Statistics for Computing MA4413

Lecture 1

Introduction / Basic Concepts, Data and Graphical Summaries

Kevin Burke

kevin.burke@ul.ie

Timetable

Time	Monday	Thursday	Friday
9 – 10	**Lec** CSG001		
10 – 11			
11 – 12			
12 – 1			
1 – 2			
2 – 3			
3 – 4	*Tut*: KBG11		**Lec**: FB028
4 – 5			
5 – 6		*Tut*: A1053	

Tutorials

- Tutorials begin in Week 2 (next week!)
- Go to your assigned tutorial - ensures manageable class sizes
- Make sure you print the tutorial sheet!
- **Try to attempt some questions before coming to your tutorial**
- Solutions will be available to *everybody* at the end of each week

Course Content: SULIS

All content (lecture slides, tutorial sheets, solutions and any other relevant material) will be available on the SULIS website:

http://sulis.ul.ie

If you have any trouble accessing the material, let me know straight away.

Assessment

There will be *two* midterm exams:

- **Friday Week 5 - October 10th.**
- **Friday Week 10 - November 14th.**
- Likely format: 15 multiple choice questions each worth 1%.

$\Rightarrow$ Midterm 1 + Midterm 2 $=$ 15% + 15% = 30% of the overall module.

$\Rightarrow$ There will be a 10% assignment during the year. (using the "R" programming language)

$\Rightarrow$ The final exam is then worth 60%.

Grading bands:

- **A1**: 90 - 100
- **A2**: 80 - 89
- **B1**: 70 - 79
- **B2**: 60 - 69
- **B3**: 55 - 59
- **C1**: 50 - 54
- **C2**: 45 - 49
- **C3**: 40 - 44
- **D1**: 35 - 39
- **D2**: 30 - 34
- **F** : 0 - 29

Maths Learning Centre

You may or may not be aware of the Maths Learning Centre (MLC) in UL (see http://www.mlc.ul.ie for details).

This is a drop-in centre where any student can come for one-to-one help with maths problems in room **A2-018** (think of it as a free grind).

This facility is open 10am-12pm and 2pm-4pm everyday during Weeks 3 - 12. It is also open in Weeks 1 and 2 on a limited basis (see website).

Final Word on the Course

- Do not make the course more difficult for yourself.
- Each week builds on the last. Stay on top of things - do not let them build up.
- As soon as you have an issue, make sure you address it (at the end of a lecture, during a tutorial class, by going to the maths learning centre or by emailing me).

R: Statistical Programming Language

"R" is a widely-used *freely-available* statistical programming language.

The R code required to perform the various statistical methods covered will be provided in the lecture slides.

Familiarity with the language will allow you (for example) to check your tutorial answers and to get a better feel for the methods.

Statistical programmers are always highly sought after (e.g., in finance, scientific research, the pharmaceutical industry, betting companies, online companies - Google, Facebook, Amazon etc.). So a basic knowledge of R is useful at this stage.

Note: Later in the year there will be an R based assignment (10%).

How to Install R

1. Go to http://cran.r-project.org/ and click on “Download R for Windows” at the top of the main page.
2. On the next page click “install R for the first time”.
3. At the top of the next page click “Download R for Windows”.
4. Run the downloaded executable file to install R on your computer.

Using R - Basic Example

Now that you have installed R, click on the R icon to open it.

Once open, click on “File” in the top left corner and then “New script”.

Copy and paste the code below into the script that you have opened:

```
x = c(1,1,2,4,3,2,1,4,5,3,6,9,1,2,15)
mean(x)
sd(x)
```

Within this script file in R, highlight the copied code. Press “Ctrl + R” to run it.

This gives the *mean* and *standard deviation* (more on these later) for the vector of numbers stored in x - you should get 3.933333 and 3.788454 in the R console.

Using R - More Information

If you wish to learn more about R, there are many options:

- Within your R script you can use the "?" command to find out more about a given function, e.g., running the code `?mean` will tell you about the `mean` function.

- At the top of the R window you will see a "help" menu. Here you can find information about various aspects of R. In particular, under the heading "Manuals (in PDF)", the "An Introduction to R" and "R Reference Manual" are useful.

- There is extensive information about R online, e.g., google "R tutorial" or "R beginners guide" etc. There are also R help forums where many solutions to common problems can be found.

Population of Interest

- Statistics is the collection and analysis of data.
- Based on our analysis we make conclusions about a **population** of interest.
- These conclusions then allow us to make *informed* decisions.

For example, let's say we are interested in the average income of a recent UL graduate (1-3 years since graduation say).

The **population** is *all* previous UL students who graduated in the last 3 years.

Can we contact every individual in the population?

Representative Sample

Can we contact every individual in the population? - **No!** This is very rarely possible. Even if it was possible, it is unnecessary, time-consuming and expensive. We can understand the population *without* seeing it in its entirety.

Instead we work with a **sample** of individuals from the population of interest; we may contact 100 recent graduates for example.

Of course, we must be careful about how we collect our sample. It must be **representative** of the population in question.

For example, if we only asked computer science graduates about their income level, our sample would not represent the specified population - *all graduates* - leading to biased results.

Random Sampling

We must use a **random sampling** method to ensure that a representative sample is selected $\Rightarrow$ **unbiased results**.

Random sampling is any method whereby all individuals in the population have an equal chance of being selected.

For example, let's assume that 10,000 students have graduated in the last 3 years. We can assign a number to each graduate (1-10,000) and then use a random number generator to select 100 numbers in the range 1-10,000. This produces a random sample of 100 graduates.

