Chapter 16

Hypothesis testing

Contents:

A Statistical hypotheses

B Student's t-test

C The two-sample t-test for comparing population means

D The χ^2 goodness of fit test

E The χ^2 test for independence

OPENING PROBLEM

Frank is a market gardener who grows tomatoes.

Last year, Frank weighed a sample of 50 tomatoes and found that the mean weight was 106.3 g with standard deviation 12.41 g.

This year, Frank used a new fertiliser to try to increase the weight of his crop. Frank collected a random sample of 65 tomatoes and found the mean weight of the sample to be 110.1 g with standard deviation 13.1 g.

Things to think about:

a How can Frank use this information to determine whether the new fertiliser was effective?

b How *significant* is Frank's evidence? Is it sufficient to conclude that the fertiliser was effective? Would it be more significant if the sample size was bigger?

We often hear claims about **population parameters** such as the **population mean** μ or a **population proportion** p.

For example, a manufacturer of insect repellent might claim that on average, their new product is effective for longer than 6 hours. In statistics, we call this a **statistical hypothesis**.

We can decide whether a statistical hypothesis is reasonable or justified using a formal procedure called a **hypothesis test**.

Hypothesis tests have three key components:

- **formulating** statistical hypotheses
- collecting data from a sample and **calculating statistics** to test our hypotheses
- **making decisions** about the population based on what we see in the sample.

A STATISTICAL HYPOTHESES

Suppose a claim is made that a population mean μ has the value μ_0. We call this the **null hypothesis** H_0, and we write

$$H_0: \ \mu = \mu_0.$$

This statement is assumed to be true unless we have enough evidence to reject it.

If H_0 is not rejected, we accept that the population mean is μ_0. So, the null hypothesis is a statement that there is *no difference* between μ and μ_0.

If H_0 is rejected, we accept that there *is a difference* between μ and μ_0. This statement is called the **alternative hypothesis** H_1.

ONE-TAILED AND TWO-TAILED ALTERNATIVE HYPOTHESES

Given the null hypothesis $H_0: \ \mu = \mu_0$, the alternative hypothesis could be:

- $H_1: \ \mu > \mu_0$ (**one-tailed hypothesis**)
- $H_1: \ \mu < \mu_0$ (**one-tailed hypothesis**)
- $H_1: \ \mu \neq \mu_0$ (**two-tailed hypothesis**, as $\mu \neq \mu_0$ could mean $\mu > \mu_0$ or $\mu < \mu_0$).

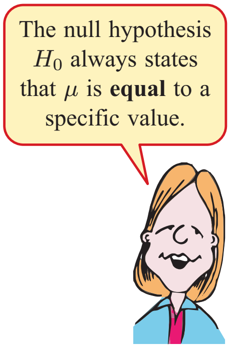
Consider the insect repellent example on the previous page:

- If the manufacturer of the new brand wants evidence that the new product is *superior* in protection time, the hypotheses would be:

  H_0: $\mu = 6$ {the new product gives the same protection as the old ones}
  H_1: $\mu > 6$ {the new product protects for longer than the old ones}.

- If a competitor wants evidence that the new product has an *inferior* protection time, the hypotheses would be:

  H_0: $\mu = 6$ {the new product gives the same protection as the old ones}
  H_1: $\mu < 6$ {the new product protects for less time than the old ones}.


The null hypothesis H_0 always states that μ is **equal** to a specific value.

- If a market researcher studying all products on the market wants to show that the new product *differs* from the old ones, but is not concerned whether the protection time is more or less, the hypotheses would be:

  H_0: $\mu = 6$ {the new product gives the same protection as the old ones}
  H_1: $\mu \neq 6$ {the new product gives a different protection time compared with the old ones}.

