Probability and Statistics

Probabilistic statements and statistical arguments are ubiquitous in the media. Consider, for example, the statements

> "There is a 70% chance of rain this afternoon."
>
> "Yesterday's polls predict, at the 95% confidence level, that candidate X will win with between 56% and 58% of the vote."
>
> "The results reported in Nature, linking smoking with the incidence of disease Y, are significant at the 99% level."

What do these statements mean? How do the writers or speakers obtain these numbers? That is, in part, the topic of this module.

Part 1 – Probability

We all know that if you roll a single unloaded, fair die, you have a chance of 1 in 6 to roll a four. That is a statement about the *probability* of an event. Understanding probabilities makes it possible to answer more complex questions, such as

- Before the Health Insurance Company (HIC) accepts to underwrite Jim's health insurance, he has to be tested for AIDS. The test that HIC practices is highly reliable: only 1% of healthy people get a "false positive" result, only 1% of infected people a "false negative." Jim tests positive, and HIC refuses to give him health insurance. Jim has taken Math Alive, and points out to HIC that there is in fact less than one chance in ten that he is truly HIV positive. How is this possible?
- What is the winning strategy for the old TV game show "Make a deal"?

To answer these questions, we need to understand notions of probability theory.

Probability—the mathematics of chance

Probability theory emerged very late in the development of mathematics; its basic notions were articulated for the first time in the 17th century. The concept of probability seems

harder to grasp for us intuitively than that of, say, arithmetic. It is not uncommon, even for trained mathematicians, to make mistakes when estimating how probable an event is, if they don't think about it carefully.

Let's quickly review some basic definitions:

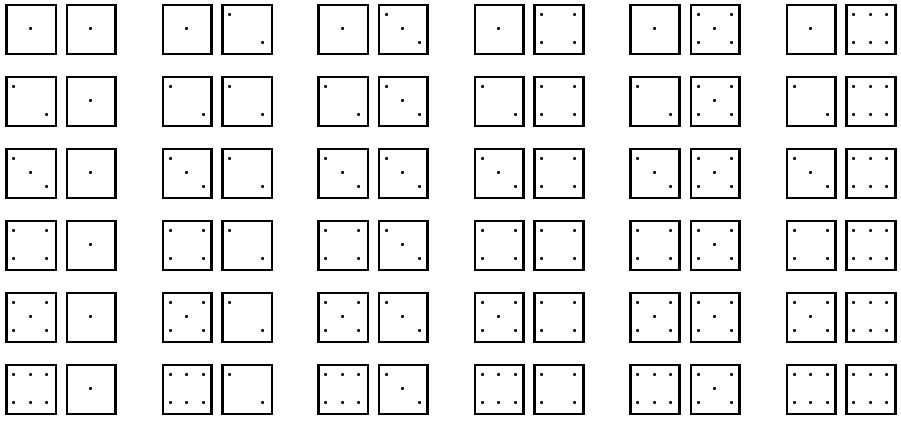
- **Random phenomenon:** the individual outcomes are uncertain, but the long-term pattern of many individual outcomes is predictable. Example: heads or tails when tossing a fair coin.
- **Probability of an outcome:** proportion of trials in which the outcome occurs in a very long run of trials. Example: the probability of getting heads when tossing a fair coin is $\frac{1}{2}$.
- **Sample space:** set of all possible outcomes.

Let's look at a few examples. If you toss a fair coin, then your sample space is $\{H, T\}$, if you roll one fair die, your sample space is $\{1, 2, 3, 4, 5, 6\}$.

Another example: 2 dice are rolled. Now the sample space depends on the game. Do both values matter? Or only the sum? (As in craps.) If only the sum matters, then the sample space is

$$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\} .$$

Not all these outcomes are equally likely now, as you can see in the figure below.



$P(\text{outcome} = 2) = \frac{1}{36}$
$P(\text{outcome} = 3) = \frac{2}{36}$
$P(4) =$
$P(5) =$
$P(6) =$
$P(7) =$

$P(8) =$
$P(9) =$
$P(10) =$
$P(11) =$
$P(12) =$
not equal probabilities!