In R this can be achieved via `sample(1:10000,size=100)`.

Of course it is also possible using C or Java.

Parameter Vs Statistic

We are interested in some feature of the population (average income of a UL graduate from our previous example).

The true value of this feature is known as the **parameter**, i.e., the value based on the *whole* population.

The parameter value is *unknown* and must be *estimated from the sample*.

Our estimate of the parameter is called the **statistic**, i.e., the value calculated using our sample. For example, the average income in our sample of 100 graduates.

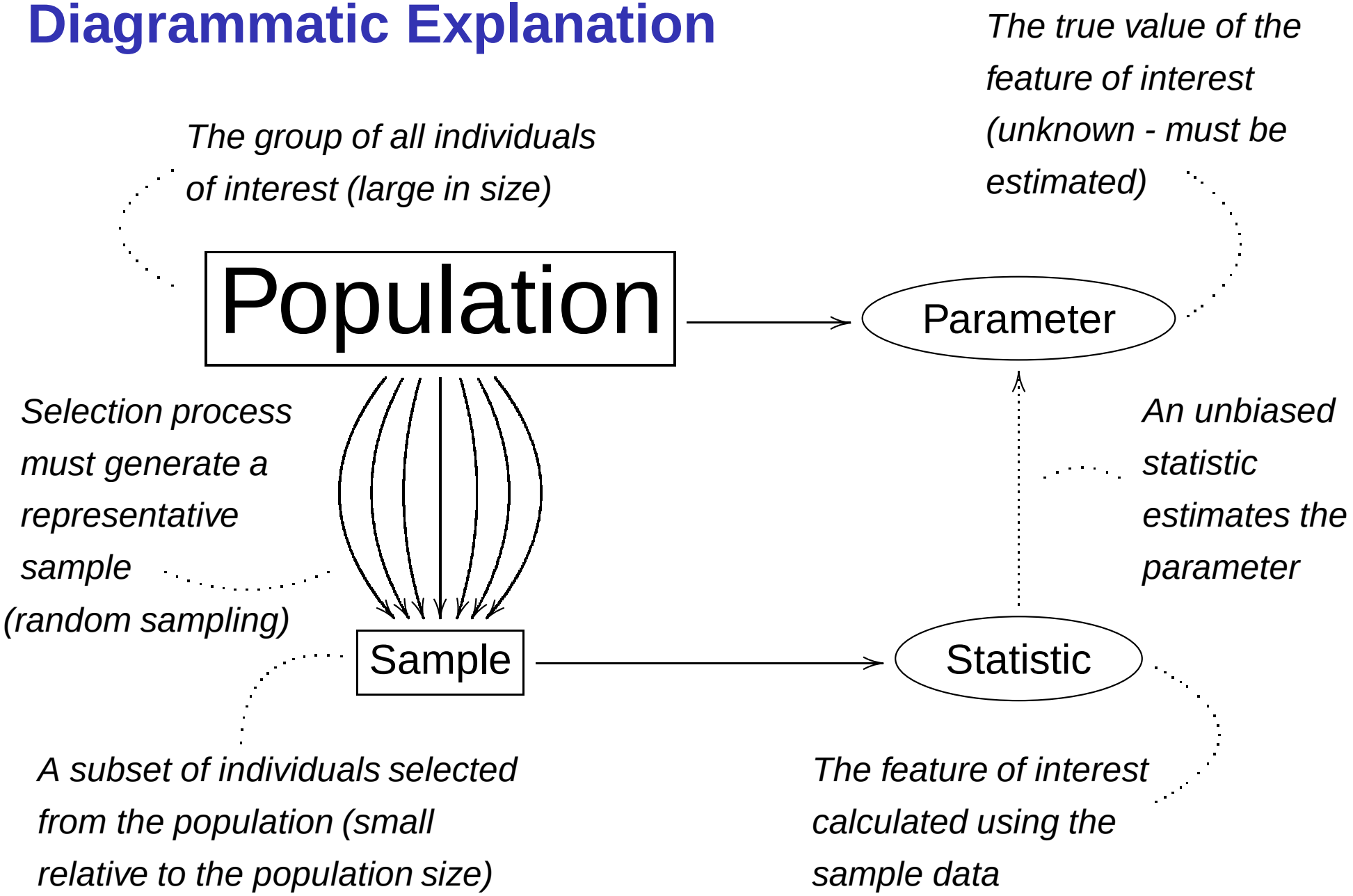
Memory Aid: *"P" is for population and parameter.*
"S" is for sample and statistic.

Parameter Vs Statistic: Symbols

It is important to know the symbols used to denote particular features of interest. In this course we deal with **proportions** and **means**.

- **Proportion**
  - *Examples*: the proportion of unemployed individuals, of individuals in favour of some government policy, of viruses classed as "high-threat", of times a user wins in online poker etc.
  - *Parameter:* The population proportion is $\boldsymbol{p}$.
  - *Statistic:* The sample proportion is $\boldsymbol{\hat{p}}$ (pronounced "p-hat").
- **Mean** (i.e., the arithmetic average)
  - *Examples:* the mean income of UL graduates, annual rainfall, lifetime of a laptop, age of users of some Android application, number bugs in some piece of software etc.
  - *Parameter:* The population mean is μ (the Greek letter "mu").
  - *Statistic:* The sample mean is $\boldsymbol{\bar{x}}$ (pronounced "x-bar").

