

YEAR 1 STATISTICS · CHAPTER 7

Hypothesis Testing

Specification: Edexcel Paper 3 §5.1–5.2 — Hypothesis testing · AQA section O

KEY FACTS — THIS CHAPTER

- A hypothesis test asks whether a sample gives **enough evidence** to reject a claim about a population proportion p .
- H_0 (null) always says p **equals** a stated value. H_1 (alternative) says $p >$, $p <$ (one-tailed) or $p \neq$ (two-tailed) that value.
- Assume H_0 is true, so the test statistic $X \sim B(n, p_0)$, and ask how likely the observed result, **or something more extreme**, is.
- **p-value method:** if $P(\text{observed or more extreme}) \leq \alpha$, reject H_0 . For a two-tailed test at α , compare each tail with $\frac{\alpha}{2}$.
- **Critical region:** the set of values that would lead to rejecting H_0 . Its total probability must not exceed α (or $\frac{\alpha}{2}$ per tail).
- The **actual significance level** is the true probability of the critical region, which is usually less than α because X is discrete.
- Always finish **in context**, and never claim H_0 is proved.

Chapter 6 asked “given p , how likely is this result?” This chapter turns the question round: given a result, is it strange enough to doubt the claimed p ? The answer is never certainty — it is a judgement, made with a stated risk of being wrong.

1. The language

SETTING UP A TEST

1. Define the variable and the parameter: “let p be the proportion of ...” and “let X be the number of ...”.
2. State the hypotheses, both about p : $H_0 : p = p_0$ against $H_1 : p > p_0, p < p_0$ or $p \neq p_0$.
3. State the **test statistic** and its distribution **assuming** H_0 : $X \sim B(n, p_0)$.
4. State the **significance level** α (usually 5%, 10% or 1%).

The alternative hypothesis comes from the question. “Is the coin biased **towards heads**?” is one-tailed ($p > 0.5$). “Is the coin biased?” is two-tailed ($p \neq 0.5$).

COMMON MISTAKE

Hypotheses are about the **population** proportion p , never about the sample. Writing $H_0 : X = 10$ or “ H_0 : the coin is fair” without the symbols loses marks. And H_0 always has an equals sign.

2. The p-value method

Assume H_0 . Work out the probability of the observed result, **or anything more extreme in the direction of H_1** . That probability is the **p-value**.

KEY RESULT

$$\text{p-value} \leq \alpha \Rightarrow \text{reject } H_0, \quad \text{p-value} > \alpha \Rightarrow \text{do not reject } H_0$$

For a two-tailed test at level α , compare the one tail you are in with $\frac{\alpha}{2}$.

WORKED EXAMPLE 7.1 • A ONE-TAILED TEST

A coin is tossed 20 times and lands heads 15 times. Test, at the 5% level, whether the coin is biased towards heads.

Let p be the probability of heads and X the number of heads in 20 tosses.

$$H_0 : p = 0.5 \quad H_1 : p > 0.5$$

Assuming H_0 , $X \sim B(20, 0.5)$. The observed value is 15, and “more extreme” means more heads:

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.9793 = 0.0207$$

Since $0.0207 < 0.05$, the result is significant: **reject** H_0 .

IN CONTEXT

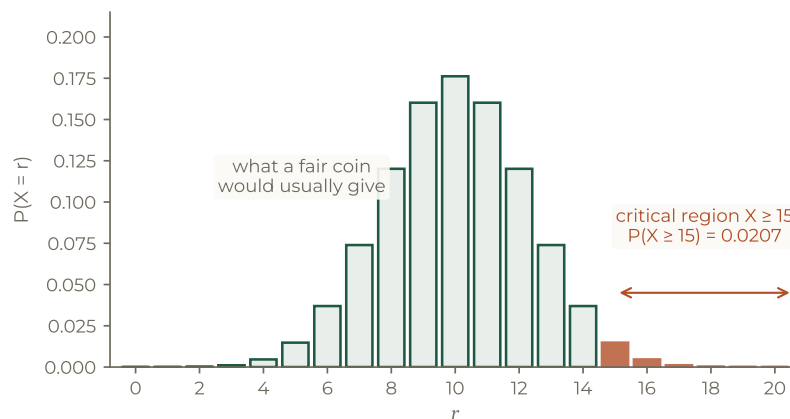
There is sufficient evidence at the 5% level to suggest that the coin is biased towards heads.

COMMON MISTAKE

$P(X = 15)$ is not the p-value. A test asks how unusual the result is, so it must include everything **at least as extreme**: $P(X \geq 15)$.

3. Critical regions

Instead of testing one result, you can work out in advance which results would lead to rejection. That set is the **critical region**.



