Edexcel A Level Further Maths: Further Statistics 1

Your notes

Geometric & Negative Binomial Distributions

Contents

Your notes

The Geometric Distribution

Conditions for Geometric Models

What is the geometric distribution?

- The **geometric distribution** models the **number of trials needed to reach the first success**
  - For example, how many times will you have to roll a dice until it lands on a '6' for the first time
- The **notation** for the geometric distribution is $\text{Geo}(p)$
  - For a random variable X that has the geometric distribution you can write $X \sim \text{Geo}(p)$
  - X is the number of trials it takes to reach the **first** success
    - For example, $X = 8$ means the first success occurred on the 8th trial
  - p is the fixed probability of success in any one trial

What are the conditions for using a geometric model?

- A **geometric** model can be used for an experiment that satisfies the following **conditions**:
  - The experiment consists of an indefinite number of successive **trials**
  - The outcome of each trial is **independent** of the outcomes of all other trials
  - There are exactly **two possible outcomes** for each trial (**success** and **failure**)
  - The **probability of success** in any one trial (p) is **constant**
- Note that these conditions are very similar to the conditions for the **binomial distribution**
  - But for a **binomial** distribution **the number of trials (n) is fixed**
    - And you count the number of successes
  - While for a **negative binomial** distribution **the experiment continues until the first success is achieved**
    - And you count the number of trials it takes to reach that first success

When might the conditions not be satisfied?

- If asked to **criticise** a geometric model, you may be able to question whether the **trials** are really **independent**
  - For example, someone may be repeating an activity until they achieve a success
    - The trials may **not** be independent because the person gets **better** from practising the activity
    - This also means the **probability** of success, p, is **not constant**
  - In order to proceed using the model, you would have to **assume** that the trials are **independent**

Examiner Tip

- Replace the word "trials" with the context (e.g. "flips of a coin") when commenting on conditions and assumptions

Your notes

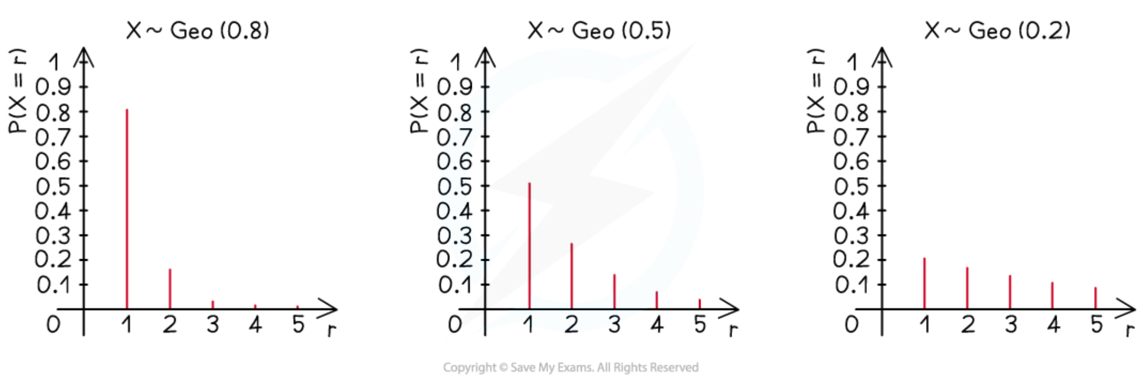
Geometric Probabilities

What are the probabilities for the geometric distribution?

- If $X \sim \text{Geo}(p)$ then X has the **probability function**:
  - $\mathrm{P}(X=x)=\mathrm{p}(x)=p(1-p)^{x-1}, \quad x=1, 2, 3, \ldots$
  - the random variable X is the number of trials needed to get the **first** success
  - p is the constant probability of success in one trial
  - $\mathrm{P}(X=x)$ is the probability that the **first** success will occur on the xth trial
- Note that that is the **product** of
  - the **probability of first getting $x-1$ failures**, $(1-p)^{x-1}$,
  - and the **probability of getting a success in the xth trial**, p
- Also note that there is no greatest possible x
  - It could require **any number** of trials to reach the **first** success
  - However $\mathrm{P}(X=x)$ gets closer and closer to zero as x gets larger
- Your **calculator** may allow you to calculate Geometric probabilities directly
  - i.e., without having to use the above formula

What are the properties of the geometric distribution?

