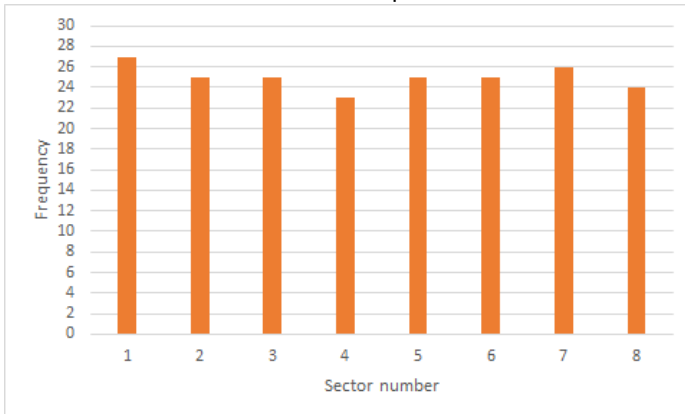


Probability distribution modelling process:

- Look at the data (hopefully in a graph) and describe the features
- Read the information about the context and identify any parameters
- Select a distribution and justify your selection (use visual features and contextual reasons)
- Check fit of model VISUALLY by calculating a few probabilities and sketching the theoretical shape (discrete distributions) OR by using the parameters to sketch the theoretical shape (continuous distributions) - it's important to go "through" the heights of the bars not just on the "outside"
- Do some probability calculations and compare what the model gives you with what the data gives you

Example 1

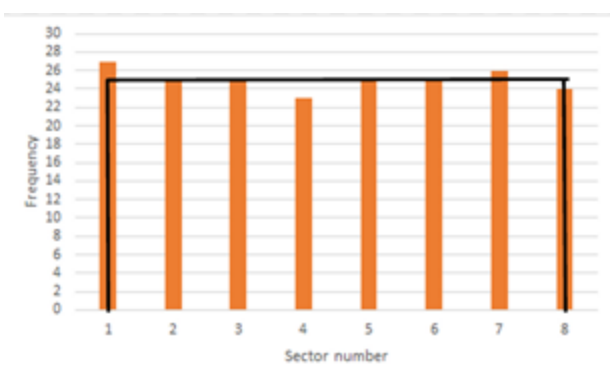
Bob has a spinner that is divided into eight equally-sized sectors. Below are shown the results of 200 spins:



State which model could be used for this probability distribution and justify your selection.

Uniform distribution as discrete variable and each outcome equally likely to happen. Min value of 1 and max value of 8. Shape also looks rectangular. Assume spinner is not biased towards any one sector.

Sketch the theoretical model on the graph and do some calculations to check the fit of the model.

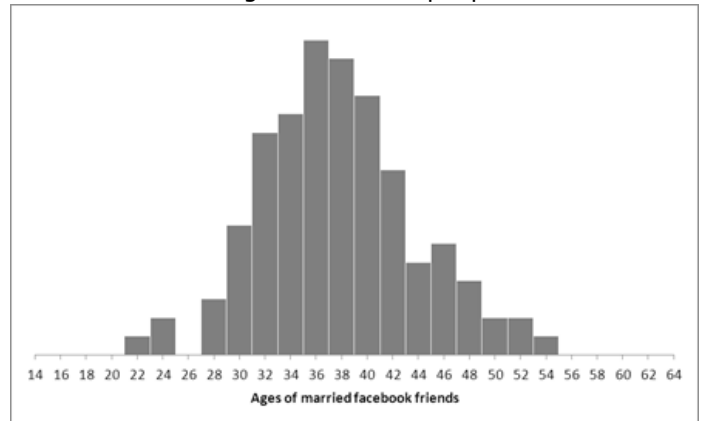


Theoretical model looks a pretty close fit.

Using data, getting $P(X \leq 2) = 52/200 = 0.26$. Using uniform model, getting $P(X \leq 2) = 0.25$, so it's pretty close.

Example 2

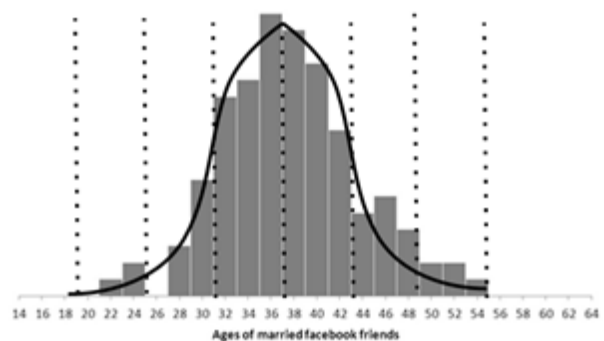
A certain Statistics teacher has generated a graph of the distribution of the ages of married people on Facebook.



State which model could be used for this probability distribution and justify your selection.

Normal, as symmetric, age is a continuous variable (years), looks like most of ages close to mean age of around 37 years (bell shaped), standard deviation around 6 years

Sketch the theoretical model on the graph and do some calculations to check the fit of the model.



Theoretical model looks a pretty close fit.

Can't compare calculations for data vs theoretical as don't have proportions/frequencies for data, but looks like most of data within two standard deviations of the mean.