Normal Approximation to the Binomial

e.g. Estimate the probability that a school of 1200 students contains more than 150 left-handed students

Q: How would you solve such a problem?

A: One approach would be to take a large sample, say 50 students, and count the number of left handed students. From that information you could estimate the probability.

Say the sample contained 8 left handed students, we would estimate the probability of being left handed as;

$$P(\text{left handed}) = \frac{8}{50} = 0.16$$

Using binomial probability;

$$\text{Let } X = \text{number of left handed students}$$

$$X \sim \text{Bin}(1200, 0.16)$$

$$P(X > 150) = 1 - P(X \leq 150)$$

$$= 1 - \binom{1200}{0}(0.84)^{1200}(0.16)^{0} - \binom{1200}{1}(0.84)^{1199}(0.16)^{1} - \dots$$

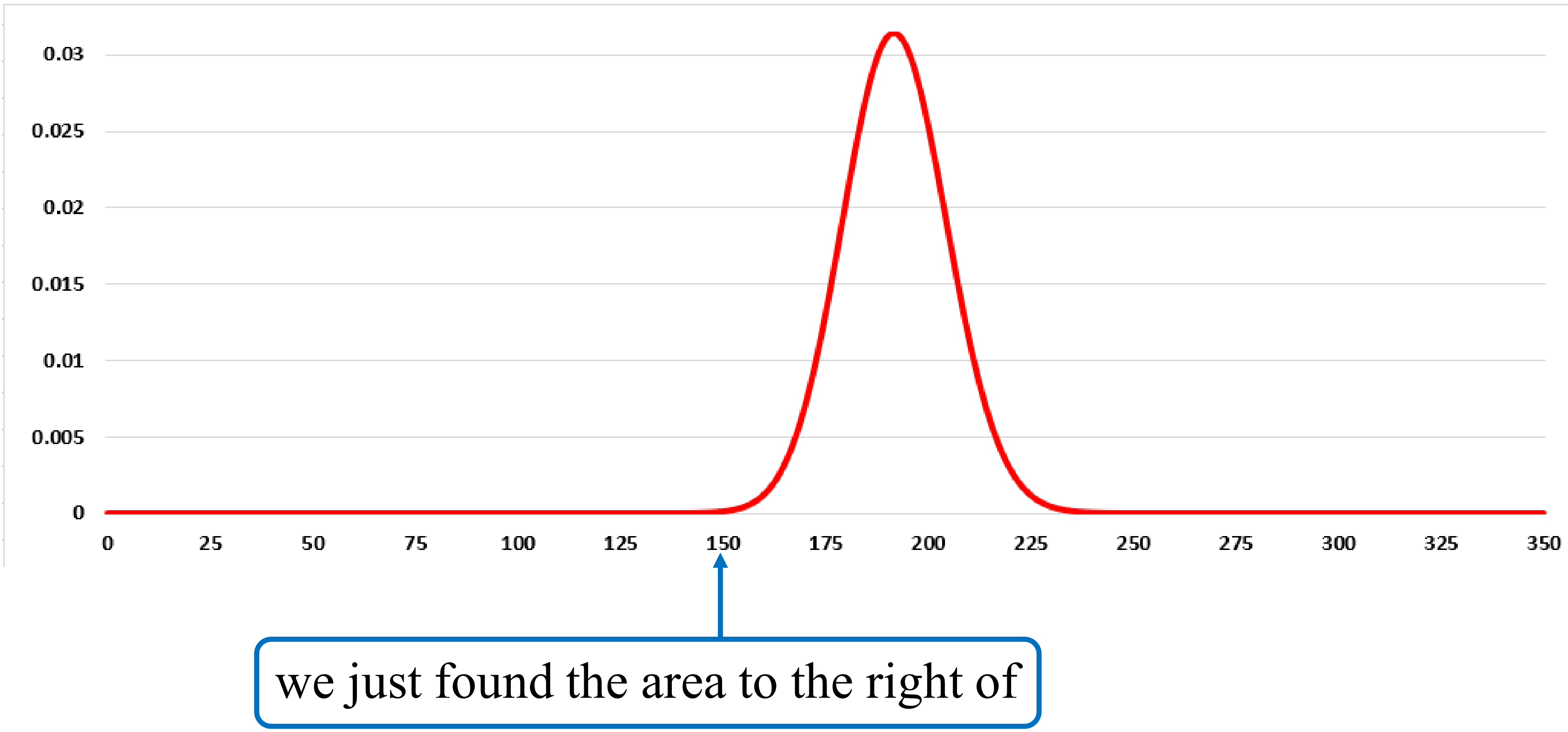
$$\dots - \binom{1200}{149}(0.84)^{1051}(0.16)^{149} - \binom{1200}{150}(0.84)^{1050}(0.16)^{150}$$

