Chapter 22

Revision of Binomial, Hypergeometric and Geometric Probability Distributions

In Chapter 20.4 and Chapter 21, we learnt about the hypergeometric, binomial and geometric distributions. Often students find it difficult to recognise which of these three probability distributions might apply to a particular question, and so do not feel confident about knowing which formula to use to calculate a probability. In this chapter, we will revise the main features of those three types of probability distributions, emphasizing their main features, and the similarities/differences between those distributions. We will see that it is useful to ask ourselves whether an example contains an n-value, and whether it contains a p-value.

In the **binomial** distribution:

- we let X be the *number of successes* in a sequence of n Bernoulli trials.
  (There must be no conditions about the *order* that the successes and failures occur in.)
- the probability of a success remains *constant*, and is usually represented by p.
  The ***p*** is often given as a ***p***ercentage of some large population.

  We have $\boxed{\Pr(X = x) \;=\; {}^{n}\mathrm{C}_{x}\, p^{x}\, (1-p)^{n-x}\,.}$

  Notice that this binomial formula contains n and p symbols, and these *parameters* give us a clue about how to recognise whether or not an example is binomial.
- Since the binomial formula contains an n, this indicates that a binomial variable *cannot* increase indefinitely (since we must have $X \leq n$).
- Since the binomial formula contains a p, this indicates that the probability of a success *must be constant* in binomial examples.

In the **hypergeometric** distribution examples:

- we let X be the *number of items with a particular property in a sample* of size n.
- we are choosing (or "selecting" or "sampling") items *without replacement*. Thus the probability of a "success" changes (and so there is *no* p-value).
- we are told *how many* of the available items *have* the particular property (rather than being told a percentage for a larger population).

$$\text{Overall group: } \begin{cases} D \text{ items have property} \quad \text{and} \quad N-D \text{ do not.} \\ N \text{ items in total} \end{cases}$$

$$\text{Sample: } \begin{cases} x \text{ items have property} \quad \text{and} \quad n-x \text{ do not.} \\ n \text{ items in sample} \end{cases}$$

Then we have $$\boxed{\Pr(X = x) = \frac{{}^{D}\mathrm{C}_{x} \times {}^{N-D}\mathrm{C}_{n-x}}{{}^{N}\mathrm{C}_{n}}}$$

(obtained directly from our summaries of the overall group and the sample).

Hypergeometric examples are usually rather easy to recognise if we remember that those examples involve ***choosing a sample without replacement****.*

Alternatively, we can notice that the hypergeometric formula (above)

- contains an n (which is a feature it shares with the binomial formula), but
- does *not* contain a p.

Noticing these features helps us to recognise whether or not an example is hypergeometric. In particular,

- since the hypergeometric formula contains an n, this indicates that a hypergeometric variable *cannot* increase indefinitely (because we *cannot* have $X > n$).
- since the hypergeometric formula does *not* contain the symbol p, this indicates that the probability of a success *changes* in hypergeometric examples.

We have seen that an important similarity between hypergeometric and binomial examples is that they both contain an n-value. However, a crucial difference between hypergeometric and binomial examples is that hypergeometric variables have a ***changing*** *probability of success, whereas binomial examples have a* ***constant*** *probability of success.*

In the **geometric** distribution examples:

- we let X be the *number of failures before a success* (not including the success) in a sequence of Bernoulli trials.
- there is **no** maximum value for X (which is reflected by the absence of an n–symbol in the geometric formulae).
- the probability of a success remains *constant*, and is usually represented by p. As with the binomial distribution, the ***p***-value is often given as a ***p***ercentage of some large population.

Then we have $\boxed{\Pr(X = x) \;=\; (1-p)^x\, p}$

and $\boxed{\Pr(X \geq x) \;=\; (1-p)^x.}$

Note that this very useful formula for $\Pr(X \geq x)$ is *not* on the Formula Sheet!

Geometric examples are usually rather easy to recognise if we remember that those examples involve counting

how many times we perform an experiment
until *something special happens ("success").*

Alternatively, we can notice that the geometric formulae (above)

- do *not* contain n (in contrast to the binomial and hypergeometric examples), but
- *do* contain p (which is a feature the geometric distribution shares with the binomial distribution).

Noticing these features helps us to recognise whether or not an example is geometric. In particular,

- since the geometric formulae do *not* contain n, this indicates that a geometric variable has *no* maximum value.
- since the geometric formulae *do* contain the symbol p, this indicates that the probability of a success remains *constant* in geometric examples.

