The Language of a Hypothesis Test

A hypothesis test asks: *is the data we observed consistent with some default assumption about the population, or is it so extreme that the assumption looks doubtful?*

Definition.

- The **null hypothesis** H_0 is the default assumption about a population parameter — always an *exact* statement, e.g. $H_0 : p = 0.4$ or $H_0 : \mu = 500$.
- The **alternative hypothesis** H_1 is what we suspect instead: $p > 0.4$, $p < 0.4$ (a **one-tail test**) or $p \neq 0.4$ (a **two-tail test**).
- The **test statistic** is the quantity computed from the sample whose distribution is known *under* H_0 (e.g. the number of successes X, or the sample mean $\overline{X}$).
- The **significance level** (e.g. 5%) is the threshold of "suspiciously extreme": the probability, calculated assuming H_0 is true, below which we reject H_0.
- The **p-value** is the probability, under H_0, of observing a result *as extreme or more extreme* than the one obtained.
- The **critical region** (or rejection region) is the set of values of the test statistic that lead to rejecting H_0; a **critical value** is a boundary of this region.

> **Remark** ("Acceptance region"). You may meet the term *acceptance region* for the complement of the critical region. Avoid the phrase: it encourages the wrong-headed conclusion "accept H_0". Think of it simply as *not in the critical region*.

The logic of the conclusion

The structure is a proof by (probabilistic) contradiction. We *assume* H_0, and compute how surprising the data would then be.

- If the p-value is *less* than the significance level, the data casts suspicion on the assumption: there is evidence to **reject** H_0.
- If not, the data is unremarkable under H_0 — but this gives no positive support to H_0! All we may say is there is **insufficient evidence to reject** H_0.

Tip (Model conclusions — memorise these)

Significant: *"Reject H_0 — there is sufficient evidence to suggest that the population mean of [context] is greater than [value] at the 5% level."*

Not significant: *"Do not reject H_0 — there is insufficient evidence to suggest that the population mean of [context] is greater than [value] at the 5% level."*

Never write "accept H_1", "accept H_0", or assertive claims like "the mean has increased". The conclusion is uncertain and must sound uncertain; being too assertive loses marks. The double negation in the second conclusion is awkward English but correct logic — like a court verdict of "not guilty", which is not a finding of innocence.

> **Remark.** Hypotheses must be stated in terms of *parameter values*, with symbols defined. OCR's own example: "$H_0 : p = 0.7$, $H_1 : p < 0.7$, where p is the population proportion in favour of the resolution." A null hypothesis is

exact even when the claim isn't: if a factory claims at most 5% of components are faulty and we wish to test the claim, we take $H_0 : p = 0.05$, $H_1 : p > 0.05$.

Recap: Testing a Binomial Proportion

You met this in L8 single maths. The test statistic is X, the number of successes, and under H_0 we have $X \sim B(n, p_0)$. Note this tests *observations from the distribution itself*, not a sample mean.

Example

A coin is suspected of bias towards heads. It is tossed 20 times and shows 15 heads. Test, at the 5% significance level, whether the coin is biased towards heads.

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

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

Under H_0, $X \sim B(20, 0.5)$. One-tail test at the 5% level; the observed value is 15.

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

Since $0.0207 < 0.05$, the result is significant. Reject H_0 — there is sufficient evidence, at the 5% level, to suggest that the coin is biased towards heads.

Tip (Discrete distributions need care)

Calculators give *cumulative* probabilities $\mathbb{P}(X \leq k)$. For an upper tail you must convert:

$$\mathbb{P}(X \geq k) = 1 - \mathbb{P}(X \leq k - 1)$$

The $k - 1$ is the classic slip. And you can **never** compare $\mathbb{P}(X = k)$ with the significance level — the p-value is the probability of k *or more extreme*, not of exactly k.

Example (In class)

A factory claims that at most 5% of its components are faulty. A buyer tests a random sample of 40 components and finds that 5 are faulty. Test the factory's claim at the 5% significance level. (Take care setting up H_0 — the null hypothesis must be an exact value.)