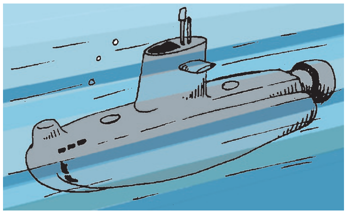
EXERCISE 16A

1 Current torch globes have a mean life of 80 hours. Globe Industries are considering mass production of a new globe they believe will last longer.

a If Globe Industries wants to demonstrate that their new globe lasts longer, what set of hypotheses should they consider?

b The new globe costs less to produce, so Globe Industries will adopt it unless it has an inferior lifespan to the old type. What set of hypotheses should they now consider?

2 The top speed of submarines currently produced by a manufacturer is 26.3 knots. When their engineers modify the design to reduce drag, they believe that the maximum speed will be increased. What set of hypotheses should they consider to test whether or not the new design is faster?



3 A machine is used to fill bags with 250 g of potato chips. A quality inspector wants to determine whether the machine is filling the bags with the *correct* weight of potato chips. What set of hypotheses should the quality inspector consider?

4 Whitex produce copy paper, and the weight of their copy paper is given as 80 g per m^2. The company wants to determine whether this information is correct. What set of hypotheses should be considered?

5 The average peak-hour travel time along a particular stretch of road is currently 27 minutes. To help reduce travel times, electronic signs displaying real-time information are erected. If the travel times improve, the signs will be more widely implemented. What set of hypotheses should be considered?

6 Brand A's muesli bars have 3 g of fat. Brand B claims that their muesli bars have 10% less fat than Brand A's muesli bars. Brand A wants evidence that Brand B's muesli bars have more fat than is claimed. What set of hypotheses should they consider?

B STUDENT'S t-TEST

In order to test our statistical hypotheses, we need *evidence* to base our decision on. The evidence comes from collecting data from a **sample** and then calculating **statistics**.

HISTORICAL NOTE

William Sealy Gosset (1876 - 1937) studied chemistry and mathematics at New College, Oxford University. In 1899 he moved to Dublin, Ireland, to work for the brewery of Arthur Guinness & Son. His desire was to improve the production process by selecting the best yielding varieties of barley. Through his study and work with **Karl Pearson** from University College, London, he devised a **test statistic**. His work was published in 1908 in the journal *Biometrika*, but using the name **Student** because of concern by Guinness that other brewers may use his work to their advantage.

Gosset's test statistic was revised by **Sir Ronald Aylmer Fisher** (1890 - 1962) who recognised the importance of Gosset's work. Fisher called the new statistic t, completing the name **Student's t-test**.

William Sealy Gosset

THE TEST STATISTIC OR t-STATISTIC

A **test statistic** summarises the information in a sample.

Consider a hypothesis test of H_0: $\mu = \mu_0$. If H_0 is true then we would expect the difference between the mean of the sample $\overline{x}$ and μ_0 to be close to 0. So a suitable test statistic should involve $\overline{x} - \mu_0$.

However, $\overline{x} - \mu_0$ alone does not take into account the *variation* of the data. If the standard deviation is also very small, a value of $\overline{x} - \mu_0$ close to 0 is likely to happen simply by chance alone. The test statistic should therefore also involve the sample standard deviation s.

Consider a hypothesis test of H_0: $\mu = \mu_0$.

Given a sample of size n with sample mean $\overline{x}$ and sample standard deviation s, the **test statistic** is:

$$\boldsymbol{t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}}$$

For example, consider the insect repellent example from the previous Section. Suppose a researcher took a random sample of 50 bottles of the new product and found that the mean protection time was $\overline{x} = 6.12$ hours with standard deviation $s = 15$ minutes $= 0.25$ hours.

In this case, the test statistic is $t = \dfrac{6.12 - 6}{\frac{0.25}{\sqrt{50}}} \approx 3.39$.

In examinations, you will not be required to calculate the test statistic by hand. You can click on the icon to obtain instructions for your **graphics calculator**.

