence of any gaps in the data
- extent of symmetry in the distribution values
- number and location of peaks
- presence of any outlying values

Stem-and-leaf plot: variations

There are many ways in which stem-and-leaf plots can be modified to meet particular needs:

- rounding or truncating the numbers to a few digits before making a stemplot to avoid too much irrelevant detail in the stems
- splitting each stem to give greater detail in distribution
- back-to-back stemplots with common stems to compare two related distributions

Stem-and-leaf plot: splitting each stem

A more informative display can be created by repeating each stem value twice, once for the low leaves 0, 1, 2, 3, 4 and again for the high leaves 5, 6, 7, 8, 9.

$\Rightarrow$For the binge-drinking data this yields (compare Slide 12)

```
0 | 4
0 |
1 | 134
1 | 5678889
2 | 12234
2 | 56666777889999
3 | 0112233344
3 | 55566667777888899999
4 | 11122222334444
4 | 5566666677788888999
5 | 001112222334
5 | 5566666777788899
6 | 011112444
6 | 55666778
```

Stem-and-leaf plot: back-to-back plots

Suppose you have two data sets, each consisting of observations on the same variable (for instance, exam scores for two different classes).

- in what ways are the two data sets similar? how do they differ?
- comparative stem-and-leaf plot, or back-to-back stemplots

The stems are common, the leaves for one data set are listed to the right and the leaves for the other to the left.

For instance, for exam scores we could observe (Class 1 | Class 2)

2588	5	9
2234578	6	01445
0225556689	7	1223567
4479	8	01334578
	9	156688

$\Rightarrow$ The right side appears to be shifted down one row from the other side (better scores in Class 2 than in Class 1)

Frequency distribution, bar charts and histograms

A natural continuation of the stemplot is to count the number of observations (that is, the **frequency**) in each row.

$\Rightarrow$ This can be done each time that proper "categories" are available

The frequency of observations in each category then forms the **frequency distribution**.

Which categories ?

- For a categorical variable, the categories are obviously defined

$\Rightarrow$ count the frequency of observations in each category, mark each category on a horizontal axis and draw rectangles whose heights are the corresponding frequencies. This is called a **bar chart**

- For a discrete numerical variable, the categories are given by the distinct values taken by the variable

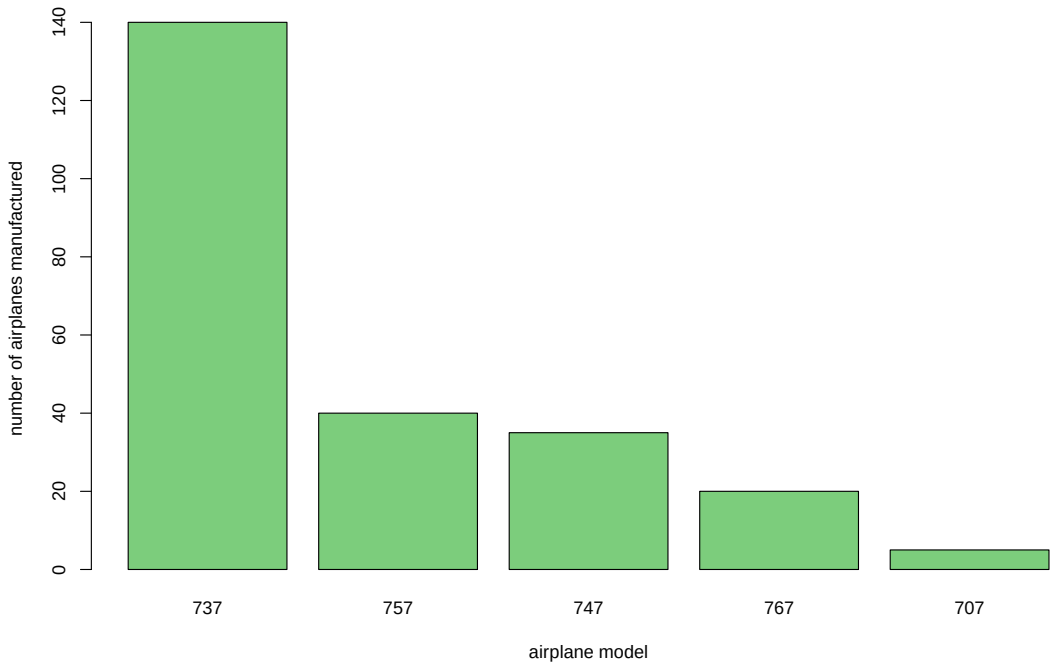
$\Rightarrow$ then, proceed as above. This is called a **histogram**

Bar chart: example

Example

In 1985 the Boeing Company published the figures for its production of transport aircraft. That year, they produced 5 Boeing 707's, 140 Boeing 737's, 35 Boeing 747's, 40 Boeing 757's and 20 Boeing 767's

The frequency distribution for this production is presented in the following bar chart



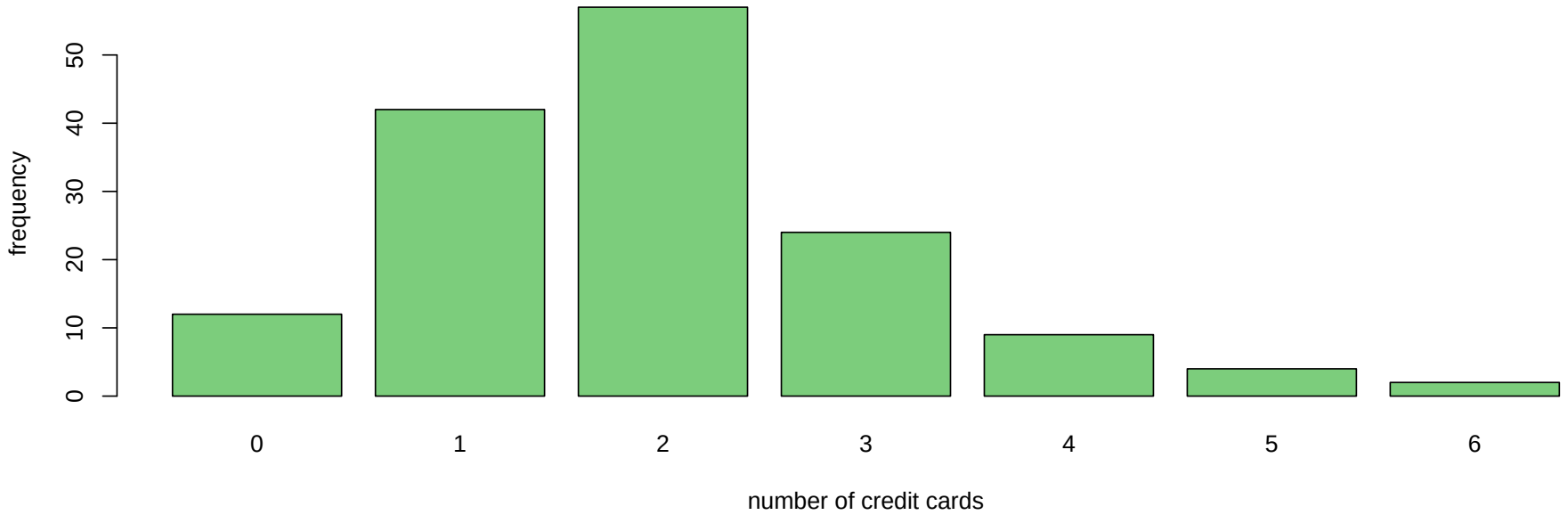
Histogram: example

Example

A sample of students from a statistics class were asked how many credit cards they carry. For 150 students, the frequency distribution is given below

Number of cards	0	1	2	3	4	5	6
Frequency	12	42	57	24	9	4	2

The frequency distribution for the number of credit cards is given in the following histogram



Histogram for a continuous numerical variable

If the variable is numerically continuous, there are no obvious categories

$\Rightarrow$ we have to decide on some categories, called classes, which will be intervals that do not overlap and accommodate all observations

Once the classes have been defined, we can proceed similarly to the above methods, that is:

- Determine the frequency of observations in each class, mark the class boundaries on a horizontal axis and draw rectangles whose heights are the corresponding frequencies
- However, important practical questions arise like how many classes to use and what are the limits for each class

Histogram for continuous numerical variable

Generally speaking, the number of classes depends on the total number of observations and the range of the data.