Critical regions and actual significance level

Definition. Because X is discrete, we usually cannot find a region with probability exactly 5%. The critical region is the largest tail region whose probability under H_0 does not exceed the significance level; its actual

probability is called the **actual significance level** of the test.

Example

A treatment is successful for 15% of patients nationally. A clinic suspects its own success rate is *lower*. It audits 30 randomly chosen patients. Using a 5% significance level, find the critical region for the test, and state the actual significance level.

Let X be the number of successes among the 30 patients and p the clinic's success probability.

$$H_0 : p = 0.15 \qquad H_1 : p < 0.15$$

Under H_0, $X \sim B(30, 0.15)$. We need the largest c with $\mathbb{P}(X \leq c) \leq 0.05$. Straddle the significance level by trying values:

$$\mathbb{P}(X \leq 1) = 0.0480 \leq 0.05$$
$$\mathbb{P}(X \leq 2) = 0.1514 > 0.05$$

So the critical region is $X \leq 1$, and the actual significance level is $\mathbb{P}(X \leq 1) = 0.0480 = 4.80\%$.

Tip

There is no "inverse binomial" on most calculators: find critical values by trial and error, *straddling* the significance level — exhibit one cumulative probability below it and the neighbouring one above it. Both calculations must be shown for full marks. (A calculator List lets you compute several at once.)

Remark (Locating critical values for large n)**.** If n is large, trial and error is painful, but $B(n,p) \approx N(np, npq)$ locates the critical value quickly. For example, for $H_1 : p < 0.3$ with $n = 400$ at 5%: $X \dot{\sim} N(120, 84)$, so the boundary is near $120 - 1.645\sqrt{84} \approx 104.9$, suggesting a critical region of roughly $X \leq 104$, which can then be confirmed exactly. This use of the approximation appeared in June 2019; trial and error, or a calculator with inverse binomial, also works.

Example (OCR S2, June 2014)

In a city the proportion of inhabitants from ethnic group Z is known to be 0.4. A sample of 12 employees of a large company in this city is obtained and it is found that 2 of them are from ethnic group Z. A test is carried out, at the 5% significance level, of whether the proportion of employees in this company from ethnic group Z is less than in the city as a whole.

(a) State an assumption that must be made about the sample for a significance test to be valid.

(b) Describe briefly an appropriate way of obtaining the sample.

(c) Carry out the test.

(d) A manager believes that the company discriminates against ethnic group Z. Explain whether carrying out the test at the 10% significance level would be more supportive or less supportive of the manager's belief.

(a) The sample must be random *— every employee equally likely to be selected.*

(b) Number a list of all the company's employees, then select 12 *using random numbers (ignoring repeats and out-of-range values).*

(c) Let X *be the number of employees in the sample from group* Z *and* p *the proportion of the company's employees from group* Z*.*

$$H_0 : p = 0.4 \qquad H_1 : p < 0.4$$

Under H_0, $X \sim B(12, 0.4)$*. One-tail (lower) test at* 5%*; observed value* 2*.*

$$p\text{-value} = \mathbb{P}(X \leq 2) = 0.0834$$

Since $0.0834 > 0.05$*, the result is not significant. Do not reject* H_0 *— there is insufficient evidence, at the* 5% *level, to suggest that the proportion of this company's employees from ethnic group* Z *is less than in the city as a whole.*

(Critical-region alternative: $\mathbb{P}(X \leq 1) = 0.0196 \leq 0.05$ *but* $\mathbb{P}(X \leq 2) = 0.0834 > 0.05$*, so the critical region is* $X \leq 1$*; the observed* 2 *is not in it.)*

(d) At the 10% *level the p-value* $0.0834 < 0.10$*, so* H_0 would *be rejected — the test would conclude there is evidence that the proportion is lower. This is* more *supportive of the manager's belief.*

