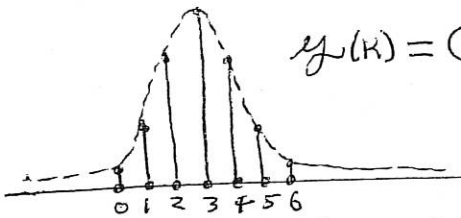
Central Limit Theorem

Notes on the Binomial Distribution

$$\frac{e^{-x^2/n}}{\sqrt{\pi n}} \sim \frac{C^{2n}_{n+x}}{2^{2n}}$$

```
                1
              1   1
            1   2   1
          1   3   3   1
        1   4   6   4   1
      1   5  10  10   5   1
    1   6  15  20  15   6   1
```



$$y(K) = C^N_K = \frac{N(N-1)\cdots(N-K+1)}{K!}$$

$$\boxed{C^N_{K+1} = \left(\frac{N-K}{K+1}\right) C^N_K}$$

$\uparrow n$ $\quad N = 2n \ (n=3)$

Note: $\{1 \xrightarrow{6/1} 6 \xrightarrow{5/2} 15 \xrightarrow{4/3} 20 \xrightarrow{3/4} 15 \xrightarrow{2/5} 6 \xrightarrow{1/6} 1\}$

$$\boxed{y(K+1) = \left(\frac{N-K}{K+1}\right) y(K)}$$

$$\Delta y = y(K+1) - y(K) = \left(\left(\frac{N-K}{K+1}\right) - 1\right) y(K)$$

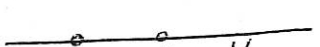
$$\Delta y = \left(\frac{N-2K-1}{K+1}\right) y(K). \quad \text{Let } x = K-n,\ N = 2n$$

(centering dist at $K=n \Leftrightarrow x=0$). Then

$$\Delta y = \left(\frac{2(n-K)-1}{n+x+1}\right) y(x) = \left(\frac{-2x-1}{n+x+1}\right) y(x)$$

So $\Delta y \approx (-2x/n)\, y(x)$ for n large and $\Delta x = 1$. Now shift the spacing so that

replaced by



, and

Vertical increments also reduced to multiples of $\frac{1}{\sqrt{n}}$. With $\Delta x = \frac{1}{\sqrt{n}}$ we get

$\Delta y/\Delta x \approx -2x\, y(x)$. So if there is a limiting smooth curve through this ($\sqrt{n}$ – readjusted) limit of binomial distribution, then

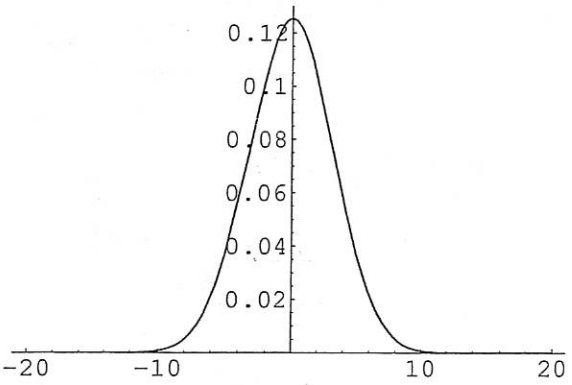
$\frac{dy}{dx} = -2x\, y(x)$. Hence $y(x) = Ae^{-x^2}$

for some constant A. This explains why the normal distribution is the limit of the binomial distribution (suitably normalized) as $n \to \infty$!

