# Lesson - Hypothesis tests: correlation and Normal mean

Year: Year 13

This chapter reviews hypothesis testing, then tests for correlation and for the mean of a Normal distribution.

**Starter activity**

Toss a coin 10 times and record the number of heads. A friend claims that the coin is biased towards heads. If it shows 8 heads, would you be convinced? What about 10 out of 10? Discuss how unlikely an outcome has to be before you stop believing the coin is fair. Then think about a sample of 25 packets of crisps from a factory: how would you decide whether the mean mass is really lower than the label says?

## Hypothesis testing basics

**Definition: hypotheses and significance level**

The **null hypothesis** $H_0$ is the claim being tested, assumed true at the start. The **alternative hypothesis** $H_1$ is what we conclude if $H_0$ is rejected. **One-tailed**: $H_1 : \mu > \mu_0$ or $\mu$. **Two-tailed**: $H_1 : \mu \neq \mu_0$. The **significance level** (5% or 1%) is the probability of rejecting $H_0$ when it is true, a **Type I error**.

**Note: wording**

We never "prove" $H_0$; we only decide whether there is sufficient evidence to reject it. Decide one- or two-tailed from the wording of the question before you look at the data.

## Testing for correlation

**Definition: product moment correlation coefficient**

$r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ measures linear association, with $-1 \leq r \leq 1$. Coding the data (adding or multiplying by positive constants) does not change $r$.

**Rule: test for correlation**

Set $H_0 : \rho = 0$ with $H_1 : \rho > 0$, $\rho$ or $\rho \neq 0$. Use the sample size $n$ and the significance level to read a **critical value** from the tables, for example: $n = 10$: 5% one-tail $0.5494$, two-tail $0.6319$; $n = 12$: one-tail $0.4973$; $n = 8$: one-tail $0.6215$, two-tail $0.7067$. If $|r| \geq$

the critical value, reject $H_0$.

**Worked example 1**

**Ten pupils' revision hours and scores give $r = 0.72$. Test at 5% for positive correlation.**

$H_0 : \rho = 0$, $H_1 : \rho > 0$. The critical value is $0.5494$. Since $0.72 > 0.5494$, reject $H_0$: there is sufficient evidence of positive correlation between revision time and score.

**Worked example 2**

**Two-tailed, 5%, $n = 10$, $r = 0.62$. Then negative: $n = 12$, $r = -0.55$, $H_1 : \rho$.**

The two-tailed critical value is $0.6319$ and $0.62$, so there is insufficient evidence to reject $H_0$. For the negative case the critical value is $-0.4973$ and $-0.55$, so reject $H_0$. (With $S_{xy} = 14.5$, $S_{xx} = S_{yy} = 17.5$, $r = 0.829$.)

**Watch out**

Significant correlation does **not** mean cause and effect: ice cream sales and drowning incidents are both higher in summer because of the weather. Compare a negative $r$ with the critical value using the modulus.

## The sampling distribution of the mean

**Theorem: distribution of the sample mean**

If $X \sim N(\mu, \sigma^2)$ and a random sample of size $n$ is taken, then $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$. The standard deviation $\dfrac{\sigma}{\sqrt{n}}$ is smaller than $\sigma$: larger samples give means closer to $\mu$.

**Worked example 3**

$X \sim N(80, 12^2)$, $n = 9$. **Find** $P(\bar{X} > 84)$.

$\bar{X} \sim N(80, 4^2)$, so $P(Z > 1) = 0.1587$.

## Testing a Normal mean

**Method: testing a Normal mean**

1. Define $\mu$ in words and state $H_0 : \mu = \mu_0$ and $H_1$.

2. State the distribution of $\bar{X}$ under $H_0$, showing the variance $\frac{\sigma^2}{n}$.
3. Calculate the $p$-value (to 4 decimal places) or the critical region.
4. Compare with the significance level (halve it for two-tailed).
5. Conclude in context.

**Worked example 4**

**Packets should have mean 500 g, standard deviation 8 g. A sample of 16 has mean 496.5 g. Test at 5% whether the mean is lower.**

$H_0 : \mu = 500$, $H_1 : \mu$. Under $H_0$, $\bar{X} \sim N(500, 2^2)$.
$P(\bar{X} \leq 496.5) = P(Z \leq -1.75) = 0.0401$, so reject $H_0$: there is evidence the mean mass is less than 500 g.

**Worked example 5**

$n = 25$, $\sigma = 10$, $H_0 : \mu = 50$, $H_1 : \mu > 50$, **5%. Find the critical region.**

$\bar{X} \sim N(50, 2^2)$; critical value $50 + 1.645 \times 2 = 53.29$. The critical region is $\bar{x} \geq 53.3$, and a sample mean of 54 lies inside it.

**Worked example 6**

$n = 36$, $\sigma = 15$, $H_0 : \mu = 100$, $H_1 : \mu \neq 100$, **sample mean 104.5.**

Standard error 2.5, $z = 1.8$; two-tailed $p$-value $2 \times 0.0359 = 0.0719$. At 5% do not reject $H_0$; at 10% reject. For a two-tailed test at 5% split the level (2.5% in each tail, critical $z = \pm 1.96$).

## Writing a conclusion

**Rule: two-part conclusion**

State the decision about $H_0$ and then the meaning in context: "There is sufficient evidence at the 5% level to suggest that the mean mass of packets is less than 500 g." If you do not reject, say "There is insufficient evidence to suggest that ...", never "the claim is proved" and never "$H_0$ is true".

**Watch out**

Use $\frac{\sigma}{\sqrt{n}}$, not $\sigma$, for the mean. Write hypotheses with $\mu$ or $\rho$, never with $\bar{x}$. Do not use a one-tailed critical value in a two-tailed test.

