Normal Approximation to the Binomial Distribution

- Sometimes when n is large, the binomial formula can be difficult to use.
- In these cases, we may be able to use the normal distribution to estimate binomial probabilities.

Example

- Approximately 15% of the US population smokes cigarettes.
- A local government commissioned a survey of 400 randomly selected individuals.
- The survey found that only 42 of the 400 participants smoke cigarettes.
- If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?

Example

First, we check that this is a binomial setting:

1. $n = 400$ community members
2. This is a random sample, so the trials are independent.
3. We define Success = `smoker` and Failure = `nonsmoker`.
4. $p = P(\texttt{smoker}) = 0.15$

So this is a binomial distribution.

We are interested in $k = 42$ *or fewer*.

Example

Let X be the number of smokers in a community. We want to know

$$P(X \leq 42)$$

which is the same as

$$\begin{aligned} &P(X = 42 \text{ or } X = 41 \text{ or } X = 40 \text{ or } \ldots \text{ or } X = 1 \text{ or } X = 0) \\ &= P(X = 42) + P(X = 41) + \cdots + P(X = 1) + P(X = 0) \end{aligned}$$

We *could* calculate each of the 43 probabilities individually by using our binomial formula and adding them together...

Example

If we were to do this, we would find

$$P(X = 42) + P(X = 41) + \cdots + P(X = 1) + P(X = 0) = 0.0054$$

That is, if the true proportion of smokers in the community is $p = 0.15$, then the probability of observing 42 or fewer smokers in a sample of $n = 400$ is 0.0054.

Normal Approximation to the Binomial Distribution

...but why would we do this if we don't have to?

- Calculating probabilities for a range of values is much easier using the normal model.
- We'd like to use the normal model in place of the binomial distribution.

Normal Approximation to the Binomial Distribution

Surprisingly, this works quite well as long as

$$np > 10$$

and

$$n(1-p) > 10$$

Note that *both of these conditions must hold*!

Normal Approximation to the Binomial Distribution

If these conditions are met, a Binomial(n, p) variable X is well-approximated by a normal distribution with

$$E(X) = \mu = np$$

and

$$\sigma = \sqrt{np(1-p)}.$$

Example

Can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

Example

Can we use the normal approximation to estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$?

We verified that the binomial model is reasonable. Now,

$$np = 400 \times 0.15 = 60$$

and

$$n(1-p) = 400 \times 0.85 = 340$$

so both are at least 10 and we may use the normal approximation.

Example

For the normal approximation,

$$\mu = np = 400 \times 0.15 = 60$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{400 \times 0.15 \times 0.85} = 7.14$$

Example

We want to find the probability of observing 42 or fewer smokers using or $N(\mu = 60, \sigma = 7.14)$ model.

We start by finding our Z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{42 - 60}{7.14} = -2.52$$

Example

- Then, using `R`, the left-tail area is 0.0059.
- When we calculated this using the binomial distribution, the true probability was 0.0054.
- So this is a pretty good approximation!

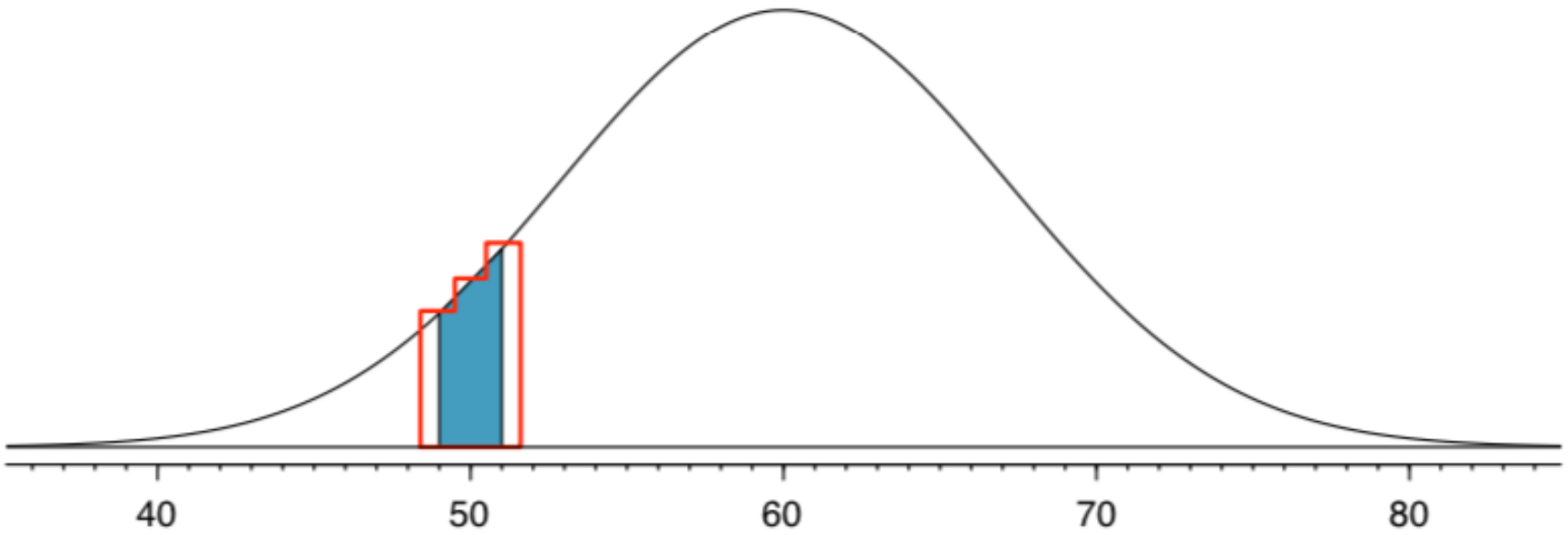
Breakdown of the Normal Approximation

- The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts.
- This is true even when $np > 10$ and $n(1-p) > 10$

Breakdown of the Normal Approximation

- Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$.
- We know that $np = 60 > 10$ and $n(1-p) = 340$, so we might want to apply the normal approximation and use the range 49 to 51.
- But this time the approximation and the binomial solution are noticeably different!
  - Binomial: 0.0649
  - Normal: 0.0421

Why Does This Breakdown Happen?



The binomial probability is shown outlined in red; the normal probability shaded in blue.

Can We Fix It? Improving the Normal Approximation for Intervals

We can usually improve this estimation by modifying our cutoff values.

- Cutoff values for the left side should be reduced by 0.5.
- Cutoff values for the right side should be increased by 0.5.

Example

- Suppose we wanted to compute the probability of observing 49, 50, or 51 smokers in 400 when $p = 0.15$.
- Let's try this again with our modification.
- For our normal distribution, we used a $N(60, 7.14)$ model.
- Our upper value is 51, adjusted to $51 + 0.5 = 51.5$.
- Our lower value is 49, adjusted to $49 - 0.5 = 48.5$.

Example

Then

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{51.5 - 60}{7.14} = -1.190476$$

and

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{48.5 - 60}{7.14} = -1.610644$$

Example

Now, using R,

$$\begin{aligned} P(z_2 < Z < z_1) &= P(Z < z_1) - P(Z < z_2) \\ &= 0.1169297 - 0.05362867 \\ &= 0.0633 \end{aligned}$$

Example

$P(49 \leq X \leq 51)$		
Binomial	Normal Approx (Adjusted)	Normal Approx (Unadjusted)
0.0649	0.0633	0.0421

Making those small adjustments makes a significant difference!