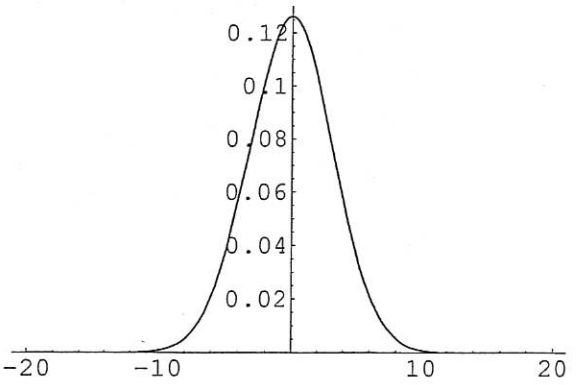
[Recall that $\sum_{K=1}^{N} C^N_K / 2^N = 1$ and that $\int_{-\infty}^{\infty} \left[e^{-x^2}/\sqrt{\pi}\right] dx = 1$.]

This result is called the Central Limit Theorem and sometimes the Law of Large Numbers.

```
n = 20;
y[x_] = Binomial[2 n, x + n] / (2^(2 n));
BIT = Table[y[i], {i, -n, n}];
ListPlot[BIT, Axes -> False];
Plot[y[x], {x, -n, n}, PlotRange → {0, .13}]
Const = N[Sqrt[n Pi]];
Plot[Exp[-(x^2) / n] / Const, {x, -n, n}, PlotRange → {0, .13}
]
```



- Graphics -



- Graphics -

```
m = 20;
t = 0;
N[Binomial[2 m, t + m] / (2^(2 m)), 10]
N[Exp[-(t^2) / m] / Sqrt[m Pi], 10]
```

0.125371

0.126157

$$\left[\int_{-\infty}^{\infty} e^{-x^2} dx\right]^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2} dx dy$$

$$= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r\, dr\, d\theta$$

$$= 2\pi \int_0^{\infty} e^{-r^2} r\, dr$$

$$= 2\pi \int_0^{\infty} -\frac{1}{2} d\left(e^{-r^2}\right)$$

$$= -\pi\left(e^{-r^2}\Big|_0^{\infty}\right) = \pi$$

$$\therefore \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

$$\frac{e^{-x^2/n}}{\sqrt{\pi n}} \sim \frac{C_{n+x}^{2n}}{2^{2n}}$$

$$\Longrightarrow \pi \sim \frac{2^{4n}(n!)^4}{n((2n)!)^2}.$$

$$\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2 = \iint_{\mathbb{R}^2} e^{-r^2} dx\,dy$$

$$= \int_0^{\infty}\int_0^{2\pi} e^{-r^2} r\,dr\,d\theta$$

$$= 2\pi \int_0^{\infty} -\frac{1}{2} d\left(e^{-r^2}\right)$$

$$= -\pi\left(e^{-r^2}\Big|_0^{\infty}\right.$$

$$= \pi$$

$$\therefore \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

$$\boxed{\frac{e^{-x^2/n}}{\sqrt{\pi n}} \sim \frac{C_{n+x}^{2n}}{2^{2n}}}$$

$$\frac{1}{\sqrt{\pi n}} \sim C_n^{2n}/2^{2n}$$

$$\frac{2^{2n}}{\sqrt{n} C_n^{2n}} \sim \sqrt{\pi}$$

$$\frac{2^{2n}}{\sqrt{n}\,\frac{(2n)!}{n!(n!)}} = \frac{2^{2n}(n!)^2}{\sqrt{n}\,(2n)!}$$

$$\boxed{\pi \sim \frac{2^{4n}(n!)^4}{n\,((2n)!)^2}}$$

Wallis Product For $\pi/2$

$$\frac{\pi}{2} = \prod_{k=1}^{\infty} \left(\frac{2k}{2k-1} \cdot \frac{2k}{2k+1} \right)$$

$$= \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\cdots$$

$P_K \to \frac{\pi}{2}$ as $K \to \infty$

$$P_K = \prod_{n=1}^{K} \left(\frac{2n}{2n-1}\right)\left(\frac{2n}{2n+1}\right)$$

$$= \left(\frac{1}{2K+1}\right) \prod_{n=1}^{K} \frac{(2n)^4}{[(2n)(2n-1)]^2}$$

$$= \frac{1}{2K+1} \quad \frac{2^{4K}(K!)^4}{((2K)!)^2}$$

$$2P_K = \left(\frac{2}{2K+1}\right) \frac{2^{4K}(K!)^4}{((2K)!)^2}$$

$$\sqrt{\frac{1}{\left(\frac{2K+1}{2}\right)(2P_K)}} = \frac{(2K)!}{2^{2K}(K!)^2} = \frac{C_K^{2K}}{2^{2K}}$$

