

“JUST THE MATHS”

SLIDES NUMBER

19.8

PROBABILITY 8
(The normal distribution)

by

A.J.Hobson

19.8.1 Limiting position of a frequency polygon

19.8.2 Area under the normal curve

19.8.3 Normal distribution for continuous variables

UNIT 19.8 - PROBABILITY 8

THE NORMAL DISTRIBUTION

19.8.1 LIMITING POSITION OF A FREQUENCY POLYGON

The distribution considered here is appropriate to examples where the number of trials is large and hence the calculation of frequencies and probabilities, using the binomial distribution, would be inconvenient

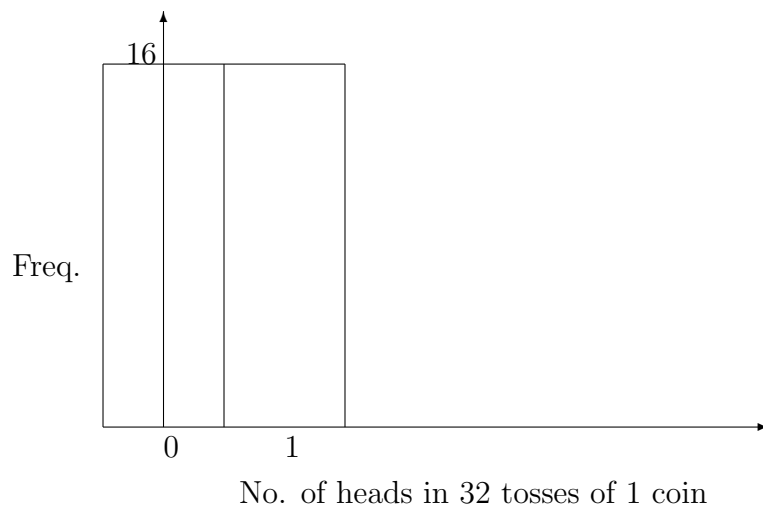
We introduce the “**normal distribution**” by considering the histograms of the binomial distribution for a toss of 32 coins as the number of coins increases.

The probability of obtaining a head is $\frac{1}{2}$ and the probability of obtaining a tail is also $\frac{1}{2}$.

(i) One Coin

$$32\left(\frac{1}{2} + \frac{1}{2}\right)^1 =$$

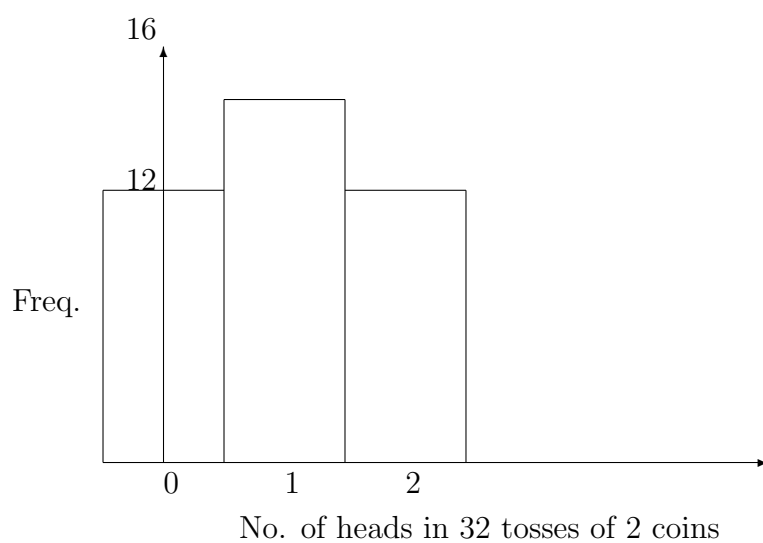
$$32\left(\frac{1}{2} + \frac{1}{2}\right) = 16 + 16.$$



