Stats Yr2 Chapter 3: Distribution-N

The Normal Gaussian

Chapter Overview

1:: Characteristics of the Normal Distribution

What shape is it? What parameters does it have?

2:: Finding probabilities on a standard normal curve.

“Given that IQ is distributed as $X \sim N(100, 15^2)$, determine the probability that a randomly chosen person has an IQ above 130.”

3:: Finding unknown means/standard deviations.

In Wales, 30% of people have a height above 1.6m. Given the mean height is 1.4m and heights are normally distributed, determine the standard deviation of heights.

4:: Binomial → Normal Approximations

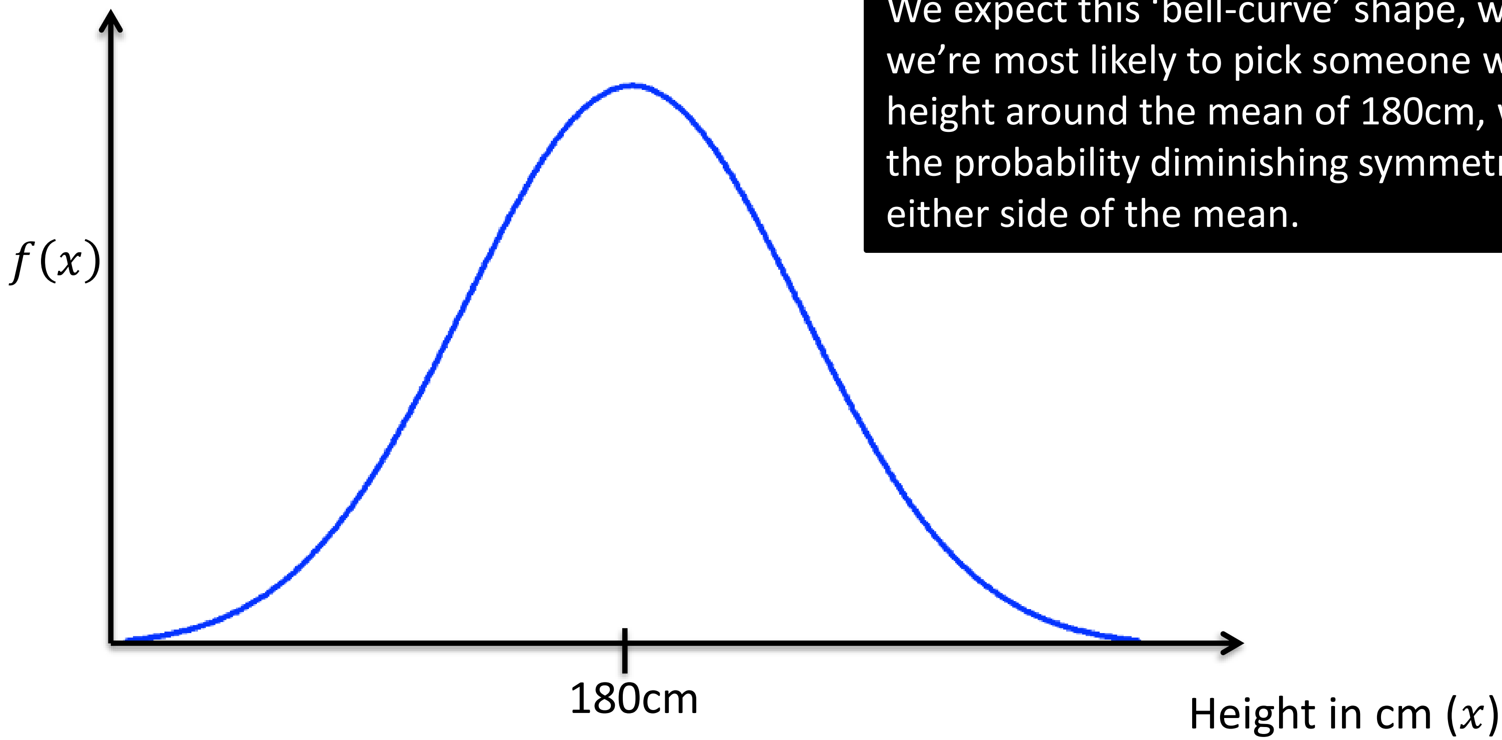
How would I approximate $X \sim B(10, 0.4)$ using a Normal distribution? Under what conditions can we make such an approximation?