Diagrammatic Explanation



Question 1

A manager wants to estimate the proportion of faulty resistors produced (in a particular week). Individual units are selected at random times during the morning shift of each day and then tested for faults. In total 1520 resistors were tested and 18 if these were found to be faulty.

a) What is the population?
b) What is the sample?
c) What is the parameter? What symbol do we use? What is its value?
d) What is the statistic? What symbol do we use? What is its value?
e) Identify any potential bias.

Question 2

ITD wish to determine the duration of time that a UL student spends on Facebook each day. They send an email of enquiry to 500 students (by randomly selecting ID numbers) - 286 students respond. The mean time spent on Facebook in this sample was found to be 1.5 hours per day.

a) What is the population?
b) What is the sample?
c) What is the parameter? What symbol do we use? What is its value?
d) What is the statistic? What symbol do we use? What is its value?
e) Identify any potential bias.

Data Types

There are two main types of data (the second subdivides further into two groups):

1. **Categorical**
   - Labels / words which define various categories.

2. **Numerical**
   - **Discrete**: Only a limited number of values (usually integers).
   - **Continuous**: Any (decimal) value in a particular range.

Question 3: Classify the Data Type

- Your age in years (20, 21, 30 etc.)
- Temperature
- Opinion of maths (dislike, indifferent, like)
- Processor speed (gigahertz)
- Number of bugs in an application
- Employment status (unemployed, employed, retired)
- Gender (male, female)
- Time taken to process some task
- Distance
- Paying attention in class (yes, no)
- File size (gigabytes)

Categorical $\Rightarrow$ Proportions. Numerical $\Rightarrow$ Means.

Recall: the main features we deal with are **proportions** and **means**.

- Categorical data: calculate the *proportion* of each category. For example, consider the variable "paying attention in class" with two categories - "yes" and "no". We calculate the proportion of individuals paying attention and the proportion not paying attention.

- Numerical data: calculate the *mean*. For example, consider the variable "income of a UL graduate". We calculate the mean income.

Note: we can also split a numerical variable by a categorical variable and *compare* the means in each group, e.g., mean income for UL graduates who got a 1.1 degree versus those who got a 2.1.

Visualising Data

We would like to "see" the data. This is more helpful than attempting to eyeball the individual values - especially if we have collected a large sample.

In particular we would like to discover the **distribution of data** which describes how various categories or values are *distributed*, i.e., how likely they are to occur.

The type of data determines the type of graph:

- Categorical data: **Bar chart**
- Numerical data: **Histogram**

Visualising Categorical Data

We first count the number of entries in each category - the frequencies - and construct a **frequency table**.

A **bar chart** is simply a graph with the frequencies (or relative frequencies) on the y-axis and the category labels on the x-axis.

Consider the following example:

In 2011 a market researcher carried out an online survey with the intention of discovering the market share of various mobile devices. Participants were asked tick a box indicating the mobile device that they use: “Android”, “Apple”, “BlackBerry”, “Windows” or “Other”. In total 500 individuals were surveyed and a frequency table was constructed (see next slide).