Textbook Exercises: [CUP.1] Ch 18 §2; [S2] Ch 5, 6

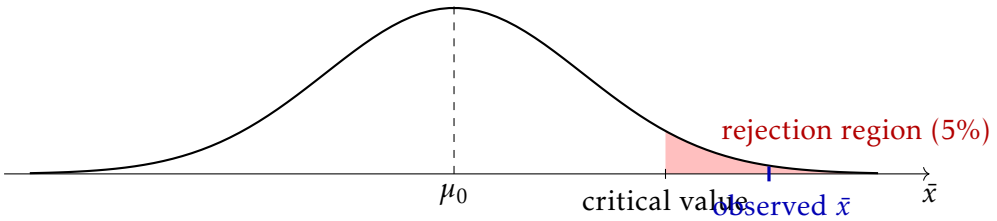
Testing the Mean of a Normal Distribution

Now the test statistic is the *sample mean*. Recall the key distributional fact from the Central Limit Theorem notes:

Fact — If $X \sim N(\mu, \sigma^2)$, then for a random sample of size n, *exactly*,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Under $H_0 : \mu = \mu_0$, therefore, $\overline{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$.



Always sketch: the distribution of $\overline{X}$ under H_0, the rejection region in the tail(s) being tested, and the observed value.

Fact (The full template) —

1. **Define variables.** "Let X be the [quantity in context] and μ its population mean."
2. **Hypotheses.** $H_0 : \mu = \mu_0$; $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ (one-tail) or $\mu \neq \mu_0$ (two-tail).
3. **Distribution.** Under H_0, $\overline{X} \sim N\left(\mu_0, \sigma^2/n\right)$.
4. **Sketch**, marking the rejection region (5% in one tail, or 2.5% in each).
5. **Calculate**: either the p-value of the observed $\bar{x}$, or the critical value(s), and compare *explicitly* with the significance level / observed value.
6. **Conclude** — "significant" or "not significant", then a model conclusion in context.

Remark. Writing $\mu = \mu_0$ in step 3 is an *assumption under H_0*, not a fact. And mind your notation: $\mathbb{P}(\overline{X} < 0.9941)$ is a statement about the random variable $\overline{X}$; the observed number is $\bar{x}$. Mixing up $\overline{X}$ and $\bar{x}$ is a recurring homework offence.

Example (One-tail, p-value method)

Bags of sugar are filled by a machine so that the mass of a bag is $N(\mu, 0.012^2)$ kg. The target is $\mu = 1$. An inspector suspects underfilling, and weighs a random sample of 10 bags, finding mean 0.9941 kg. Test the inspector's suspicion at the 5% significance level.

Let X be the mass of a bag (kg) and μ the population mean mass.

$$H_0 : \mu = 1 \qquad H_1 : \mu < 1$$

Under H_0, $\overline{X} \sim N\left(1, \frac{0.012^2}{10}\right)$, *i.e. standard deviation* $0.012/\sqrt{10} = 0.003795$. *One-tail (lower) test at* 5%.

$$p\text{-value} = \mathbb{P}(\overline{X} \leq 0.9941) = \mathbb{P}\left(Z \leq \frac{0.9941 - 1}{0.003795}\right) = \mathbb{P}(Z \leq -1.555) = 0.0600$$

Since $0.0600 > 0.05$, *the result is not significant. Do not reject* H_0 *— there is insufficient evidence, at the* 5% *level, to suggest that the population mean mass of the bags of sugar is less than* 1 *kg.*

(Note what we have not *concluded: not that the machine is fine, only that this sample fails to demonstrate underfilling.)*

Example (Two-tail, both methods)

A drinks machine dispenses volumes distributed $N(\mu, 5^2)$ ml and is set so that $\mu = 250$. After a service, the machine is checked: a random sample of 20 cups has mean 252.7 ml. Test, at the 5% level, whether the mean volume has changed.

(a) Carry out the test using a p-value.

(b) Find the critical region for $\overline{X}$ and confirm the conclusion.

Let X be the volume dispensed (ml) and μ the population mean volume.

$$H_0 : \mu = 250 \qquad H_1 : \mu \neq 250$$

Under H_0, $\overline{X} \sim N\left(250, \frac{5^2}{20}\right)$, standard deviation $5/\sqrt{20} = 1.1180$. Two-tail test at 5% (2.5% in each tail).