This is a trade-off between

1. choosing too few classes at a cost of losing information about actual values
2. choosing too many classes will result in the frequencies of each class to be too small for a pattern to be discernible

An empirical rule is

$$\text{number of classes} \simeq \sqrt{\text{number of observations}}$$

$\Rightarrow$ between 5 and 20 classes will be satisfactory for most data sets

Note: it is common, although not essential, to choose classes of equal width

Histogram: example 1.8 in the textbook

Example

Power companies need information about customer usage to obtain accurate forecasts of demand. Here we consider the energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes

The sample is

```
10.04 13.47 13.43 9.07 11.43 12.31 4.00 9.84 10.28 8.29
6.94 10.35 12.91 10.49 9.52 12.62 11.09 6.85 15.24 18.26
11.21 11.12 10.28 8.37 7.15 9.37 9.82 9.76 8.00 10.21
6.62 12.69 13.38 7.23 6.35 5.56 5.98 6.78 7.73 9.43 9.27
8.67 15.12 11.70 5.94 11.29 7.69 10.64 12.71 9.96 13.60
16.06 7.62 2.97 11.70 13.96 8.81 12.92 12.19 16.90 9.60
9.83 8.26 8.69 6.80 9.58 8.54 7.87 9.83 10.30 8.61 7.93
13.11 7.62 10.95 13.42 6.72 10.36 12.16 10.40 5.20 10.50
8.58 14.24 14.35 8.47 7.29 12.28 11.62 7.16
```

Histogram: example

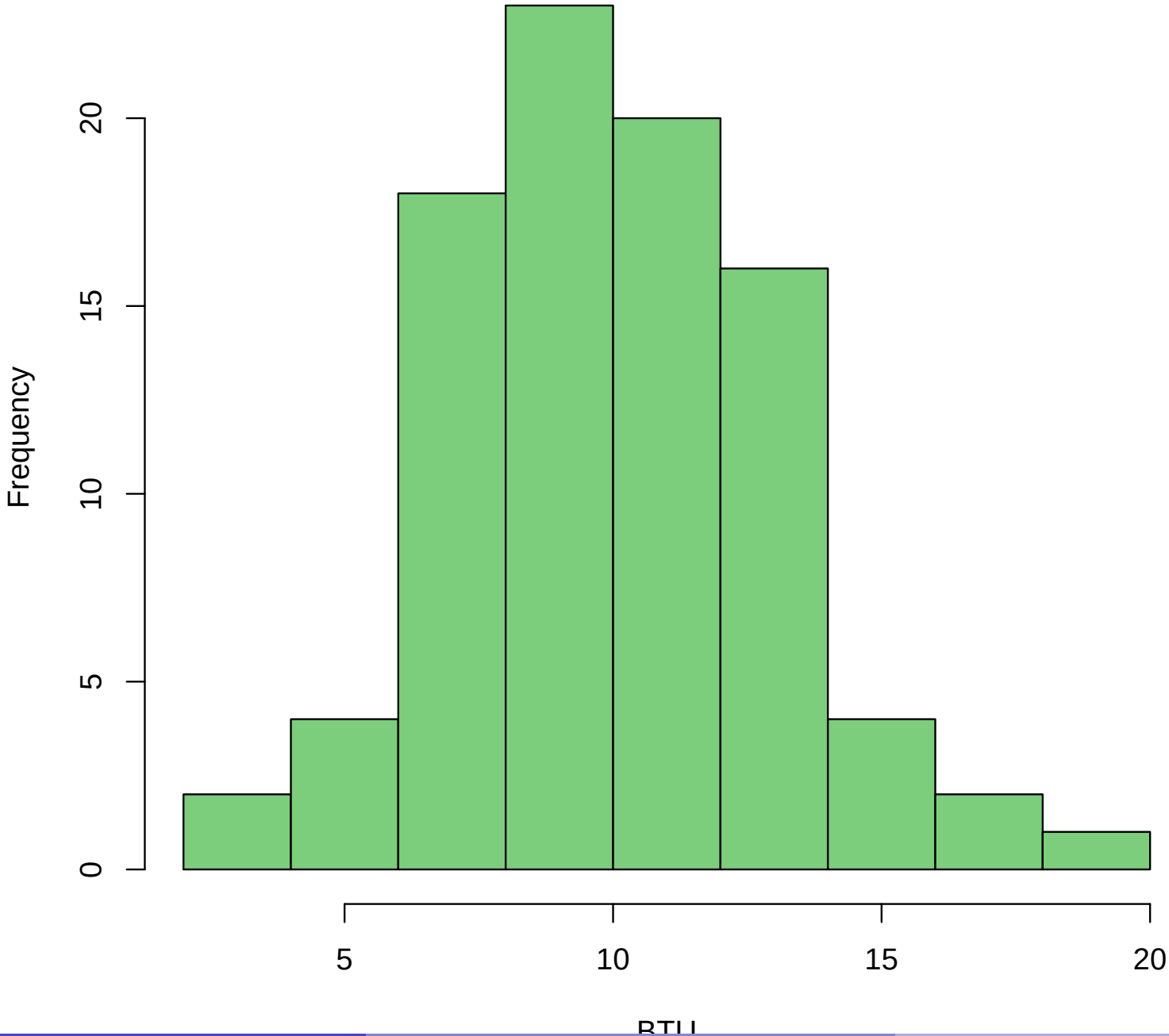
- The data set contains 90 observations, and since $\sqrt{90} \simeq 9.48$, we suspect that about nine classes will provide a satisfactory frequency distribution
- The smallest and largest data values are 2.97 and 18.26, so the classes must cover a range of at least 15.29 BTU
- As $15.29/9 \simeq 1.7$, we take the common class width equal to 2 (for simplicity), and we start at 2 (again for simplicity)

Counting the frequencies in the so-defined classes, we get

$[2,4)$	$[4,6)$	$[6,8)$	$[8,10)$	$[10,12)$	$[12,14)$	$[14,16)$	$[16,18)$	$[18,20)$
1	5	18	23	20	16	4	2	1

Note: we adopt the left-end inclusion convention, i.e. a class contains its left-end but not its right-end boundary point (interval $[a, b)$, left-closed right-open)