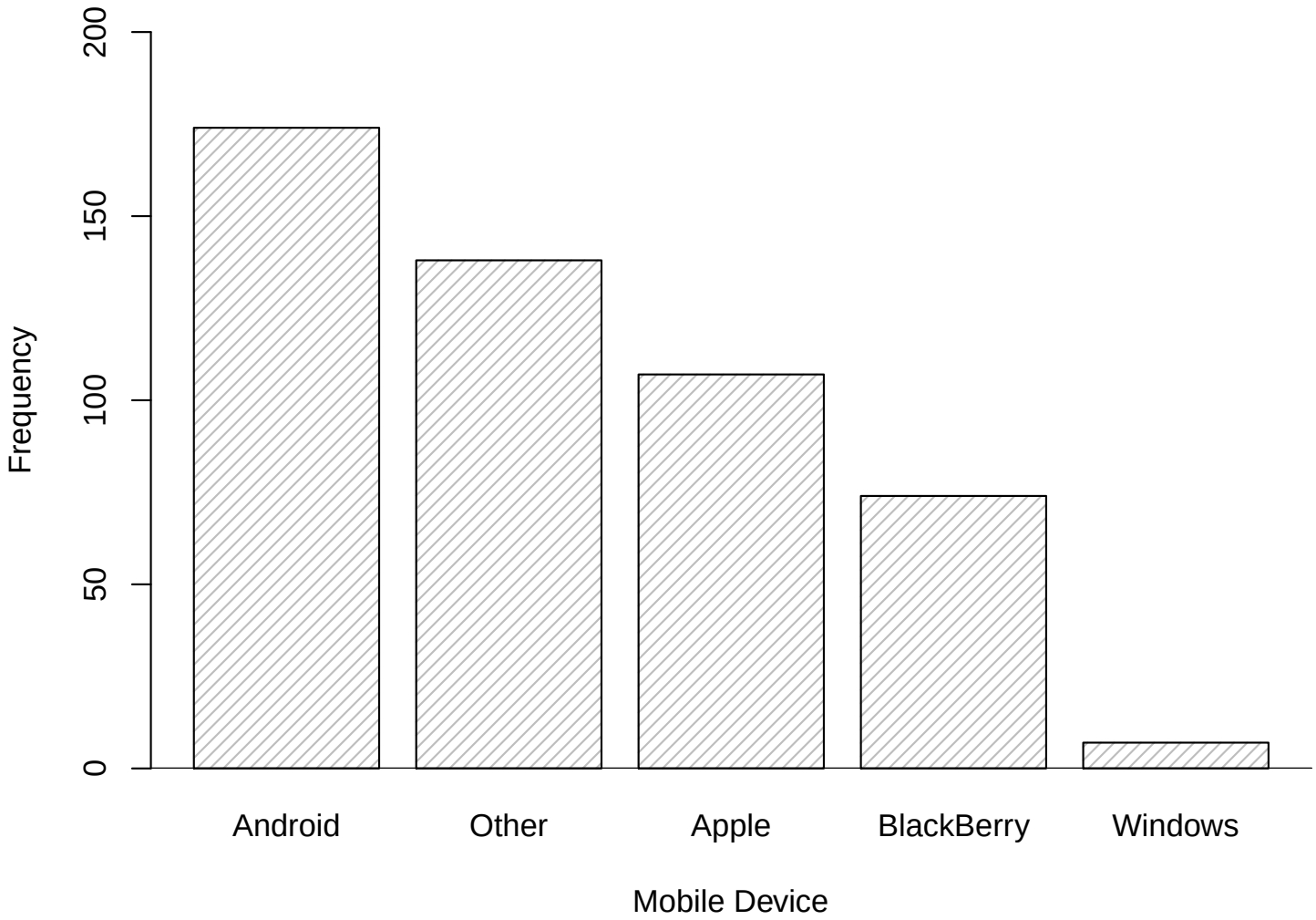
Categorical Data: Frequency Table

Market Share 2011: Ordered highest to lowest frequency

Category	Frequency	Relative Frequency
Android	174	$\frac{174}{500} = 0.348$
Other	138	$\frac{138}{500} = 0.276$
Apple	107	$\frac{107}{500} = 0.214$
BlackBerry	74	$\frac{74}{500} = 0.148$
Windows	7	$\frac{7}{500} = 0.014$
Total:	$n = 500$	$\frac{500}{500} = 1.000$

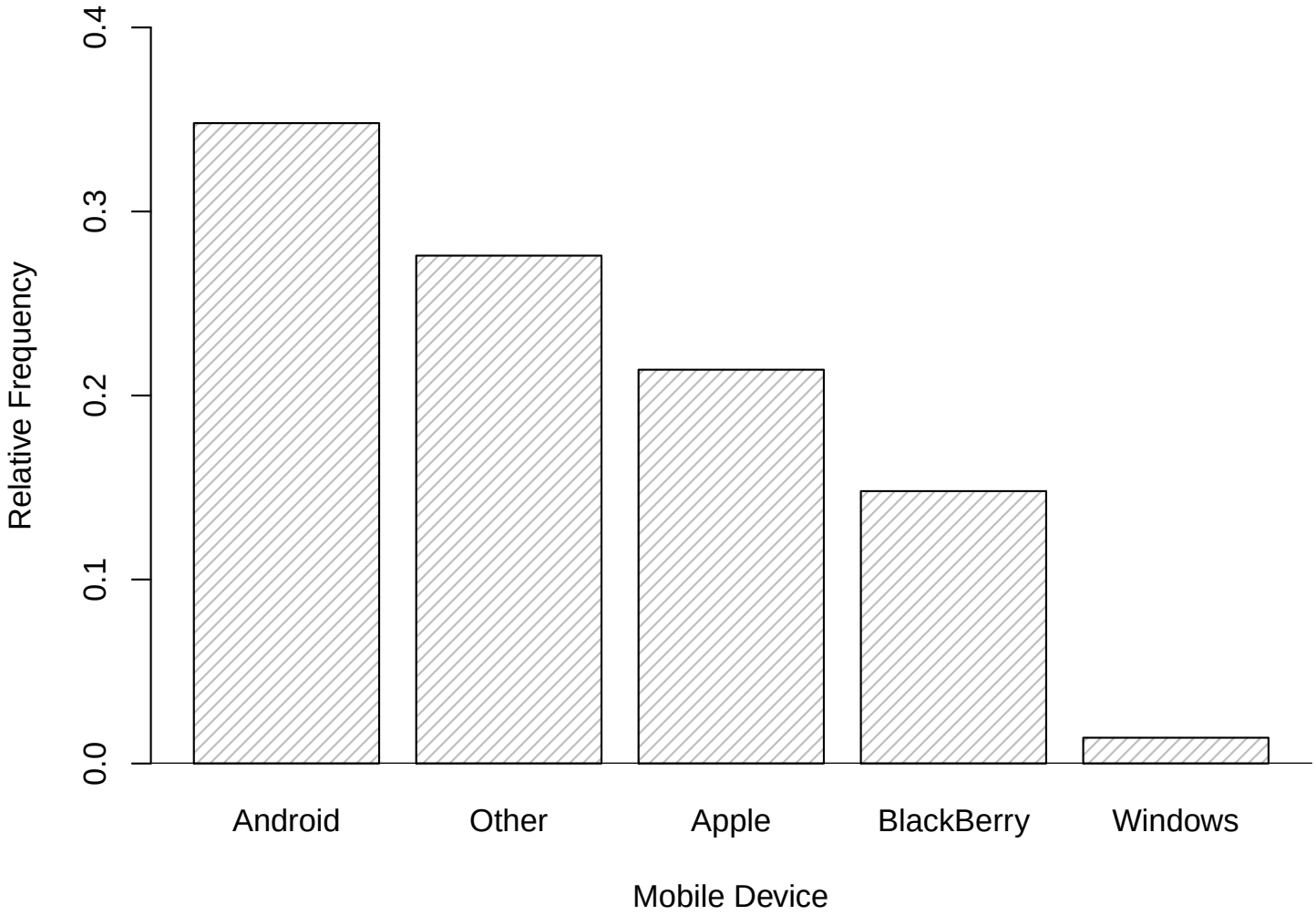
- The symbol for the total sample size is ***n*** - we will use this throughout the course.
- The relative frequencies (or proportions) add to 1.00. Also, these serve as estimates of the true *population proportions*.