even using the idea of complimentary events, it still involves 151 calculations

$$= 1 - 0.0003838$$

$$= 0.9996 \quad (\text{to 4 dp})$$

The distribution's polygon would look like



- The distribution has a modal class somewhere in the middle of the range of values
- The distribution is symmetrical
- The frequency density tails off fairly rapidly as values move further away from the modal class

These are the features of a normal distribution

So what would have happened if we assumed our distribution was normal;

first we need to find the mean and variation

$$X \sim \text{Bin}(1200, 0.16) \qquad \begin{aligned} \mu &= np \\ &= (1200)(0.16) \\ &= 192 \end{aligned} \qquad \begin{aligned} \sigma^2 &= np(1-p) \\ &= (1200)(0.16)(0.84) \\ &= 161.28 \end{aligned}$$

$$X \sim N(192, 161.28) \Leftrightarrow Z \sim N(0,1) \text{ where } Z = \frac{X-\mu}{\sigma}$$

we need to standardise the data

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) \\ &= 1 - P\left(Z \leq \frac{150-192}{\sqrt{161.28}}\right) \\ &= 1 - \Phi(-3.31) \\ &= 1 - 0.0046654 \\ &= 0.9995 \text{ (to 4 dp)} \end{aligned}$$

% difference to binomial answer

$$\frac{0.999616326 - 0.99953346}{0.999616326} \times \frac{100}{1}$$

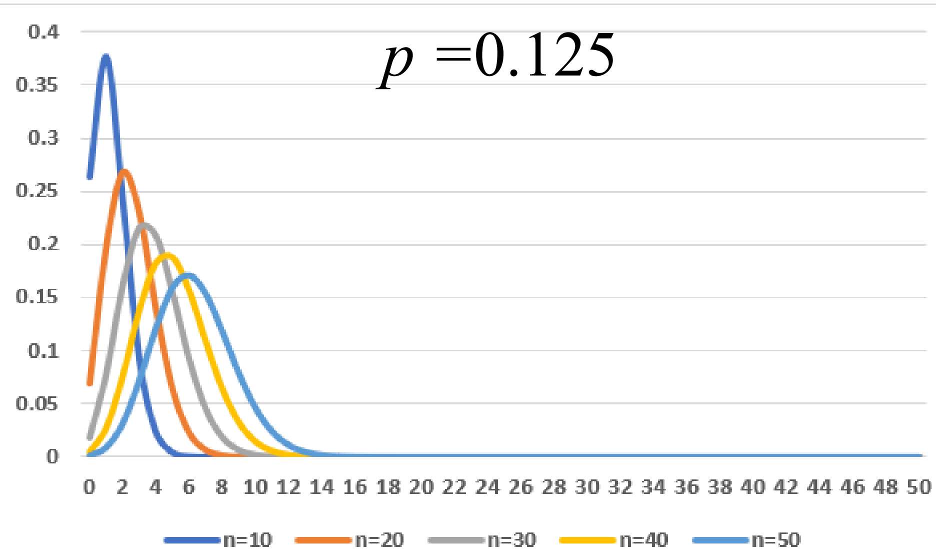
$$= 0.008\%$$

When is it okay to approximate a binomial distribution with a normal distribution?

