Lesson - Conditional probability and the Normal distribution

Year: Year 13

This chapter covers conditional probability, tree diagrams and Venn diagrams, then the Normal distribution and its use as a model.

Starter activity

Roll a fair die twice. What is the probability that the total is 8? Now suppose you are told that the first die showed a 6. How does the probability change? Discuss why the extra information changes the answer. Next, think of the heights of the students in your class. Are most of them close to the average? Sketch a graph of how often each height might occur.

Conditional probability

Definition: conditional probability

A **conditional probability** is the probability of an event given that another event has already happened. We write $P(A \mid B)$, read "the probability of A given B".

Rule: conditional probability and independence

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(A \cap B) = P(A \mid B)\,P(B).$$

Events are **independent** if $P(A \mid B) = P(A)$, equivalently $P(A \cap B) = P(A)P(B)$.

Worked example 1

$P(A) = 0.5$, $P(B) = 0.4$, $P(A \cup B) = 0.7$. **Are A and B independent?**

$P(A \cap B) = 0.5 + 0.4 - 0.7 = 0.2$ and $P(A)P(B) = 0.2$, so A and B are independent.

Tree diagrams

Method: tree diagrams

1. Draw each stage; probabilities on a branch are conditional on the previous branches.
2. **Multiply along** the branches to get the probability of a path.
3. **Add** the paths that make up the event.

4. For "given that", divide the probability of both events by the probability of the given event.

Worked example 2

A bag has 5 red and 3 blue counters; two are taken without replacement. Find $P(\text{both red})$, $P(\text{different})$ and $P(2\text{nd red} \mid 1\text{st blue})$.

$\frac{5}{8} \times \frac{4}{7} = \frac{5}{14}$. Different: $\frac{5}{8} \times \frac{3}{7} + \frac{3}{8} \times \frac{5}{7} = \frac{15}{28}$. Given the first is blue, 5 of the remaining 7 are red: $\frac{5}{7}$.

Worked example 3

Machine A makes 60% of items, 2% defective; machine B makes 40%, 5% defective. Find $P(\text{defective})$ and $P(A \mid \text{defective})$.

$P(D) = 0.6(0.02) + 0.4(0.05) = 0.032$. Reversing the conditioning, $P(A \mid D) = \dfrac{0.012}{0.032} = 0.375$.

Venn diagrams and two-way tables

Note: the denominator

In a Venn diagram or two-way table the denominator is only the circle (or row) for the event that is "given". From 40 students choosing drama, 28 of them girls: $P(\text{girl} \mid \text{drama}) = \frac{28}{40} = 0.7$.

Worked example 4

A test has prevalence 1%, detection rate 95% and false positive rate 4%. Find $P(\text{disease} \mid \text{positive})$.

$P(\text{positive}) = 0.01(0.95) + 0.99(0.04) = 0.0491$, so $P(\text{disease} \mid \text{positive}) = \dfrac{0.0095}{0.0491} \approx 0.194$. The value is low because the condition is rare: a classic exam context.

Watch out

For $P(A \mid B)$ the denominator is $P(B)$. Without replacement the second-branch probabilities change.

The Normal distribution

Definition: Normal distribution

$X \sim N(\mu, \sigma^2)$ is a symmetrical bell-shaped curve centred on the mean μ with standard deviation σ. The total area is 1. About 68% of values lie within one standard deviation of the mean, 95% within two and 99.7% within three. $Z = \dfrac{X - \mu}{\sigma}$ follows $N(0, 1)$.

Worked example 5

$X \sim N(50, 4^2)$. **Find** $P(X > 56)$, $P(46$ **and** a **with** $P(X$.

$P(Z > 1.5) = 0.0668$. $P(-1$. Inverse Normal gives $z = 1.2816$, so $a = 50 + 4(1.2816) = 55.13$.