In our probability distribution examples, it is important to know *how to define the random variable* in each example. That is especially important when trying to decide *which* distribution is relevant for a particular example. Recall that

- in the binomial distribution, we define X to be
  the number of successes in a sequence of n Bernoulli trials.
- in the hypergeometric distribution, we define X to be
  the number of items with a particular property ("successes") in a sample of size n.
- in the geometric distribution, we define X to be
  the number of failures before the first success in a sequence of n Bernoulli trials.

In particular, notice that each of these definitions starts with the words ***"the number of"***. (Also notice that each of these random variables is *discrete*.) Thus, when we are trying to decide how to define a random variable, it can be very helpful to rearrange the words so that we *start with that particular phrase* ***"the number of"***.

In the following examples, blue text has been used to indicate the thought-processes that could help us to decide ***which*** *distribution is relevant. Hopefully this helps to give an idea of* ***how*** *to make that decision in examples where the distribution does not seem obvious.*

Note that, if it is ***already clear*** *to you which distribution is relevant, then it is* ***not*** *necessary to include the ideas expressed in blue text as part of the solution.*

Example 1. Suppose that 30% of adults have high blood–pressure. Consider a group of 25 adults. Find the probability that exactly 12 of these adults have high blood–pressure. Give your answer to 5 decimal places.

Solution:

We want to find
$\Pr($***the number of*** *adults in the group with high blood–pressure* $= 12)$.

So we let $X =$ ***the number of*** adults in the group with high blood–pressure. Our aim is to find $\Pr(X = 12)$.

- *We must have* $X \leq 25$ *(since there were 25 adults in the group).*
  Since this X *cannot increase indefinitely, this indicates that* **we have an n–value**. *Therefore, this* X *is definitely* ***not*** *geometric.*

To decide whether X is binomial or hypergeometric, we should think about whether or not there is a p-value. That is, we should decide whether or not

$$\Pr(\textit{an adult has high blood–pressure})$$

remains ***constant*** *(when we consider different adults).*

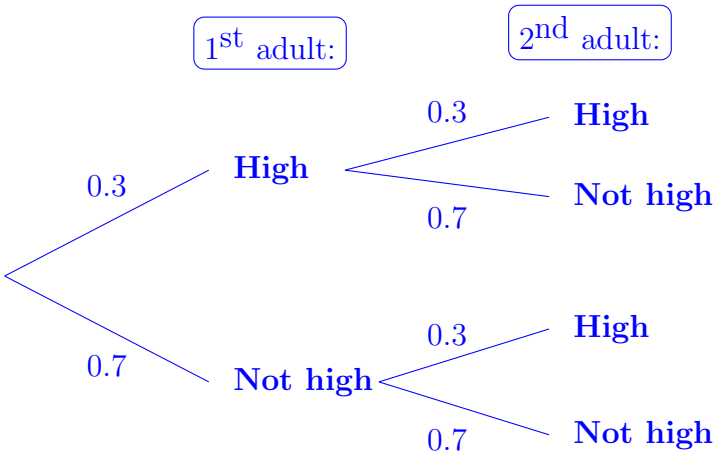
Notice that we are told that 30% of adults have high blood–pressure. This percentage of a large population gives us our p-value. We have

$$\boldsymbol{p = 0.3}\,.$$

However, if we ***did not recognize*** *that the percentage was telling us the p-value, then we could consider whether or not*

$$\Pr(\textit{an adult has high blood–pressure})$$

remains ***constant*** *(when we consider different adults), by considering just a small part of a tree diagram:*



We see that $\Pr(\textit{an adult has high blood–pressure})$ *is* ***constant*** *in this tree diagram, and is equal to* 0.3.

We have noticed that ***X* has an *n*-value and a *p*-value**, and so we conclude that X is *binomial*. We have $p = 0.3$ and $n = 25$, and so, using the probability formula for the binomial distribution, we find that

$$\begin{aligned}
\Pr(X = 12) &= {}^{n}\mathrm{C}_{12} \times p^{12} \times (1-p)^{n-12} \\
&= {}^{25}\mathrm{C}_{12} \times 0.3^{12} \times 0.7^{13} \\
&= 0.02678 \quad (5 \text{ d.p.})
\end{aligned}$$

□

Example 2. Consider a group of 25 adults, of whom 8 have high blood–pressure. If 10 of the adults from the group are tested (without replacement), find the probability that exactly 4 of those tested adults have high blood–pressure. Give your answer to 4 decimal places.

Solution:

We want to find

$$\Pr(\textit{the number of tested adults with high blood–pressure} = 4)\,.$$

So we let $Y =$ ***the number of*** tested adults with high blood–pressure.
Our aim is to find $\Pr(Y = 4)$.

