17C Normal approximations to a binomial

A pollster asks a sample of 10 000 people whether they intend to vote for the Working Together Party in the next election. The WTP won $\frac{1}{3}$ of the vote in the last election, and using the assumption that this will be the case in the next election, he wants to know the probability that his survey will give him a result within [3300, 3400].

Using binomial probability, the calculation is

$$P(3300 \leq X \leq 3400) = {}^{10\,000}C_{3300} \left(\tfrac{1}{3}\right)^{3300} \left(\tfrac{2}{3}\right)^{6700} + {}^{10\,000}C_{3301} \left(\tfrac{1}{3}\right)^{3301} \left(\tfrac{2}{3}\right)^{6699} + \cdots + {}^{10\,000}C_{3399} \left(\tfrac{1}{3}\right)^{3399} \left(\tfrac{2}{3}\right)^{6601} + {}^{10\,000}C_{3400} \left(\tfrac{1}{3}\right)^{3400} \left(\tfrac{2}{3}\right)^{6600}.$$

This is a dreadful calculation because of all the huge factorials and massive sizes of the numbers. Even with a computer, there are serious problems about controlling all the errors in the calculation. And if the calculation involves the whole population of Australia — over 25 000 000 — the situation is even more difficult.

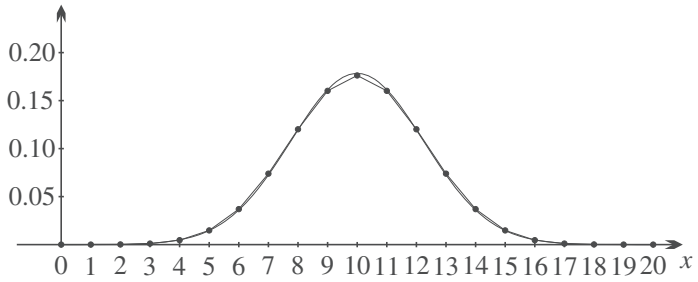
Fortunately, binomial distributions can be approximated very well by the normal distribution, and the larger the numbers are, the better the approximation. This section is about approximating a binomial distribution by a normal distribution.

An example of approximations

In Section 16E, we took the binomial distribution of the number of heads in 20 coin tosses, where $n = 20$ and $p = q = \frac{1}{2}$, so that

$$\mu = np = 20 \times \tfrac{1}{2} = 10 \quad \text{and} \quad \sigma^2 = npq = 20 \times \tfrac{1}{2} \times \tfrac{1}{2} = 5,$$

and $\sigma = \sqrt{5}$. We graphed the frequency polygon, then we superposed a normal curve with the same mean $\mu = 10$ and the same standard deviation $\sigma = \sqrt{5}$. The result looks reasonably good,



Thus the normal PDF approximates the frequency polygon of the binomial, and taking cumulative frequencies, it follows that the normal CDF approximates the ogive of the binomial. We cannot in this course give theoretical arguments why such approximations work — we can only look at the pictures and confirm the results. Then we can use these results in problems.

For example, the standard deviation is $\sqrt{5} \doteq 2.236$. Suppose that we want to find the probability that when we toss 20 coins, the number of heads will be within one standard deviation of the mean. These numbers are small enough for us to calculate it two ways, as in the next worked example.

Notation for normal distributions

The notation $N(\mu, \sigma^2)$ is used for the normal distribution with mean μ and variance σ^2. Thus in the diagram above, we are claiming that the binomial distribution $B\left(20, \frac{1}{2}\right)$ is approximated by the normal distribution $N(10, 5)$. In general, $B(n, p)$ is approximated by $N(np, npq)$, where $q = 1 - p$.

ISBN 978-1-108-76630-2

Example 11 17C

A coin is tossed 20 times, and X is the number of heads. Find the probability that the number of heads is greater than 12:

a using a binomial distribution,

b using the normal approximation.

SOLUTION

a $P(\text{at least 13 heads}) = {}^{20}C_{13}\left(\frac{1}{2}\right)^{20} + {}^{20}C_{14}\left(\frac{1}{2}\right)^{20} + \cdots + {}^{20}C_{20}\left(\frac{1}{2}\right)^{20}$

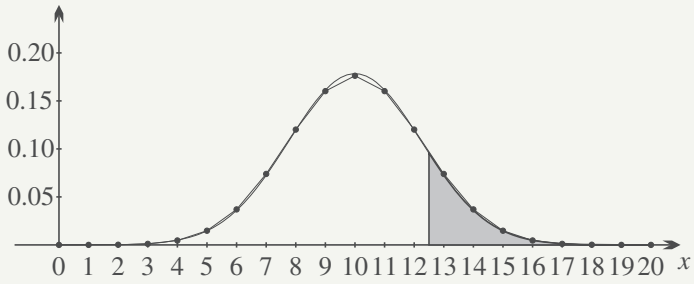
$= \left(\frac{1}{2}\right)^{20}\left({}^{20}C_{13} + {}^{20}C_{14} + \cdots + {}^{20}C_{20}\right)$

$= \left(\frac{1}{2}\right)^{20} \times 137\,980$

$\doteq 0.132.$

b To approximate the binomial by the normal distribution $N(10, 5)$, we treat the discrete random variable X *as if it were a continuous normal variable*. Because we are using a continuous density function, we need to follow the boundaries of the histograms and find the probability, not that $X \geq 13$, but that $X \geq 12.5$, which is the left-hand boundary of the first histogram.

$P(\text{at least 13 heads}) \doteq P(X \geq 12.5).$



This adjustment from 13 to 12.5 is called the *continuity correction* — it is the correction needed when a discrete distribution is treated as if it were continuous.

Using z-scores, $z = \dfrac{x - \mu}{\sigma}$

$\doteq 1.118,$

so $P(\text{at least 13 heads}) \doteq P(Z > 1.118)$

$\doteq 1 - \Phi(1.118)$

$\doteq 1 - 0.868$

$\doteq 0.132.$

Continuity corrections

We are approximating a discrete binomial variable by a continuous normal variable. We thus integrate from halfway between the values of the discrete distribution. In the example above, that means from 12.5.

In the next worked example, we are adding the five values at $x = 8,\ 9,\ 10,\ 11$ and 12, so we integrate over the interval $[7.5, 12.5]$.

ISBN 978-1-108-76630-2

Example 12 17C

A coin is tossed 20 times, and X is the number of heads. Find the probability that the number of heads is within one standard deviation of the mean:

a using a binomial distribution, **b** using the normal approximation.

Comment on the results of the calculations.

SOLUTION

a $P(X \text{ is within one standard deviation of the mean}) = P(7.764 \leq X \leq 12.236)$

$$= P(X = 8,\ 9,\ 10,\ 11 \text{ or } 12)$$

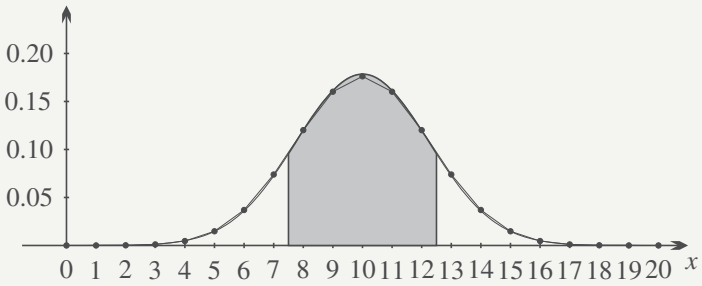
$$= {}^{20}C_8 \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_9 \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{10} \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{11} \left(\tfrac{1}{2}\right)^{20} + {}^{20}C_{12} \left(\tfrac{1}{2}\right)^{20}$$

$$= \frac{{}^{20}C_8 + {}^{20}C_9 + {}^{20}C_{10} + {}^{20}C_{11} + {}^{20}C_{12}}{2^{20}}$$

$$= \frac{772\,616}{1024 \times 1024}$$

$$\fallingdotseq 0.737.$$

b We are approximating the binomial distribution by the normal distribution $N(10, 5)$, so we treat the discrete random X as if it were a continuous normal variable, and calculate $P(7.5 \leq X \leq 12.5)$.



Using z-scores, the two limits of integration are

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \mu}{\sigma}$$

$$= \frac{12.5 - 10}{\sqrt{5}} \qquad = \frac{7.5 - 10}{\sqrt{5}}$$

$$= 1.118, \qquad = -1.118,$$

so $P(X \text{ is within one standard deviation of the mean}) \fallingdotseq P(-1.118 \leq Z \leq 1.118)$

$$\fallingdotseq 0.736.$$

The results in parts **a** and **b** are in excellent agreement.

ISBN 978-1-108-76630-2

5 NORMAL APPROXIMATION TO A BINOMIAL DISTRIBUTION

- A binomial distribution can be approximated by the normal distribution with the same mean and standard deviation.
- Thus the binomial distribution $B(n, p)$ or $\text{Bin}(n, p)$ can be approximated by the normal distribution $N(np, npq)$, where $q = 1 - p$.
- $N(\mu, \sigma^2)$ means the normal distribution with mean μ and variance σ^2.
- When approximating, we treat the discrete binomial variable X as if it were a continuous normal variable.
- For small values of n, apply the *continuity correction*. This means integrating between half-intervals, corresponding to the boundaries of the cumulative frequency histogram. For example, to approximate $P(X = 8,\ 9,\ 10,\ 11 \text{ or } 12)$, we treat X as a continuous normal variable and find $P(7.5 \leq X \leq 12.5)$.

The continuity correction for large numbers

In the example in the introduction to this chapter, the numbers are large enough so that the continuity correction is negligible, as the next worked example shows.

Example 13 17C

Compute $P(3300 \leq X \leq 3400)$ in the example in the introduction:

a ignoring the continuity correction, **b** using the continuity correction.

Then comment on the results of the calculations.

SOLUTION

Here $n = 10\,000$ and $p = \frac{1}{3}$, so $\mu = 3333\frac{1}{3}$ and $\sigma^2 = 2222\frac{2}{9}$ and $\sigma \doteqdot 47.1416$.
We will treat the discrete random variable X as if it were a continuous normal variable with the same mean $3333\frac{1}{3}$ and standard deviation 47.1416.

a Without the continuity correction, we are finding $P(3300 \leq X \leq 3400)$.
The z scores are -0.7071 and 1.4142,
so $P(3300 \leq X \leq 3400) \doteqdot P(-0.7071 \leq Z \leq 1.4142)$
$\doteqdot 0.6816$.

b With the continuity correction, we are finding $P(3299.5 \leq X \leq 3400.5)$.
The z scores are -0.7177 and 1.4248,
so $P(3299.5 \leq X \leq 3400.5) \doteqdot P(-0.7177 \leq Z \leq 1.4248)$
$\doteqdot 0.6859$.

No pollster would ever need the third decimal place of the probabilities, so there is no problem whatsoever ignoring the continuity correction.

ISBN 978-1-108-76630-2

Using the normal approximation to calculate a binomial coefficient

One of the consequences of the normal approximation to the binomial is that for large parameters, an individual binomial coefficient can be approximated. These calculations demonstrate very well how vital the continuous correction can be.

Example 14 17C

Approximate ${}^{100}C_{53}$ using the normal approximation to the binomial distribution $B\left(100, \frac{1}{2}\right)$.
Check how good the approximation is.

SOLUTION

We know that ${}^{100}C_{53} \times \left(\frac{1}{2}\right)^{100} = P(X = 53)$.
We treat the discrete binomial variable X as if it were a continuous normal variable,

$$ {}^{100}C_{53} \times \left(\frac{1}{2}\right)^{100} = P(52.5 \leq X \leq 53.5). $$

The binomial distribution has mean $\mu = np = 50$ and SD $\sigma = \sqrt{npq} = \sqrt{25} = 5$.
So also does the normal approximation, and the limits have z-scores 0.5 and 0.7.
Hence

$$ \begin{aligned} {}^{100}C_{53} \times \left(\frac{1}{2}\right)^{100} &\doteq P(0.5 \leq Z \leq 0.7) \\ &\doteq 0.066574 \\ {}^{100}C_{53} &\doteq 2^{100} \times 0.066574 \\ &\doteq 8.438 \times 10^{28}. \end{aligned} $$

This compares with ${}^{100}C_{53} \doteq 8.441 \times 10^{28}$ (all according to Excel).

Sampling with replacement — when is a survey a binomial experiment?

Suppose that I choose ten cards in succession *without replacement* from a pack of 52 cards and count the number of aces. This is not a binomial experiment because the probabilities of getting an ace change as each card is removed from the pack. If I *replace the card each time*, however, then I have a binomial experiment.

More generally, a survey that questions n people chosen at random from N people without replacement is not a binomial experiment. A survey that chooses each person independently of all previous choices, however, thus allowing the same person possibly to be questioned more than once, is a binomial experiment because the stages are identical and independent.