- Note that
  $\mathrm{P}(X=1), \mathrm{P}(X=2), \mathrm{P}(X=3), \mathrm{P}(X=4), \ldots = p, p(1-p), p(1-p)^2, p(1-p)^3, \ldots$
  - This means that the geometric probabilities form a **geometric sequence**
    - The **first term** is p
    - The **common ratio** is $(1-p)$
  - This is where the geometric distribution gets its name!
- Assuming that $0<p<1$, then it is also true that $0<(p-1)<1$
  - This means that $\mathrm{P}(X=1)>\mathrm{P}(X=2)>\mathrm{P}(X=3)>\mathrm{P}(X=4)>\ldots$
    - i.e., the probabilities form a decreasing sequence
    - and $\mathrm{P}(X=1)$ is the largest probability in the sequence
    - Therefore $X=1$ is the **mode** of the distribution



Your notes

- The geometric distribution has no 'memory'
  - It doesn't matter what has happened previously, or how many 'failures' in a row there have been
  - The probability of getting a 'success' in **any** trial is always p
  - This means that the number of additional trials needed for the first success is not dependent on the number of trials that have already occurred
    - e.g. if 5 (failed) trials have already occurred, the probability of the first success happening after 7 trials is simply the probability of success happening after 2 trials in the first place, i.e. $p(1-p)$

What are the cumulative probabilities for the geometric distribution?

- If $X \sim \mathrm{Geo}(p)$ then X has the **cumulative geometric distribution**:
  - $\mathrm{P}(X \leq x) = 1 - (1-p)^x, \quad x = 1, 2, 3, \ldots$
  - the random variable X is the number of trials needed to get the **first** success
  - p is the constant probability of success in one trial
  - $\mathrm{P}(X \leq x)$ is the probability that the **first** success will occur **on or before** the xth trial
- Your **calculator** may allow you to calculate Geometric probabilities directly
  - i.e., without having to use the above formula
- The formula can be **proved** as follows
  - If the **first** success occurs on or before the xth trial, that means that **the first x trials have not all been failures**
    - The probability of getting x failures in a row is $(1-p)^x$
    - So the probability of that **not** happening is $1-(1-p)^x$
- Alternatively, it can be proved **algebraically**
  - The geometric probabilities form a **geometric sequence** with first term p and common ratio $1-p$
  - Putting that into the **geometric series** formula $S_n = \dfrac{a(1-r^n)}{1-r}$ gives

$$\mathrm{P}(X \leq x) = \sum_{r=1}^{x} P(X=r) = \frac{p(1-(1-p)^x)}{1-(1-p)} = \frac{p(1-(1-p)^x)}{p} = 1-(1-p)^x$$

- Because $\lim_{x \to \infty} (1-p)^x = 0$ (assuming $0 < p < 1$), it follows that

$$\sum_{r=1}^{\infty} \mathrm{P}(X=r) = 1 - 0 = 1$$

  - So the **sum** of **all probabilities** is equal to **1**
    - This is a requirement of any probability distribution

Examiner Tip

- If you forget the formulae in the exam, you can often still do questions using basic probability concepts and geometric series

Your notes

Your notes

Worked example

Joshua is an inspector in a factory. His job is to randomly sample widgets produced by a particular machine, until he finds a widget that has a defect. If he finds a widget with a defect, then the machine must be stopped until a repair procedure has been completed. Given that the probability of a widget being defective is 0.002, find the probability that:

a) the 10th widget that Joshua inspects is the first one that is defective

Let X = number of widgets inspected until first defective one

$X \sim \text{Geo}(0.002)$

Use $P(X=x) = p(1-p)^{x-1}$

$P(X=10) = 0.002(0.998)^9 = 0.00196428...$

$= \boxed{0.00196 \text{ (3 s.f.)}}$

Note: Your calculator may have a Geometric Distribution function that can calculate these probabilities directly

b) the 250th widget that Joshua inspects is the first one that is defective

$P(X=250) = 0.002(0.998)^{249} = 0.00121488...$

$= \boxed{0.00121 \text{ (3 s.f.)}}$

c)