(ii) Two Coins

$$32 \left(\frac{1}{2} + \frac{1}{2} \right)^2 =$$

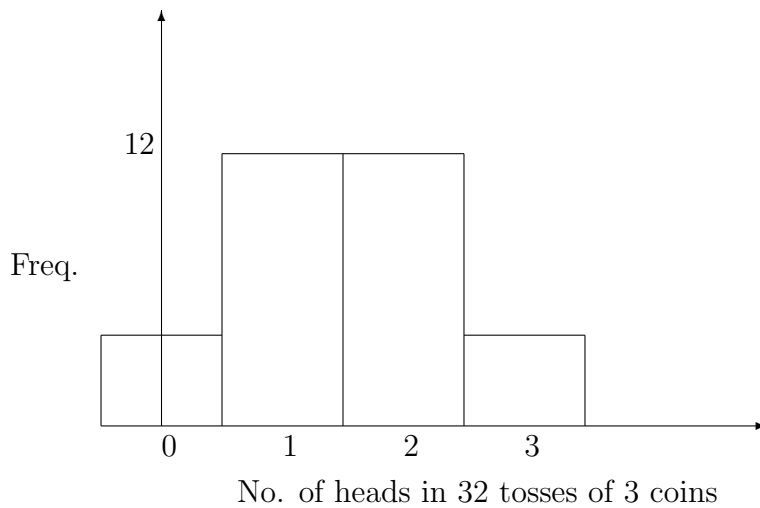
$$32 \left(\left[\frac{1}{2} \right]^2 + 2 \left[\frac{1}{2} \right] \left[\frac{1}{2} \right] + \left[\frac{1}{2} \right]^2 \right) = 8 + 16 + 8.$$



(iii) Three Coins

$$32\left(\frac{1}{2} + \frac{1}{2}\right)^3 =$$

$$32\left(\left[\frac{1}{2}\right]^3 + 3\left[\frac{1}{2}\right]^2\left[\frac{1}{2}\right] + 3\left[\frac{1}{2}\right]\left[\frac{1}{2}\right]^2 + \left[\frac{1}{2}\right]^3\right) = 4 + 12 + 12 + 4.$$

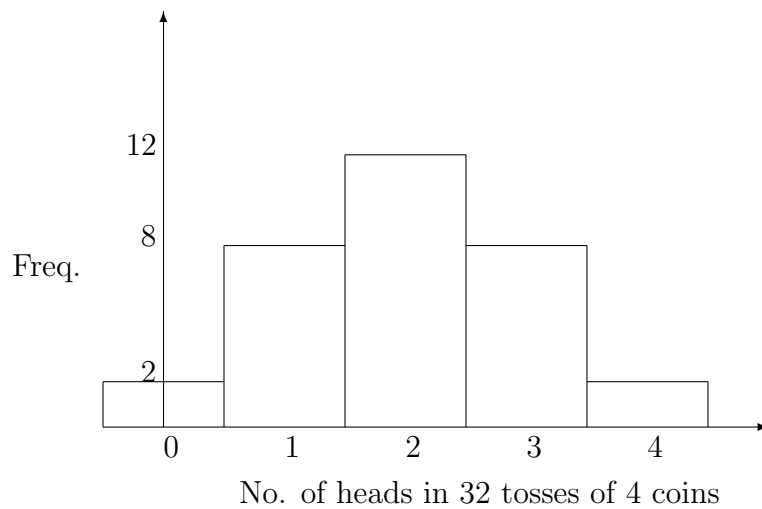


(iv) Four Coins

$$32\left(\frac{1}{2} + \frac{1}{2}\right)^4 =$$

$$32\left(\left[\frac{1}{2}\right]^4 + 4\left[\frac{1}{2}\right]^3\left[\frac{1}{2}\right] + 6\left[\frac{1}{2}\right]^2\left[\frac{1}{2}\right]^2 + 4\left[\frac{1}{2}\right]\left[\frac{1}{2}\right]^3 + \left[\frac{1}{2}\right]^4\right)$$

$$= 2 + 8 + 12 + 8 + 2.$$



As the number of coins increases, the frequency polygon approaches a symmetrical bell-shaped curve.

This is true only when the histogram itself is either symmetrical or nearly symmetrical.

DEFINITION

As the number of trials increases indefinitely, the limiting position of the frequency polygon is called the “**normal frequency curve**”.