So $$\boxed{\frac{1}{\sqrt{\pi K}} \sim \frac{C_K^{2K}}{2^{2K}}}$$

Also: $\sin(x) = x\prod_{n=1}^{\infty}\left(1 - x^2/\pi^2 n^2\right)$ (Euler)

$x = \pi/2 \Rightarrow$ Wallis prod.

$$\Rightarrow \frac{2}{\pi} = \prod_{n=1}^{\infty}\left(1 - \frac{\pi^2/4}{\pi^2 n^2}\right) = \prod_{n=1}^{\infty}\left(1 - \frac{1}{4n^2}\right)$$

$$= \prod_{n=1}^{\infty}\left(\frac{4n^2-1}{4n^2}\right) = \prod_{n=1}^{\infty}\left(\frac{(2n-1)(2n+1)}{(2n)(2n)}\right)$$

Elements of the Central Limit Theorem (Excerpt from "Mathematics for the Million")

The frequency terms of the foregoing distribution corresponds to those of the expansion of $(\frac{1}{2} + \frac{1}{2})^6$. If we ask what is the area included by deviations of 2 from the mean on either side of it, the appropriate expression is:

$$\sum_{X=-2}^{X=+2} y_x \Delta X \qquad (\Delta x = 1 = \Delta X)$$

The derivation of the normal curve for a binomial distribution involving different values of p and q is very laborious. Here we shall consider only the case (e.g. r-fold spin of a coin) when $p = \frac{1}{2} = q$, and it simplifies our arithmetic if we write $r = 2n$

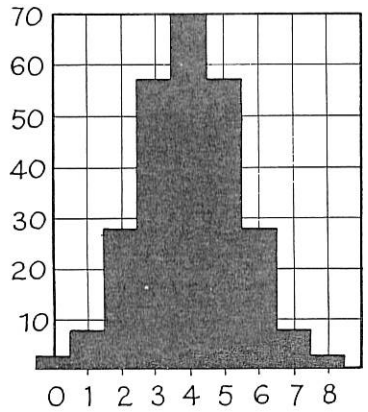


Fig. 193. Probability as the Ratio of Two Areas

If $p = \frac{1}{2} = q$ meaning equal probability that a newly born will be a boy or a girl, we can represent the probability that a family of eight will consist of 0, 1, 2 . . . 8 girls (or boys) as indicated on the horizontal axis. The vertical unit is $(\frac{1}{2})^8$. The area of the first four columns is then the probability that there will not be more than three of the same sex, that of the remaining five that there will be at least four.

in the expression $(\frac{1}{2} + \frac{1}{2})^r$. The mean score is then $M = n$. Without ambiguity, we may use the symbol y_x to signify both the frequency of a score x and of its corresponding score deviation $X = x - n$, so that $x = X + n$. As we have seen

$$y_x = 2^{-2n} \frac{(2n)!}{(2n-x)!x!} = \frac{2^{-2n}(2n)!\ (x+1)}{(2n-x)!\ (x+1)!}$$

$$y_{x+1} = 2^{-2n} \frac{(2n)!}{(2n-x-1)!(x+1)!} = \frac{2^{-2n}(2n)!(2n-x)}{(2n-x)!(x+1)!}$$

$$\therefore\ y_{x+1} = \frac{2n-x}{x+1} y_x$$

$$\therefore\ \Delta y_x = y_{x+1} - y_x = \left(\frac{2n-x}{x+1} - 1\right) y_x$$

$$\therefore\ \frac{1}{y_x} \Delta y_x = \frac{2n-2x-1}{x+1}$$

Since $x = n + X$

$$\frac{1}{y_x} \Delta y_x = \frac{-2X-1}{n+X+1} = \frac{-2X}{n+X+1} - \frac{1}{n+X+1}$$