P (outcome is less than or equal to 6) =
P (outcome is odd) =
P (both dice have even outcome) =
P (same number on both dice) =

OBSERVATIONS:

- every probability is a number between 0 and 1
- sum of probabilities over all possible different outcomes is 1.

More examples:

- What is the probability that a 4-digit number does not contain a zero? (Leading zeros are allowed)

  number $\underbrace{\cdot \quad \cdot \quad \cdot \quad \cdot}$
  each entry is a number, from 0 to 9.
  $\Rightarrow$ number of elements in sample space =
  number of elements in sample space that do not have a zero =

  What is the probability that of three 4-digit PIN numbers that are assigned to you randomly by the different institutions, exactly two have one or more zeros, and the third doesn't?
  What is the probability that such a randomly assigned number is "funny", meaning that it strikes you as very improbable that you would get this number by chance?

- Birthday coincidences
  Assume 1 year = 365 days. No leap year birthdays allowed.
  - If you meet a stranger, what is the probability that she has the same birthday as you?
  - If two people meet, what is the probability that they have the same birthday?
  - Same for three people: what is the probability that at least two have same birthday? Easier: What is the probabilty that all their birthdays are different?
  - Four people: probability that at least two have same birthday?
  - How many people would you guess that you need for this probability to reach $\frac{1}{2}$?

Conditional probability

Back to the dice; roll two dice, and let's compute

$$P \text{ (one of the outcomes is 3, given that total outcome is 7) } =$$

Note that we have here a restricted sample space: we only look at those outcomes for which the total outcome is 7.

$P(A|B) =$ probability that A happens, given that B occurs.

Note:

$$P(A \text{ and } B) = P(A|B) \times P(B)$$

Example: Six men out of 100 are color-blind; one woman out of 100 is color-blind. In a population where men and women occur in the same proportion, find

- the probability that a person chosen at random is color-blind:
- the probability that a random color-blind person in this population is female?

Testing for infection (or for drug use)

Population at large:

$$P(\text{not infected}) = N$$
$$P(\text{infected}) = I = 1 - N$$

Effectiveness of test:

$$P(\text{test } ni|ni) = 0.99$$
$$P(\text{test } i|ni) = 0.01$$
$$P(\text{test } ni|i) = 0.01$$
$$P(\text{test } i|i) = 0.99$$

Question: $P(ni|\text{test } i) =?$

$$= \frac{P(ni \text{ and test } i)}{P(\text{test } i)}$$

$$= \frac{P(ni \text{ and test } i)}{P(\text{test } i|ni)P(ni)+P(\text{test } i|i)P(i)}$$

$$= \frac{P(\text{test } i|ni)P(ni)}{P(\text{test } i|ni)P(ni)+P(\text{test } i|i)P(i)}$$

$$= \frac{0.01\times N}{0.01\times N+0.99\times(1-N)}$$

Example:

$$N = 0.99 \rightarrow$$
$$N = 0.5 \rightarrow$$
$$N = 0.999 \rightarrow$$

The TV game show "Make a Deal"

In the game show three doors are shown; behind one of them, there is a car; behind the other two, a goat. The car could be sitting behind any of the 3 doors, with equal chances for each door. You pick one door, but it does not get opened right away. The game show host Monty Hall next opens one of the two doors that you did **not** pick, showing the goat behind it, and then offers you the possibility to switch. Should you, or not?

At the start: $P(\text{car is at A}) = P(\text{car is at B}) = P(\text{car is at C}) = \frac{1}{3}$.

You pick A.

Then Monty opens door C, and shows you the goat sitting there. Which is now highest, the probability that the car is behind door A or behind door B?

At the start we had:

$$P(\text{car at A}) = \tfrac{1}{3}$$
$$P(\text{car at B or C}) = \tfrac{2}{3}$$

This does not change! Monty's action shows you that there is no car at C, so now

$$P(\text{car at B}) = P(\text{car at B or at C}) = \frac{2}{3}\ .$$

$\Rightarrow$ In order to optimize you chances, you should switch!

But this is only true if Monty **has** to open a goat-door every time. Imagine that he tries to make you switch **only** if he knows you picked the right door from the start. Should you switch then?

And what if he is sometimes "good" