# Lesson - Hypothesis testing

Year: Year 12

A hypothesis test uses sample data to decide whether a claim about a population is believable. In Year 12 you test a claim about the probability $p$ of success in a binomial setting. Learn the language and the exact layout of a full answer.

**Starter activity**

A friend gives you a coin and says it is fair. You toss it 10 times and get 9 heads. Would you still believe the coin was fair? What about 6 heads? 7 heads? 8 heads? Where would you draw the line between "could easily happen by chance" and "very unlikely if the coin were fair"? Write down a rule, then think about how you could use probabilities of the binomial distribution to turn your rule into a precise calculation.

## The language of testing

**Definition: null and alternative hypotheses**

The **null hypothesis** $H_0$ is the claim being tested, written $H_0 : p =$ value, and assumed true during the test. The **alternative hypothesis** $H_1$ is what you suspect instead: $H_1 : p >$ value or $p$ value (**one-tailed**), or $H_1 : p \neq$ value (**two-tailed**).

**Definition: test statistic, significance level and critical region**

The **test statistic** is the sample result, here $X$, the number of successes; under $H_0$, $X \sim B(n, p)$. The **significance level** $\alpha$ (often 5% or 1%) is the threshold for rejecting $H_0$. The **critical region** is the set of values of $X$ that lead to rejecting $H_0$, and its boundary is the **critical value**. The **p-value** is the probability, assuming $H_0$, of a result at least as extreme as the one observed.

**Key fact: choosing the hypotheses**

Use a one-tailed test when the question says "increased", "decreased", "better" or "biased towards". Use a two-tailed test for "different", "changed" or "biased" with no direction. Always define $p$ in words, and write the hypotheses with $p$, not with $X$ or the sample.

## One-tailed test

**Method: a full hypothesis test**

1. State $H_0$ and $H_1$ in terms of $p$, defining $p$ in words.
2. State the distribution under $H_0$: $X \sim B(n, p_0)$.

3. Calculate the probability of a result at least as extreme as the observed one.
4. Compare with the significance level as a number sentence.
5. Decide: reject or do not reject $H_0$.
6. Write a conclusion in context.

**Worked example 1**

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

$H_0 : p = 0.5$, $H_1 : p > 0.5$, where $p$ is the probability of heads. Under $H_0$, $X \sim B(20, 0.5)$. $P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.9793 = 0.0207$. Since $0.0207$, reject $H_0$. There is sufficient evidence, at the 5% level, that the coin is biased towards heads.

**Watch out**

If the p-value were larger than 0.05 you would say "insufficient evidence to reject $H_0$". Never say "$H_0$ is true", "the coin is fair" or "we have proved".

**Worked example 2**

**A gardener claims 40% of seeds germinate. With new compost 11 of 15 germinate. Test at the 1% level whether the compost improves germination.**

$H_0 : p = 0.4$, $H_1 : p > 0.4$. Under $H_0$, $X \sim B(15, 0.4)$ and $P(X \geq 11) = 0.0093$, so reject $H_0$. There is evidence that the new compost improves germination.

**Worked example 3**

**A student guesses a 40-question test with four options and gets 15 right. Test at 5% whether she is better than guessing.**

$H_0 : p = 0.25$, $H_1 : p > 0.25$. $P(X \geq 15) = 0.0544 > 0.05$, so do not reject $H_0$, although the result is very close. There is insufficient evidence that she is better than guessing, which is not the same as saying she is guessing.

## Two-tailed test

**Rule: two-tailed tests**

For $H_1 : p \neq p_0$, split the significance level: use $\frac{\alpha}{2}$ in each tail (2.5% each for a 5% test). Compare the observed result with the nearer tail.

**Worked example 4**

**The coin gives 4 heads in 20 tosses. Test at 5% whether it is biased.**

$H_0 : p = 0.5$, $H_1 : p \neq 0.5$. $P(X \leq 4) = 0.0059$. Since $0.0059$, reject $H_0$: there is evidence the coin is biased.

## Critical regions and actual significance level

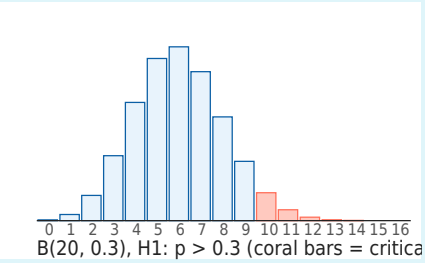
### Definition: actual significance level

The **actual significance level** is the true probability of the critical region: the probability of wrongly rejecting a true $H_0$. To find a critical region take the largest region whose probability is at most the significance level.

### Worked example 5

**$X \sim B(20, 0.3)$, $H_1 : p > 0.3$ at 5%. Find the critical region.**

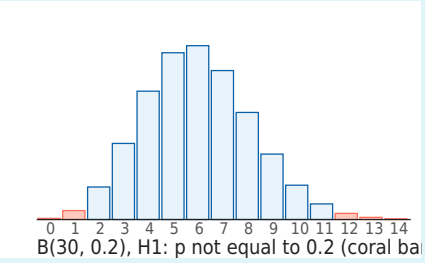
$P(X \geq 9) = 0.1133$ and $P(X \geq 10) = 0.0480$. The critical region is $X \geq 10$ and the actual significance level is $4.80\%$.