Our assumption in seeking a curve cutting very closely the mid-points at the top of each column of the histogram is that n is very large. So the overwhelming majority of values of X will cluster round the mean, i.e. $X = 0$ since $(x - n) = X$. We may thus plausibly explore the possibility of getting a good fit by writing:

$$(n + X + 1) \simeq n \quad \text{and} \quad \frac{1}{n+X+1} \simeq 0$$

$$\therefore\ \frac{1}{y_x} \Delta y_x \simeq \frac{-2X}{n}$$

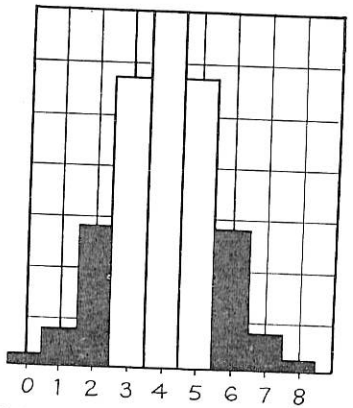


Fig. 194. Probability as the Ratio of Two Areas

Units as in Fig. 193. The unblackened area represents the probability that a family of eight will have not less than three or more than five of the same sex.

Since $\Delta X = 1$:

$$\frac{1}{y_x}\Delta y_x \simeq \frac{-2}{n} X \Delta X$$

When, as we assume, n is very large, we may write:

$$\frac{1}{y}dy \simeq \frac{-2X}{n}dX$$

$$\therefore \quad \int\frac{1}{y}dy \simeq \frac{-2}{n}\int X\,dX$$

$$\therefore \quad \log_e y = \frac{-X^2}{n} + C$$

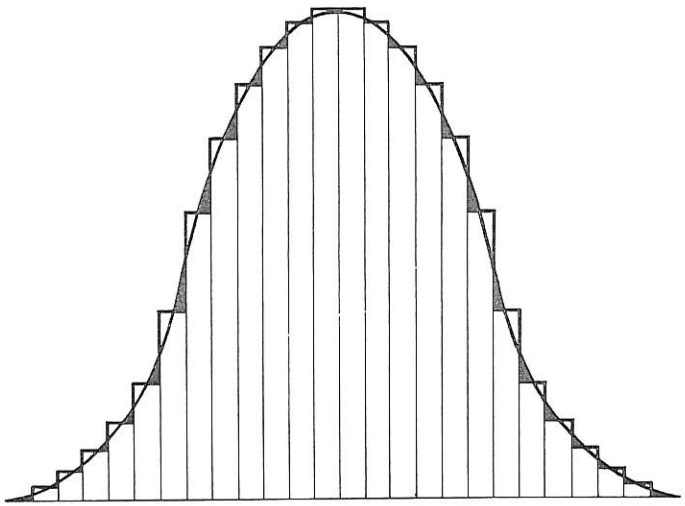


Fig. 195. The so-called Normal Curve as an Approximation to the Binomial Distribution

To tailor the last expression so that X is the independent variable, we may put $C = \log_e A$, so that

$$\log_e y - \log_e A = \frac{-X^2}{n} = \log_e\left(\frac{y}{A}\right)$$

Bearing in mind that $log_e\, a = b$ means $a = e^b$, we may therefore write:

$$\frac{y}{A} = e^{\frac{-X^2}{n}} \quad \text{and} \quad y = Ae^{\frac{-X^2}{n}}$$

