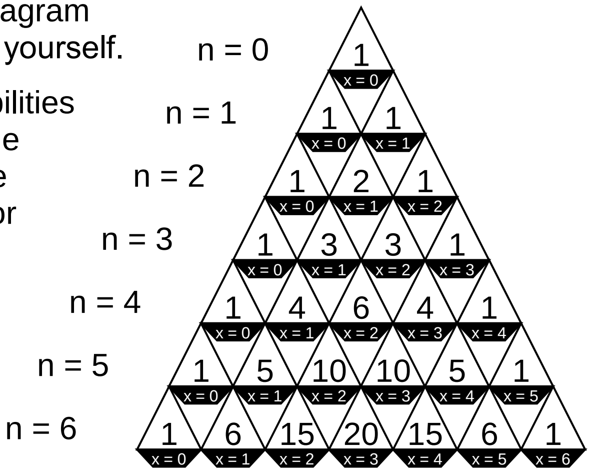
Math 1111 Learning Centre

VANCOUVER COMMUNITY COLLEGE

The Normal Approximation to the Binomial Distribution

We've seen that for a binomially distributed variable, we can calculate a mean and a standard deviation, and we know that the values of $_nC_x$ rise and fall, so that when x is close to 0, or when x is close to n, the value of $_nC_x$ is least, and when x is close to ½n, $_nC_x$ is greatest. The values of $_nC_x$ are collected in a diagram called Pascal's triangle, so that you can see them for yourself.

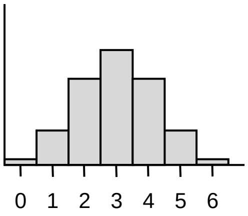


In fact, if you were to create histograms of the probabilities of various values of x for a fixed n and p, the larger the values of n you used, the closer to a normal curve the histogram would become. This is convenient, since for larger values of n, it becomes harder and harder to answer questions like, "What is the probability that in 38 trials, we get fewer than 12 successes?" the sample space for this question would be {x = 0, 1, 2, 3, 4, …, 11} and each of those possibilities would require its own calculation. The process would become unwieldy quite quickly. For large enough problems, we can use the similarity to the normal curve to approximate answers quite closely, and using a single calculation.

Recall that if X is a binomially distributed variable, then the mean for X is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$ where n is the number of trials, p is the probability of success in each trial, and q is the probability of failure, $1 - p$. We can use these values of μ and σ directly in calculating a z-score, but we'll need to do a little extra work to adapt the x-value.

Let's look at the histogram for the n = 6 line of Pascal's triangle, and letting p = q = 0.5 for simplicity's sake. The probabilities and histogram are:

x	0	1	2	3	4	5	6
P(X = x)	1/64	6/64	15/64	20/64	15/64	6/64	1/64



It's a technical thing, but the bars in the histogram are centered over the numbers they represent. This makes sense if you think about it this way: the bars in a histogram are supposed to touch. The regions on the number line covered by each bar include the decimal numbers that round to the value of x represented by the bar. (Anything from 1.5 to just under 2.5 rounds to 2, and that's the range on the horizontal axis that the bar for 2 occupies.) In approximating the discrete binomial distribution with the continuous normal distribution, this technicality with the rounding matters. If we need $P(X \leq 2)$ in the original binomial problem, then we want to include all the area in the bar for 2. So what we really want is $P(X < 2.5)$ on the normal curve.

This adjustment to account for the differences between discrete and continuous distributions is called the **continuity correction**. In practice, the continuity correction has the following effects:

- If h is *included* as the least number for X sought in a question, then $h - 0.5$ is the least number for X in the approximation.
- If h is *excluded* as the least number for X sought in a question, then $h + 0.5$ is the least number for X in the approximation.

- If k is *included* as the greatest number for X sought in a question, then $h + 0.5$ is the greatest number for X in the approximation.
- If k is *excluded* as the greatest number for X sought in a question, then $h - 0.5$ is the greatest number for X in the approximation.