Histogram: example



Histogram: comments

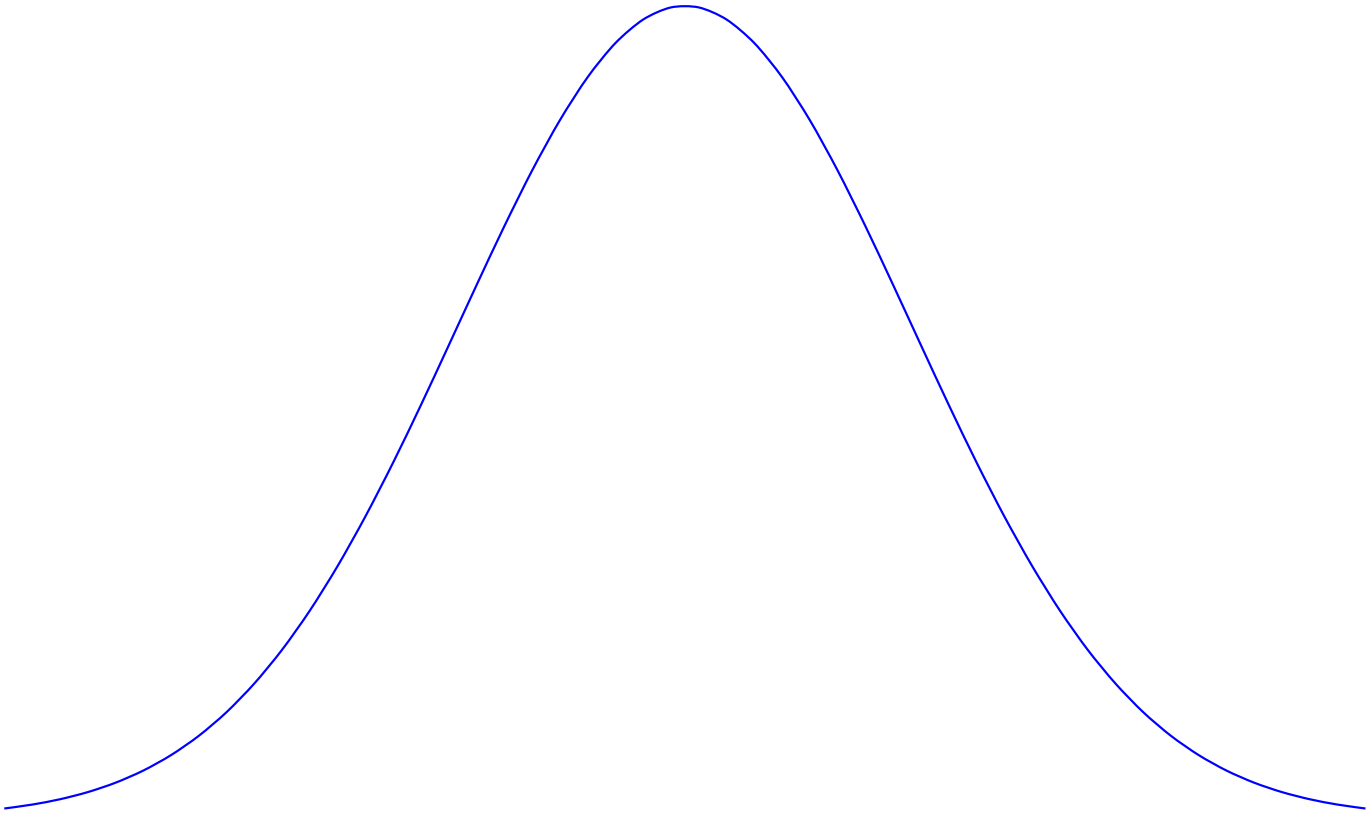
Unlike the histogram for discrete numerical variables, the histogram for continuous numerical variables consists of adjacent rectangles. This reflects the continuity of the underlying variable.

Like the stemplot, the histogram (discrete or continuous) provides a visual impression of the shape of the distribution of the observations, as well as information about the central tendency and dispersion in the data, that may not be immediately apparent from the data themselves.

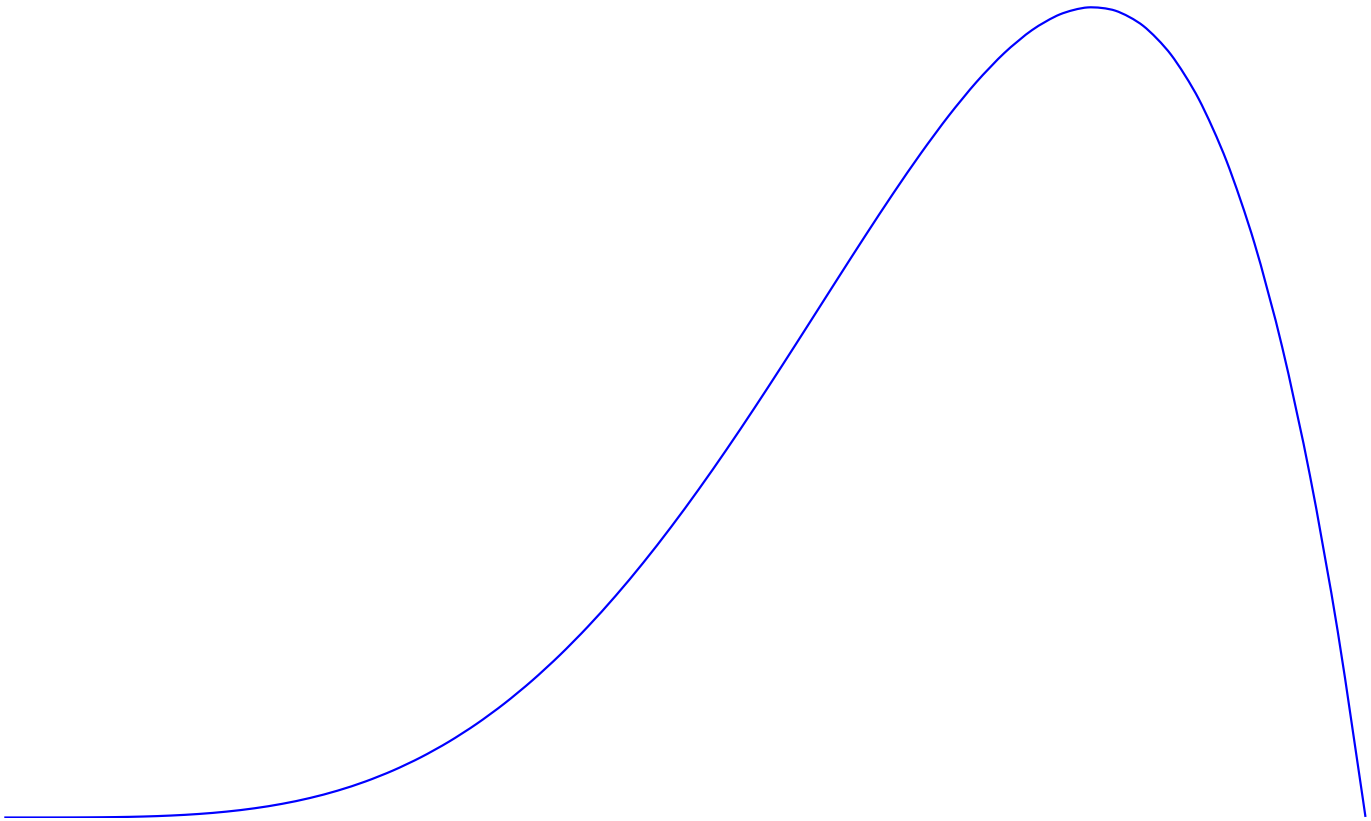
Typical words/phrases used to describe histograms:

- symmetric, or skewed to the right/left ($\Rightarrow$ with right/left tail);
- unimodal (one peak), or bimodal/multimodal;
- bell-shaped (if symmetric & unimodal);
- there are possible outliers around..., or there are no outliers;
- typical value of the data is..., the range of the data is...