For a 5% one-tailed test on $B(20, 0.5)$, the critical region is $X \geq 15$: the largest tail whose probability stays under 5%.

KEY RESULT

The **critical region** is the set of values of the test statistic for which H_0 is rejected. Choose it so its probability is **at most** α (or $\frac{\alpha}{2}$ in each tail of a two-tailed test).

The **critical value** is the first value inside the region, and everything else is the **acceptance region**.

WORKED EXAMPLE 7.2 · FINDING A CRITICAL REGION

Find the critical region for a one-tailed test of $H_0 : p = 0.5$ against $H_1 : p > 0.5$ at the 5% level, using $X \sim B(20, 0.5)$, and state the actual significance level.

Work down from the top until the tail first drops below 5%:

$$P(X \geq 14) = 0.0577 > 0.05, \quad P(X \geq 15) = 0.0207 < 0.05$$

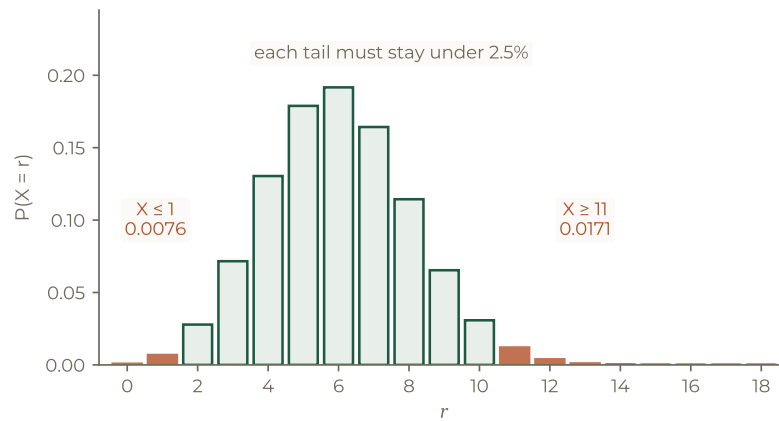
So the critical region is $X \geq 15$, and the critical value is 15.

IN CONTEXT

The **actual significance level** is 0.0207, or 2.07%: because X is discrete, no region has probability exactly 5%, so the test is more cautious than 5% suggests.

4. Two-tailed tests

A two-tailed test asks whether p has changed **in either direction**, so the significance level is split between the two tails.



A 5% two-tailed test allows 2.5% in each tail: here $X \leq 1$ (0.0076) and $X \geq 11$ (0.0171), an actual significance level of 2.48%.

WORKED EXAMPLE 7.3 • A TWO-TAILED TEST

A machine is claimed to produce faulty items 30% of the time. In a sample of 20 items, 2 are faulty. Test at the 5% level whether the rate has changed.

Let p be the proportion of faulty items and X the number faulty in 20.

$$H_0 : p = 0.3 \quad H_1 : p \neq 0.3$$

Assuming H_0 , $X \sim B(20, 0.3)$. The observation, 2, is low, so use the lower tail and compare with $\frac{0.05}{2} = 0.025$:

$$P(X \leq 2) = 0.0355 > 0.025$$

The result is not significant: **do not reject** H_0 .

IN CONTEXT

There is insufficient evidence at the 5% level to suggest that the rate of faulty items has changed from 30%.

WHY THE SAME NUMBER CAN PASS ONE TEST AND FAIL ANOTHER

$P(X \leq 2) = 0.0355$ is below 5% but above 2.5%. So 2 faulty items **would** be significant in a one-tailed test of $p < 0.05$, but is not significant in a two-tailed test at the same level. Asking a vaguer question (“has it changed?”) demands stronger evidence, which is why the tails must be decided *before* you see the data.

COMMON MISTAKE

Don’t halve the p-value; halve the **significance level**. And never choose the direction of H_1 after seeing which way the data went — that quietly doubles your chance of a false positive.

5. Writing the conclusion

Every test ends with two sentences: a statistical one and a contextual one.

THE TWO-SENTENCE ENDING

1. “The result is significant / not significant at the $\alpha\%$ level, so reject / do not reject H_0 .”
2. “There is sufficient / insufficient evidence at the $\alpha\%$ level to suggest that ...” — in the words of the question.

COMMON MISTAKE

Never write “ H_0 is true” or “the coin is fair”. Failing to reject H_0 only means this sample wasn’t unusual enough — which is not the same as proof. Keep the hedged wording: *insufficient evidence to suggest*.

IN CONTEXT

A 5% significance level means that, even when H_0 is true, about 1 test in 20 will reject it by chance. That is the risk you accept when you choose α .

Practice

Twelve questions in three tiers. Try each one properly before opening the solution. Every test needs the same skeleton: define p and X , state H_0 and H_1 , state the distribution under H_0 , compare with the significance level, and finish in context.