(a) $\mathbb{P}(\overline{X} \geq 252.7) = \mathbb{P}\left(Z \geq \frac{252.7-250}{1.1180}\right) = \mathbb{P}(Z \geq 2.415) = 0.00787$. *The test is two-tailed, so the p-value is the probability of a result this extreme* in either direction*:*

$$\textit{p-value} = 2 \times 0.00787 = 0.0157$$

Since $0.0157 < 0.05$, the result is significant. Reject H_0 — there is sufficient evidence, at the 5% level, to suggest that the population mean volume dispensed has changed.

(b) The critical values are $250 \pm 1.96 \times 1.1180 = 250 \pm 2.191$, so the critical region is

$$\overline{X} \leq 247.8 \quad \textit{or} \quad \overline{X} \geq 252.2 \ (4 \textit{ s.f.})$$

The observed $\bar{x} = 252.7$ lies in the critical region, confirming: reject H_0, as before.

Example (In class)

The yield of a chemical process is distributed $N(\mu, 12^2)$ grams. A modification is claimed to increase the yield from its current mean of 200 g. The process is run 36 times after modification. Using a 5% significance level, find the critical region for the sample mean $\overline{X}$, i.e. find c such that $\mathbb{P}(\overline{X} \geq c) = 0.05$ under H_0.

Tip (Two-tail tests: double the p-value)

For a two-tail test the p-value is *twice* the one-tail probability, compared against the full significance level. Halving the significance level instead gives the same decision and is accepted in exams, but then the number you computed is not a p-value — doubling is the better habit.

Remark. The critical-region method shines when a question has several parts (e.g. "would the conclusion change if $\bar{x} = 251.9$?") — you avoid running a whole new test. Note also that for a *continuous* test statistic the probability of incorrectly rejecting H_0 (when it is true) is exactly the significance level, with no "actual significance level" subtlety as in the binomial case.

The Three Further Maths Cases

So far the population was normal with known variance. We need hypothesis tests for a population mean in three situations:

Fact (When may we use a normal test for the mean?) —

1. **A sample drawn from a normal population of known, given or assumed variance.** Then $\overline{X} \sim N(\mu_0, \sigma^2/n)$ *exactly*, for any n.
2. **A large sample drawn from any population with known, given or assumed variance.** By the Central Limit Theorem, $\overline{X} \approx N(\mu_0, \sigma^2/n)$ for large n (roughly $n > 25$); proceed as before, noting the approximation.
3. **A large sample drawn from any population with unknown variance.** Estimate σ^2 by the unbiased estimate
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$$
and use $\overline{X} \approx N(\mu_0, s^2/n)$. (Why $n-1$? See the Estimation and Confidence Intervals notes.)

Remark (Case 3 is mathematically shabby)**.** For large n it *is* true that $\overline{X} \approx N(\mu, \sigma^2/n)$ — but σ^2 is unknown, so we quietly substitute the estimate s^2 and hope. For large n the estimate is good and the procedure is sound; for small n it is not. The honest fix for small samples is Student's t-distribution, which builds the extra uncertainty of estimating σ into the test. You will meet it at university (or in the S3 textbook).

Remark (Continuity correction — rarely examined)**.** If the parent population is *discrete*, the CLT approximation in cases 2–3 strictly wants a continuity correction: ± 0.5 on the total $\sum X_i$, equivalently $\pm\frac{1}{2n}$ on $\overline{X}$. This is very rarely in exams, the correction is tiny for large n, and OCR will not penalise its omission. Know it exists; then relax.

Example (Case 2: large sample, known variance)

A manufacturer claims its bulbs last 1200 hours on average, with standard deviation 200 hours; the distribution of lifetimes is not known to be normal. A consumer group tests a random sample of 100 bulbs and finds a mean lifetime of 1158 hours. Test, at the 1% significance level, whether the mean lifetime is less than the manufacturer claims.