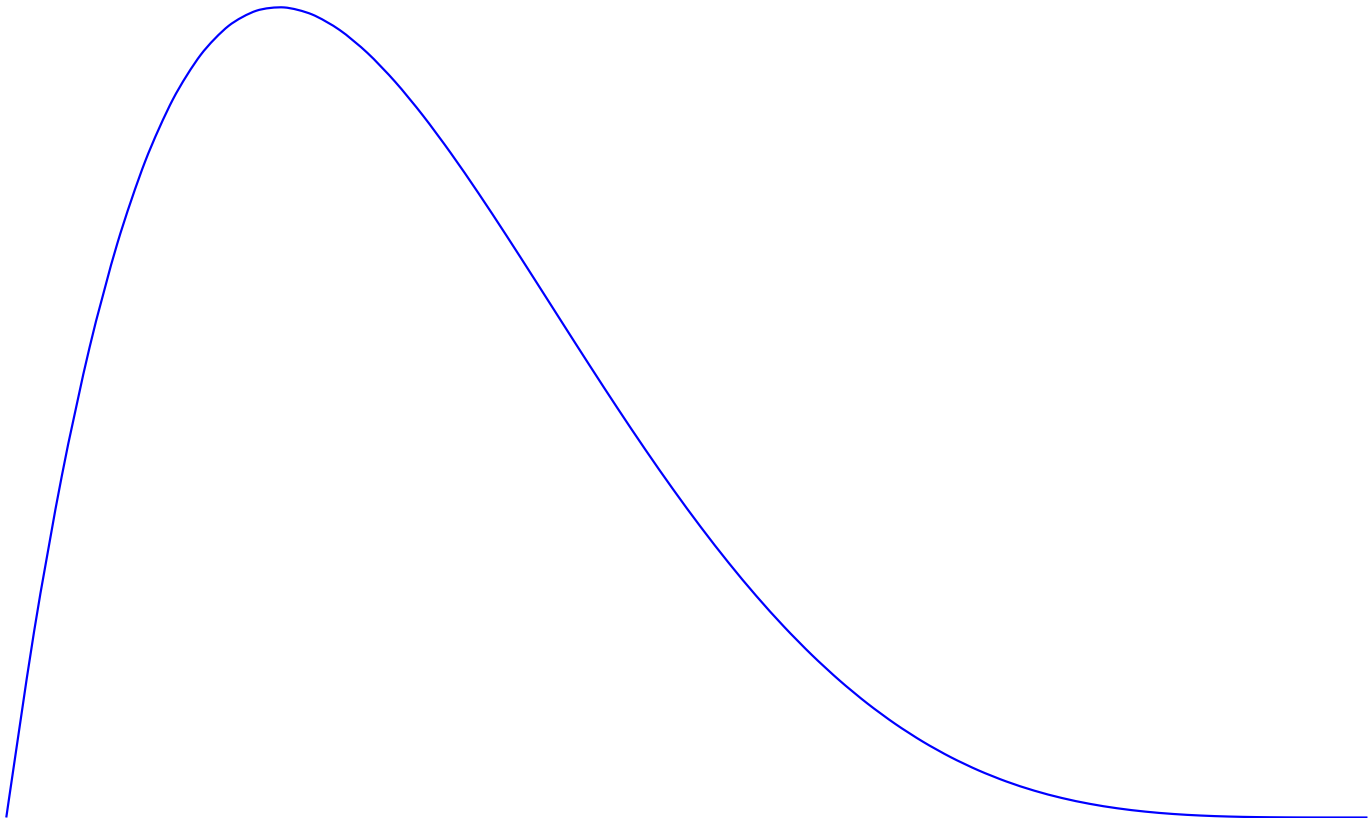
Example: Symmetric shape



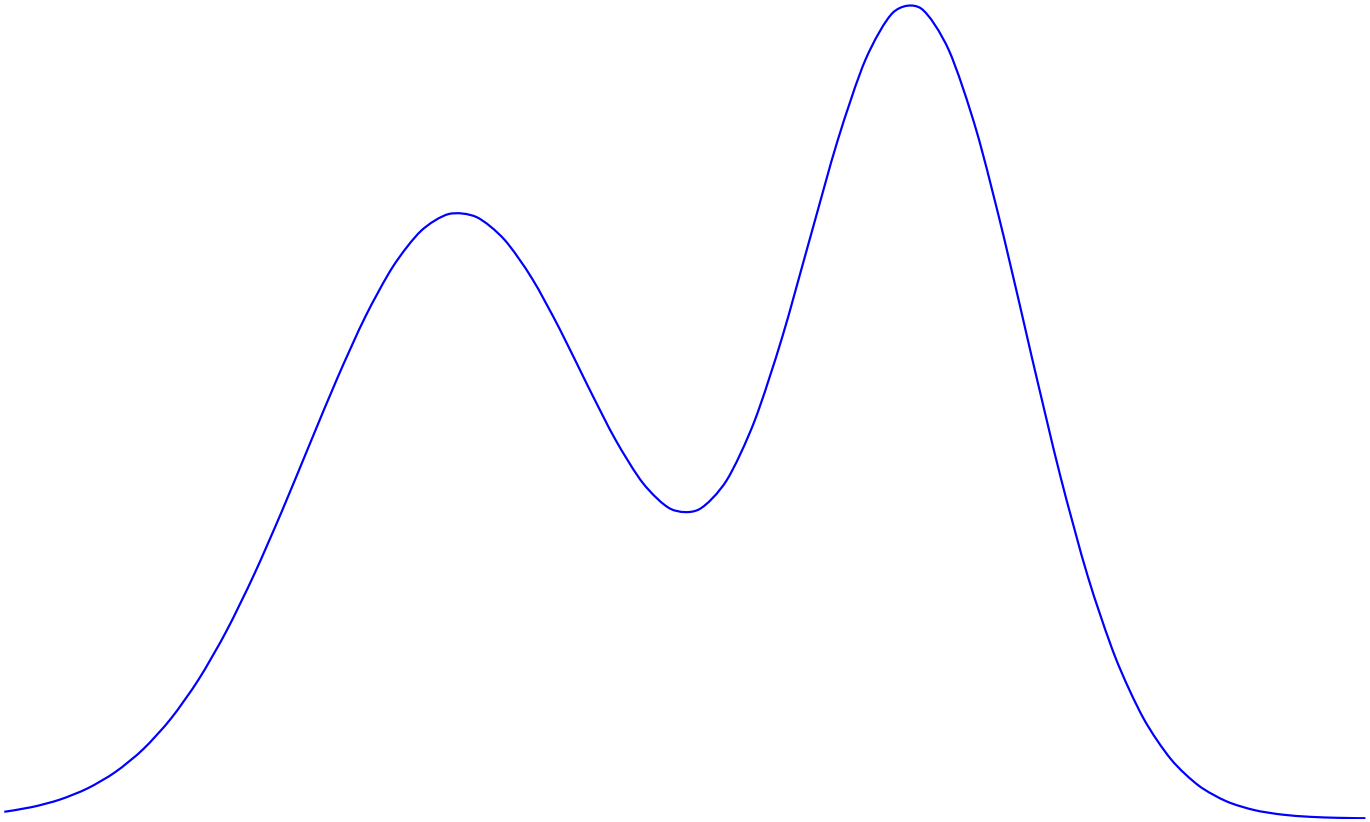
Example: Skewed to the left



Example: Skewed to the right



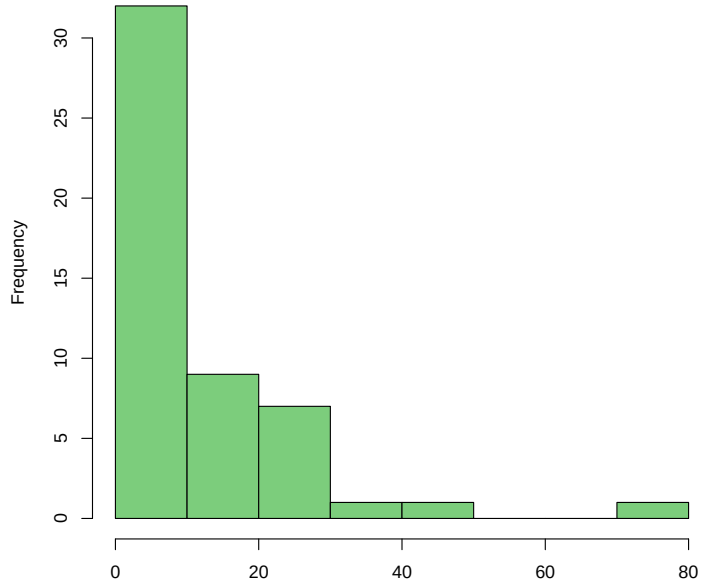
Example: Bimodal



Unequal class width

- Sometimes, equal-width classes may not be a sensible choice
- For instance, if the data have several extreme observations or outliers, nearly all observations will fall in just a few of the classes

Consider the following histogram:



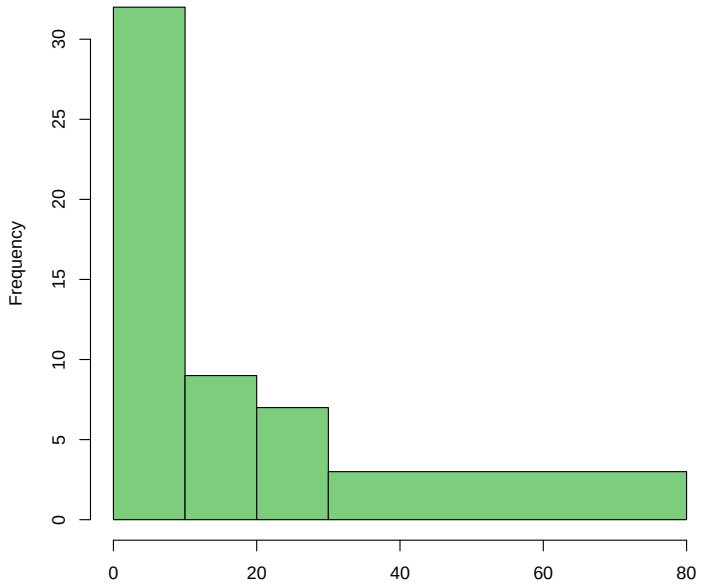
$\Rightarrow$ The last 5 classes together have only 3 observations!

$\Rightarrow$ Might preferable to regroup the observations > 30 into a wider class

Unequal class width

However, wider classes are likely to include more observations than narrower ones!

$\Rightarrow$ when class widths are unequal, the frequencies will give a distorted representation of reality



$\Rightarrow$ the rectangular areas (not their heights) should be proportional to the frequencies : this is what a **density histogram** achieves

Density histogram

- The relative frequency of a class is the proportion of observations in that class, ie, the frequency of the class divided by the total number of observations
- We call the **density** of a class the relative frequency of the class divided by the class width
- A **density histogram** is a histogram whose rectangle heights are the densities of each class (no longer the frequencies)