Notice that we are told that we are testing the adults *without replacement*. Those words *"without replacement"* are an important signal that Y is *hypergeometric*.

However, if we did not realize the significance of the words "without replacement", then a useful way to decide which distribution we have is to think about whether or not there is an n-value and/or a p-value.

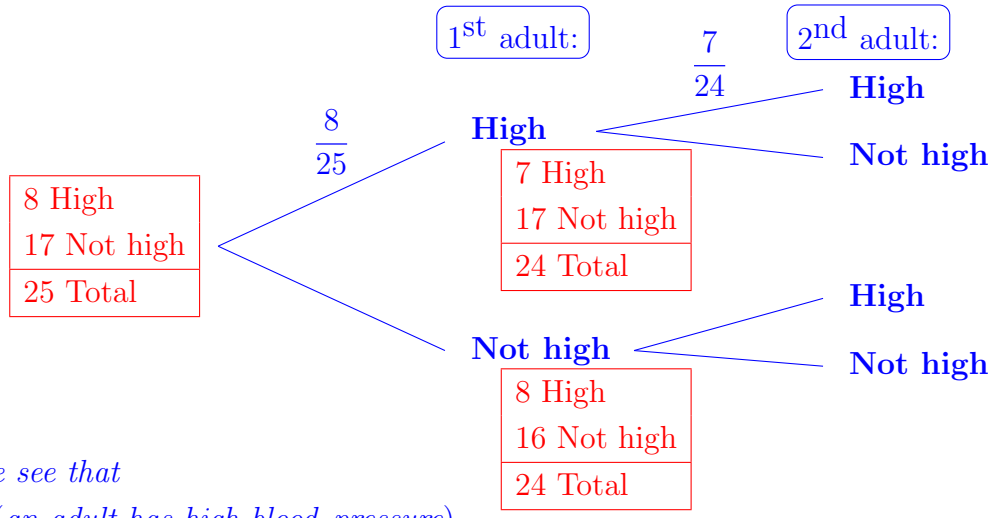
Since we are only testing 10 adults, we realize that we need $Y \leq 10$.

Alternatively, we are told that there are only 8 adults with high blood-pressure in the overall group, which implies that we must have $Y \leq 8$.

Using either reasoning, we see that this Y *cannot increase indefinitely, which indicates that* **we have an *n*–value**. *(Thus this* Y *is definitely* ***not*** *geometric.) Next we think about whether*

$$\Pr(\textit{an adult has high blood–pressure})$$

remains ***constant*** *or* ***changes*** *(when we consider different adults). We can do this by considering just a small part of a tree diagram:*



We see that
$\Pr(\textit{an adult has high blood–pressure})$
is ***changing*** *in this tree diagram. Thus there is* ***no*** *p-value.*

We have noticed that **Y has an n-value, but does *not* have a p-value**, and so we conclude that Y is *hypergeometric*.

For hypergeometric examples, it is useful to consider summaries of the overall group and the sample:

We are told that 8 of the 25 adults have high blood–pressure
*(which means that the other 17 adults do **not** have high blood–pressure).*

$$\text{Overall group: } \begin{cases} 8 \text{ adults have high blood–pressure} & \text{and} \quad 17 \text{ do not.} \\ 25 \text{ adults in total} \end{cases}$$

$$\text{Sample: } \begin{cases} 4 \text{ adults have high blood–pressure} & \text{and} \quad 6 \text{ do not.} \\ 10 \text{ tested adults} \end{cases}$$

$$\text{Thus we find that} \quad \Pr(Y = 4) = \frac{{}^{8}\mathrm{C}_{4} \times {}^{17}\mathrm{C}_{6}}{{}^{25}\mathrm{C}_{10}}$$

$$= 0.2650 \text{ (4 d.p.)}$$

□

Example 3. Suppose that I throw a fair die until I get a 6.

(a) What is the probability that I throw the die 10 times (including the throw which gives the 6)? Give your answer to 3 decimal places.

(b) What is the probability that I have between 3 and 20 **non**–sixes (inclusive) before I get a 6? Give your answer to 3 decimal places.

Solution:

We are throwing the die *until* we get a 6, and we can state this example in terms of how many non-sixes occur *before* a 6 occurs. Furthermore, Pr(get a 6) *remains constant*. This is the style of question we have with the *geometric* distribution.

Let $X =$ ***the number of*** $\underbrace{\text{non–sixes}}_{\textit{failures}}$ before I $\underbrace{\text{get a 6}}_{\textit{success}}$ (*not* including the 6).