Note: calculator skills

Learn which entry is the standard deviation (not the variance). Most calculators give the area to the left, so subtract from 1 for an upper tail. Always sketch the curve and shade the area. For a continuous distribution $P(X = a) = 0$.

Finding the mean and standard deviation

Method: unknown μ and σ

1. Convert each probability to a z-value with the inverse Normal.
2. Write $\dfrac{x - \mu}{\sigma} = z$ twice.
3. Subtract to find σ, then substitute to find μ.

Worked example 6

$P(X$ **and** $P(X > 50) = 0.1$**. Find** μ **and** σ**.**

$z = -0.8416$ for 30 and $z = 1.2816$ for 50 (since $P(X$). Subtracting, $\dfrac{20}{\sigma} = 2.1232$, so $\sigma = 9.42$ and $\mu = 30 + 0.8416(9.42) = 37.9$.

Using the Normal distribution as a model

Worked example 7

Egg masses $X \sim N(60, 5^2)$ g. Find the proportion between 50 g and 68 g, and the probability that all of five eggs are in this range.

$P(-2;\, 0.9224^5 = 0.668.$

Note: is Normal appropriate?

The Normal distribution suits data that is symmetrical and bell-shaped, such as heights or masses, but not strongly skewed data such as waiting times or incomes.

Rule: Normal approximation to the binomial

If $X \sim B(n, p)$ with n large, use $Y \sim N(np,\, np(1-p))$ with a **continuity correction**.

Worked example 8

$X \sim B(100, 0.4)$. **Estimate** $P(X \geq 45)$.

$Y \sim N(40, 24)$. $P(X \geq 45) \approx P(Y > 44.5) = P\left(Z > \dfrac{4.5}{\sqrt{24}}\right) = 0.179$; the exact binomial value is 0.1789.

Choosing a method

Method: choosing a tool

1. Percentages of groups and "given that": tree diagram.
2. Overlapping counts: Venn diagram or two-way table.
3. $P(A)$, $P(B)$, $P(A \cup B)$ given: addition law first.
4. Write the formula, substitute, and check every probability is between 0 and 1.

Watch out

Do not enter the variance for the standard deviation. Apply the continuity correction the right way. Always sketch the Normal curve.

Worked example 9: A two-way table with a given condition

In a college, 150 students study maths. 90 are girls. Of the girls, 54 also study physics; of the boys, 36 study physics. Find the probability that a student chosen at random studies physics given that they are a boy.

There are $150 - 90 = 60$ boys. Of these, 36 study physics. The word "given" tells us the denominator is the number of boys, not 150. So $P(\text{physics} \mid \text{boy}) = \dfrac{36}{60} = 0.6$. For comparison, $P(\text{physics} \mid \text{girl}) = \dfrac{54}{90} = 0.6$ as well, so in this college studying physics does not depend on sex: the events are independent.

Worked example 10: Reading a probability from the Normal curve

The time spent on homework by students is modelled by $H \sim N(95, 20^2)$ minutes. Find the probability that a student spends more than 2 hours.

Two hours is 120 minutes. Standardise: $z = \dfrac{120 - 95}{20} = 1.25$. Sketch the curve and shade the right tail. $P(H > 120) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$. About one student in ten spends over two hours.

Worked example 11: Both parameters unknown with a shortcut

$X \sim N(\mu, \sigma^2)$ and the quartiles of X are 40 and 60. Find μ and σ.

The Normal curve is symmetrical, so the mean lies halfway between the quartiles: $\mu = 50$. The upper quartile has $z = 0.6745$, so $\dfrac{60 - 50}{\sigma} = 0.6745$ and $\sigma = 14.83$. Always look for symmetry before setting up simultaneous equations.

Worked example 12: At least one in a sample

The mass M of a parcel is modelled by $N(2.5, 0.6^2)$ kg. A courier will not carry parcels over 3.5 kg. Find the probability that, of 4 parcels, at least one is over the limit.