GRAPHICS CALCULATOR INSTRUCTIONS

INVESTIGATION 1 THE DISTRIBUTION OF THE TEST STATISTIC

Every random sample will be different. The sample mean $\overline{x}$ and sample standard deviation will vary between samples, so the value of the test statistic t will be different for every sample.

In this Investigation, we will explore the *distribution* of the t-values for samples from a normal distribution.

What to do:

1 Click on this icon to access a simulation that generates random samples of size n from a normal distribution. The test statistic t is calculated for each sample.

SIMULATION

a Set $n = 10$ and use the sliders to change the mean μ and standard deviation σ of the normal distribution. Describe the distribution of the t-values.

b Use the slider to increase the value of n. What do you notice?

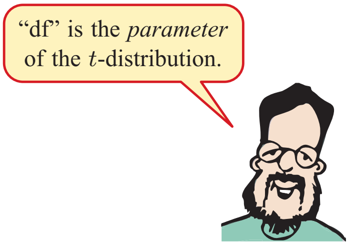
2 Describe how likely it is to obtain a t-value in the interval:

a $-1 \leqslant t \leqslant 1$ **b** $-2 \leqslant t \leqslant 2$ **c** $-3 \leqslant t \leqslant 3$

The characteristics you observed in the **Investigation** might suggest that the distribution of the t-values is a normal distribution with mean 0. However, the distribution of the t-values actually has a different curve.

Suppose a population is approximately normally distributed. The distribution of t-values for samples of size n is a ***t*-distribution** with $n-1$ **degrees of freedom** (df).

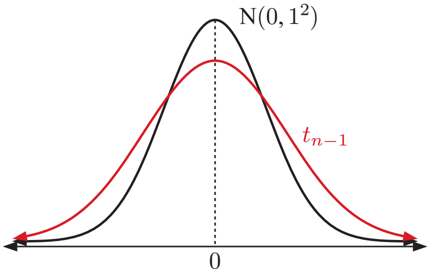
We write $\boldsymbol{T \sim t_{n-1}}$.



The t-distribution is slightly "flatter" than the normal $\text{N}(0,\,1^2)$ distribution.

As the sample size and therefore the degrees of freedom increases, the shape of the t-distribution approaches that of $\text{N}(0,\,1^2)$.

THE t-DISTRIBUTION



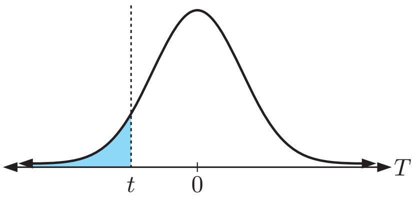
OBTAINING EVIDENCE

"Extreme" values of the test statistic are unlikely, so observing such a value is evidence against the null hypothesis. However, we need a measure of just *how* extreme a value is.

The ***p*-value** of a test statistic is the probability of that result being observed if H_0 is true.

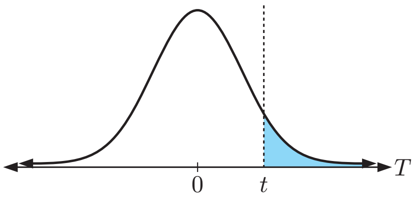
The meaning of “extreme” depends on whether the alternative hypothesis is one-tailed or two-tailed:

- If H_1 is a **one-tailed** alternative hypothesis, we use *one* tail of the t-distribution to calculate the p-value.
  - If H_1: $\mu < \mu_0$, we use the lower tail.

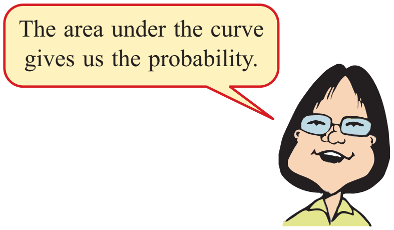


p-value $= \mathrm{P}(T \leqslant t)$

  - If H_1: $\mu > \mu_0$, we use the upper tail.