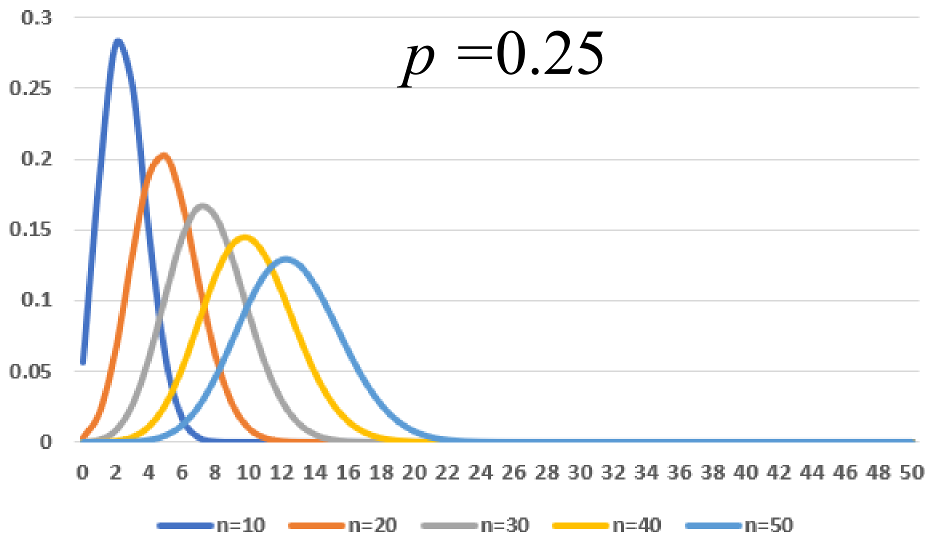
Keep in mind the three key features of a normal distribution

- The distribution has a modal class somewhere in the middle of the range of values
- The distribution is symmetrical
- The frequency density tails off fairly rapidly as values move further away from the modal class

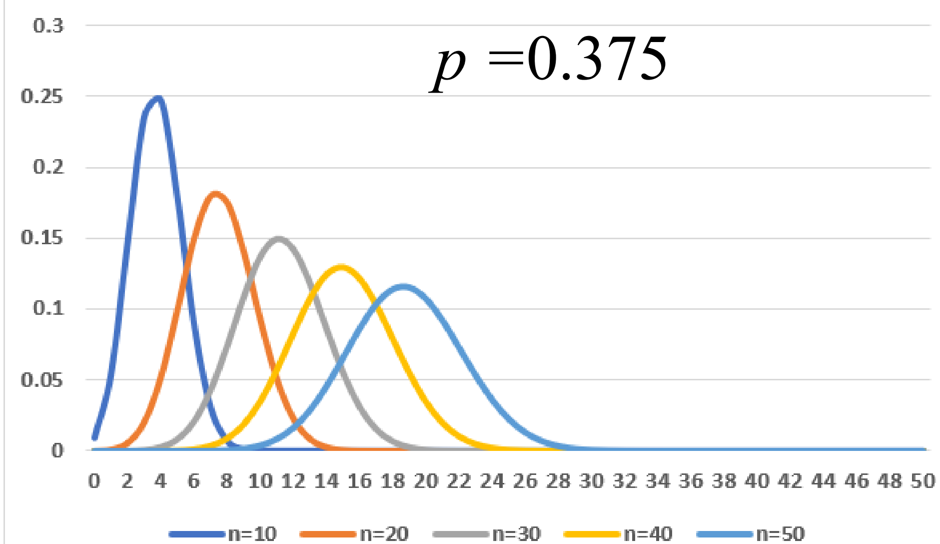
Let's take a look at some binomial distributions;