The value of A is obtainable by putting $X = 0$ in which event $y = y_n$, i.e. the ordinate of the mean, so that:

$$y_n = \frac{2^{-2n}(2n)!}{n!\,n!} = A$$

We have already dealt with the properties of this curve in chapter 10 (page 493). For tabulation of values of the definite integral cited below, it is useful to invoke a formula which gives good values for $n!$ in A, so long as n is larger than 10, i.e.

$$n! \simeq e^{-n}n^n\sqrt{2\pi n}$$

so that $$A = \frac{2^{-2n}e^{-2n}2^{2n}\sqrt{4\pi n}}{e^{-2n}2^{2n}2\pi n} = \frac{1}{\sqrt{\pi n}}$$

We cannot say in advance how big n must be to justify the approximations we have made. Actually, the fit is very close in the range $X = \pm 2$ when n is as small as 8 so that $2n = 16$.

Our aim in this derivation is to be able to state the probability that a score will lie in a certain range expressed as an area on the assumption that the total area is unity, i.e.

$$\int_{\infty}^{+\infty} y\,dx = 1$$

To use it as such we have to remember that the curve approximately cuts the midpoint of the top of each column of Fig. 195, i.e. for the column score value $X = \pm a$ there is a deficiency of $\pm\frac{1}{2}$ in the range enclosed by the particular values of y. If n is relatively small ($n < 100$), we should perform our calculations on the assumption that

$$\sum_{X=-a}^{X=a} y_X \simeq \int_{-a+\frac{1}{2}}^{a-\frac{1}{2}} y\,dX$$

The reader who hopes to gain an intelligent grasp of how to use such tables should verify the meaning of the half interval correction $(a \pm \frac{1}{2})$ above by drawing the histogram of $(\frac{1}{2} + \frac{1}{2})^{10}$ and the curve which cuts the tops of each column approximately at its midpoint.

To perform the calculation, we can refer to tabulated values of the definite integral on the right. This implies that it is possible to evaluate the indefinite integral, as can be done by series integration (p. 584) which shows that the value of the definite integral over its whole range is unity, i.e.:

$$\frac{1}{\sqrt{\pi n}}\int_{\infty}^{+\infty} e^{\frac{-X^2}{n}}\, dx = 1$$

Probability and the Real World. Aside from its earliest use in connection with games of chance and its later misuse in connection with life insurance, we may broadly distinguish three domains in which the algebraic calculus of probability has staked a claim. One is the *theory of error*, largely due to Gauss. What prompts the question to which it seeks an answer is the circumstance that successive observations on the same phenomenon, especially ones made with instruments having a large assemblage of cog-wheels as in an astronomical observatory, are never exactly the same. Such an application of the theory of probability seeks to side-step the dilemma which arises when the investigator finds himself confronted with two sets of repeated measurements or counts respectively clustering around different mean values and to some extent overlapping. Here the decision sought is whether the difference between the means may or may not betoken two *true* values rather than one. In common parlance: is an observed difference between two means consistent with errors of observation? We need not dismiss the considerable plausibility of the assumption that the distribution of errors is comparable to what we obtain when we toss a die; but the practical issue is mathematically intractable unless we can justifiably invoke a correct value for p and q.

A second use of the algebraic theory in scientific investigation is wholly above criticism. One may call it a Calculus of Aggregates. This term embraces scientific hypotheses which deal with populations of particles, atoms, molecules, genes, and chromosomes, whose properties the observer can study in large assemblages. Such is the situation when we explore the Brownian movement of microscopic solid particles bombarded by molecules in solution, or when we explore the effect on the proportions of progeny of a particular type when an extra chromosome turns up in the parental stock. In such situations, the investigator can use a card pack, die, roulette wheel, or lottery 'urn' as a model of

interpretation. The justification for this, as for reliance on any other model, depends solely on *whether it works*. In the reflected glory of the spectacular success of scientific hypotheses of this class, e.g. the Kinetic Theory of gases in physical chemistry and the Theory of the Gene in biology, there has however proliferated an overgrowth of statistical theory which involves assumptions which one cannot hope to justify indisputably by whether they do work. One may distinguish between two broad categories: (*a*) techniques for *exploration*; (*b*) tests for the *validity of judgments*. The writer has elsewhere set forth at length the exceptionable assumptions these involve. Here he will content himself with the task of introducing the reader to an example of each category without critical comment.