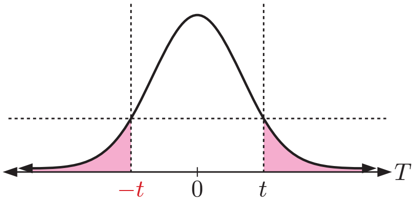
p-value $= \mathrm{P}(T \geqslant t)$



- If H_1 is the **two-tailed** alternative hypothesis H_1: $\mu \neq \mu_0$, then we must consider *both* tails of the t-distribution. We define “extreme” values as those which have probability less than or equal to that of the test statistic.

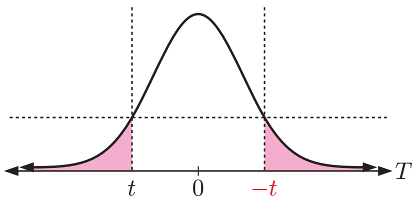
If $t \geqslant 0$,

p-value $= \mathrm{P}(T \geqslant t \;\text{ or }\; T \leqslant -t)$
$= 2 \times \mathrm{P}(T \geqslant t)$ {symmetry}



If $t < 0$,

p-value $= \mathrm{P}(T \geqslant -t \;\text{ or }\; T \leqslant t)$
$= 2 \times \mathrm{P}(T \geqslant -t)$ {symmetry}



So, for a two-tailed alternative hypothesis, **p-value $= 2 \times \mathrm{P}(T \geqslant |\,t\,|)$** .

Click on the icon for instructions on how to calculate probabilities for the t-distribution.

GRAPHICS CALCULATOR INSTRUCTIONS

MAKING DECISIONS

Although “extreme” values are unlikely to occur, it is still *possible* to observe such values in a sample. We therefore need a rule which defines how much evidence is required to reject the null hypothesis.

The **significance level** α of a statistical hypothesis test is the largest p-value that would result in rejecting H_0. Any p-value less than or equal to α results in H_0 being rejected.

If a statistical hypothesis test has significance level α, the probability of *incorrectly* rejecting H_0 is α.

The significance level may be given as a decimal or as a percentage.

In our insect repellent example, suppose the researcher was concerned about any change in the average protection time. The researcher decides to test the hypothesis H_0: $\mu = 6$ against H_1: $\mu \neq 6$ at a significance level $\alpha = 0.01$.

Using the test statistic $t \approx 3.39$ and $T \sim t_{50-1}$,

$$\text{the } p\text{-value} \approx 2 \times \mathrm{P}(T \geqslant |\,3.39\,|)$$
$$\approx 0.001\,37$$

Since the p-value is less than the significance level α, the researcher has enough evidence to reject H_0.

The smaller the p-value, the more evidence there is against H_0.

Important:

We **must** choose a significance level **before** we test the hypothesis. Otherwise, we can be tempted to choose a significance level to give the test outcome that we desire. For example, it is *not appropriate* to calculate a p-value and then select a value of α so that H_0 will be rejected.

SUMMARY OF STEPS FOR STUDENT'S t-TEST FOR A POPULATION MEAN

Step 1: State the **null hypothesis** H_0: $\mu = \mu_0$ and **alternative hypothesis** H_1.

Step 2: State the **significance level** α.

Step 3: Calculate the value of the **test statistic** $t = \dfrac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$.

Step 4: Calculate the ***p*-value** using $T \sim t_{n-1}$ as follows:

- If H_1: $\mu > \mu_0$, p-value $= \mathrm{P}(T \geqslant t)$.
- If H_1: $\mu < \mu_0$, p-value $= \mathrm{P}(T \leqslant t)$.
- If H_1: $\mu \neq \mu_0$, p-value $= 2 \times \mathrm{P}(T \geqslant |\,t\,|)$.

Step 5: Reject H_0 if p-value $\leqslant \alpha$.

Step 6: Interpret your decision in the context of the problem. Write your conclusion in a sentence.