### Worked example 7: A one-tailed test of correlation with a calculator

**The heights and arm spans (cm) of 6 students are given. Test, at the 5% level, for positive correlation if $r = 0.93$ and the critical value for $n = 6$ is 0.7293.**

$H_0 : \rho = 0$, $H_1 : \rho > 0$, where $\rho$ is the population correlation between height and arm span. Since $0.93 > 0.7293$ we reject $H_0$. There is sufficient evidence, at the 5% level, of positive correlation between height and arm span. Notice that with only 6 points the critical value is large, because a small sample can show a strong correlation by chance.

### Worked example 8: Standard error and sample size

**$X \sim N(100, 20^2)$. Compare the standard deviation of $\bar{X}$ for samples of size 4, 25 and 100.**

$\sigma_{\bar{X}} = \dfrac{20}{\sqrt{n}}$. For $n = 4$: 10. For $n = 25$: 4. For $n = 100$: 2. To halve the standard deviation of the sample mean you must multiply the sample size by 4. A larger sample gives a sample mean that is more tightly clustered around $\mu$.

### Worked example 9: Decision using a critical region for a lower tail

**$X \sim N(\mu, 5^2)$, $n = 25$, test $H_0 : \mu = 60$ against $H_1 : \mu$ at the 5% level. Find the critical region and decide for $\bar{x} = 58.7$.**

$\bar{X} \sim N(60, 1^2)$. Critical value $= 60 - 1.6449 \times 1 = 58.355$. Critical region: $\bar{x} \leq 58.36$. Since $58.7 > 58.36$, we do not reject $H_0$: there is insufficient evidence at the 5% level that the mean is less than 60.

### Worked example 10: A two-tailed test using the p-value

**$\sigma = 8$, $n = 64$, $H_0 : \mu = 70$, $H_1 : \mu \neq 70$, $\bar{x} = 71.8$. Test at the 5% level.**

$\bar{X} \sim N(70, 1^2)$ since $8/\sqrt{64} = 1$. $z = 1.8$. $P(\bar{X} > 71.8) = 0.0359$, so the two-tailed $p$-value is $0.0719$. Since $0.0719 > 0.05$ we do not reject $H_0$. (For a one-tailed test with $H_1 : \mu > 70$ the $p$-value 0.0359 would lead to rejection: the form of $H_1$ matters.)

### Worked example 11: A test written out in full

**A bakery claims its loaves have mean mass 800 g with standard deviation 20 g. A customer weighs 10 loaves and finds a mean of 787 g. Test the claim that the mean is lower, at 5%.**

Let $\mu$ be the mean mass of a loaf. $H_0 : \mu = 800$, $H_1 : \mu$. Under $H_0$, $\bar{X} \sim N\left(800, \dfrac{400}{10}\right) = N(800, 6.325^2)$. $z = \dfrac{787 - 800}{6.325} = -2.055$ and $p = P(Z$. Since

$0.0199$ we reject $H_0$. There is sufficient evidence at the 5% level to suggest that the mean mass of the bakery loaves is less than 800 g. Notice the four steps: hypotheses, distribution, comparison, conclusion in context.

### Worked example 12: Choosing between one-tailed and two-tailed

**For each claim, state whether a one-tailed or two-tailed test is needed: (a) "the new recipe makes the biscuits heavier"; (b) "the mean has changed since the machine was serviced"; (c) "there is a negative correlation between age and speed".**

(a) One-tailed, $H_1 : \mu > \mu_0$, because the claim gives a direction. (b) Two-tailed, $H_1 : \mu \neq \mu_0$, because "changed" allows an increase or a decrease. (c) One-tailed, $H_1 : \rho$. The wording of the question, not the data, decides. A common mistake is to look at the sample, see that the mean is higher, and then pick a one-tailed test: this makes the result look more significant than it is.

### Worked example 13: Using a sample to estimate how often a result happens

**A machine produces rods of mean length 20.0 cm and standard deviation 0.3 cm. Find the probability that a sample mean of 9 rods is more than 20.2 cm, and the probability that a single rod is more than 20.2 cm.**

For a single rod, $P(X > 20.2) = P(Z > 0.667) = 0.2525$. For the mean of 9 rods, the standard error is $0.3/3 = 0.1$, so $P(\bar{X} > 20.2) = P(Z > 2) = 0.0228$. The same difference from the mean is far less likely for a sample mean than for a single value, which is why large samples give good evidence.

### Method: the five steps of every test

1. Define the parameter in words ($\mu$ or $\rho$) and state $H_0$ and $H_1$ before looking at the data.
2. State the distribution of the test statistic under $H_0$ (for a mean: $\bar{X} \sim N(\mu_0, \sigma^2/n)$ ).
3. Find the $p$-value or the critical region, halving the significance level for a two-tailed test.
4. Compare and decide: reject or do not reject $H_0$.
5. Write a conclusion in the context, with a measure of caution such as "sufficient evidence" or "insufficient evidence".

### Note: significance and Type I errors

The significance level is the probability of rejecting $H_0$ when $H_0$ is true. At 5%, about 1 test in 20 of true null hypotheses will be rejected by chance alone. This is why a result needs a sensible context and, ideally, repeating before it is trusted. A small $p$-value tells you the data would be unusual if $H_0$ were true; it does not tell you the

probability that $H_0$ is true.

**Did you know?**

The idea of a significance test was developed by the statistician Ronald Fisher, who in the 1920s proposed the 5% level as a convenient benchmark. A famous story says Fisher designed a test to check whether a lady could really tell whether tea or milk was poured into a cup first. Karl Pearson, who gave his name to the product moment correlation coefficient, worked on correlation in the 1890s.

**Otto's tip**

Decide whether the test is one-tailed or two-tailed before you look at the data. The wording of the question tells you.