This X is *geometric*, with $p = \Pr(\text{success}) = \Pr(\text{get a 6}) = \frac{1}{6}$.

(a) Throwing the die 10 times (including the throw that shows a "6" (success)), corresponds to having 9 "non-sixes" (failures) appear before the 6 appears:

$$\overbrace{\underbrace{F\,F\,F\,F\,F\,F\,F\,F\,F}_{9\text{ non-sixes}}\,S}^{10\text{ throws}}$$

Thus we need to find

$$\begin{aligned}
\Pr(10\text{ throws}) &= \Pr(\text{we get 9 non–sixes followed by a 6}) \\
&= \Pr(X = 9) \\
&= (1-p)^9\, p \\
&= \left(\frac{5}{6}\right)^9 \times \left(\frac{1}{6}\right) \\
&= 0.032 \ \ (3\text{ d.p.}).
\end{aligned}$$

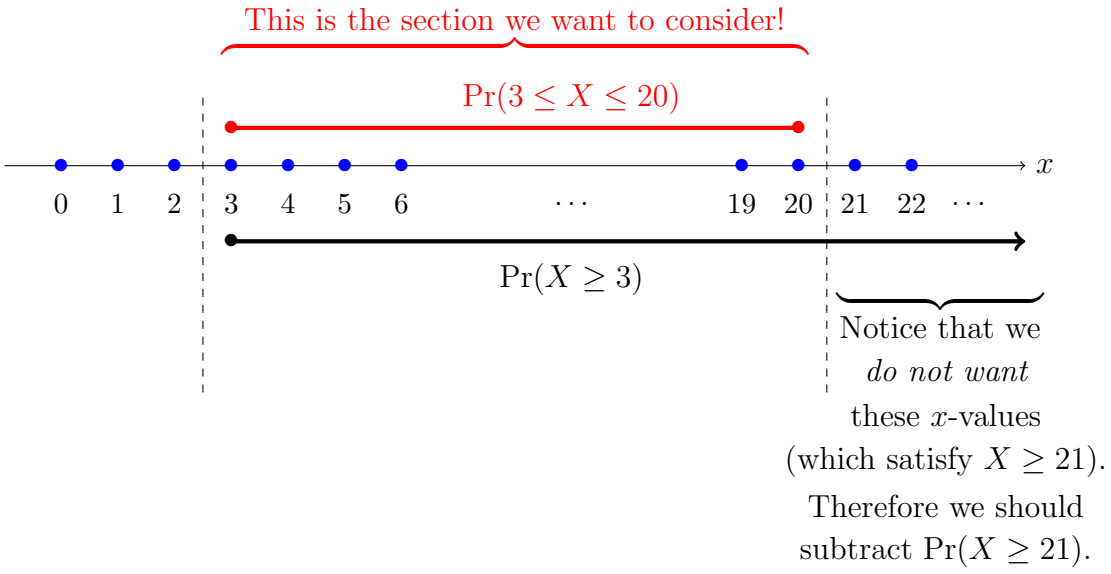
(b) We want to find $\mathbf{Pr(3 \leq X \leq 20)}$.

Notice that it would be *rather slow* to do this by calculating

$$\Pr(X = 3) + \Pr(X = 4) + \ldots + \Pr(X = 20).$$

So we look for a quicker approach, making use of the formula

$$\Pr(X \geq x) = (1-p)^x.$$



From this number-line, we see that

$$\Pr(3 \leq X \leq 20) = \Pr(X \geq 3) - \Pr(X \geq 21).$$

Alternatively, this can be seen by using the following reasoning:

$$\begin{aligned}
&\Pr(X \geq 3) \\
&= \Pr(X = 3) + \Pr(X = 4) + \ldots + \Pr(X = 20) + \Pr(X = 21) + \ldots \\
&= \Pr(3 \leq X \leq 20) \ + \ \Pr(X \geq 21).
\end{aligned}$$

Rearranging this gives

$$\Pr(3 \leq X \leq 20) \ = \ \Pr(X \geq 3) \ - \ \Pr(X \geq 21)$$

(as seen previously).

We calculate

$$\begin{aligned}
\Pr(3 \leq X \leq 20) &= \Pr(X \geq 3) \ - \ \Pr(X \geq 21) \\
&= (1-p)^3 \ - \ (1-p)^{21} \\
&= \left(\frac{5}{6}\right)^3 - \left(\frac{5}{6}\right)^{21} \\
&= 0.557 \ \ (3 \text{ d.p.}).
\end{aligned}$$

□