Example 1

Self Tutor

The manager of a restaurant chain goes to a seafood wholesaler and inspects a large catch of over 50 000 prawns. It is known that the population is normally distributed. He will buy the catch if the mean weight exceeds 55 grams per prawn.

A random sample of 60 prawns is taken. The sample mean weight is 56.2 grams with standard deviation 4.2 grams.

Conduct a one-tailed hypothesis test with significance level $\alpha = 0.05$ to determine whether the manager should purchase the catch.

Step 1: Let μ be the population mean weight per prawn.
The hypotheses that should be considered are:
H_0: $\mu = 55$ {the mean weight is 55 grams per prawn}
H_1: $\mu > 55$ {the mean weight exceeds 55 grams per prawn}

Step 2: The significance level is $\alpha = 0.05$.

Step 3: $\overline{x} = 56.2$ grams, $s = 4.2$ grams, $n = 60$

The value of the test statistic is

$$t = \frac{56.2 - 55}{\frac{4.2}{\sqrt{60}}} \approx 2.21$$

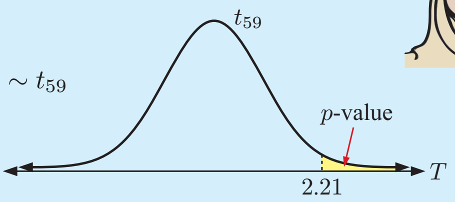
In examinations you can calculate t using your calculator.

Step 4: Since $H_1\colon\ \mu > 55$,

p-value $= \mathrm{P}(T \geqslant t)$ where $T \sim t_{59}$

$\approx \mathrm{P}(T \geqslant 2.21)$

≈ 0.0154



Step 5: Since p-value $< 0.05 = \alpha$, we have enough evidence to reject H_0 in favour of H_1 on a 5% significance level.

Step 6: Since we have accepted H_1, we conclude that the mean weight exceeds 55 grams per prawn. The manager should purchase the catch.

Example 2 **Self Tutor**

The fat content (in grams) of 30 randomly selected pasties at the local bakery was recorded:

15.1	14.8	13.7	15.6	15.1	16.1	16.6	17.4	16.1	13.9
17.5	15.7	16.2	16.6	15.1	12.9	17.4	16.5	13.2	14.0
17.2	17.3	16.1	16.5	16.7	16.8	17.2	17.6	17.3	14.8

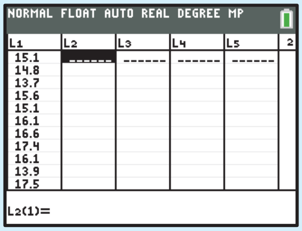
For a mean fat content of pasties made at this bakery μ, conduct a two-tailed t-test of $H_0\colon\ \mu = 16$ grams on a 10% level of significance.

Step 1: $H_0\colon\ \mu = 16$ {the mean fat content is 16 grams}

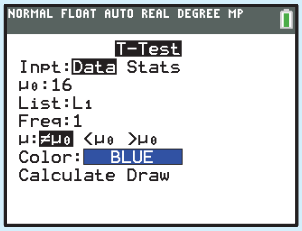
$H_1\colon\ \mu \neq 16$ {the mean fat content is *not* 16 grams}

Step 2: The significance level is $\alpha = 0.1$.

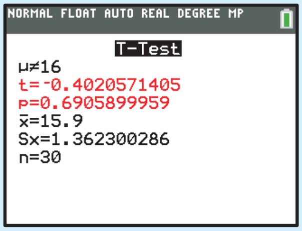
Steps 3 and 4:



```
NORMAL FLOAT AUTO REAL DEGREE MP
L1    L2      L3      L4      L5      2
15.1  ------  ------  ------  ------
14.8
13.7
15.6
15.1
16.1
16.6
17.4
16.1
13.9
17.5
L2(1)=
```