5:: Hypothesis Testing

Teacher Notes: This is a combination of all the old S1 content combined with aspects of S2 (Normal approximations) and S3! (hypothesis testing on the mean of a normal distribution)

What does it look like?

The following shows what the probability distribution might look like for a random variable X, if X is the height of a randomly chosen person.

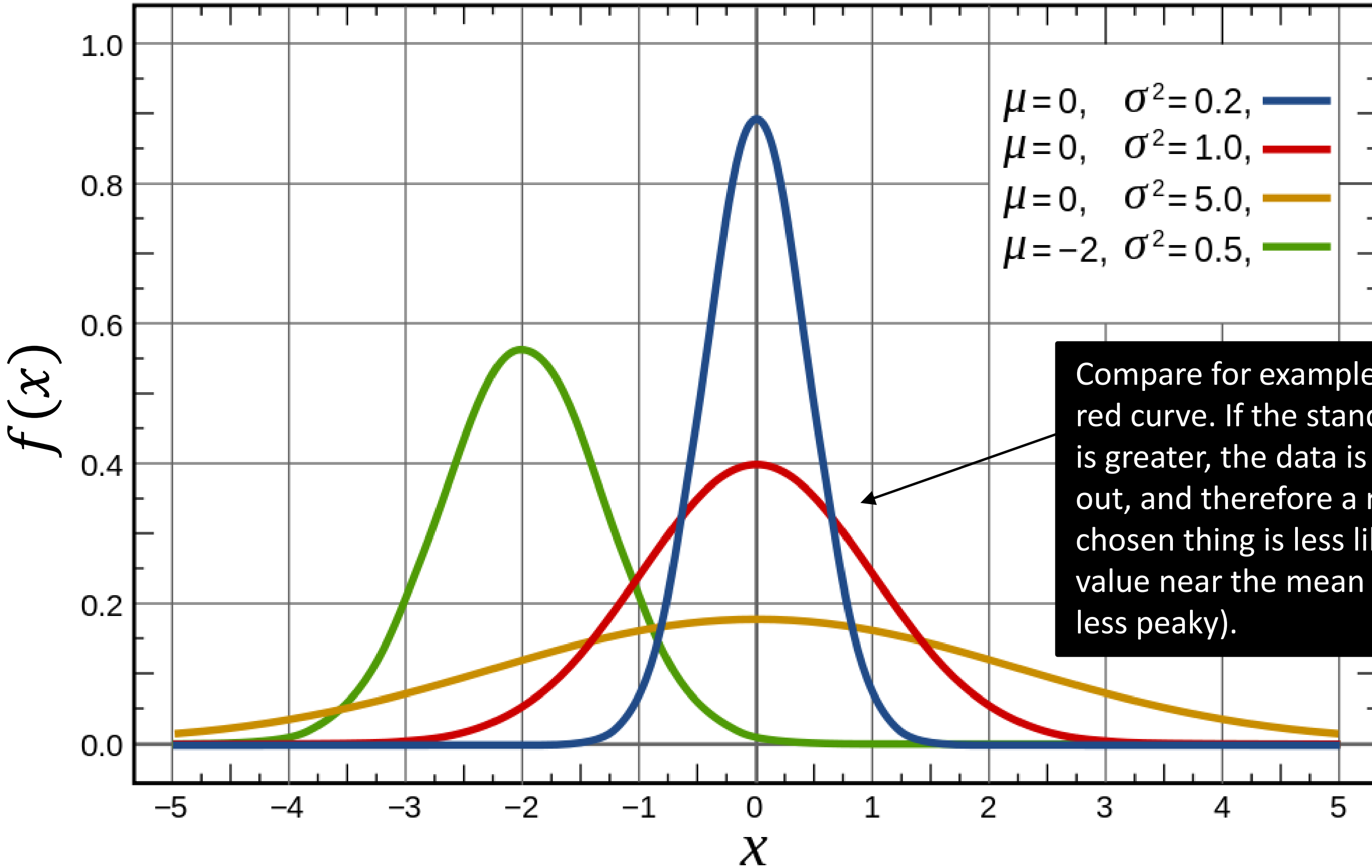


We expect this 'bell-curve' shape, where we're most likely to pick someone with a height around the mean of 180cm, with the probability diminishing symmetrically either side of the mean.