$P(M > 3.5) = P(Z > 1.667) = 0.0478$. "At least one" is best found as $1 - P(\text{none})$. $P(\text{none}) = (1 - 0.0478)^4 = 0.9522^4 = 0.822$. So $P(\text{at least one}) = 1 - 0.822 = 0.178$.

Worked example 13: A tree diagram for a sequence of events

Whether a student passes a driving test at the first attempt is 0.5. A student who fails passes at the second attempt with probability 0.7. Find the probability that a student passes within two attempts, and that a student who passes needed two attempts.

Pass on first attempt: 0.5. Fail then pass: $0.5 \times 0.7 = 0.35$. So $P(\text{pass within two}) = 0.5 + 0.35 = 0.85$. Given that the student passed, the probability that two attempts were needed is $\dfrac{0.35}{0.85} = 0.412$.

Worked example 14: Judging whether a Normal model is sensible

A researcher records the number of text messages sent in a day by 200 students. The mean is 12 and the standard deviation is 15. Comment on a Normal model.

If the data were Normal, about 95% would lie within two standard deviations, so values between $12 - 30 = -18$ and $12 + 30 = 42$. A count cannot be negative. The mean is smaller than the standard deviation, which signals strong positive skew. So

the Normal model is **not** appropriate: $P(X$ would give a 21% chance of an impossible negative number.

Worked example 15: Using symmetry and the Normal curve in context

The lengths of screws made by a machine are $L \sim N(40, 0.5^2)$ mm. Screws shorter than 39 mm or longer than 41 mm are rejected. Find the proportion rejected, and the proportion rejected when the machine is reset so that the standard deviation is 0.3 mm.

The limits are 1 mm either side of the mean. With $\sigma = 0.5$ they are ± 2 standard deviations, so $P(\text{accepted}) = P(-2$ and $P(\text{rejected}) = 0.0455$. With $\sigma = 0.3$ the limits are ± 3.33 standard deviations, so $P(\text{rejected}) = 2P(Z > 3.333) = 2 \times 0.000429 = 0.00086$. A smaller standard deviation makes the production far more consistent, even though the mean has not changed.

Note: reading the question for the model

Questions about modelling often ask you to *state an assumption* or *comment on the model*. Typical answers: the heights of adults are symmetrical and bell-shaped, so a Normal distribution is reasonable; the probability that a ticket-holder turns up may not be the same for every person, so the binomial assumption of a constant probability may not hold; two events in a tree diagram might not be independent, for example if the weather affects both the bus and the walk. Always link your comment to the context.

Method: working with the Normal distribution on a calculator

1. Write the model in full, e.g. $X \sim N(\mu, \sigma^2)$, and note whether you were given the variance or the standard deviation.
2. Sketch a bell curve, mark the mean and shade the required area.
3. Use the calculator to find areas, or inverse Normal for a boundary value; use symmetry to turn upper tails into lower tails.
4. Round sensibly (3 s.f.) and write a conclusion in the context of the question.

Watch out: common errors

- Dividing by the wrong total: in $P(A \mid B)$ the denominator is $P(B)$, never the grand total.
- Forgetting to update the probabilities on the second set of branches when objects are not replaced.
- Entering the variance instead of the standard deviation (for $N(50, 16)$ the standard deviation is 4).
- Mixing up the direction of the continuity correction: $P(X \leq 40)$ becomes $P(Y$.

Did you know?

The Normal distribution is sometimes called the Gaussian distribution after Carl Friedrich Gauss, who used it in 1809 to analyse the errors made when measuring the orbit of the dwarf planet Ceres. Conditional probability has a long history too: the Reverend Thomas Bayes wrote the rule linking $P(A \mid B)$ and $P(B \mid A)$ in the 1740s, and it is now used in medical testing, spam filters and machine learning.

Otto's tip

In a conditional probability question, underline the word "given". The event after it goes underneath the fraction.