Example 1: For each of the following set-ups for binomial questions, determine the equivalent set-up for the appropriate normal approximation:

a) $P(X \geq 7)$
b) $P(X > 7)$
c) $P(X < 24)$
d) $P(13 < X \leq 19)$
e) $P(X = 21)$

Solution: a) This question wants the total area for the bars {7, 8, 9, …, n}. To capture all the area for bar 7, we start back at 6.5: $P(X > 6.5)$.

b) For questions involving the normal distribution, there's no difference between the expression in part (a) and part (b) because the probability that X = 7.000… is essentially 0. Here, it makes a big difference!

This question wants the total area for the bars {8, 9, 10, …, n}. It doesn't include bar 7, since 7 is not greater than 7. To leave out bar 7, we start at 7.5: $P(X > 7.5)$.

c) We want X to be strictly less than 24. The question wants {0, …, 21, 22, 23}. To leave out bar 24, we stop at 23.5: $P(X < 23.5)$.

d) The inequality signs here don't match, but that's okay. We always treat both boundaries separately. This question wants {14, 15, 16, 17, 18, 19}. We start at 13.5 and stop at 19.5: $P(13.5 < X < 19.5)$.

e) This question just wants bar 21. The region of the horizontal axis that the bar occupies is 20.5 to 21.5: $P(20.5 < X < 21.5)$.

The binomial problem must be "large enough" that it behaves like something close to a normal curve. The histogram illustrated on page 1 is too chunky to be considered normal. We may only use the normal approximation if $np > 5$ and $nq > 5$.

Once we have the correct x-values for the normal approximation, we can find a z-score and determine the probability from there, as we always have. When we do so, we do *not* use a sampling distribution; we use the original formula for z based on a single observation. The reason for this is that, while we are compiling multiple observations, we are not taking the mean of a set of quantitative observations. We classify binomial observations as successes and failures, and counting those generates just one number, not a batch of numbers.

Example 2: An escape room advertises that only 40% of teams who attempt the room get out in time. What is the probability that of the next 38 teams to attempt the room, fewer than 12 get out?

Solution: To solve this problem using the binomial formula would require 12 separate calculations. We should use the normal approximation if possible.

For this question, $n = 38$, $p = 0.4$ and so $q = 0.6$. We need to test whether we can use the normal approximation. We calculate $np = 15.2$ and $nq = 22.8$. Both these values are greater than 5, so we can proceed.

We need the mean and standard deviation: $\mu = np = 15.2$, and $\sigma = \sqrt{npq} \approx 3.02$. The binomial question wants $P(X < 12)$ since it says "fewer than 12". Using the continuity correction this becomes $P(X < 11.5)$. We calculate a z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 15.2}{3.02} = -1.2252\ldots \approx -1.23 \xRightarrow{\text{A-2}} 0.1093$$

The probability that of the next 38 teams, fewer than 12 of them will succeed at the escape room is just under 11%. (The actual value, calculated with Excel, is 0.108880….)

EXERCISES

A. Reframe each of these probability expressions for a binomially distributed variable X_B to refer to a normally distributed variable approximating it, X_N.

1) $P(X_B > 57)$

2) $P(X_B \leq 41)$

3) $P(X_B < 28)$

4) $P(X_B \geq 60)$

5) $P(29 < X_B < 83)$

6) $P(102 < X_B \leq 112)$

7) $P(6 \leq X_B \leq 19)$

8) $P(X_B = 26)$

B. Calculate the following probabilities using the normal approximation to the binomial distribution, if possible.

1) A bored security guard opens a new deck of playing cards (including two jokers and two advertising cards) and throws them one by one at a wastebasket. He's done this every night for years, and he makes the shot 62% of the time. What is the probability that he throws at least 40 cards into the wastebasket?