Let X be the lifetime of a bulb (hours) and μ the population mean lifetime.

$$H_0 : \mu = 1200 \qquad H_1 : \mu < 1200$$

The population is not known to be normal, but $n = 100$ is large, so by the Central Limit Theorem, under H_0,

$$\overline{X} \approx N\left(1200, \frac{200^2}{100}\right) = N(1200, 20^2)$$

One-tail (lower) test at 1%.

$$p\text{-value} = \mathbb{P}(\overline{X} \le 1158) = \mathbb{P}\left(Z \le \frac{1158 - 1200}{20}\right) = \mathbb{P}(Z \le -2.1) = 0.0179$$

Since $0.0179 > 0.01$, the result is not significant. Do not reject H_0 — there is insufficient evidence, at the 1% level, to suggest that the population mean lifetime of the bulbs is less than 1200 hours. (At the 5% level the conclusion would differ — the significance level matters and should be fixed before *looking at data.)*

Example (OCR S2, January 2013)

Gordon is a cricketer. Over a long period he knows that his population mean score, in number of runs per innings, is 28, and the population standard deviation is 12. In a new season he adopts a different batting style and he finds that in 30 innings using this style his mean score is 28.98.

(a) Stating a necessary assumption, test at the 5% significance level whether his population mean score has increased.

(b) Explain whether it was necessary to use the Central Limit Theorem in part (a).

(a) Assume the 30 innings form a random sample *of scores with the new style. Let X be his score in an innings and μ the population mean score with the new style.*

$$H_0 : \mu = 28 \qquad H_1 : \mu > 28$$

Scores are not known to be normally distributed, but $n = 30$ is large, so by the Central Limit Theorem, under H_0 (with $\sigma = 12$ given),

$$\overline{X} \approx N\left(28, \frac{12^2}{30}\right), \quad \textit{standard deviation } \frac{12}{\sqrt{30}} = 2.191$$

One-tail (upper) test at 5%.

$$p\textit{-value} = \mathbb{P}(\overline{X} \geq 28.98) = \mathbb{P}\left(Z \geq \frac{28.98 - 28}{2.191}\right) = \mathbb{P}(Z \geq 0.447) = 0.327$$

Since $0.327 > 0.05$, the result is not significant. Do not reject H_0 — there is insufficient evidence, at the 5% level, to suggest that his population mean score has increased.

(b) Yes. The distribution of scores per innings is not known to be normal, so the (approximate) normality of $\overline{X}$ rests entirely on the Central Limit Theorem, valid here because $n = 30$ is large. This is case 2 of the three cases above.

Example (Case 3: large sample, unknown variance)

The mean time for a particular journey is claimed to be 20 minutes. A commuter believes it is longer. Over 50 randomly chosen days she records her journey times x (minutes), finding

$$\sum x = 1060 \qquad \sum x^2 = 23\,000$$

Test her belief at the 5% significance level.

Let X be the journey time (minutes) and μ the population mean journey time.

$$H_0 : \mu = 20 \qquad H_1 : \mu > 20$$

From the data, $\bar{x} = \frac{1060}{50} = 21.2$ and the unbiased estimate of variance is

$$s^2 = \frac{1}{49}\left(23\,000 - \frac{1060^2}{50}\right) = \frac{528}{49} = 10.78 \text{ (4 s.f.)}$$

The variance is unknown, but $n = 50$ is large, so under H_0 (using the CLT and s^2 in place of σ^2):

$$\overline{X} \approx N\left(20, \frac{10.78}{50}\right), \quad \text{standard deviation } \sqrt{10.78/50} = 0.4642$$

One-tail (upper) test at 5%.

$$p\text{-value} = \mathbb{P}(\overline{X} \geq 21.2) = \mathbb{P}\left(Z \geq \frac{21.2 - 20}{0.4642}\right) = \mathbb{P}(Z \geq 2.585) = 0.00489$$

Since $0.00489 < 0.05$, the result is significant. Reject H_0 — there is sufficient evidence, at the 5% level, to suggest that the population mean journey time is greater than 20 minutes.

Example (OCR S2, June 2010)

A machine is designed to make paper with mean thickness 56.80 micrometres. The thicknesses, x micrometres, of a random sample of 300 sheets are summarised by

$$n = 300, \qquad \sum x = 17\,085.0, \qquad \sum x^2 = 973\,847.0$$

Test, at the 10% significance level, whether the machine is producing paper of the designed thickness.