Your notes

the 250th widget that Joshua inspects is the first one that is defective, given that the first 240 were not defective

$P(X = 250 \mid X > 240) = P(X = 10)$

The geometric distribution has no 'memory'

$= \boxed{0.00196 \text{ (3 s.f.)}}$

i.e., same as answer to part (a)

d) Joshua will inspect 250 or fewer widgets before finding the first one that is defective

Use $P(X \le x) = 1 - (1 - p)^x$

$P(X \le 250) = 1 - (0.998)^{250} = 0.393772...$

$= \boxed{0.394 \text{ (3 s.f.)}}$

Note: Your calculator may have a Cumulative Geometric Distribution function that can calculate these probabilities directly

e) Joshua will need to inspect more than 250 widgets before finding the first one that is defective.

$P(X > 250) = 1 - P(X \le 250) = 0.998^{250}$

$= 0.606227... = \boxed{0.606 \text{ (3 s.f.)}}$

f) State an assumption you have used in calculating the above probabilities.

It has been assumed that any one widget being defective is independent of any other widget being defective.

Your notes

Your notes

Geometric Mean & Variance

What are the mean and variance of the geometric distribution?

- If $X \sim \text{Geo}(p)$, then
  - The **mean** of X is $\text{E}(X) = \mu = \frac{1}{p}$
  - The **variance** of X is $\text{Var}(X) = \sigma^2 = \frac{1-p}{p^2}$
- You need to be able to **use these formulae** to answer questions about the geometric distribution

Examiner Tip

- If a question gives you the value of the mean or variance, form an equation in p and solve it

Your notes

Worked example

Palamedes is rolling a biased dice for which the probability of the dice landing on a '6' is p. The random variable X represents the number of times he needs to roll the dice until a '6' appears for the first time.

Given that the standard deviation of X is $2\sqrt{3}$, find:

a) the value of p.

$X \sim \text{Geo}(p)$

Use $\text{Var}(X) = \sigma^2 = \dfrac{1-p}{p^2}$

$$\frac{1-p}{p^2} = (2\sqrt{3})^2 = 12$$

$$12p^2 + p - 1 = (4p-1)(3p+1) = 0$$

$$p = \frac{1}{4} \text{ or } -\frac{1}{3}$$

But p is a probability, so it can't be negative.

$$\boxed{p = \frac{1}{4}}$$

b) the mean of X

Your notes

Use $E(X) = \mu = \frac{1}{p}$

$E(X) = \frac{1}{1/4} = \boxed{4}$

c) $\mathrm{P}(X \le 3)$.

$X \sim \text{Geo}\left(\frac{1}{4}\right)$

Use $P(X \le x) = 1 - (1-p)^x$

$P(X \le 3) = 1 - \left(1 - \frac{1}{4}\right)^3 = \boxed{\frac{37}{64}}$

Note: Your calculator may have a Cumulative Geometric Distribution function that can calculate these probabilities directly

Your notes

The Negative Binomial Distribution

Conditions for Negative Binomial Models

What is the negative binomial distribution?

- The negative binomial distribution models the **number of trials needed to reach a fixed number of successes**, r
  - For example, how many times will you have to roll a dice until it lands on a '6' for the third time
- There is no one standard form of **notation** for the negative binomial distribution
  - But for a random variable X that has the negative binomial distribution you could write either:
    - $X \sim \text{NB}(r, p)$ or $X \sim \text{Negative B}(r, p)$
  - X is the number of trials that will be required to reach a **total** of r **successes**
  - p is the fixed probability of success in any one trial

What are the conditions for using a negative binomial model?