THEOREM

In a binomial distribution for N samples of n trials each, where the probability of success in a single trial is p , it may be shown that, as n increases indefinitely, the frequency polygon approaches a smooth curve, called the “**normal curve**”, whose equation is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}},$$

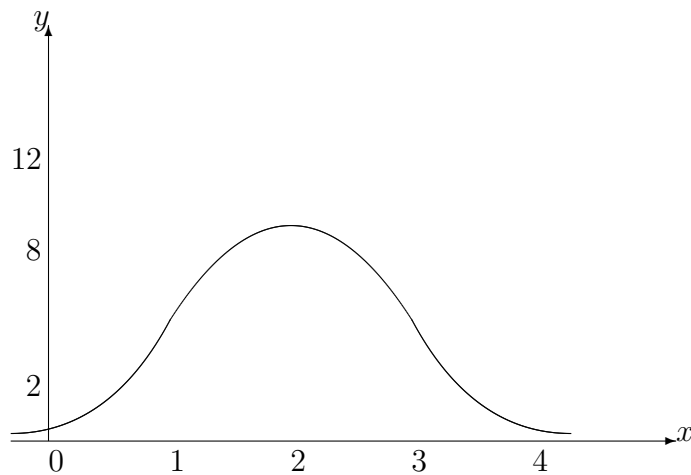
where

\bar{x} is the mean of the binomial distribution = np ;

σ is the standard deviation of the binomial distribution = $\sqrt{np(1-p)}$;

y is the frequency of occurrence of the value, x .

For example, the histogram for 32 tosses of 4 coins approximates to the following normal curve:



Notes:

- (i) We omit the proof of the Theorem.
- (ii) The larger the value of n , the better is the level of approximation.
- (iii) The normal curve is symmetrical about the straight line $x = \bar{x}$, since the value of y is the same at $x = \bar{x} \pm h$ for any number, h .
- (iv) If the relative frequency (or probability) with which the value, x , occurs is denoted by P , then $P = y/N$ and the relationship can be written

$$P = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}.$$

The graph of this equation is called the “**normal probability curve**”.

(v) Symmetrical curves are easier to deal with if the vertical axes of co-ordinates is the line of symmetry.

The normal probability curve can be simplified if we move the origin to the point $(\bar{x}, 0)$ and plot $P\sigma$ on the vertical axis instead of P .

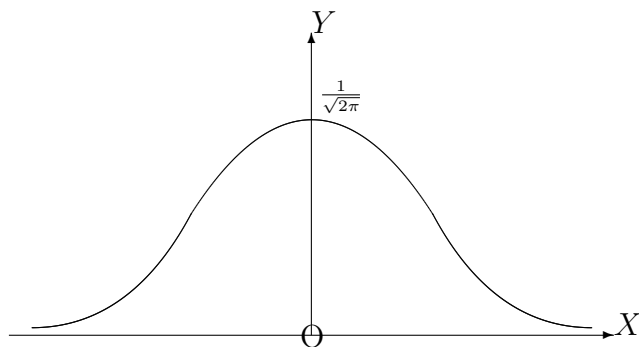
Letting $P\sigma = Y$ and $\frac{x-\bar{x}}{\sigma} = X$ the equation of the normal probability curve becomes

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}}.$$

This equation represents the “**standard normal probability curve**”.

From any point on the standard normal probability curve, we may obtain the values of the original P and x values by using the formulae

$$x = \sigma X + \bar{x} \quad \text{and} \quad P = \frac{Y}{\sigma}.$$



(vi) If the probability of success, p , in a single trial is **not** equal to or approximately equal to $\frac{1}{2}$, then the distribution given by the normal frequency curve and the two subsequent curves will be a poor approximation and is seldom used for such cases.

19.8.2 AREA UNDER THE NORMAL CURVE

For the histogram of a binomial distribution corresponding to values of x , suppose that $x = a$ and $x = b$ are the values of x at the base-centres of two particular rectangles, where $b > a$ and all rectangles have width 1.

The area of the histogram from $x = a - \frac{1}{2}$ to $x = b + \frac{1}{2}$ represents the number of times which we can expect values of x , between $x = a$ and $x = b$ inclusive, to occur.

For a large number of trials, we may use the area under the normal curve between $x = a - \frac{1}{2}$ and $x = b + \frac{1}{2}$.

