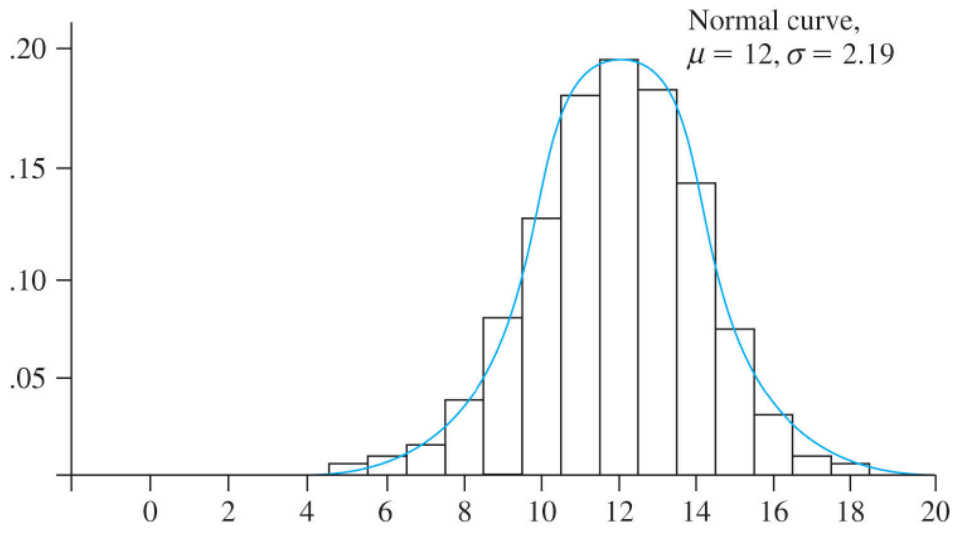
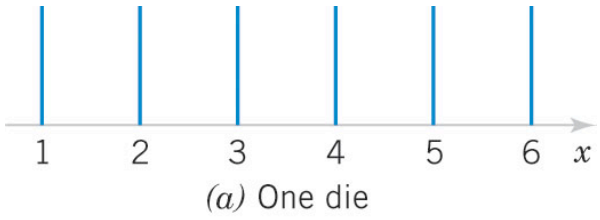
Lecture 12

Normal Distribution to Binomial

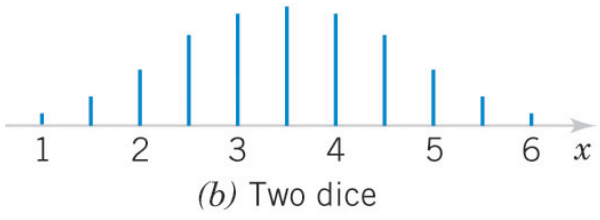


Normal arises when adding many similar independent quantities. Binomial is sum of independent Bernoulli random variables. Graph shows histogram of $Binomial(20, 0.6)$. Height of each small rectangle centred at a value x is Binomial p.m.f. $f(x)$ and base =1 so area is $f(x)$. Superimpose normal density with same $\mu = np = 20(0.6) = 12$ and same $\sigma = \sqrt{np(1-p)} = \sqrt{20(0.6)(0.4)} = 2.19$.

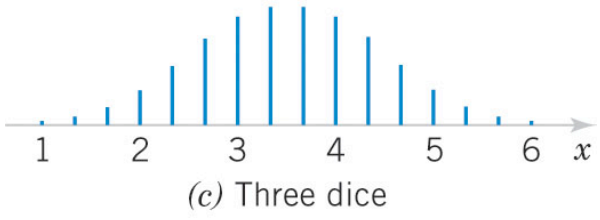
Another example. Average of n fair dice rolls as n increases.



(a) One die

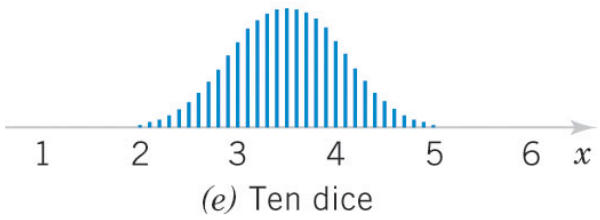


(b) Two dice



(c) Three dice

(d) Five dice



(e) Ten dice

Let $X \sim Binomial(n,p)$. Know $\mu = np, \sigma^2 = np(1-p)$. Normal approximation says that $P(X \leq x)$ can be approximated by Normal

$$X \sim N\left(np, np(1-p)\right)$$

and evaluated using Normal tables by standardization if n is sufficiently large.

$$P(X \leq x) \;=\; P(Z \leq \frac{x - np}{\sqrt{np(1-p)}})$$

Requires both $np > 5, n(1-p) > 5$.