Now suppose that the pool of N people from which the sample is chosen is very large. This allows us to ignore these distinctions, because whether replacement is used or not, the probabilities will be virtually the same in all stages. This is certainly the case with a sample chosen from all Australians, but in a college of say 3000 people and a sample that is small relative to 3000, the effect is also almost negligible.

How good an approximation is the normal?

This depends on how much accuracy is needed. The diagrams below show that n should be larger for a skewed distribution, meaning that p is near 0 or 1, than when the distribution is reasonably symmetric, meaning that p is near 0.5. A common rule of thumb therefore is to require

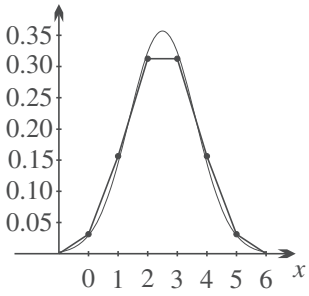
$$ np > 5 \quad \text{and} \quad nq > 5, $$

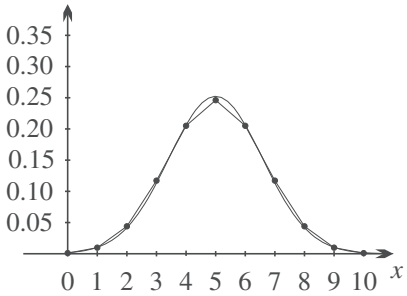
but numbers other than 5 are often stated, and there are other tests.

ISBN 978-1-108-76630-2

When $p = \frac{1}{2}$, we have already drawn the approximation when $n = 20$. Here are the ogive and the normal approximation, still for $p = \frac{1}{2}$, when $n = 5$ and $n = 10$.

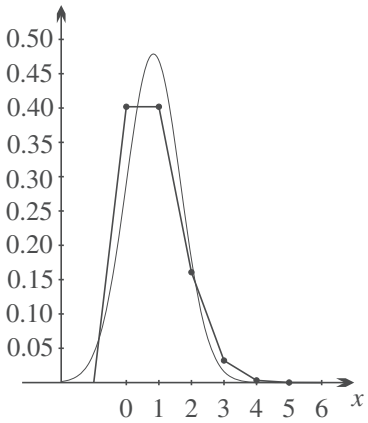
a

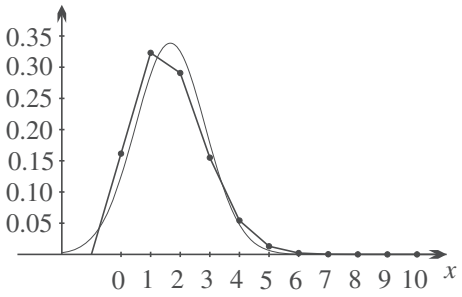


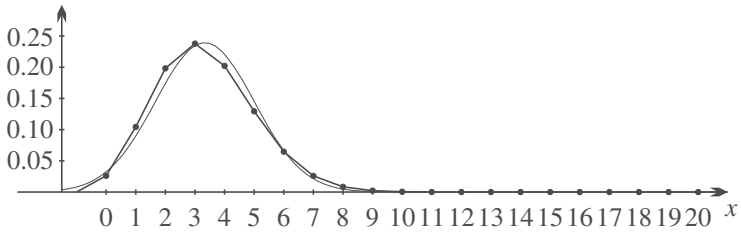


When $p = \frac{1}{6}$, which arises when throwing a die and recording if it is a six, here are the graphs for $n = 5$, $n = 10$ and $n = 20$

b







6 A RULE OF THUMB FOR USING THE NORMAL APPROXIMATION

- For a fixed number n of Bernoulli trials, the normal approximation to a binomial distribution is less satisfactory when the distribution is skewed.
- For a fixed probability p, the normal approximation to a binomial distribution is more satisfactory when the number of trials is large.
- A common rule of thumb — one amongst many — for using the normal approximation to a binomial distribution is to require

  $np > 5$ and $nq > 5$,

The normal approximation to a binomial distribution is a special case

We remarked at the start of Section 16G in the last chapter that the normal approximation to a binomial distribution was one of the first examples of a very general theorem in statistics called the *central limit theorem*. This would be a good occasion to read again the short account of the central limit theorem given there.

ISBN 978-1-108-76630-2