So we have

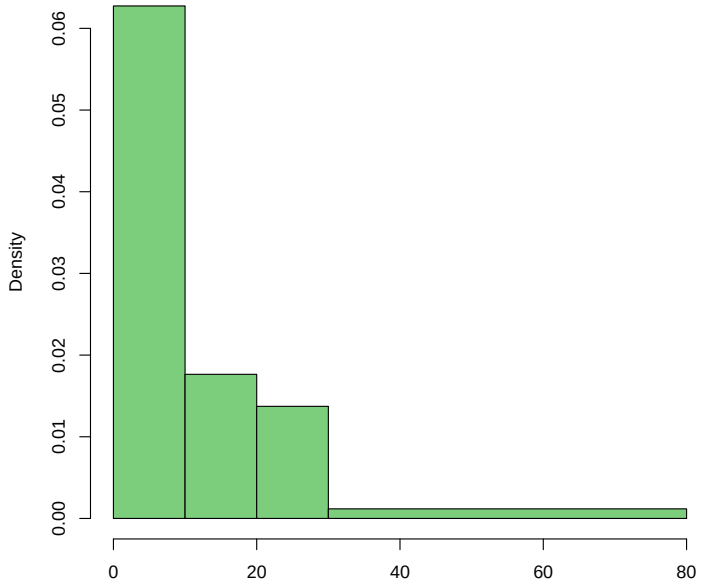
$$\begin{aligned} \text{relative frequency} &= \text{density} \times \text{class width} \\ &= \text{rectangle height} \times \text{rectangle width} \\ &= \text{rectangle area} \end{aligned}$$

Note: the total area of the rectangles must be equal to 1, as the sum of all relative frequencies must be 1.

This property will play an important role in the sequel, so that it is always preferable to represent a density histogram instead of a (frequency) histogram, even when the classes have equal width.

Density histogram

For the previous case, the density histogram would be



which is much more faithful to reality than histogram on Slide 31.

$\Rightarrow$We can check that the area of the first rectangle is $10 \times 0.063 = 0.63$, that is the first class $[0, 10)$ includes 63% of the observations

$\Rightarrow$We could calculate the areas of the other rectangles the same way, and check that their sum is equal to 1

Density histogram: Example

The accompanying specific gravity values for various wood types used in construction appeared in the article "Bolted Connection Design Values Based on European Yield Model" (J. of Structural Engr., 1993: 2169-2186):

```
0.36 0.45 0.66 0.66 0.44 0.40 0.48 0.75 0.51 0.67 0.42
0.35 0.47 0.38 0.37 0.41 0.41 0.46 0.54 0.62 0.48 0.42
0.43 0.42 0.31 0.54 0.42 0.48 0.42 0.40 0.40 0.68 0.55
0.36 0.58 0.46
```