A variable with this kind of distribution is said to have a **normal distribution**.

For normal distributions we tend to draw the y axis at the mean for symmetry.

What does it look like?



Compare for example the blue and red curve. If the standard deviation is greater, the data is more spread out, and therefore a randomly chosen thing is less likely to have a value near the mean (hence it being less peaky).

We can set the mean μ and the standard deviation σ of the Normal Distribution. If a random variable X is normally distributed, then we write

$$\boldsymbol{X \sim N(\mu, \sigma^2)}$$

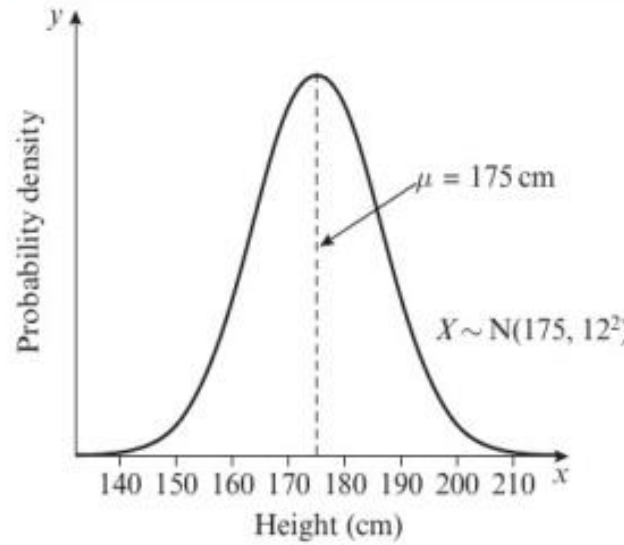
Key Facts

- **The normal distribution**
  - **has parameters μ, the population mean and σ^2, the population variance**
  - **is symmetrical (mean = median = mode)**
  - **has a bell-shaped curve with asymptotes at each end**
  - **has total area under the curve equal to 1**
  - **has points of inflection at $\mu + \sigma$ and $\mu - \sigma$**

For a normally distributed variable:

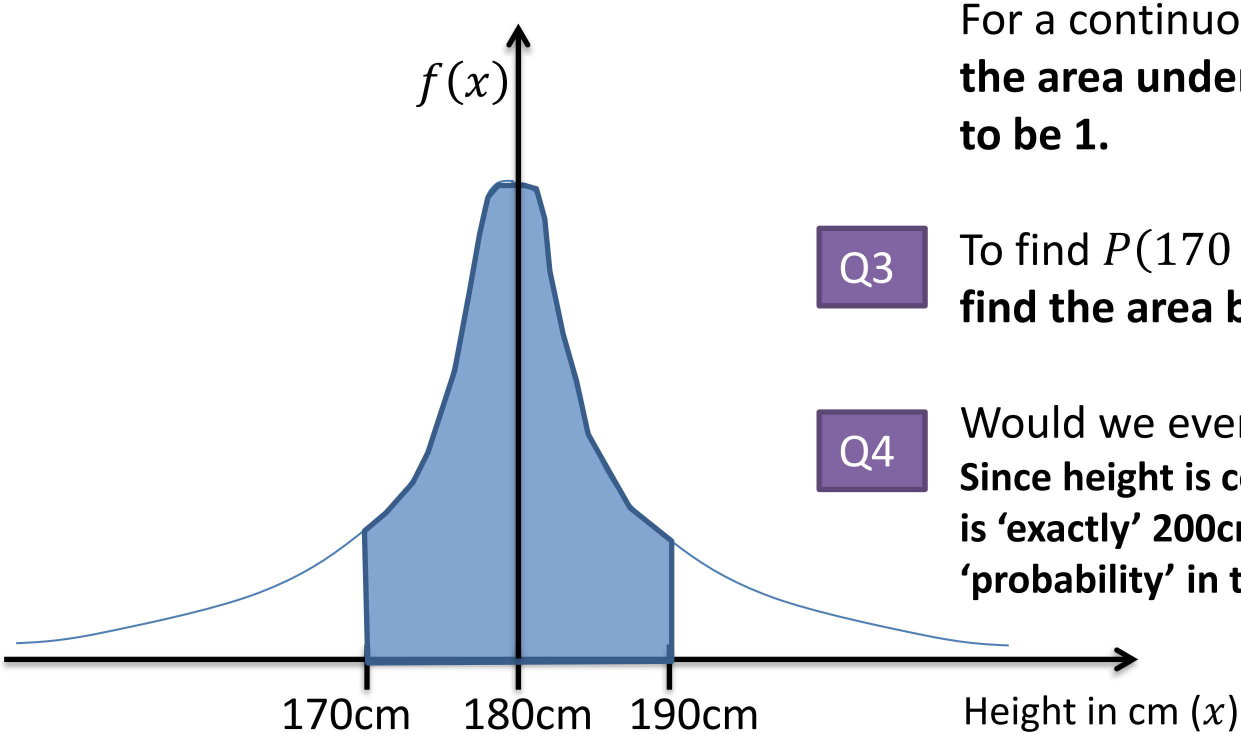
- approximately 68% of the data lies within one standard deviation of the mean
- 95% of the data lies within two standard deviations of the mean
- nearly all of the data (99.7%) lies within three standard deviations of the mean