### Worked example 6

**$X \sim B(30, 0.2)$, $H_1 : p \neq 0.2$, 5% level. Find the critical region.**

Lower tail: $P(X \leq 1) = 0.0105 \leq 0.025$ but $P(X \leq 2) = 0.0442 > 0.025$, so $X \leq 1$.
Upper tail: $P(X \geq 12) = 0.0095 \leq 0.025$ but $P(X \geq 11) = 0.0256 > 0.025$, so $X \geq 12$.
Critical region: $X \leq 1$ or $X \geq 12$. Actual significance level $= 0.0105 + 0.0095 = 0.0200$, that is 2.00%.




### Note: exam layout

Show the probability for the boundary value and for the next value outside the region to justify the critical region. Write the comparison as a number sentence such as $0.0207$.

**Watch out**

Do not use the full 5% in each tail of a two-tailed test. Calculate $P(X \geq 15)$, not $P(X = 15)$. A conclusion with no reference to the context loses the last mark.

**Worked example 7: Writing a complete one-tailed test**

**A shop claims that 15% of its customers use a loyalty card. A sample of 40 customers includes 11 with cards. Test at the 5% level whether the proportion is higher.**

$H_0 : p = 0.15$, $H_1 : p > 0.15$. Under $H_0$, $X \sim B(40, 0.15)$.
$P(X \geq 11) = 1 - P(X \leq 10) = 1 - 0.9701 = 0.0299$ (using the cumulative table). Since $0.0299$ we reject $H_0$. There is sufficient evidence at the 5% level that more than 15% of customers use a loyalty card. The sample proportion $\frac{11}{40} = 27.5\%$ is much higher than 15%, which agrees with the test. At the 1% level the result would not be significant, because $0.0299 > 0.01$.

**Worked example 8: Why the critical region excludes the boundary**

$X \sim B(10, 0.5)$**. Find the critical region for** $H_1 : p > 0.5$ **at 5%, and explain why the actual level is below 5%.**

$P(X \geq 9) = P(X = 9) + P(X = 10) = 0.0098 + 0.0010 = 0.0107$;
$P(X \geq 8) = 0.0547 > 0.05$. So the critical region is $X \geq 9$ and the actual level is 1.07%. It is well below 5% because the probability jumps from 5.47% to 1.07% when moving from 8 to 9. The critical region must be the largest set whose probability does not exceed the stated level, so it is usually smaller than the nominal level.

**Worked example 9: Choosing hypotheses from the question**

**Decide the form of** $H_1$**: (a) "a drug is claimed to cure 60%; is it less effective?" (b) "a die is thought to be biased" (c) "has the pass rate improved from 70%?"**

(a) The phrase "less effective" is directional, so $H_1 : p$ (one-tailed, lower). (b) "Biased" does not give a direction, so $H_1 : p \neq \frac{1}{6}$ (two-tailed). (c) "Improved" is directional upwards, so $H_1 : p > 0.7$ (one-tailed, upper). The null hypothesis always has the form $H_0 : p = p_0$ with the equals sign, where $p_0$ is the value stated in the claim.

**Worked example 10: The meaning of the significance level**

**In 100 independent tests at the 5% level, each with a true null hypothesis, how many false rejections would you expect?**

The significance level is the probability of rejecting $H_0$ when it is true. With a 5% level, the expected number of false rejections in 100 tests is $100 \times 0.05 = 5$. That is why a single unexpected "significant" result should be repeated and why scientists decide on

the significance level before collecting data. A stricter 1% level would give an expected 1 false rejection in 100.

### Worked example 11: A test with a small sample

**A coin is tossed 8 times and 7 heads occur. Is there evidence at the 5% level that it is biased towards heads?**

$X \sim B(8, 0.5)$. $P(X \geq 7) = P(X = 7) + P(X = 8) = \dfrac{8+1}{256} = 0.0352$. Reject $H_0$: there is sufficient evidence of bias towards heads. With only 8 tosses a result of 7 or 8 heads is needed, so the critical region is $X \geq 7$ with actual significance 3.52%.

### Method: carrying out a hypothesis test

1. Define the parameter $p$ and write $H_0 : p = p_0$ and $H_1$ (one- or two-tailed).
2. State the distribution of the test statistic under $H_0$, for example $X \sim B(n, p_0)$.
3. Find the probability of the observed value or more extreme (or find the critical region).
4. Compare with the significance level (halve it for each tail in a two-tailed test).
5. State the conclusion: reject or do not reject $H_0$, and say what this means in context using non-definitive language.

### Watch out: common hypothesis testing errors

- Writing hypotheses in terms of the sample statistic ($X$) instead of the population parameter ($p$).
- Using $P(X > \text{observed})$ instead of $P(X \geq \text{observed})$ for the upper tail.
- Not halving the significance level for a two-tailed test.
- Saying "accept $H_0$" or "prove" instead of "do not reject".
- Concluding without context; always refer to the situation in words.

### Note: errors in testing

Two kinds of mistake are possible. A **Type I error** rejects $H_0$ although it is true; its probability is the significance level. A **Type II error** fails to reject $H_0$ although it is false; its probability depends on the true value of $p$ and on the sample size. Lowering the significance level reduces Type I errors but makes Type II errors more likely, and the only way to reduce both is to increase the sample size. Good experimental design therefore starts by deciding how large a sample is needed.

### Did you know?

Modern hypothesis testing was developed in the 1920s and 1930s by Ronald Fisher, Jerzy Neyman and Egon Pearson. Fisher introduced the 5% level after a famous

experiment in which a lady claimed she could tell whether milk or tea had been poured into a cup first. He designed a test using 8 cups, and the lady was right in all 8 (a chance of $\frac{1}{70}$ of doing so by guessing). Today hypothesis tests are used to approve medicines, check manufacturing standards and test new website designs.

**Otto's tip**

"Reject or fail to reject, and always say it in the story." Use the words of the question in your final sentence.