Categorical Data: Bar Chart (Frequency)



- Note that there are **gaps** between the various categories.

Categorical Data: Bar Chart (Relative Frequency)



- Same picture but using relative frequency (see y-axis).

R Code: Bar Chart

The R code used to draw a bar chart is:

```
freq = c(174,138,107,74,7)
mobile = c("Android","Other","Apple","BlackBerry",
           "Windows")
barplot(freq, names=mobile)
```

You should *always* label the axes:

```
barplot(freq, names=mobile, xlab="Mobile Device",
        ylab="Frequency")
```

Some aesthetic improvements:

```
barplot(freq, names=mobile, xlab="Mobile Device",
        ylab="Frequency",density=20)
abline(h=0)
```

Run `?barplot` for more details.

Question 4

The following year (2012) a survey found that 359 individuals used Android, 81 used Apple, 18 used BlackBerry, 18 used Windows and 24 used other devices.

a) What is the value of n?

b) Construct a frequency table (ordered highest to lowest frequency) and include a column with relative frequencies.

c) Estimate the proportion of individuals who use either Android or Apple devices.

d) Estimate the proportion of individuals who use other devices. What symbol would we use for this proportion?

e) What is the *true* proportion of individuals who use other devices? What symbol would we use for this proportion?

f) Draw the bar chart.

g) Comment on how the market has changed since 2011.

Visualising Numerical Data

We first group the values into *classes* (effectively converting the data into categorical data) which allows us to construct a **frequency table**.

A **histogram** is simply a graph with the frequencies (or relative frequencies) on the y-axis and the class *breakpoints* on the x-axis.

Let the following set of numerical data represent the ages of $n = 30$ customers of a particular service:

43	42	62	29	28	29	44	44	56	21
32	29	33	61	43	27	53	32	35	39
47	51	50	33	38	34	42	37	21	35

We will group the above into the following classes:

$$19 - 27.9,\ 28 - 36.9,\ 37 - 45.9,\ 46 - 54.9 \text{ and } 55 - 63.9.$$

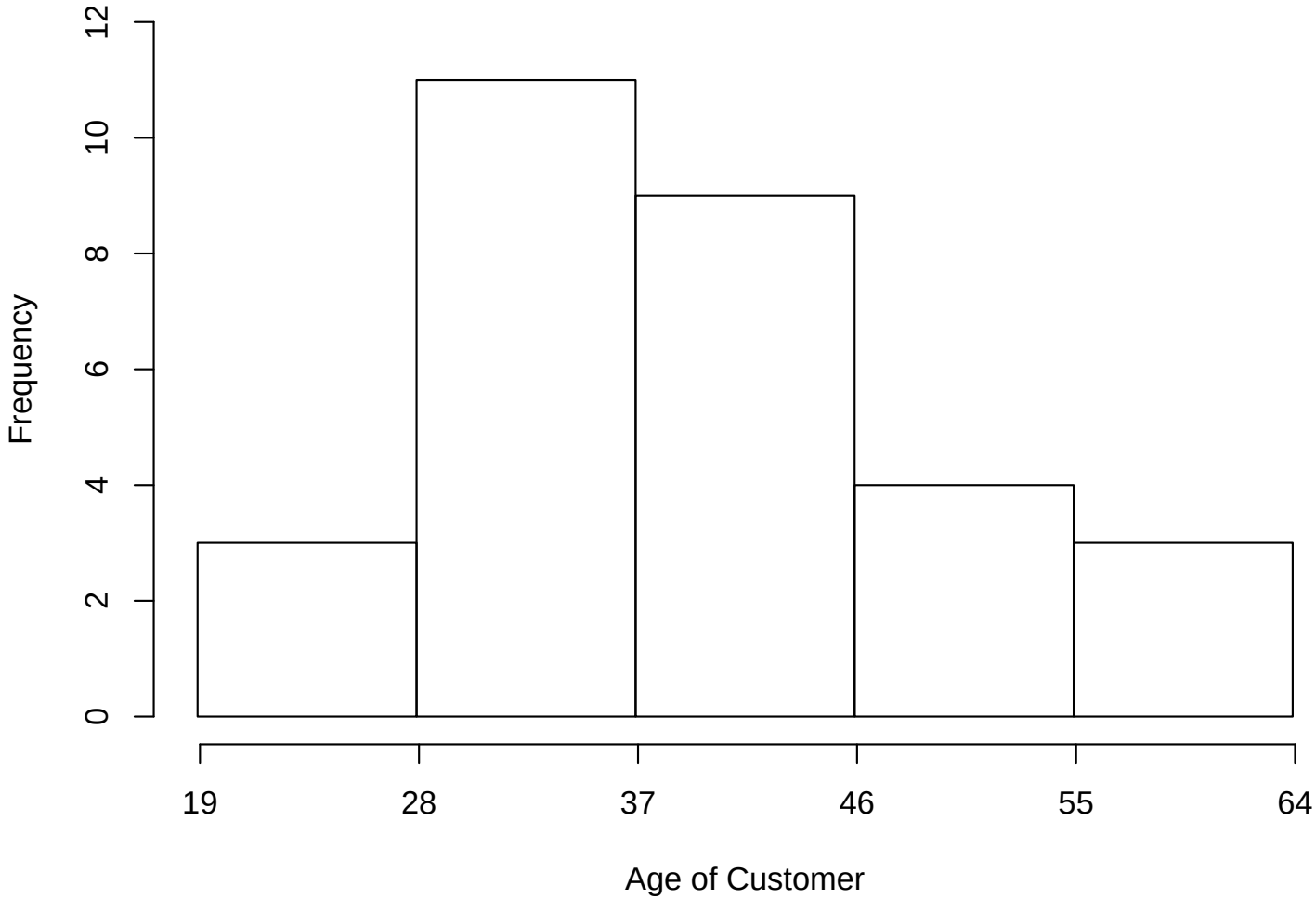
We then simply count the number of values contained in each class.