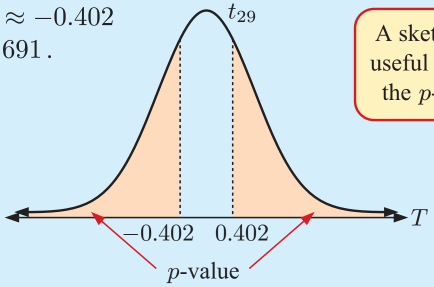


```
NORMAL FLOAT AUTO REAL DEGREE MP
          T-Test
 Inpt:Data Stats
 μ0:16
 List:L1
 Freq:1
 μ:≠μ0 <μ0 >μ0
 Color: BLUE
 Calculate Draw
```

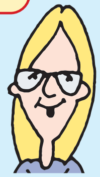


```
NORMAL FLOAT AUTO REAL DEGREE MP
          T-Test
 μ≠16
 t=-0.4020571405
 p=0.6905899959
 x̄=15.9
 Sx=1.362300286
 n=30
```

Using technology, $t \approx -0.402$ and the p-value ≈ 0.691.



A sketch of the curve is useful to remind us what the p-value represents.



Step 5: Since p-value $> 0.1 = \alpha$, we do not have enough evidence to reject H_0 in favour of H_1 on a 10% significance level. We therefore accept H_0.

Step 6: Since we have accepted H_0, we cannot conclude that the mean fat content is appreciably different from 16 grams.

EXERCISE 16B

1 A sample of size 36 is taken. The sample mean $\overline{x} = 23.75$ and the sample standard deviation $s = 3.97$. We are required to test the hypothesis H_0: $\mu = 25$ against H_1: $\mu < 25$.

a Find: **i** the test statistic **ii** the p-value.

b What decision should be made at a 5% level?

2 A statistician believes that a population has a mean μ that is greater than 80. To test this he takes a random sample of 200 measurements, and finds the sample mean is 83.1 and the sample standard deviation is 12.9. He then performs a hypothesis test with significance level $\alpha = 0.01$.

a Write down the null and alternative hypotheses.

b Find the value of the test statistic. **c** Calculate the p-value.

d Make a decision to reject or not reject H_0.

e State the conclusion for the test.

3 Bags of salted cashew nuts state their net contents is 100 g. A customer claims that the bags have been lighter in recent purchases, so the factory quality control manager decides to investigate. He samples 40 bags and finds that their mean weight is 99.4 g with standard deviation 1.6 g.

Perform a hypothesis test at the 5% level of significance to determine whether the customer's claim is valid.

4 An alpaca breeder wants to produce fleece which is extremely fine. In 2015, his herd had mean fineness 20.3 microns. In 2019, a sample of 80 alpacas from the herd was randomly selected, and the mean fineness was 19.2 microns with standard deviation 2.89 microns.

Perform a two-tailed hypothesis test at the 5% level of significance to determine whether the herd fineness has changed.

5 The length of screws produced by a machine is known to be normally distributed. The machine is supposed to produce screws with mean length $\mu = 2.00$ cm. A quality controller selects a random sample of 15 screws. She finds that the mean length of the 15 screws is $\overline{x} = 2.04$ cm with sample standard deviation $s = 0.09$ cm. Does this justify the need to adjust the machine on a 2% level of significance?

6

A machine packs sugar into 1 kg bags. It is known that the masses of the bags of sugar are normally distributed. A random sample of eight filled bags was taken and the masses of the bags measured to the nearest gram. Their masses in grams were:

1001, 998, 999, 1002, 1001, 1003, 1002, 1002.

Perform a test at the 1% level, to determine whether the machine under-fills the bags.

7 A market gardener claims that the carrots in his field have a mean weight of more than 50 grams. A prospective buyer will purchase the crop if the market gardener's claim is true. To test this she pulls 20 carrots at random, and finds that their individual weights in grams are:

57.6	34.7	53.9	52.5	61.8	51.5	61.3	49.2	56.8	55.9
57.9	58.8	44.3	58.3	49.3	56.0	59.5	47.0	58.0	47.2

a Explain why it is reasonable to assume that the carrots' weights are normally distributed.

b Determine whether the buyer will purchase the crop using a 5% level of significance.