Let X be the thickness of a sheet (micrometres) and μ the population mean thickness. "Of the designed thickness" cuts both ways, so the test is two-tailed:

$$H_0 : \mu = 56.80 \qquad H_1 : \mu \neq 56.80$$

From the data, $\bar{x} = \frac{17\,085.0}{300} = 56.95$ and

$$s^2 = \frac{1}{299}\left(973\,847.0 - \frac{17\,085.0^2}{300}\right) = 2.864 \text{ (4 s.f.)}$$

The variance is unknown but $n = 300$ is large, so under H_0 (CLT, with s^2 for σ^2), $\overline{X} \approx N\left(56.80, \frac{2.864}{300}\right)$, standard deviation 0.09770. Two-tail test at 10% (5% in each tail).

$$\mathbb{P}(\overline{X} \geq 56.95) = \mathbb{P}\left(Z \geq \frac{56.95 - 56.80}{0.09770}\right) = \mathbb{P}(Z \geq 1.535) = 0.0624$$

so the p-value is $2 \times 0.0624 = 0.125$. *Since* $0.125 > 0.10$, *the result is not significant. Do not reject* H_0 *— there is insufficient evidence, at the* 10% *level, to suggest that the population mean thickness differs from the designed* 56.80 *micrometres.*

(Critical-value alternative: $56.80 + 1.645 \times 0.09770 = 56.96 > 56.95$, *so the observed mean is not in the critical region.)*

Textbook Exercises: [CUP.2] Ch 18 §2; [CUP.S] Ch 8 §3, 5, Ch 9 §1; [S2] Ch 5

Extension: Testing a Poisson Mean

Remark (Poisson rate tests)**.** A hypothesis test on the rate of a Poisson distribution is excellent consolidation: the set-up is identical, and the discrete tail-handling has all the same "binomial-type" traps. One can also practise the Poisson-to-normal approximation for locating critical values.

Example

Calls arrive at a helpdesk at a mean rate of 4.5 per hour. After an advertising campaign, 9 calls arrive in a single (randomly chosen) hour. Test, at the 5% significance level, whether the mean rate of calls has increased.

Let X be the number of calls in an hour and λ the population mean rate per hour. Assume calls occur singly, independently and at constant average rate, so that $X \sim Po(\lambda)$.

$$H_0 : \lambda = 4.5 \qquad H_1 : \lambda > 4.5$$

Under H_0, $X \sim Po(4.5)$. One-tail (upper) test at 5%; observed value 9.

$$p\text{-value} = \mathbb{P}(X \geq 9) = 1 - \mathbb{P}(X \leq 8) = 1 - 0.9597 = 0.0403$$

Since $0.0403 < 0.05$, the result is significant. Reject H_0 — there is sufficient evidence, at the 5% level, to suggest that the mean rate of calls has increased.

Exercise. For the test above, find the critical region and the actual significance level. Then check, using the normal approximation $\text{Po}(4.5) \approx N(4.5, 4.5)$ with a continuity correction, that the approximation locates the same critical value.

Example (OCR S2, June 2013)

The number of floods in a certain river plain is known to have a Poisson distribution. It is known that up until 10 years ago the mean number of floods per year was 0.32. During the last 10 years there were 6 floods. Test at the 1% significance level whether there is evidence of an increase in the mean number of floods per year.

The observation is a count over ten *years, so rescale the rate first: let X be the number of floods in a 10-year period and λ its population mean, so the historical rate 0.32 per year corresponds to $\lambda = 3.2$.*

$$H_0 : \lambda = 3.2 \qquad H_1 : \lambda > 3.2$$

Under H_0, $X \sim Po(3.2)$. *One-tail (upper) test at* 1%; *observed value* 6.

$$p\text{-value} = \mathbb{P}(X \geq 6) = 1 - \mathbb{P}(X \leq 5) = 0.105$$

Since $0.105 > 0.01$, *the result is not significant. Do not reject* H_0 — *there is insufficient evidence, at the* 1% *level, to suggest that the mean number of floods per year has increased.*

(Critical-region alternative: $\mathbb{P}(X \geq 9) = 0.0057 \leq 0.01$ *but* $\mathbb{P}(X \geq 8) = 0.0168 > 0.01$, *so the critical region is* $X \geq 9$, *and* 6 *is not in it.)*

Textbook Exercises: [S2] Ch 6