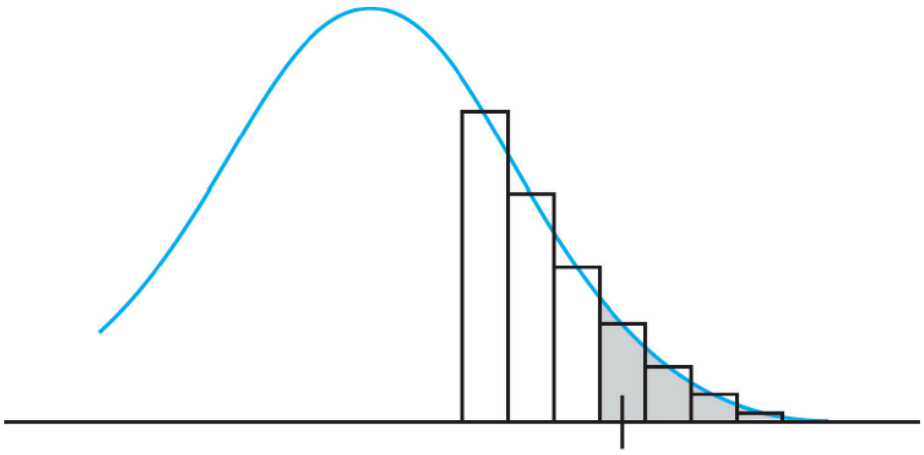
Example 4-17. Digital communication channel. Number of bits X received in error modelled as $Binomial(16000000, 0.00001)$. Hence

$$\begin{aligned} P(X \leq 150) &= \sum_{x=0}^{150} \left(\begin{array}{c} 16000000 \\ x \end{array} \right) (0.00001)^x (0.99999)^{16000000} \\ &= 0.2280 \end{aligned}$$

The answer was obtained using matlab. It take a long time to compute by hand calculator.

Continuity Correction

Look at histogram shown earlier. We are approximating discrete Binomial with continuous Normal. For the latter $P(X = x) = 0$ for any x. All the probability at x is spread over the interval $(x - 0.5, x + 0.5)$. The Normal p.d.f $f(x)$ crosses Binomial histogram near midpoint of each interval around x on the x-axis. Need to correct by adding half of area of rectangle at end of interval around x.



This means that to find $P(X \leq x)$ using the normal approximation we get more accuracy if we replace x with $x + 0.5$.

Likewise to find $P(X \geq x)$ replace x with $x - 0.5$ in the Normal approximation. If you forget whether to add or subtract, always pick whatever gives the larger region. Therefore

$$P(X \leq x) = P(X \leq x + 0.5) = P(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}})$$

$$P(x \leq X) = P(x - 0.5 \leq X) = P(\frac{x - 0.5 - np}{\sqrt{np(1-p)}} \leq Z)$$

for both $np > 5, n(1-p) > 5$.

Example 4-18. Refers Example 4-17.
$n = 16000000, p = 0.00001, np = 160,$
$np(1-p) = 159.9984, \sqrt{np(1-p)} = 12.6490.$
Without correction.

$$P(X \leq 150) = P(Z \leq \frac{150 - 160}{12.6490})$$

$$= P(Z \leq -0.7906) = 0.2146$$

With correction.

$$P(X \leq 150) = P(Z \leq \frac{150 + 0.5 - 160}{12.6490})$$

$$= P(Z \leq -0.7510) = 0.2263$$

The exact answer from Matlab is 0.228.

Example 4-19. Refers Example 4-18. $n = 50, p = 0.1, np = 5, np(1-p) = 4.5, \sqrt{np(1-p)} = 2.12$. Exact value

$$\begin{aligned} P(X \leq 2) &= \sum_{x=0}^{2} \binom{50}{x} 0.1^x (0.9)^{50-x} \\ &= 0.0042 + 0.0286 + 0.0779 = 0.1117 \end{aligned}$$

Normal approximation

$$\begin{aligned} P(X \leq 2) &= P(\frac{X-5}{2.12} \leq \frac{2-5}{2.12}) \\ &= P(Z \leq \frac{2+0.5-5}{2.12}) \quad \text{continuity correction} \\ &= P(Z \leq -1.1792) = 0.1192 \end{aligned}$$

So what is the purpose of the Normal approximation if software is available? Three answers. (1) Historical. Useful when modern computer capabilities were not easily available and it was much easier to calculate one Normal probability than very many Binomial probabilities. (2) Theoretical. Central Limit Theorem is a more general version. (3) Inverse problems. Given a specified cumulative probability$F(x)$ find corresponding x.

Normal Approximation to Hypergeometric

Recall that Binomial approximates Hypergeometric when sample size n is much smaller than population size N. Text suggests $\frac{n}{N} < 0.1$. More conservative range requires $\frac{n}{N} < 0.05$. Here $p = \frac{K}{N}$. So if $\frac{n}{N} < 0.05$ and n is large, then we can also use the Normal approximation for the Hypergeometric.