Numerical Data: Frequency Table

Class	Frequency	Relative Frequency
19 - 27.9	3	$\frac{3}{30} = 0.100$
28 - 36.9	11	$\frac{11}{30} = 0.367$
37 - 45.9	9	$\frac{9}{30} = 0.300$
46 - 54.9	4	$\frac{4}{30} = 0.133$
55 - 63.9	3	$\frac{3}{30} = 0.100$
Total:	$n = 30$	$\frac{30}{30} = 1.000$

- Note that we **do not reorder the table** from highest to lowest frequency here because the *classes have a natural order* already - going from smallest to largest ages.

Numerical Data: Histogram



- Note that there are **no gaps** between the classes.
  (this differs from a bar chart where the groups *are* separated)

Constructing the Classes

1. Decide on the number of classes:
   - Typically between 5 and 20 classes.
   - $\sqrt{n}$ is often a good choice.

In our example $n = 30$, so $\sqrt{30} = 5.48$ (we chose 5 classes).

2. Calculate the class width:
   - Formula: $\boxed{\text{width} = \dfrac{\max(x) - \min(x)}{\text{number of classes}}}$.
   - Always **round up** this value (if it is not a whole number).

In our example $\max(x) = 62$ and $\min(x) = 21$. So width is $(62 - 21)/5 = 41/5 = 8.2 \Rightarrow$ rounded up to 9.

Constructing the Classes

3. Calculate the total class range and choose the first breakpoint:
   - total class range $=$ number of classes $\times$ class width.
   - We choose the first breakpoint such that the minimum and maximum data values are covered by this total class range.

In our example the number of classes $= 5$ and width $= 9$. So the total class range $= 5 \times 9 = 45$.

If we chose the value 0 as the first breakpoint then the last breakpoint is $0 + 45 = 45$ giving a span of 0 - 45. Or we could choose 10 - 55. Or 15 - 60. None of these work as the span must contain the minimum and maximum data values: 21 and 62.

Choices that work: 18 - 63, 19 - 64, 20 - 65, 21 - 66.

In our example we chose 19 - 64.

Constructing the Classes

4. Construct the classes and count the number of data points contained in each.
   - **Every data point belongs to *only one* class.**

In our example we have the first breakpoint $=$ 19 and class width $=$ 9. So the first class goes from 19 up to 19 + 9 = 28. This interval means 19 up to but *not including* 28. So we say 27.9 to make this clear. The next class is then 28 up to 28 + 9 = 37 $\Rightarrow$ 36.9.

Thus, the classes are:

- 19 - 27.9
- 28 - 36.9
- 37 - 45.9
- 46 - 54.9
- 55 - 63.9

Counting the number of data points contained in these classes gives the frequency table previously shown.

R Code: Histogram

The R code used to draw the histogram is:

```
x = c(43, 42, 62, 29, 28, 29, 44, 44, 56, 21, 32,
      29, 33, 61, 43, 27, 53, 32, 35, 39, 47, 51,
      50, 33, 38, 34, 42, 37, 21, 35)
breakpts = c(18.9, 27.9, 36.9, 45.9, 54.9, 63.9)
hist(x, breaks=breakpts)
```

We can retrieve the frequencies for each class (to create the frequency table) as follows:

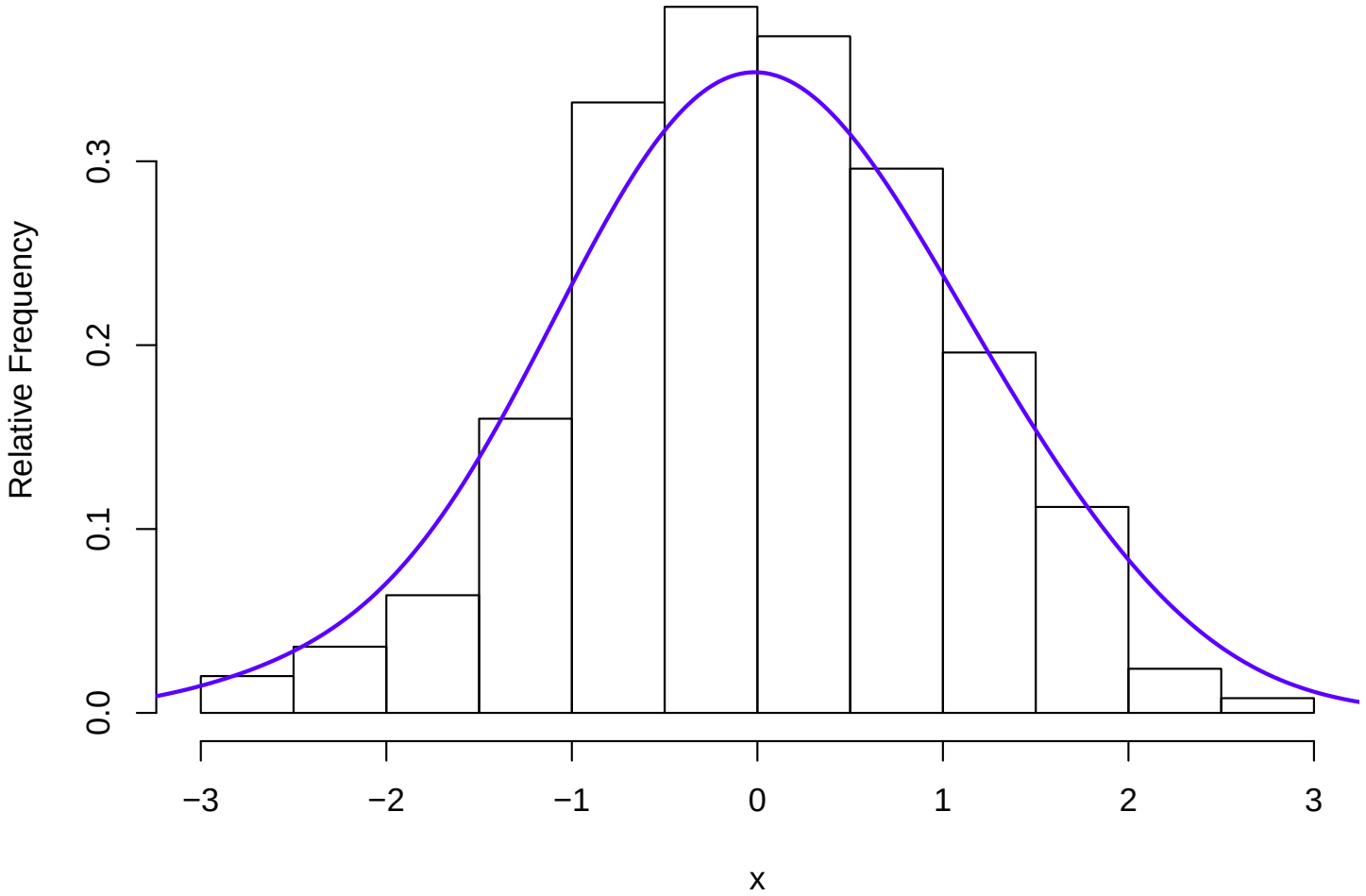
```
hist(x, breaks=breakpts)$counts
```

If we do not specify breakpoints, R does it automatically:

```
hist(x)
```

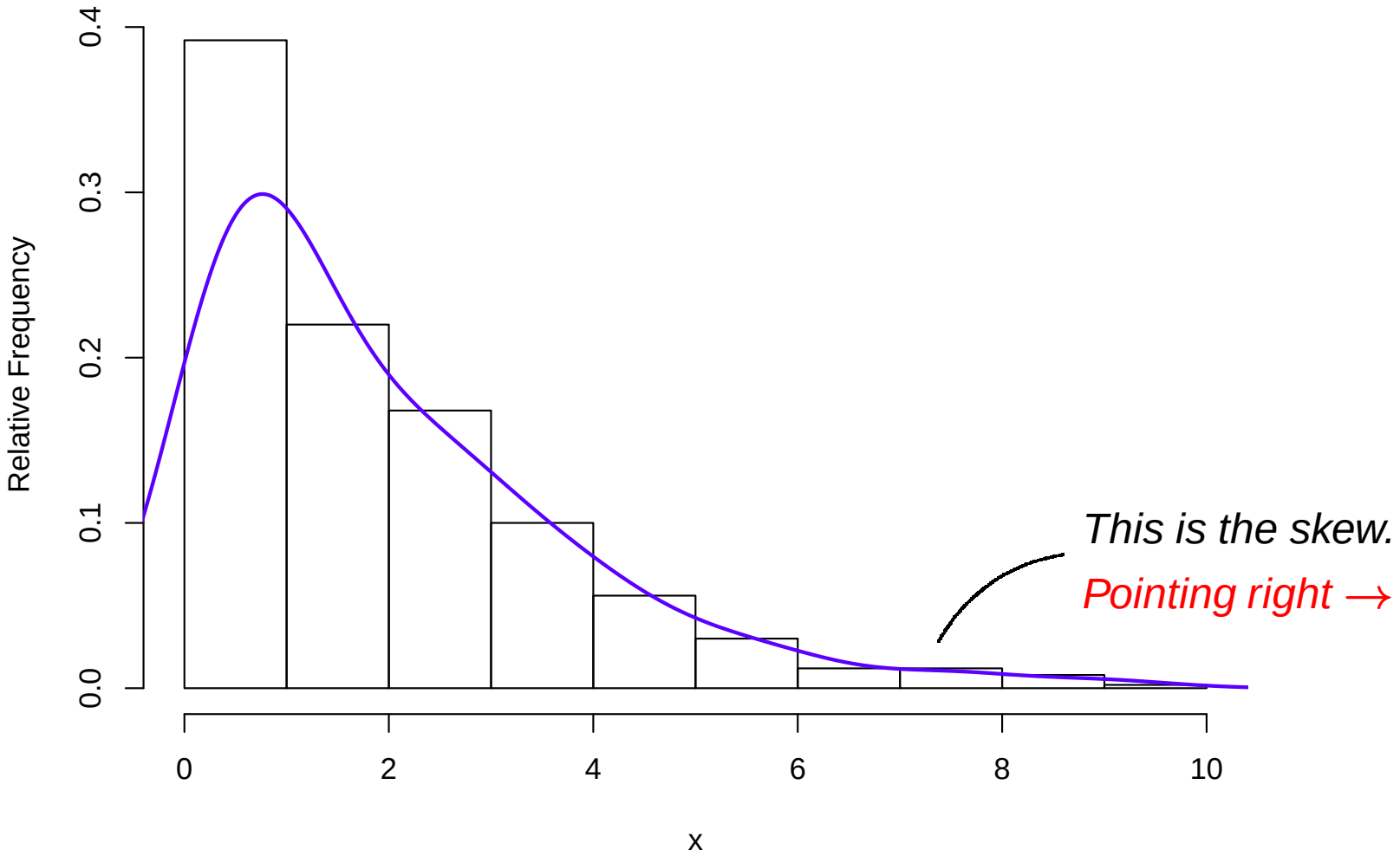
Run `?hist` for more details.