Notation If X is a normally distributed random variable, you write $X \sim N(\mu, \sigma^2)$ where μ is the population mean and σ^2 is the population variance.



Watch out Although a normal random variable could take any value, in practice observations a long way (more than 5 standard deviations) from the mean have probabilities close to 0.

Normal Distribution Q & A



Q1 For a Normal Distribution to be used, the variable has to be:
continuous

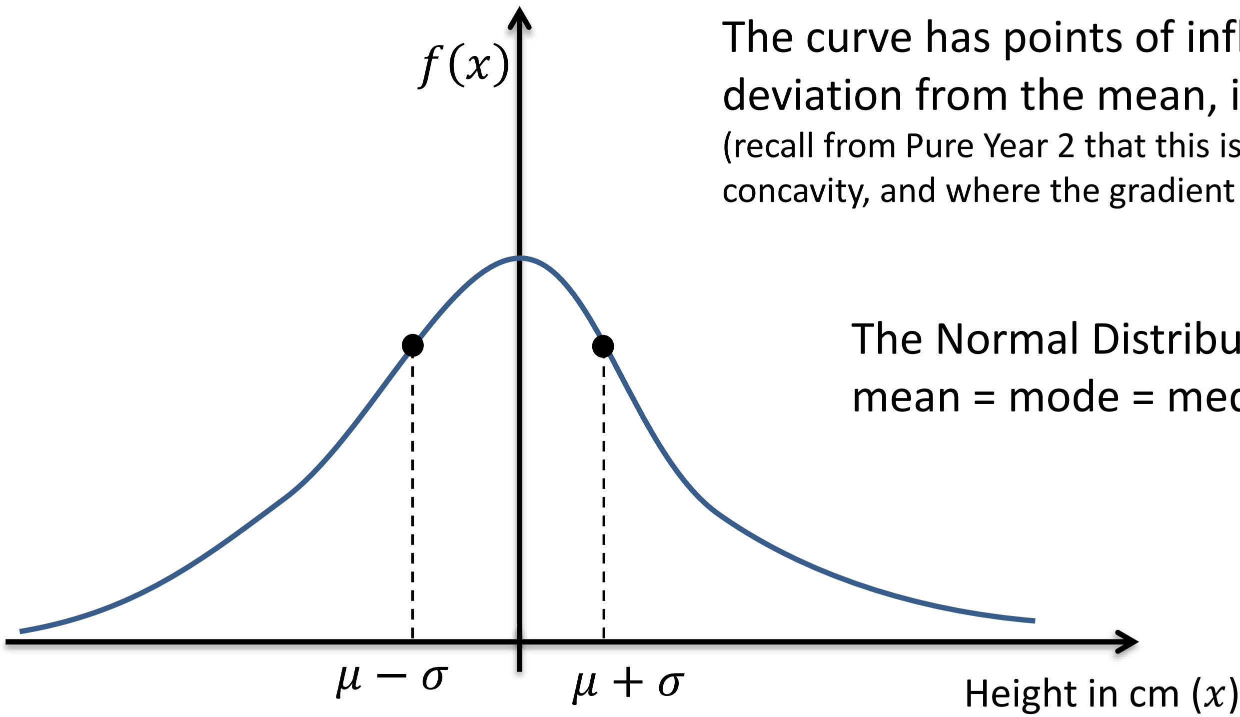
Q2 With a discrete variable, all the probabilities had to add up to 1.
For a continuous variable, similarly:
the area under the probability graph has to be 1.