- A **negative binomial** model can be used for an experiment that satisfies the following **conditions**:
  - The experiment consists of an indefinite number of successive **trials**
  - The outcome of each trial is **independent** of the outcomes of all other trials
  - There are exactly **two possible outcomes** for each trial (**success** and **failure**)
  - The **probability of success** in any one trial (p) is **constant**
- Note that these conditions are very similar to the conditions for the **binomial distribution**
  - But for a **binomial** distribution **the number of trials (n) is fixed**
    - And you count the number of successes
  - While for a **negative binomial** distribution **the number of successes (r) is fixed**
    - And you count the number of trials it takes to reach that number of successes

When might the conditions not be satisfied?

- If asked to **criticise** a negative binomial model, you may be able to question whether the **trials** are really **independent**
  - For example, someone may be repeating an activity until they achieve the rth success
    - The trials may **not** be independent because the person gets **better** from practising the activity
    - This also means the **probability** of success, p, is **not constant**
  - In order to proceed using the model, you will have to **assume** trials are **independent**

Examiner Tip

- Replace the word "trials" with the context (e.g. "flips of a coin") when commenting on conditions and assumptions

Your notes

Negative Binomial Probabilities

What are the probabilities for the negative binomial distribution?

- If $X \sim \text{Negative B}(r, p)$, then X has the **probability function**:
  - $\mathrm{P}(X=x)=\mathrm{p}(x)=\binom{x-1}{r-1}p^r(1-p)^{x-r}, \quad x=r, r+1, r+2, \ldots$
  - the random variable X is the number of trials needed to get r **successes**
  - p is the constant probability of success in one trial
  - $\mathrm{P}(X=x)$ is the probability that the $r\text{th}$ **success** will occur on the $x\text{th}$ trial
- Note that that is the **product** of
  - the **binomial probability of getting $r-1$ successes in $x-1$ trials**, $\binom{x-1}{r-1}p^{r-1}(1-p)^{x-r}$,
  - and the **probability of getting a success in the $x\text{th}$ trial**, p
- $\binom{x-1}{r-1}$ is the **binomial coefficient**
  - i.e., $\binom{x-1}{r-1}={}^{(x-1)}C_{(r-1)}$
- Also note that there is no greatest possible value of X
  - It could require **any number** of trials to reach the $r\text{th}$ success
  - However for any given r, $\mathrm{P}(X=x)$ gets closer and closer to zero as x gets larger

Where does the formula come from?

- Consider **rolling a fair dice** and wanting to know the **probability that it would take 12 rolls for the dice to land on '6' a total of 3 times**
  - This is the same as saying that the third '6' occurs on the 12th roll
- You can **model** this situation using the random variable $X \sim \text{Negative B}\left(3, \frac{1}{6}\right)$
  - 'success' here is defined as 'roll a 6'
  - $r=3$ because we're interested in the number of trials required to reach a total of 3 successes
  - $p=\frac{1}{6}$ is the probability of rolling a '6' on a fair dice
  - X is the number of trials that will be required to reach 3 successes
- The **probability** you are looking for is therefore $\mathrm{P}(X=12)$

Your notes

- For the third '6' to occur on the 12th roll, the following things need to happen:
  - '6' must occur **exactly 2 times in the first 11 rolls**
    - It doesn't matter which rolls those two '6's occur on
    - The probability of that happening is $\mathrm{P}(Y=2)$, for $Y \sim \mathrm{B}\left(11, \frac{1}{6}\right)$
    - So that part of the answer is a binomial probability
  - Then a '6' must **also occur on the 12th roll**
    - The probability of that happening is $\frac{1}{6}$
  - So the probability of both those things happening is

$$\mathrm{P}(Y=2) \times \frac{1}{6} = \left(\binom{11}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^9\right) \times \frac{1}{6} = \binom{11}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^9 = 0.0493489...$$

- That same logic can be used to find the **general formula** for negative binomial probabilities

Is there a connection between negative binomial probabilities and geometric probabilities?