Histogram Shape: Symmetrical



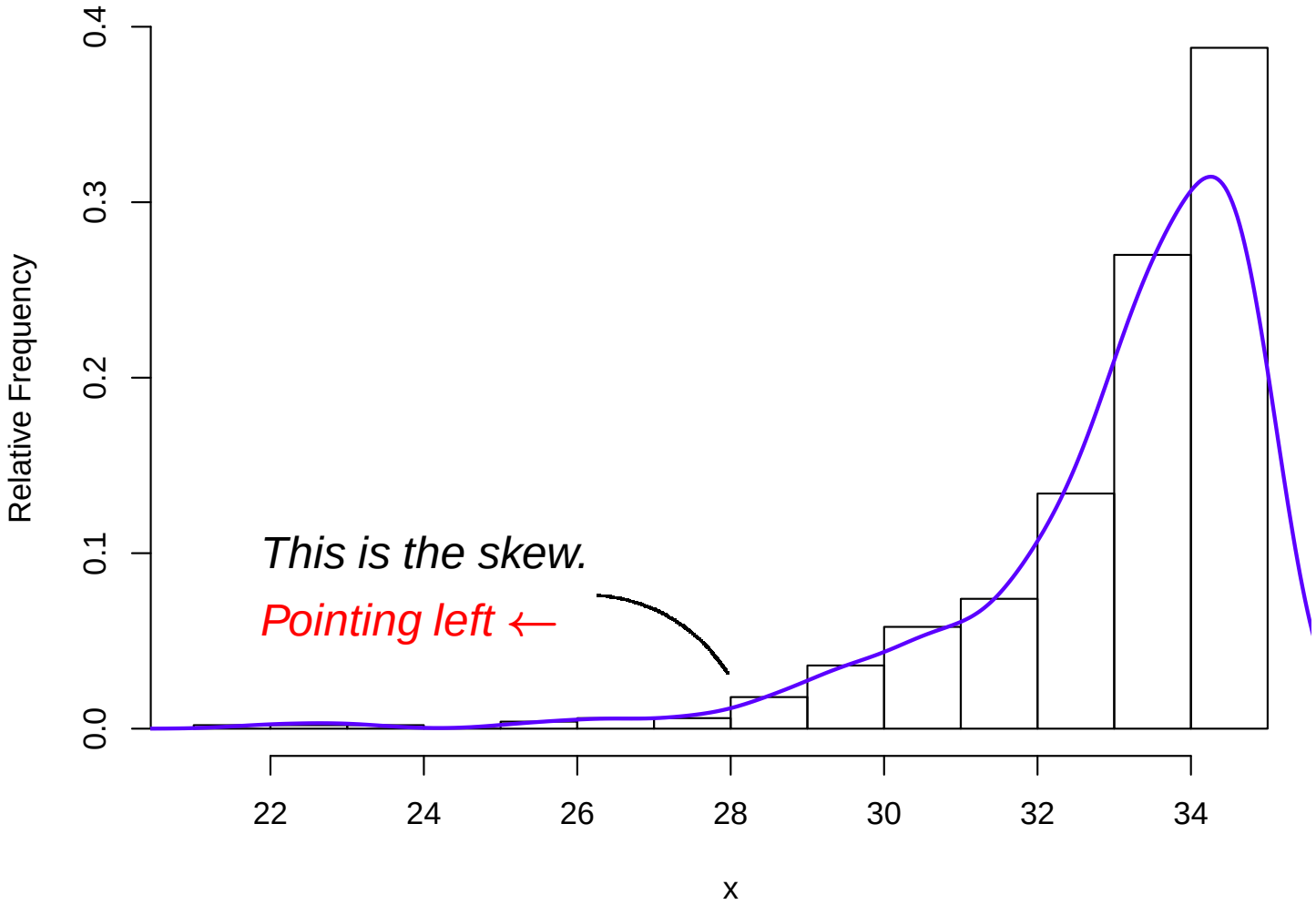
- Data symmetrical about the centre.

Histogram Shape: Skewed to the Right



- A few values **larger** than the main body of data.

Histogram Shape: Skewed to the Left



- A few values **smaller** than the main body of data.

Question 5

25 individuals were asked how long their laptop lasts on a full charge. The recorded times (measured in hours) are as follows:

2.2	0.4	4.2	12.9	1.5	3.0	5.7	0.7	1.0	3.3
0.2	0.2	5.6	1.6	3.0	0.1	14.3	3.4	0.9	6.1
1.4	1.0	0.7	5.4	2.3					

a) What is the value of n? What is the value of $\bar{x}$?

b) Construct a frequency table with 5 classes and let zero be the first breakpoint. (Note: the fact that the number of classes and first breakpoint are given simplifies the question)

c) Include a column with relative frequencies.

d) Estimate the proportion of laptops that last more than 6 hours.

e) This estimated proportion is called a statistic - what is the true proportion called? What is its value?

f) Comment on the shape of the histogram.