Q3 To find $P(170 < X < 190)$, we could:
find the area between these values.

Q4 Would we ever want to find $P(X = 200)$ say?
Since height is continuous, the probability someone is 'exactly' 200cm is infinitesimally small. So not a 'probability' in the normal sense.

Side Notes: You might therefore wonder what the y-axis actually is. It is **probability density**, i.e. "the probability per unit cm". This is analogous to frequency density with histograms, where the y-value is frequency density area under the graph gives frequency. We use $f(x)$ rather than $p(x)$, to indicate probability density.

Further Facts



The curve has points of inflection one standard deviation from the mean, i.e. $\mu \pm \sigma$
(recall from Pure Year 2 that this is where the curve changes concavity, and where the gradient is not changing)

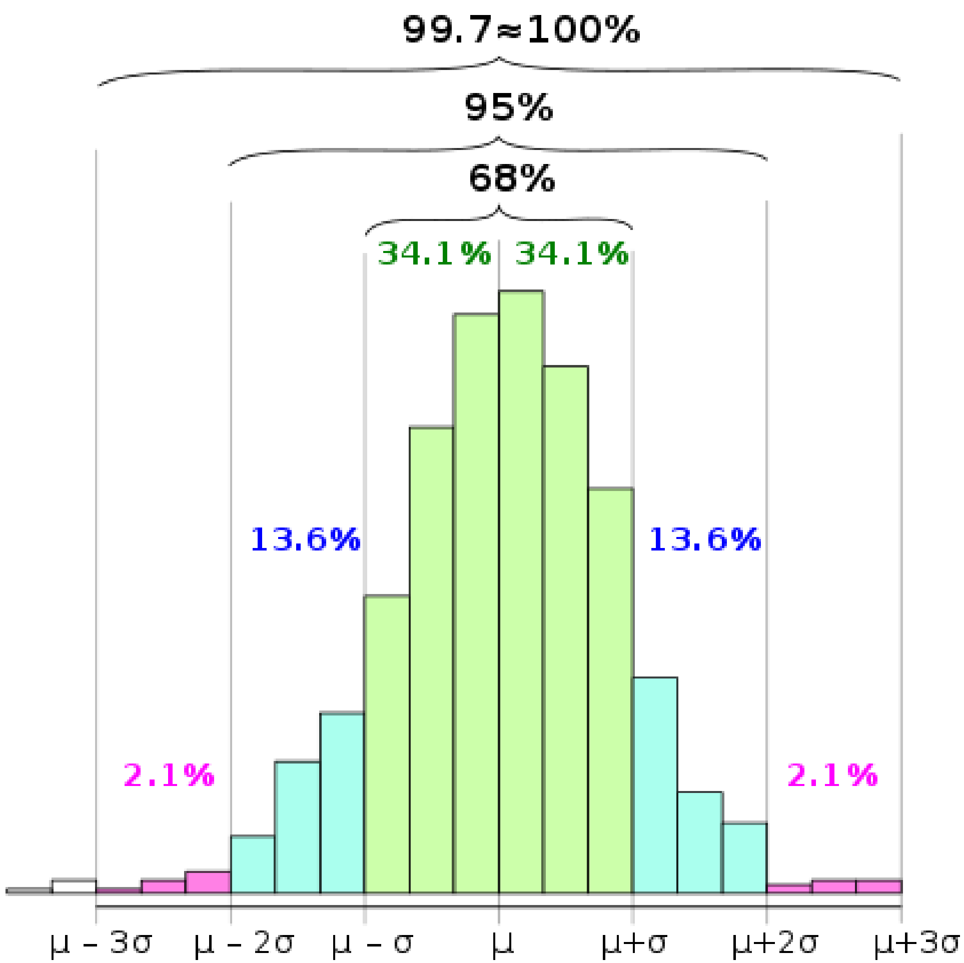
The Normal Distribution is symmetrical, i.e. mean = mode = median