Warm-up

Question 1

[3 marks]

$X \sim B(10, 0.5)$ under H_0 . An experiment gives $X = 8$, and $H_1 : p > 0.5$.

- (a) Write down the probability needed for the test.
- (b) Find it, and state whether the result is significant at the 5% level.

SOLUTION

- (a) $P(X \geq 8)$ — the observed value or anything more extreme.
- (b) $P(X \geq 8) = 1 - P(X \leq 7) = 1 - 0.9453 = 0.0547$. Since $0.0547 > 0.05$, the result is **not significant** at the 5% level.

Question 2

[3 marks]

Under H_0 , $X \sim B(30, 0.2)$, and $H_1 : p < 0.2$. The observed value is 2. Find the p-value and test at the 5% level.

SOLUTION

$P(X \leq 2) = 0.0442$. Since $0.0442 < 0.05$, the result is significant: **reject** H_0 .

Question 3

[4 marks]

Find the critical region for a one-tailed test of $H_0 : p = 0.4$ against $H_1 : p > 0.4$ at the 5% level, using $X \sim B(15, 0.4)$, and state the actual significance level.

SOLUTION

$P(X \geq 9) = 0.0950 > 0.05$ and $P(X \geq 10) = 0.0338 < 0.05$.

The critical region is $X \geq 10$, and the actual significance level is 0.0338, or 3.38%.

Question 4

[3 marks]

For each situation, state H_0 and H_1 , and whether the test is one- or two-tailed.

- (a) A manufacturer claims 20% of its sweets are red; a shopper thinks there are fewer.
- (b) A drug is known to work for 60% of patients; a researcher asks whether a new version has a different success rate.

SOLUTION

(a) $H_0 : p = 0.2$, $H_1 : p < 0.2$. One-tailed.

(b) $H_0 : p = 0.6$, $H_1 : p \neq 0.6$. Two-tailed.

Standard**Question 5**

[6 marks]

A coin is tossed 25 times and lands heads 18 times. Test at the 5% level whether the coin is biased towards heads.

SOLUTION

Let p be the probability of heads and X the number of heads in 25 tosses.

$H_0 : p = 0.5$, $H_1 : p > 0.5$. Assuming H_0 , $X \sim B(25, 0.5)$.

$$P(X \geq 18) = 1 - P(X \leq 17) = 1 - 0.9784 = 0.0216$$

Since $0.0216 < 0.05$, reject H_0 .

IN CONTEXT

There is sufficient evidence at the 5% level to suggest that the coin is biased towards heads.

Question 6

[6 marks]

A garage claims that 15% of cars fail their first MOT. An inspector checks 40 cars and finds 2 failures. Test at the 1% level whether the failure rate is lower than claimed.

SOLUTION

Let p be the proportion of cars that fail and X the number failing out of 40.

$H_0 : p = 0.15$, $H_1 : p < 0.15$. Assuming H_0 , $X \sim B(40, 0.15)$.

$$P(X \leq 2) = 0.0486$$

Since $0.0486 > 0.01$, do not reject H_0 .

IN CONTEXT

There is insufficient evidence at the 1% level to suggest that the failure rate is lower than 15%. (At the 5% level the same data *would* have been significant, which shows how much the chosen level matters.)

Question 7

[6 marks]

$X \sim B(20, 0.25)$ under $H_0 : p = 0.25$, tested against $H_1 : p \neq 0.25$ at the 10% level.

- (a) Find the critical region.
- (b) Find the actual significance level of the test.

SOLUTION

(a) Each tail must stay under 5%.

Lower: $P(X \leq 1) = 0.0243 < 0.05$ but $P(X \leq 2) = 0.0913 > 0.05$, so the lower part is $X \leq 1$.

Upper: $P(X \geq 9) = 0.0409 < 0.05$ but $P(X \geq 8) = 0.1018 > 0.05$, so the upper part is $X \geq 9$.

The critical region is $X \leq 1$ or $X \geq 9$.

(b) $0.0243 + 0.0409 = 0.0652$, so the actual significance level is 6.52%.

Question 8

[4 marks]

Explain, in the context of a hypothesis test:

- (a) what the significance level means;
- (b) why “do not reject H_0 ” is not the same as “ H_0 is true”;
- (c) why the actual significance level is usually less than the stated one.

SOLUTION

(a) It is the probability of rejecting H_0 when H_0 is actually true — the risk of a false positive you are willing to accept. At 5%, about 1 test in 20 would reject a true H_0 by chance.

(b) The test only asks whether the data are unusual enough. A result inside the acceptance region is consistent with H_0 , but it is also consistent with values of p close to p_0 , so nothing is proved.