- Note that when $r=1$,the negative binomial probability function becomes:

$$\mathrm{P}(X=x) = \mathrm{p}(x) = \binom{x-1}{0} p^1(1-p)^{x-1} = p(1-p)^{x-1}, \quad x = 1, 2, 3, ...$$

  - That is the same as the **geometric distribution** probability function
- The geometric distribution is the 'special case' of the negative binomial distribution when $r=1$

Examiner Tip

- Make sure you are clear about what x, r, p and $\mathrm{P}(X=x)$ refer to in the formula!
- Read the question carefully to determine whether binomial, geometric or negative binomial probabilities are required

Your notes

Worked example

Emanuel is playing in a chess tournament, where his probability of winning any one game is 0.55. Find the probability that:

a) his first win is in the third game he plays

$X \sim \text{Negative B}(1, 0.55)$

Use $P(X=x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$

$P(X=3) = \binom{2}{0}(0.55)(0.45)^2 = 0.111375$

$= \boxed{0.111 \ (3 \text{ s.f.})}$

You could also do this as $X \sim \text{Geo}(0.55)$

Then $P(X=3) = (0.55)(0.45)^2 = 0.111375$

b) he wins exactly 4 of his first 7 games

This is just a regular binomial question

$X \sim B(7, 0.55)$

$P(X=4) = \binom{7}{4}(0.55)^4(0.45)^3 = 0.291847\ldots$

$= \boxed{0.292 \ (3 \text{ s.f.})}$

c) he wins for the fourth time in his seventh game

$X \sim \text{Negative B}(4, 0.55)$

$P(X=7) = \binom{6}{3}(0.55)^4(0.45)^3 = 0.166770\ldots$

$= \boxed{0.167 \ (3 \text{ s.f.})}$

d) he wins for the fourth time in his seventh game, given that he won his first game

This means he will need 6 more games to get his next 3 wins.

$X \sim \text{Negative B}(3, 0.55)$

$P(X=6) = \binom{5}{2}(0.55)^3(0.45)^3 = 0.151609\ldots$

$= \boxed{0.152 \ (3 \text{ s.f.})}$

e) his fourth win occurs in or before his seventh game.

Your notes

Your notes

This means he's won at least 4 times in his first 7 games. So it's another regular binomial question.

$X \sim B(7, 0.55)$

$P(X \geq 4) = 1 - P(X \leq 3) = 0.608287...$

$= \boxed{0.608 \text{ (3 s.f.)}}$

f) Criticise the model used in this question.

The results of the games may not be independent, and the probability of winning may not always be the same. For example, he may play against different opponents, or he may improve with practice as the tournament goes on.

Your notes

Negative Binomial Mean & Variance

What are the mean and variance of the negative binomial distribution?

- If $X \sim \text{Negative B}(r, p)$, then
  - The **mean** of X is $\mathrm{E}(X) = \mu = \dfrac{r}{p}$
  - The **variance** of X is $\mathrm{Var}(X) = \sigma^2 = \dfrac{r(1-p)}{p^2}$
- You need to be able to **use these formulae** to answer questions about the negative binomial distribution.

Examiner Tip

- If a question gives you the mean and/or variance with one known parameters (r or p), form an equation to find the other

Your notes

Worked example

Croesus is tossing a biased coin for which the probability of the coin landing on 'heads' is p. The random variable X represents the number of times he needs to flip the coin until it has landed on heads four times. Given that the mean of X is 10, find:

a) the value of p

$X \sim \text{Negative B}(4, p)$

Use $E(X) = \mu = \frac{r}{p}$

$10 = \frac{4}{p} \Rightarrow \boxed{p = 0.4}$

b) the standard deviation of X

Use $\text{Var}(X) = \sigma^2 = \frac{r(1-p)}{p^2}$

$\sigma = \sqrt{\frac{4(1-0.4)}{0.4^2}} = \boxed{\sqrt{15}}$

c) the probability that the fourth 'head' will occur on the seventh toss.

Your notes

$X \sim \text{Negative B}(4, 0.4)$

Use $P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$

$P(X = 7) = \binom{6}{3}(0.4)^4(0.6)^3 = 0.110592$

$= \boxed{0.111 \text{ (3 s.f.)}}$