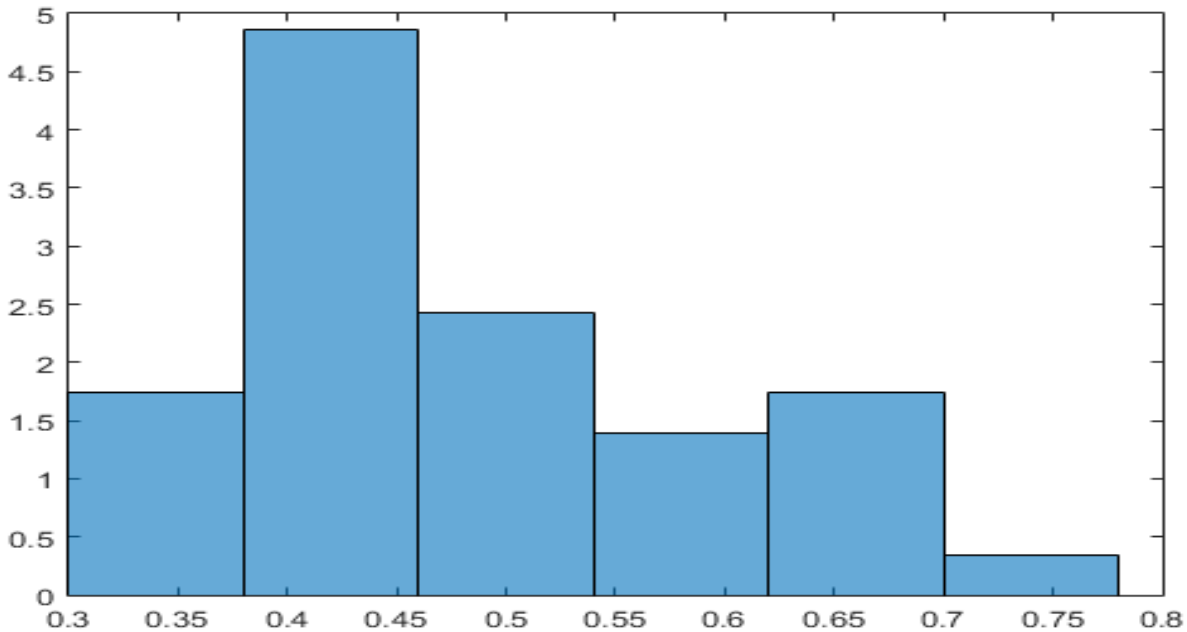
Frequency/Density table:

	$[0.3, 0.38)$	$[0.38, 0.46)$	$[0.46, 0.54)$	$[0.54, 0.62)$	$[0.62, 0.7)$	$[0.7, 0.78)$
Freq.	5	14	7	4	5	1
Relative Freq.	5/36	14/36	7/36	4/36	5/36	1/36
Density	$\frac{5/36}{0.08}$	$\frac{14/36}{0.08}$	$\frac{7/36}{0.08}$	$\frac{4/36}{0.08}$	$\frac{5/36}{0.08}$	$\frac{1/36}{0.08}$
	1.7361	4.8611	2.4306	1.3889	1.7361	0.3472

Density histogram: Example

The accompanying specific gravity values for various wood types used in construction appeared in the article "Bolted Connection Design Values Based on European Yield Model" (J. of Structural Engr., 1993: 2169-2186):

```
0.36 0.45 0.66 0.66 0.44 0.40 0.48 0.75 0.51 0.67 0.42
0.35 0.47 0.38 0.37 0.41 0.41 0.46 0.54 0.62 0.48 0.42
0.43 0.42 0.31 0.54 0.42 0.48 0.42 0.40 0.40 0.68 0.55
0.36 0.58 0.46
```



Descriptive measures

- Dotplots, stem-and-leaf plots, histograms and density histograms summarise a data set graphically so we can visually discern the overall pattern of variation
- It is also useful to numerically describe the data set

$\Rightarrow$ Summary measures that tell where a sample is centred (measures of centre or location), and what is the extent of spread around its centre (measures of variability)

- Usual notation:

$$x_1, x_2, \ldots, x_n$$

for a sample consisting of n observations of the variable X

Note: except when indicated otherwise, we assume that X is a numerical variable.

Measure of centre: the sample mean

The most frequently used measure of centre of a sample is simply the arithmetic mean (or average) of the n observations.

It is usually denoted $\bar{x}$ and is given by

Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note: the unit of $\bar{x}$ is the same as that of X.

Example

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

```
1792; 1666; 1362; 1614; 1460; 1867; 1439
```

$$\Rightarrow \bar{x} = \tfrac{1}{7}(1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439) = 1600$$

(calories/24h)

Measure of centre: the sample median

The **median** is another descriptive measure of the centre of sample.

Sample median

The median, usually denoted m (or $\tilde{x}$), is the value which divides the data into two equal parts, half below the median and half above.

$\Rightarrow$ the sample median m is the "middlemost" value of the sample

Denote the ordered sample (smallest to largest observation)

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

If n is odd, the median is the middle observation in the ordered data series, that is, $m = x_{\left(\frac{n+1}{2}\right)}$

If n is even, the median is *defined as* the average of the middle two observations in the ordered data series, that is

$$m = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

Measure of centre: the sample median

Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

$$1792;\ 1666;\ 1362;\ 1614;\ 1460;\ 1867;\ 1439$$

Find the mean and the median.

First, order the sample

$$1362 \leq 1439 \leq 1460 \leq 1614 \leq 1666 \leq 1792 \leq 1867$$

Here, $n = 7$ (odd), so the median is the middle value $x_{(7+1)/2} = x_{(4)} = 1614$ (calories/24h).

Note: since the order of the observations matters, the median can also be defined for an ordinal categorical variable

Measure of centre: the sample median

Sometimes, it is preferable to use the median instead of the mean, as it is resistant/robust to outliers.

Example

A small company employs four young engineers, who each earn \$70,000, and the owner (also an engineer), who gets \$160,000. The latter claims that on average, the company pays \$88,000 to its engineers and, hence, is a good place to work.

The mean of the five salaries is indeed

$$\bar{x} = \frac{1}{5}(4 \times 70,000 + 160,000) = 88,000 \$$$

but this hardly describes the situation.

On the other hand, the median is the middle observation in

$$70,000 = 70,000 = 70,000 = 70,000 < 160,000$$

that is, $\$70,000$: much more representative of what a young engineer earns with the firm.

Quartiles and Percentiles

- We can also divide the sample into more than two parts
- When a sample is divided into four equal parts, the division points are called sample **quartiles**

$\Rightarrow$ The first or lower quartile q_1 is the value that has 25% of the observations below (or equal to) it and 75% of the observations above (or equal to) it

$\Rightarrow$ The third or upper quartile q_3 is the value that has 75% of the observations below (or equal to) it and 25% of the observations above (or equal to) it

$\Rightarrow$ The second quartile q_2 would split the sample into two equal halves (50% below - 50% above): that is the median ($m = q_2$)

Quartiles and Percentiles

More generally, the sample $(100 \times p)$th **percentile** (or quantile) is the value such that $100 \times p\%$ of the observations are below this value (or equal to it), and the other $100 \times (1 - p)\%$ are above this value (or equal to it).

$\Rightarrow q_1$ is the 25th, the median is the 50th and q_3 is the 75th percentile

Quartiles and Percentiles

Practically,

$$\text{lower quartile} = \text{median of the lower half of the data}$$

$$\text{upper quartile} = \text{median of the upper half of the data}$$

Note: if n is an odd number, include the median in each half

Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

$$1792;\ 1666;\ 1362;\ 1614;\ 1460;\ 1867;\ 1439$$

Find the quartiles.

The median is 1614 (calories/24h). The lower half of the observations is thus

$$1362 \leq 1439 \leq 1460 \leq 1614,$$

whose median is $q_1 = \frac{1}{2}(1439 + 1460) = 1449.5$ (calories/24h).

Similarly, the third quartile $q_3 = 1729$ (calories/24h).

Five number summary

Quartiles give more detailed information about location of a data set.