2) A market research company determines that 32% of the population say that red is their favourite colour for cars. On a day when 221 cars are sold across the country, assuming all these purchases are independent and all customers had the option of getting red cars, what is the probability that no more than 75 red cars were sold?

3) It is estimated that 18% of voters feel strongly enough about their candidates during any given election that they put a sign on their lawn showing which candidate they support. In one town 35 voters are chosen at random. What is the probability that fewer than 10 of them put a sign on their lawns?

4) Assuming children are equally likely to be born in any month, what is the probability that in a class of 42 students, more than 17 of them were born in a month starting with the letter J?

5) Eight out of every nine patrons of Neighbour’s Gym sign up for a fitness class in any given year. If 40 of the current patrons are randomly selected, what is the probability that more than 30 of them signed up for a class last year?

6) An airline estimates that 0.087 of its customers are willing to pay a $50 upgrade for a seat with extra legroom. There are only 12 such seats on a plane, and the scheme is only profitable if at least half of the seats get sold. If each flight gets 132 customers maximum, what is the probability that the upgrade scheme is profitable and doesn’t oversell?

C. Find $P(X \leq 2)$ for each of the following values of n and p, using the binomial formula, and the normal approximation (whether it’s appropriate or not).

1) $n = 5$, $p = 0.4$

2) $n = 12$, $p = 0.2$

3) $n = 50$, $p = 0.03$

4) $n = 15$, $p = 0.5$

SOLUTIONS

A: (1) $P(X_N > 57.5)$ (2) $P(X_N < 41.5)$ (3) $P(X_N < 27.5)$ (4) $P(X_N > 59.5)$
(5) $P(29.5 < X_N < 82.5)$ (6) $P(102.5 < X_N < 112.5)$ (7) $P(5.5 < X_N < 19.5)$
(8) $P(25.5 < X_N < 26.5)$

B: (1) $n = 56$, $p = 0.62$, $q = 0.38$; $\mu = 34.72$, $\sigma = 3.6323\ldots$; $P(X_B \geq 40) \Rightarrow P(X_N > 39.5)$;
$z = 1.32$; $P = 1 - 0.9066 = 0.0934$
(2) $n = 221$, $p = 0.32$, $q = 0.68$; $\mu = 70.72$, $\sigma = 6.9347\ldots$;
$P(X_B \leq 75) \Rightarrow P(X_N < 75.5)$; $z = 0.69$; $P = 0.7549$
(3) $n = 35$, $p = 0.18$, $q = 0.82$; $\mu = 6.3$, $\sigma = 2.2729\ldots$; $P(X_B < 10) \Rightarrow P(X_N < 9.5)$;
$z = 1.41$; $P = 0.9207$
(4) $n = 42$, $p = \frac{3}{12}$, $q = \frac{9}{12}$; $\mu = 10.5$, $\sigma = 2.8062\ldots$; $P(X_B > 17) \Rightarrow P(X_N > 17.5)$;
$z = 2.49$; $P = 1 - 0.9936 = 0.0064$
(5) $n = 40$, $p = \frac{8}{9}$, $q = \frac{1}{9}$; $nq = 4.44\ldots$ so the normal approximation cannot be used.
(6) $n = 132$, $p = 0.087$, $q = 0.913$; $\mu = 11.484$, $\sigma = 3.2380\ldots$;
$P(6 \leq X_B \leq 12) \Rightarrow P(5.5 < X_N < 12.5)$; $z_1 = -1.85$, $z_2 = 0.31$;
$P = 0.6217 - 0.0322 = 0.5895$

C: (1) $P(X_B \leq 2) = 0.68256$, $P(X_N < 2.5) = 0.6772$
(2) $P(X_B \leq 2) = 0.55835\ldots$, $P(X_N < 2.5) = 0.5279$
(3) $P(X_B \leq 2) = 0.81080\ldots$, $P(X_N < 2.5) = 0.7967$
(4) $P(X_B \leq 2) = 0.0036926\ldots$, $P(X_N < 2.5) = 0.0049$