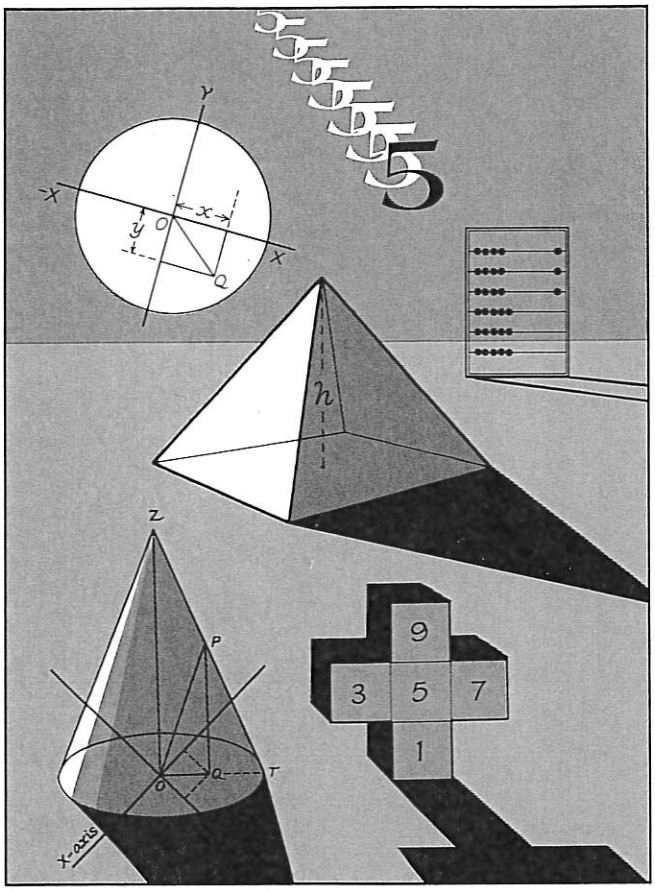
Correlation. As an example of a *technique of exploration* which invokes algebraic probability, we may instructively examine Spearman's *Rank Coefficient of Correlation*. The end in view is the search for connected characteristics, as if one asks: is ability to do mental arithmetic associated with parental income? If we arrange a class of boys and girls first in the order of the marks obtained in an arithmetical test and find the same or reverse order when we again arrange them according to the annual income of their parents, we should conclude that arithmetical facility and economic prosperity are connected in some way. Such complete correspondence would occur only if the effect of all contributory factors were perfectly standardized or negligible, and in practice we should not be discouraged from drawing a positive conclusion if a few names appeared out of place. Drawing the conclusion that a correspondence exists thus depends on adopting some standard for the amount of displacement which can occur when two such 'arrays' are compared.

A fundamental measure of displacement when we compare two arrays in this way is called *rank gain* or *loss*. Suppose the arithmetic marks of three boys A, B, C are 75, 52, and 39. Then A, B, C respectively have the ranks 1, 2, 3 in descending order of proficiency. If their parents' incomes are \$5000, \$3200, and \$4500 their ranks are 1, 3, 2, the rank of A is the same (1) in each array. The rank of B has decreased by 1 and that of C has increased by 1. The total rank gains and losses must always be the same, and the total number of either can be used as a criterion of correspondence.

To explore such a criterion, let us set out the number of all possible ways of arranging three things, i.e. $3! = 6$, viz.:

Mathematics for the Million

How to Master the Magic of Numbers



Lancelot Hogben