Just For Your Interest™: The distribution, with a given mean μ and given standard deviation σ, that '**assumes the least**' (i.e. has the **maximum possible 'entropy'**) is… the Normal Distribution!

Difficult Proof: https://en.wikipedia.org/wiki/Differential_entropy#Maximization_in_the_normal_distribution

Extra Context: This is important in something called *Bayesian Statistics*. We often have to choose a suitable distribution for the 'prior' in the model (i.e. some 'hidden' variable). When making inferences based on observed data, we want to assume *as little as possible* about any hidden variable, so using a Normal distribution therefore is the most mathematically appropriate choice.

The 68-95-99.7 rule



The histogram above is for a quantity which is approximately normally distributed.
Source: Wikipedia

The 68-95-99.7 rule is a shorthand used to remember the percentage of data that is within 1, 2 and 3 standard deviations from the mean respectively.
<u>You need to memorise this!</u>

$\approx$ 68% of data is within one standard deviation of the mean.
$\approx$ 95% of data is within two standard deviations of the mean.
$\approx$ 99.7% of data is within three standard deviations of the mean.

For practical purposes we consider all data to lie within $\mu \pm 5\sigma$

Only one in 1.7 million values fall outside $\mu \pm 5\sigma$. CERN used a "5 sigma level of significance" to ensure the data suggesting existence of the Higgs Boson wasn't by chance: this is a 1 in 3.5 million chance (if we consider just one tail).

Examples

[Textbook] The diameters of a rivet produced by a particular machine, X mm, is modelled as $X \sim N(8, 0.2^2)$. Find:

a) $P(X > 8)$

b) $P(7.8 < X < 8.2)$

Fro Tip: Draw a diagram!

a

?

b

?

Examples

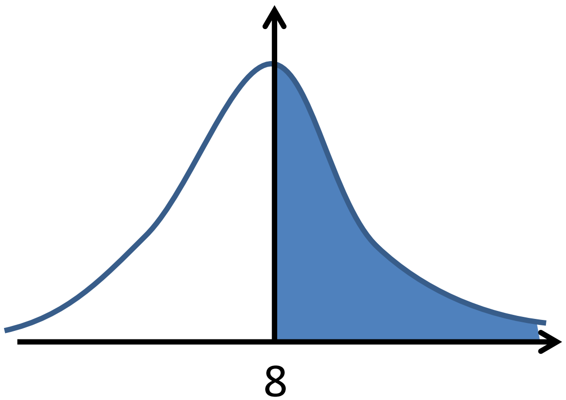
[Textbook] The diameters of a rivet produced by a particular machine, X mm, is modelled as $X \sim N(8, 0.2^2)$. Find:

a) $P(X > 8)$

b) $P(7.8 < X < 8.2)$

Fro Tip: Draw a diagram!

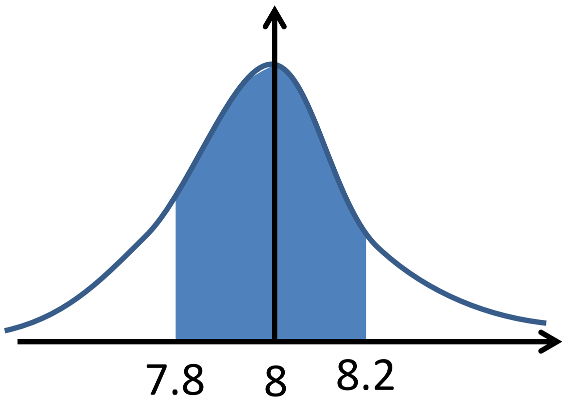
a



8 is the mean, so by the symmetry of the normal distribution, 50% of the area lies above the mean.