The **probability** that x will lie between $x = a$ and $x = b$ is represented by the area under the normal probability curve from $x = a - \frac{1}{2}$ and $x = b + \frac{1}{2}$.

We note that the total area under this curve must be 1, since it represents the probability that **any** value of x will occur (a certainty).

To make use of a standard normal probability curve for the same purpose, the conversion formulae from x to X and P to Y must be used.

Note:

Tables are commercially available for the area under a standard normal probability curve.

In using such tables, the conversion formulae will usually be necessary.

EXAMPLE

If 12 dice are thrown, determine the probability, using the normal probability curve approximation, that 7 or more dice will show a 5.

Solution

For this example, we use $p = \frac{1}{6}$, $q = \frac{5}{6}$, $n = 12$.

We need the area under the normal probability curve from $x = 6.5$ to $x = 12.5$

The mean of the binomial distribution, in this case, is $\bar{x} = 12 \times \frac{1}{6} = 2$.

The standard deviation is $\sigma = \sqrt{2 \times \frac{1}{6} \times \frac{5}{6}} \simeq \sqrt{1.67} \simeq 1.29$

The required area under the standard normal probability curve will be that lying between

$$X = \frac{6.5 - 2}{1.29} \simeq 3.49 \quad \text{and} \quad X = \frac{12.5 - 2}{1.29} \simeq 8.14$$

In practice, we take the whole area to the right of $X = 3.49$, since the area beyond $X = 8.14$ is negligible.

Also, the total area to the right of $X = 0$ is 0.5; and, hence, the required area is 0.5 minus the area from $X = 0$ to $X = 3.49$

From tables, the required area is $0.5 - 0.4998 = 0.0002$ and this is the probability that, when 12 dice are thrown, 7 or more will show a 5.

Note:

If we had required the probability that 7 or fewer dice show a 5, we would have needed the area under the normal probability curve from $x = -0.5$ to $x = 7.5$

This is equivalent to taking the whole of the area under the standard normal probability curve which lies to the left of

$$X = \frac{7.5 - 2}{1.29} \simeq 4.26$$

19.8.3 NORMAL DISTRIBUTION FOR CONTINUOUS VARIABLES

So far, the variable, x , has been able to take only the specific values 0,1,2,3.....etc.

Here, we consider the situation when x is a continuous variable.

That is, it may take any value within a certain range appropriate to the problem under consideration.

For a large number of observations of a continuous variable, the corresponding histogram need not have rectangles of class-width 1, but of some other number, say c .

In this case, it may be shown that the normal curve approximation to the histogram has equation

$$y = \frac{Nc}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}.$$

The smaller is the value of c , the larger is the number of rectangles and the better is the approximation supplied by the curve.

If we wished to calculate the number of x -values lying between $x = a$ and $x = b$ (where $b > a$), we would need to calculate the area of the histogram from $x = a$ to $x = b$ inclusive, then **divide by c** , since the base-width is no longer 1.

We conclude that the number of these x -values approximates to the area under the normal curve from $x = a$ to $x = b$.

Similarly, the area under the normal probability curve, from $x = a$ to $x = b$ gives an estimate for the probability that values of x between $x = a$ and $x = b$ will occur.

EXAMPLE

A normal distribution of a continuous variable, x , has $N = 2000$, $\bar{x} = 20$ and $\sigma = 5$.

Determine

- (a) the number of x -values lying between 12 and 22;
- (b) the number of x -values larger than 30.

Solution

(a) The area under the normal probability curve between $x = 12$ and $x = 22$ is the area under the standard normal probability curve from

$$X = \frac{12 - 20}{5} = -1.6 \quad \text{to} \quad X = \frac{22 - 20}{5} = 0.4$$

From tables, this is $0.4452 + 0.1554 = 0.6006$

Hence, the required number of values is approximately $0.6006 \times 2000 \simeq 1201$.

(b) The total area under the normal probability curve to the right of $x = 30$ is the area under the standard normal probability curve to the right of

$$X = \frac{30 - 20}{5} = 2;$$

and, from tables, this is 0.0227

Hence, the required number of values is approximately $0.0227 \times 2000 \simeq 45$.