MDM4U: Mathematics of Data Management

# When Close Enough Is Good Enough

## Approximations to Binomial Probability Distributions

J. Garvin





## Probabilities for Discrete Data

**Recap**

A spinner made of four equally-sized sectors labelled $1-4$ is spun 20 times. Use a binomial distribution to determine the probability of spinning a 3 at most twice.

We must add three cases, for $P(0)$, $P(1)$ and $P(2)$.

$$P(0) = {}_{20}C_0\left(\tfrac{1}{4}\right)^0\left(\tfrac{3}{4}\right)^{20} \approx 0.003 \qquad P(1) = {}_{20}C_1\left(\tfrac{1}{4}\right)^1\left(\tfrac{3}{4}\right)^{19} \approx 0.021$$

$$P(2) = {}_{20}C_2\left(\tfrac{1}{4}\right)^2\left(\tfrac{3}{4}\right)^{18} \approx 0.067$$

The probability is about $0.003 + 0.021 + 0.067 \approx 0.091$.





## Approximating the Binomial PD

In some cases, discrete data may be modelled fairly accurately using a normal distribution.

In particular, a normal probability distribution provides a good approximation to a binomial distribution, given certain conditions.

**Normal Approximation to the Binomial PD**

A binomial probability distribution with $n$ trials and probability of success $p$ can usually be approximated using a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$, where $q = 1-p$, given that both $np \geq 5$ and $nq \geq 5$.

These limits involving $np$ and $nq$ were empirically determined and are statistical rules-of-thumb. Some sources use higher values (e.g. $np > 10$) to ensure a more accurate result.





## Approximating the Binomial PD

**Example**

Use a normal distribution to approximate the probability of spinning a 3 at most twice on the spinner described earlier.

Since $np = 20\left(\frac{1}{4}\right) = 5$ and $nq = 20\left(\frac{3}{4}\right) = 15$, a normal approximation should be reasonable.

For the normal distribution, $\mu = np = 5$ and $\sigma = \sqrt{npq} = \sqrt{3.75}$.

Since the data are discrete, use continuity correction to find $P(x \leq 2.5)$.

$z = \dfrac{2.5-5}{\sqrt{3.75}} \approx -1.29$, corresponding to a probability of approximately 0.0985.

The approximation is close – just over half of a percent off.





## Approximating the Binomial PD

In the previous example, it was not too difficult to calculate the three cases for the binomial distribution, so using a normal approximation is probably not the best idea – it introduces inaccuracies that could be easily avoided.

In some cases, however, using a normal approximation to the binomial can save a considerable amount of time and will be "close enough" for all practical purposes.

Consider the scenario where a die is rolled 100 times. What is the probability that a 5 is rolled at most 20 times?

Using a binomial distribution would be time-consuming: 21 cases to consider!

A normal approximation, however, would require only one lookup in a table.





## Approximating the Binomial PD

**Example**

A die is rolled 100 times. What is the probability that a 5 is rolled at most 20 times?

Since $np = 100\left(\frac{1}{6}\right) \approx 16.67$ and $nq = 100\left(\frac{5}{6}\right) \approx 83.33$, a normal approximation should be appropriate.

For the normal distribution, $\mu = np \approx 16.67$ and $\sigma = \sqrt{npq} \approx 3.73$.

Since the data are discrete, use continuity correction to find $P(x \leq 20.5)$.

$z \approx \dfrac{20.5-16.67}{3.73} \approx 1.03$, corresponding to a probability of approximately 0.8485.

For comparison, the actual value is approximately 0.8481.

## Approximating the Binomial PD

**Example**

7% of all watches assembled at a factory need some amount of readjustment. In a random selection of 80 watches, what is the probability that 10 or more need readjustment?

Since $np = 80(0.07) = 5.6$ and $nq = 80(0.93) = 74.4$, a normal approximation can be used.

For the normal distribution, $\mu = np = 5.6$ and $\sigma = \sqrt{npq} \approx 2.28$.

Using continuity correction to find $P(x \geq 9.5)$, $z \approx \dfrac{9.5 - 5.6}{2.28} \approx 1.71$, so $P(x \leq 9.5) \approx 0.9564$.

Therefore, the probability that 10 or more watched need readjustment is approximately $1 - 0.9564 \approx 0.0436$.





## Approximating the Binomial PD

**Example**

Show that a normal approximation to a binomial distribution is not appropriate to determine the probability of successfully guessing the values of five random numbers between 1 and 100 at least once.

Since $np = 5 \times \frac{1}{100} = \frac{1}{20} < 5$, a normal approximation is not appropriate.

To verify this, use a normal distribution with $\mu = \frac{1}{20} = 0.05$ and $\sigma = \sqrt{5 \times \frac{1}{100} \times \frac{99}{100}} \approx 0.2225$.

$z \approx \dfrac{0.5 - 0.05}{0.2225} \approx 2.02$, with an associated probability of approximately 97.8%. Thus, the chance of guessing a number at least once is approximately $100 - 97.8 \approx 2.2\%$.





## Approximating the Binomial PD

Using a binomial distribution, the actual probability is $1 - {}_5C_0\left(\frac{1}{100}\right)^0\left(\frac{99}{100}\right)^5 \approx 0.049$, or roughly 5%, which is quite different from our approximation.

Therefore, it is always important to verify that the conditions $np \geq 5$ and $nq \geq 5$ are met before using a normal distribution to approximate a binomial distribution.





## Questions?