$$\therefore P(X > 8) = 0.5$$

b



The standard deviation is 0.2, so the data lies within $\mu \pm \sigma$

$$\therefore P(7.8 < X < 8.2) = 0.68$$

Test Your Understanding

IQ ("Intelligence Quotient") for a given population is, by definition, distributed using $X \sim N(100, 15^2)$. Find:

a) $P(70 < X < 130)$

b) $P(X > 115)$

Fro Tip: Draw a diagram!

a

?

b

?

Test Your Understanding

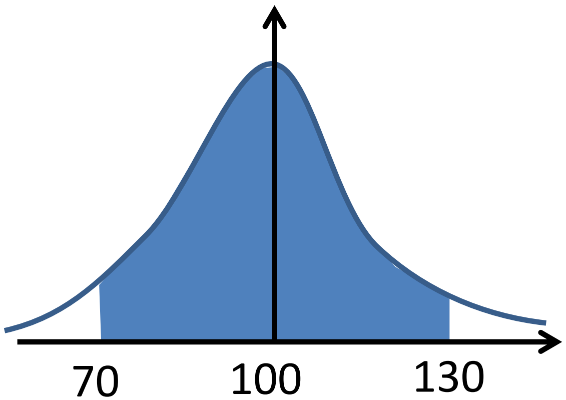
IQ ("Intelligence Quotient") for a given population is, by definition, distributed using $X \sim N(100, 15^2)$. Find:

a) $P(70 < X < 130)$

b) $P(X > 115)$

Fro Tip: Draw a diagram!

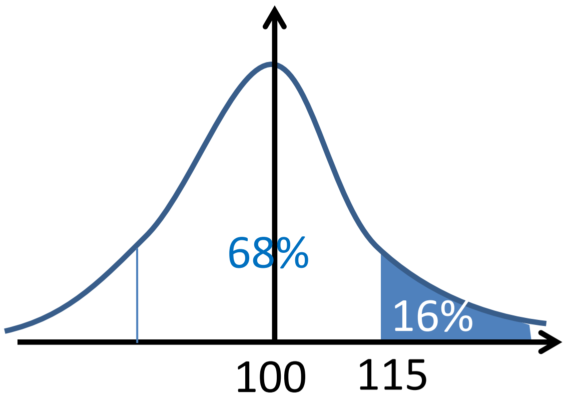
a



We know that 95% of data lies within 2 standard deviations of the mean.

$\therefore P(70 < X < 130) = 0.95$

b



68% lies within one standard deviation, so there must be 16% at each tail.

$\therefore P(X > 115) = 0.16$

Exercise 3.1

Pearson Stats/Mechanics Year 2
Pages 21-22

Homework Exercise

1 State, with a reason, whether these random variables are discrete or continuous:

a X, the lengths of a random sample of 100 sidewinder snakes in the Sahara desert

b Y, the scores achieved by 250 students in a university entrance exam

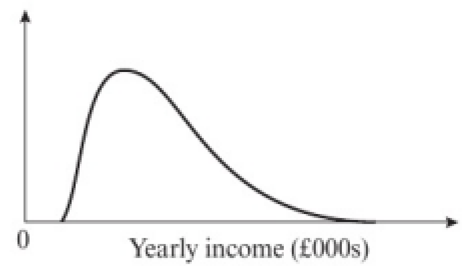
c C, the masses of honey badgers in a random sample of 1000

d Q, the shoe sizes of 200 randomly selected women in a particular town.

2 The lengths, X mm, of a bolt produced by a particular machine are normally distributed with mean 35 mm and standard deviation 0.4 mm. Sketch the distribution of X.

3 The distribution of incomes, in £000s per year, of employees of a bank is shown on the right.

State, with reasons, why the normal distribution is not a suitable model for this data.



4 The armspans of a group of Year 5 pupils, X cm, are modelled as $X \sim N(120, 16)$.

a State the proportion of pupils that have an armspan between 116 cm and 124 cm.

b State the proportion of pupils that have an armspan between 112 cm and 128 cm.

