Models in probability and statistics

What a statistical model is, how to describe a model's assumptions, and how to comment on whether a model is suitable.

Mathematical models in probability and statistics

A **model** is a simplified description of a real situation, built so that probabilities and predictions can be calculated. It is judged by how well its predictions match what really happens.

Building and testing a model

1. Observe the real situation and decide what matters.
2. Make **assumptions** that simplify it.
3. Use the model to work out predictions.
4. Compare the predictions with real data.
5. If they do not agree, change the assumptions and try again.

Common modelling assumptions

Situation	Typical model	Assumption made
A fair die or coin	Each outcome equally likely	The die or coin is unbiased
A random selection	Every item has the same chance of being chosen	The selection really is random
Repeated trials	The result of one trial does not affect the next	Trials are independent
A measured quantity such as height	Roughly bell-shaped	The values are symmetrical about a mean
A linear relationship	A straight line fits the points	The change in one variable is steady as the other changes

Commenting on a model

A common question gives a model and asks for a comment, or asks why it might not be suitable. A good answer has three parts.

Note Academy

1. Name the assumption.
2. Say whether it is likely to be true in this situation.
3. Say what difference it makes to the result.

> **TRY IT**
>
> The probability that a bus is late each day is modelled as 0.2, independent from day to day. Give one reason why this model may not be realistic.

Independence is unlikely. If there is a long road closure or bad weather, the bus is more likely to be late on one day if it was late the day before. The probability of being late would be higher on those days than 0.2, so the model would underestimate the chance of a run of late days.

> **TRY IT**
>
> A shop owner uses the mean daily sales over a week to predict sales for the next year. Comment on the model.

The model assumes sales stay at the weekly mean. It ignores differences between days, seasons and holidays. A single week is also a small sample. The prediction could be well away from the real total, so a longer record, split by season, would improve it.

Samples and populations

A **population** is all the items of interest and a **sample** is the part of it that is measured. A model built from a sample is only reliable if the sample represents the population.

> **A sample is random when every member of the population has an equal chance of being chosen. A biased sample, such as asking only people at a gym how often they exercise, makes the model unreliable.**

Discrete and continuous

Data from counting (the number of goals) are **discrete** and take separate values. Data from measuring (a time or a mass) are **continuous** and can take any value in a range. The model and the diagram used depend on which it is.

> **NOTE**

Note Academy

Do not invent assumptions. Use the ones in the question, and say what the model would predict if the assumption were false.

Note Academy