Often, the three quartiles (i.e., including the median) together with the minimum and maximum observation give a good insight into the data set $\Rightarrow$ this is known as the **five number summary**

Five number summary

$$\{x_{(1)}, q_1, m, q_3, x_{(n)}\}$$

Note:

- $m = q_2$ is the 50th percentile
- q_1 is the 25th percentile, q_3 is the 75th percentile
- $x_{(1)}$ is the "0th percentile", $x_{(n)}$ is the "100th percentile"

Example: find the 5-number summary for the calories data

$$\{1362, 1449.5, 1614, 1729, 1867\}$$

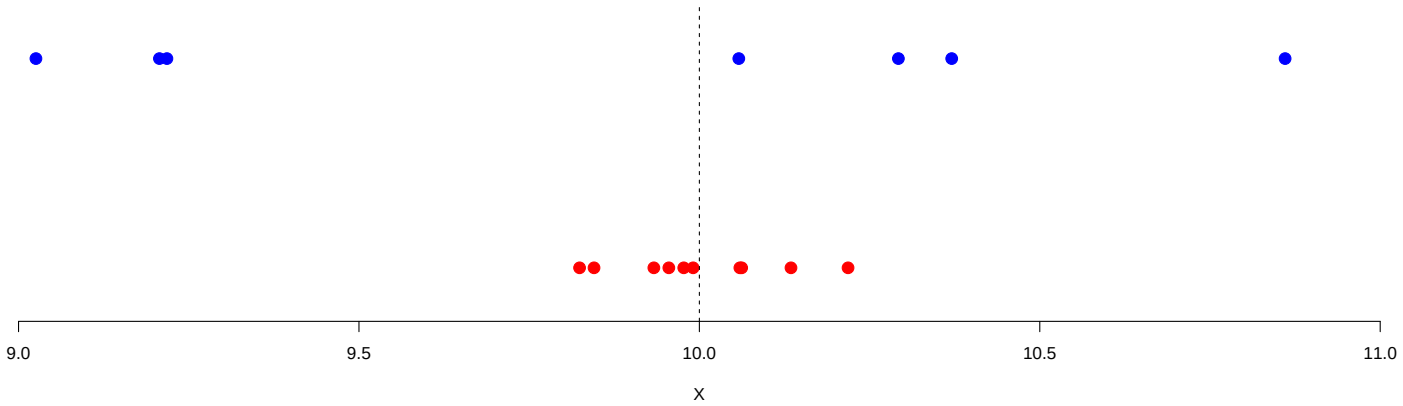
Measures of variability

One of the most important characteristics of any data set is that the observations are not all alike.

$\Rightarrow$ Mean / median describe the central location of a data set, but tell us nothing about the spread or variability of the observations

$\Rightarrow$ Different samples may have identical measures of centre, yet differ from one another in other important ways

We observe that the dispersion of a set of observations is small if the values are closely bunched about their mean (red sample), and that it is large if the values are scattered widely about their mean (blue sample) – both samples have mean 10.



Measures of variability

$\Rightarrow$It would seem reasonable to measure the variability in a data set in terms of the amounts by which the values deviate from their mean

Define the deviations from the mean

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$$

We might then think of using the average of those deviations as a measure of variability in the data set.

Unfortunately, this will not do, because this average is always 0:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = \bar{x} - \bar{x} = 0$$

$\Rightarrow$ We need to remove the signs of those deviations, so that positive and negative ones do not cancel each other out

$\Rightarrow$ Taking the square is a natural thing to do

Measure of variability: the sample variance

The most common measure of variability in a sample is the **sample variance**, usually denoted s^2.

The sample variance s^2 is *essentially* the average of the squared deviations from the mean $\bar{x}$.

Sample variance

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

See that the divisor for the sample variance is $n-1$, not n

$\Rightarrow$ s^2 is based on the n quantities $x_1 - \bar{x}$, $x_2 - \bar{x}$, $\ldots$, $x_n - \bar{x}$. But $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$\Rightarrow$ Thus, specifying the values of any $(n-1)$ of the quantities determines value of the remaining one. The number $n-1$ is the **number of degrees of freedom** for s^2

Measure of variability: the sample variance

It is not difficult to see that, expanding the square, we can write

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}\bar{x}^2-2\sum_{i=1}^{n}x_i\bar{x}\\ &= \sum_{i=1}^{n}x_i^2+n\bar{x}^2-2\bar{x}\sum_{i=1}^{n}x_i\\ &= \sum_{i=1}^{n}x_i^2+n\bar{x}^2-2n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2\end{aligned}$$

Hence,

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

This often makes the computation of the variance easier.

Measure of variability: the sample standard deviation

Notice that the unit of the variance is not that of the original observations:

$$\text{unit of } s^2 = (\text{unit of } X)^2$$

$\Rightarrow$ difficult to interpret

Consequently, one often works with the **sample standard deviation** s.

Sample standard deviation

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

The unit of s is the same as the original unit of X

$\Rightarrow$ ease of interpretation in measuring spread about the mean in the original scale

The standard deviation s has a *rough* interpretation as the average distance from an observation to the sample mean.

Measure of variability: example

Example (ctd.)

Metabolic rate is rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

$$1792;\ 1666;\ 1362;\ 1614;\ 1460;\ 1867;\ 1439$$

Find the variance and the standard deviation.

Here, $n = 7$ and the sample mean is $\bar{x} = 1600$ calories/24h. It follows

$$\begin{aligned} s^2 &= \frac{1}{6}\left(1792^2 + 1666^2 + \ldots + 1439^2\right) - \frac{7}{6} \times 1600^2 \\ &= 35811.87 \text{ (calories/24h)}^2 \end{aligned}$$

The standard deviation is $s = \sqrt{35811.87} = 189.24$ calories/24h

Measure of variability: iqr

- The sample variance s^2 is a variability measure related to the central tendency measured by the sample mean. $\bar{x}$

$\Rightarrow$ The sample **Interquartile Range** (iqr) is a measure of variability related to the sample median and the quartiles.

As the name suggests, iqr is given by the difference between the upper and the lower quartiles.

Sample Interquartile Range

$$\text{iqr} = q_3 - q_1$$

The interquartile range describes the amount of variation in the middle half of the observations.

It enjoys the properties of the quartiles, mainly the fact that it is less sensitive to outliers than the sample variance.

Detecting outliers from iqr

- We have defined an **outlier** as an observation which is too different from the bulk of the data

$\Rightarrow$ How much different should an observation be to be an outlier?

An empirical rule is the following:

an outlier is an observation farther than
1.5 $\times$ iqr from the closest quartile

Besides, we say that

- an outlier is extreme if it is more than 3 $\times$ iqr from the nearest quartile
- it is a mild outlier otherwise

$\Rightarrow$It is important to examine the data for possible outliers as those abnormal observations may affect most of the statistical procedures

Outliers: example

Example

Consider the energy consumption data on Slide 22. Find possible outliers.

Exercise: show that the five number summary is

$$\{2.97,\ 7.9475,\ 9.835,\ 12.045,\ 18.26\}$$

Specifically, $q_1 = 7.9475$ (BTUs) and $q_3 = 12.045$ (BTUs).

Hence, $\text{iqr} = 12.045 - 7.9475 = 4.0975$ (BTUs).

The limits for not being an outlier are thus

$$[q_1 - 1.5 \times \text{iqr}, q_3 + 1.5 \times \text{iqr}] = [1.80125,\ 18.19125]$$

$\Rightarrow$ only one observation (check in the data set!) falls outside that interval:

the largest value 18.26 (mild outlier)

Boxplots

A **boxplot** is a graphical display showing the five number summary and any outlier value.