Example. A fair coin is tossed $n = 100, 1000, 10000$ times. $p = 0.5, q = 1 - p = 0.5$. Find the probability that the percentage of heads lies within 2% of 50%, that is between 48% and 52%.

$$\begin{aligned} P(a \le X \le b) &= P(X \le b) - P(X \le a - 1) \\ &= \Phi(\frac{b + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{a - 0.5 - np}{\sqrt{npq}}) \end{aligned}$$

$n = 100$: $np = 50, \sqrt{npq} = 5, a = 48, b = 52$

$$\begin{aligned} P(48 \le X \le 52) &= \Phi(\frac{52 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{48 - 0.5 - np}{\sqrt{npq}}) \\ &= \Phi(0.5000) - \Phi(-0.5000) = 0.3829. \end{aligned}$$

$n = 1000$: $np = 500, \sqrt{npq} = 15.81, a = 480, b = 520$

$$\begin{aligned} P(480 \le X \le 520) &= \Phi(\frac{520 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{480 - 0.5 - np}{\sqrt{npq}}) \\ &= \Phi(1.2965) - \Phi(-1.2965) = 0.8052. \end{aligned}$$

$n = 10000$: $np = 5000, \sqrt{npq} = 50, a = 4800, b = 5200$

$$\begin{aligned} P(4800 \le X \le 5200) &= \Phi(\frac{5200 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{4800 - 0.5 - np}{\sqrt{npq}}) \\ &= \Phi(4.0100) - \Phi(-4.0100) = 0.9999 \end{aligned}$$

Example. Similar to last Example except that coin is not fair. Suppose chance of heads is $p = 0.3$. Now we would expect that the percentage of heads in repeated tosses should get close to 30%. Suppose the coin is tossed 100, 1000, 10000 times. Find the probability that the percentage of heads lies within 2% of 30%, that is between between 28% and 32%.

Now we set:
$q = 1 - p$. $n = 100$: $np = 30, \sqrt{npq} = 4.5826, a = 28, b = 32$

$$P(28 \leq X \leq 32) = \Phi(\frac{32 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{28 - 0.5 - np}{\sqrt{npq}})$$
$$= \Phi(0.5455) - \Phi(-0.5455) = 0.4146.$$

$n = 1000$: $np = 300, \sqrt{npq} = 14.4914, a = 280, b = 320$

$$P(280 \leq X \leq 320) = \Phi(\frac{320 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{280 - 0.5 - np}{\sqrt{npq}})$$
$$= \Phi(1.4146) - \Phi(-1.4146) = 0.8428.$$

$n = 10000$:$np = 3000, \sqrt{npq} = 45.8258, a = 2800, b = 3200$

$$P(2800 \leq X \leq 3200) = \Phi(\frac{3200 + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{2800 - 0.5 - np}{\sqrt{npq}})$$
$$= \Phi(4.3753) - \Phi(-4.3753) = 1.0000$$

We see that as the number of tosses increases, the proportion of heads $\widehat{p} = \frac{X}{n}$ gets closer to the true probability p and lies with increasing probability inside a decreasing width interval around p. This result is the basis of an estimation procedure called *Confidence Intervals* to be discussed later in this course.

Example. $X \sim Binomial(16, 0.5)$. Use normal approximation to find $P(X = 7)$.

$\mu = np = 16(0.5) = 8,$

$\sigma = \sqrt{np(1-p)} = \sqrt{16(0.5)(0.5)} = 2.$

$$\begin{aligned} P(X = 7) &= P(6.5 < X \leq 7.5) \\ &\approx P(\frac{6.5 - 8}{2} < Z \leq \frac{7.5 - 8}{2}) \\ &= \Phi(-0.25) - \Phi(-0.75) \\ &= 0.4013 - 0.2266 = 0.1747 \end{aligned}$$

Exact value.

$$\begin{aligned} P(X = 7) &= \begin{pmatrix} 16 \\ 7 \end{pmatrix} 0.5^7 (1 - 0.5)^9 \\ &= 0.1740 \end{aligned}$$

Exercise. Use the Normal approximation to find the probability that the proportion of heads is *exactly* 0.5 for $n = 100, 1000, 10000$ when a fair coin is tossed. Compare with exact value.

$n = 100$: $p = 0.5, np = 50, \sqrt{npq} = 5$
Normal approximation

$$\begin{aligned} P(X = 50) &= P(49.5 < X \leq 50.5) \\ &\approx P(\frac{50.5 - 50}{5} < Z \leq \frac{49.5 - 50}{5}) \\ &= \Phi(0.1) - \Phi(-0.1) \\ &= 0.5398 - 0.4602 = 0.0797 \end{aligned}$$

Exact

$$P(X = 50) = \left(\begin{array}{c} 100 \\ 50 \end{array} \right) 0.5^{50} 0.5^{50} = 0.0796$$

N	Exact	Normal Approximation
100	0.079589	0.079656
1000	0.025225	0.025227
10000	0.007979	0.007979

Here the continuity correction is essential.