INVESTIGATION 2 MULTIPLE TESTING AND STATISTICAL FALLACY

In many applications of hypothesis testing, it is common to conduct multiple identical or very similar hypothesis tests simultaneously. For example, in genetics an experiment called a **DNA microarray** is used to measure expression levels of thousands of genes, each with their own set of hypotheses to test.

In this Investigation, we will explore the effects of conducting multiple hypothesis tests on the interpretation of individual outcomes.

What to do:

1 A normally distributed population has mean μ and standard deviation $\sigma = 5$.
Consider the following hypotheses for this population: H_0: $\mu = 2$
H_1: $\mu \neq 2$

Click on the icon to run a computer simulation which generates samples of size 10 from the $\mathrm{N}(2,\, 5^2)$ distribution. The above hypotheses are tested for each sample at a significance level of α, and a p-value is calculated.

SIMULATION

a Write down the formula used to calculate the test statistic given a sample mean $\overline{x}$ and sample standard deviation s.

b Set $\alpha = 0.05$. Copy and complete the following table by generating m samples and counting the number of times H_0 is rejected.

m	*Number of times H_0 is rejected*	*Proportion of samples where H_0 was rejected*
20		
50		
100		
500		
1000		
5000		
10 000		

c Repeat **b** for:

i $\alpha = 0.1$ **ii** $\alpha = 0.025$ **iii** $\alpha = 0.01$.

Comment on your results.

2 For the simulation in **1**, explain why:

a H_0 is true in every sample that the simulation generates

b $\mathrm{P}(\text{incorrectly reject } H_0) = \alpha$ for each sample

c the *expected number* of samples where H_0 is incorrectly rejected is $m\alpha$.

3 Sabeen is a psychologist. She is writing a journal article about the effect of diet cola on a person's ability to concentrate. To account for possible confounding factors, Sabeen divides her data into 10 different age groups for each gender. For each age group and gender, she conducts a hypothesis test and obtains a p-value.

Sabeen's results are shown in the table below:

Age group	10 - 14	15 - 19	20 - 24	25 - 29	30 - 34
Male	0.296	0.143	0.305	0.378	0.169
Female	0.814	0.022	0.125	0.301	0.432

Age group	35 - 39	40 - 44	45 - 49	50 - 54	55+
Male	0.699	0.221	0.078	0.790	0.423
Female	0.987	0.643	0.448	0.672	0.789

a Which age group and gender do you think Sabeen is most likely to report on in her journal article? Explain your answer.

b Sabeen's colleague Mysha repeats Sabeen's experiment. She samples people exclusively from the group you identified in **a**. Do you think Mysha is likely to replicate Sabeen's results for this group? Explain your answer.

THEORY OF KNOWLEDGE

Sir Ronald Aylmer Fisher was an English statistician and biologist. He was known for his work in both agriculture and statistics, combining the disciplines with his work in classical statistics and significance testing.

In 1952 Fisher published a book titled *Statistical Methods for Research Workers* which is best known for the following statement about p-values:

> "*The value for which* $p = 0.05$, *or* 1 *in* 20 [....] *it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not.*"

Today, a significance level of $\alpha = 0.05$ is still widely quoted in scientific journals when testing for significance.

Sir Ronald Fisher

1 If someone tells you that they are 95% confident in something, do you normally stop to consider the 5% chance that the person is wrong? Do you think you *should*?

2 Suppose a researcher is writing a report for a medical journal.

a Is it realistic to expect the researcher to be 100% confident in their findings?

b What is *your* expectation of how confident a researcher should be, in order that they publish their results?

c Do you think it is important that the researcher tells you how confident they are?