5 The lengths of a colony of adders, Y cm, are modelled as $Y \sim N(100, \sigma^2)$. If 68% of the adders have a length between 93 cm and 107 cm, find σ^2.

Homework Exercise

6 The weights of a group of dormice, D grams, are modelled as $D \sim N(\mu, 25)$. If 97.5% of dormice weigh less than 70 grams, find μ.

Problem-solving

Draw a sketch of the distribution. Use the symmetry of the distribution and the fact that 95% of the data lies within 2 standard deviations of the mean.

7 The masses of the pigs, M kg, on a farm are modelled as $M \sim N(\mu, \sigma^2)$. If 84% of the pigs weigh more than 52 kg and 97.5% of the pigs weigh more than 47.5 kg, find μ and σ^2.

8 The percentage scores of a group of students in a test, S, are modelled as a normal distribution with mean 45 and standard deviation 15. Find:

a $P(S > 45)$ **b** $P(30 < S < 60)$ **c** $P(15 < S < 75)$

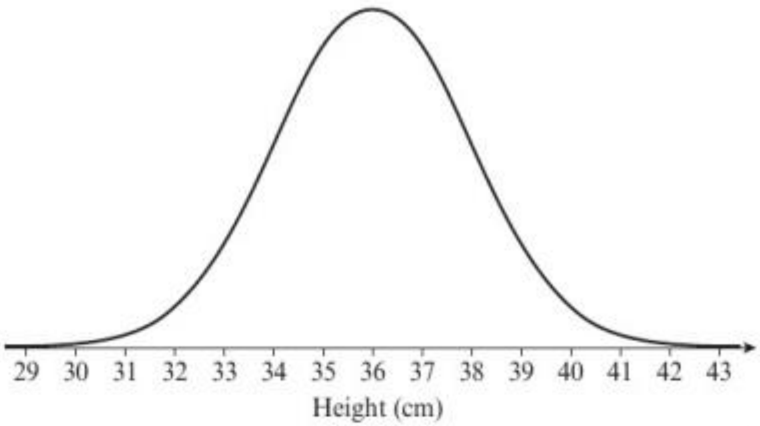
Alexia states that since it is impossible to score above 100%, this is not a suitable model.

d State, with a reason, whether Alexia is correct.

9 The diagram shows the distribution of heights, in cm, of barn owls in the UK.

An ornithologist notices that the distribution is approximately normal.

Hint The points of inflection on a normal distribution curve occur at $\mu \pm \sigma$.



a State the value of the mean height. **(1 mark)**

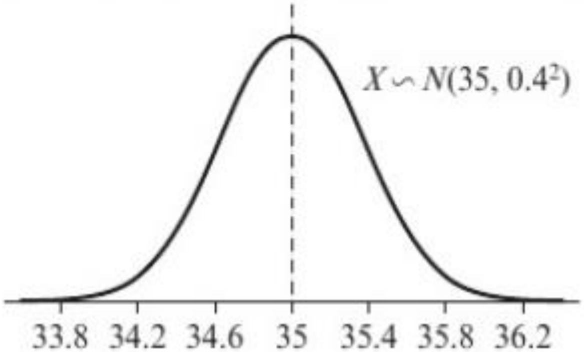
b Estimate the standard deviation of the heights. **(2 marks)**

Homework Answers

For Chapter 3, student answers may differ slightly from those shown here when calculators are used rather than table values.

1 **a** Continuous – lengths can take any value
b Discrete – scores can only take certain values
c Continuous – masses can take any value
d Discrete – show sizes can only take certain values

2



3 The distribution is not symmetrical.

4 **a** 0.68 **b** 0.95

5 49

6 60g

7 $\mu = 56.7$ (3 s.f.), $\sigma^2 = 4.69^2$ (3 s.f.)

8 **a** 0.5 **b** 0.683 (3 s.f.) **c** 0.954 (3 s.f.)
d Incorrect: although $P(X > 100) > 0$, it is very small since 100 is more than 3 standard deviations away from the mean, so the model as a whole is still reasonable.

9 **a** 36 **b** Between 2 and 3