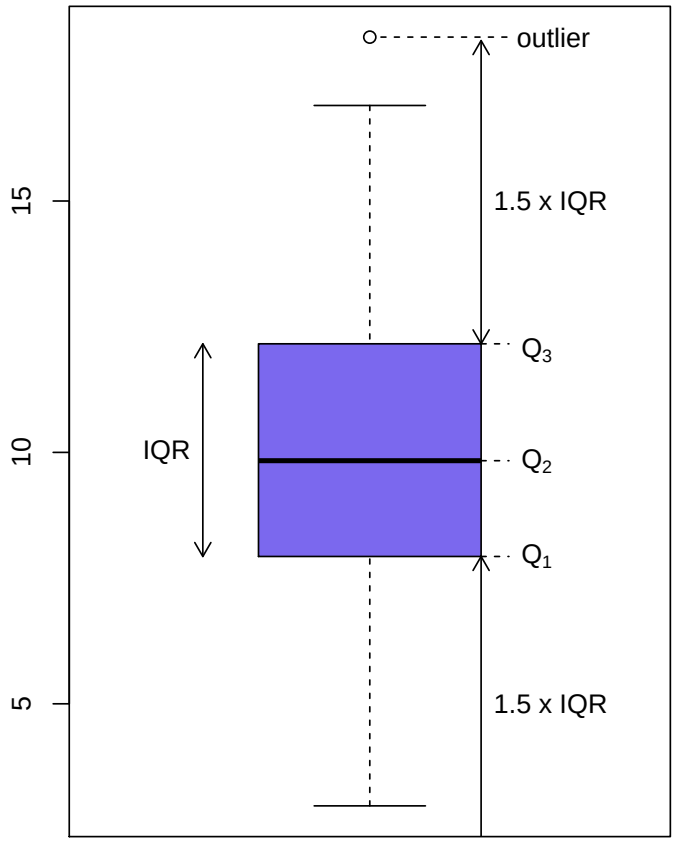
It is sometimes called **box-and-whisker** plot.

A central box spans the quartiles.

A line in the box marks the median.

Lines extend from the box out to the smallest and largest observations which are not suspected outliers.

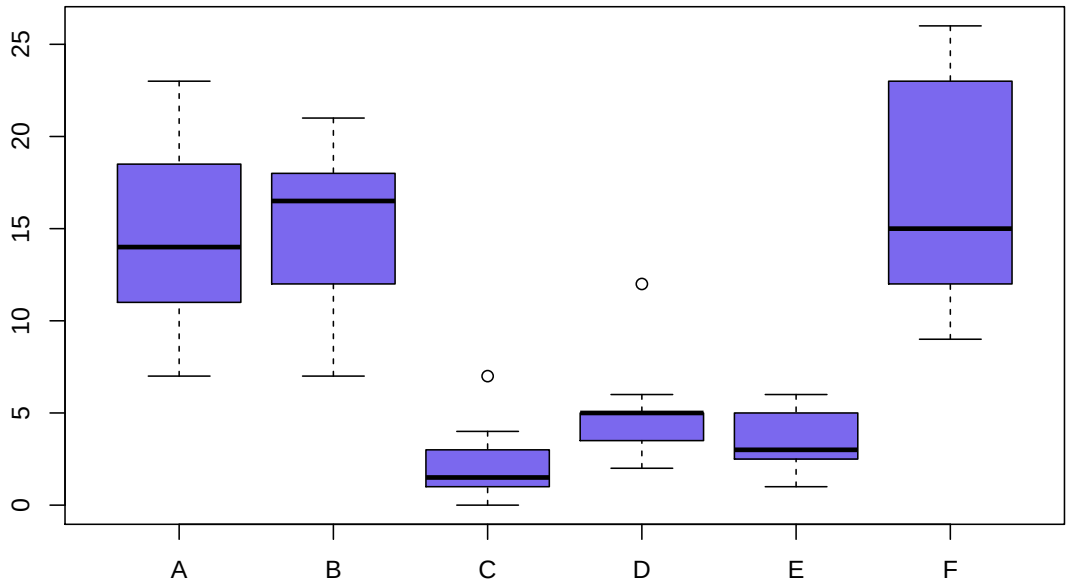
Observations more than 1.5 $\times$ IQR outside the central box are plotted individually as outliers.



Boxplots

Boxplots are very useful graphical comparisons among data sets, because they have high visual impact and are easy to understand.

The following boxplots refer to the counts of insects in agricultural experimental units treated with six different insecticides (A to F).



At a glance you can tell that Insecticide C is the most effective (but outlier $\rightarrow$ need to investigate), that D and E are also doing well unlike A, B and F, F is especially unreliable (largest variability).

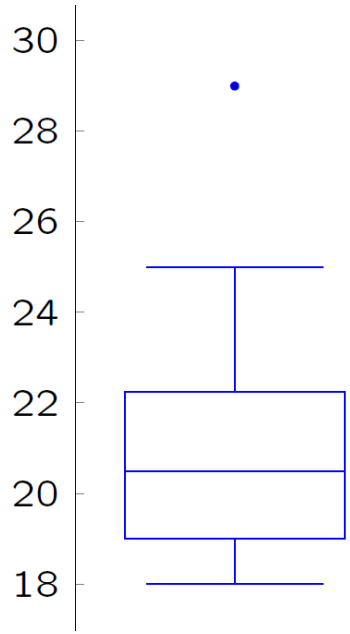
Example

A poll of age in years of 20 randomly chosen students led to the data:
22, 18, 20, 29, 21, 24, 21, 19, 19, 23, 19, 19, 25, 19, 19, 21, 24, 18, 21, 20
Determine the five-number summary. Draw a boxplot.

Five number summary:

$$\{\ 18\ 19\ 20.5\ 22.5\ 29\ \}$$

$iqr = 22.5 - 19 = 3.5,$
$q_1 - 1.5iqr = 13.75,$
$q_3 + 1.5 * iqr = 27.75,$
hence there is one outlier: 29



Quiz

Adapted from a 2011 exam.

In an air-pollution study, ozone concentrations were taken in a large California city at 5.00 p.m. The eight readings (in parts per million) were

$$7.9,\ 11.3,\ 6.9,\ 12.7,\ 13.2,\ 8.8,\ 9.3,\ 10.6$$

1. Calculate the sample mean and standard deviation.
2. Determine the five-number summary for this sample.
3. Draw a boxplot of the data and comment on its feature.

Answers:

Objectives

Now you should be able to:

- understand the importance of graphical representations, construct and interpret a dotplot ☐
- compute an interpret the sample mean, sample variance, sample standard deviation, sample median and sample quartiles ☐
- construct and interpret visual data displays, including the stem-and-leaf plot, the histogram and the boxplot ☐
- comment and assess the overall pattern of data from visual displays ☐
- explain how to use the boxplots to visually compare two or more samples of data ☐

Recommended exercises (from the textbook):

→ Q5 p.20, Q17 p.23, Q3 p.69, Q15 p.77, Q35 p.86, Q37 p.86, Q39 p.86, Q53 (a,c,d) p.95, Q59 p.96, Q67 p.98, Q69 p.98 (2nd edition)

→ Q5 p.24, Q17 p.27, Q3 p.70, Q15 p.78, Q38 p.88, Q40 p.88, Q54 (a,c,d) p.97, Q60 p.98, Q68 p.100, Q70 p.100 (3rd edition)