(c) The test statistic is discrete, so the tail probabilities jump in steps. The critical region is the largest tail whose probability stays **at or below** the stated level, and that is normally some way below it.

Exam style**Question 9**

[7 marks]

A seed company claims that 10% of its seeds fail to germinate. A gardener plants 50 seeds and 9 fail.

- (a) Test at the 5% level whether the failure rate is higher than claimed.
- (b) Find the critical region for this test, and state the actual significance level.
- (c) Explain how your answers to (a) and (b) are consistent.

SOLUTION

(a) Let p be the proportion that fail and X the number failing out of 50.

$H_0 : p = 0.1$, $H_1 : p > 0.1$. Assuming H_0 , $X \sim B(50, 0.1)$.

$$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.9421 = 0.0579$$

Since $0.0579 > 0.05$, do not reject H_0 : there is insufficient evidence at the 5% level that more than 10% of seeds fail.

(b) $P(X \geq 9) = 0.0579 > 0.05$ and $P(X \geq 10) = 0.0245 < 0.05$, so the critical region is $X \geq 10$, with actual significance level 2.45%.

(c) The observed value 9 is **not** in the critical region $X \geq 10$, which matches the p-value conclusion: both say the evidence is not strong enough. Nine failures is unlucky but not rare enough to doubt the claim.

Question 10

[6 marks]

A revision app claims that 70% of its users improve their grade. A school tries it with 30 students, and 16 improve. Test at the 5% level whether the app's claim is too high.

SOLUTION

Let p be the proportion of users who improve and X the number improving out of 30.

$H_0 : p = 0.7$, $H_1 : p < 0.7$. Assuming H_0 , $X \sim B(30, 0.7)$.

$$P(X \leq 16) = 0.0401$$

Since $0.0401 < 0.05$, reject H_0 .

IN CONTEXT

There is sufficient evidence at the 5% level to suggest that fewer than 70% of users improve their grade, so the app's claim looks too high. (The school's students may not be typical users, so the conclusion applies to this population.)

Question 11

[8 marks]

A spinner is supposed to land on red one third of the time. It is spun 24 times and lands on red 14 times. A student tests at the 5% level whether the spinner is fair.

- State the hypotheses and explain why the test is two-tailed.
- Find the critical region and the actual significance level.
- Carry out the test using the p-value method, and show it agrees with the critical region.

SOLUTION

(a) Let p be the probability of red and X the number of reds in 24 spins. $H_0 : p = \frac{1}{3}$, $H_1 : p \neq \frac{1}{3}$. The question asks whether the spinner is **fair**, not whether it favours red, so a change in either direction matters.

(b) Each tail must stay under 2.5%. Assuming H_0 , $X \sim B(24, \frac{1}{3})$.

Lower: $P(X \leq 3) = 0.0199 < 0.025$, while $P(X \leq 4) = 0.0594 > 0.025$, so $X \leq 3$.

Upper: $P(X \geq 14) = 0.0103 < 0.025$, while $P(X \geq 13) = 0.0284 > 0.025$, so $X \geq 14$.

The critical region is $X \leq 3$ or $X \geq 14$, and the actual significance level is $0.0199 + 0.0103 = 0.0302$, or 3.02%.

(c) The observed value is high, so use the upper tail: $P(X \geq 14) = 0.0103$. Since $0.0103 < 0.025$, reject H_0 . The value 14 also lies in the critical region $X \geq 14$, so the two methods agree.

IN CONTEXT

There is sufficient evidence at the 5% level to suggest that the spinner does not land on red one third of the time.

Question 12

[7 marks]

A student tosses a coin 12 times, gets 10 heads, and writes: " $P(X \geq 10) = 0.0193$, which is less than 0.05, so the coin is biased towards heads."

- The student had planned to test whether the coin was biased in either direction. Redo the comparison properly and state the conclusion.

- (b) Give two criticisms of the student's written conclusion.
- (c) Explain why choosing the tail after seeing the data is poor practice.

SOLUTION

(a) For a two-tailed test at 5%, compare the upper tail with $\frac{0.05}{2} = 0.025$. Since $0.0193 < 0.025$, the result **is** significant, so reject H_0 : there is sufficient evidence at the 5% level that the coin is biased.

(Equivalently, the two-tailed p-value is $2 \times 0.0193 = 0.0386 < 0.05$.)

(b) First, the conclusion claims the coin **is** biased, when a test only gives evidence; it should say “there is sufficient evidence to suggest”. Second, the student compared a one-tailed probability with the full 5% even though the planned test was two-tailed.

(c) Choosing the direction after seeing the data means you would have tested the lower tail had the result gone the other way. That gives two chances to find significance, so the real false-positive rate is about double the stated level.