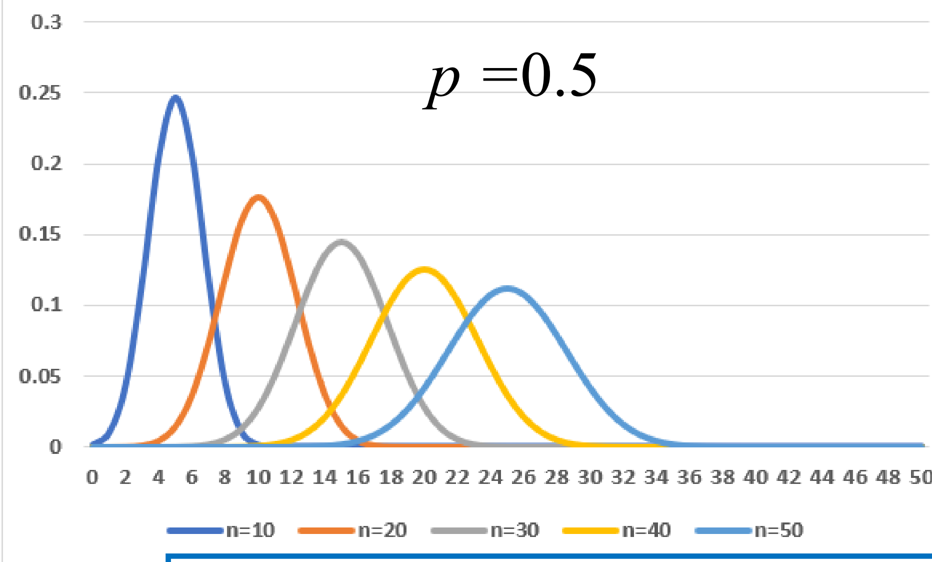
n	p	q	np	nq	☑ or ☒
10	0.125	0.875	1.25	8.75	☒
20	0.125	0.875	2.50	17.50	☒
30	0.125	0.875	3.75	26.25	☒
40	0.125	0.875	5.00	35.00	☑ ☒
50	0.125	0.875	6.25	43.75	☑



n	***p***	***q***	***np***	***nq***	☑ **or** ☒
10	0.25	0.75	2.5	7.5	☒
20	0.25	0.75	5.0	15.0	☑ ☒
30	0.25	0.75	7.5	22.5	☑
40	0.25	0.75	10.0	30.0	☑
50	0.25	0.75	12.5	37.5	☑



n	***p***	***q***	***np***	***nq***	☑ **or** ☒
10	0.375	0.625	3.75	6.25	☒
20	0.375	0.625	7.50	12.50	☑
30	0.375	0.625	11.25	18.75	☑
40	0.375	0.625	15.00	25.00	☑
50	0.375	0.625	18.75	31.25	☑



n	p	q	np	nq	☑ or ☒
10	0.5	0.5	5.00	5.00	☒ ☑
20	0.5	0.5	10.00	10.00	☑
30	0.5	0.5	15.00	15.00	☑
40	0.5	0.5	20.00	20.00	☑
50	0.5	0.5	25.00	25.00	☑

As a general rule; for a satisfactory approximation
$np > 5$ and $nq > 5$

We are approximating the area under the polygon with bins 1 unit apart, so for the interval $a \le x \le b$, the area under the polygon is actually

$$\int_{a-0.5}^{b+0.5} f(x)dx$$

(this might be more obvious if you think of the histogram)

continuity correction for small samples

for small n; $P(a \le X \le b)$ use $P(a - 0.5 \le X \le b + 0.5)$

e.g. 2021 Extension 2 HSC Question 12b)

When a particular biased coin is tossed, the probability of obtaining a head is $\frac{3}{5}$. The coin is tossed 100 times.

Let X be the random variable representing the number of heads obtained. This random variable would have a binomial distribution.

(i) Find the expected value, $E(X)$.

$$\begin{aligned} E(X) &= np \\ &= 100 \times \frac{3}{5} \\ &= \underline{60} \end{aligned}$$

(ii) By finding the variance, Var(X), show that the standard deviation of X is approximately 5

$$\begin{aligned} \text{Var}(X) &= np(1-p) \\ &= 100 \times \frac{3}{5} \times \frac{2}{5} \\ &= 24 \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{\text{Var}(X)} \\ &= \sqrt{24} \\ &= 4.898979\ldots \underline{\approx 5} \end{aligned}$$

(iii) By using a normal approximation, find the approximate probability that X is between 55 and 65.

$X \sim \text{Bin}(100, 0.6)$

$X \sim \text{N}(60, 24) \;\Leftrightarrow\; Z \sim \text{N}(0,1)$

$$P(55 \leq X \leq 65) \approx P\left(\frac{55 - 60}{5} \leq Z \leq \frac{65 - 60}{5}\right)$$

$$= P(-1 \leq Z \leq 1)$$

$$\underline{= 0.68}$$

Exercise 17C;
1, 2adg, 3, 4, 5, 7, 9, 11, 12

https://www.desmos.com/calculator/qmt2h6n8cm
link for comparing approximations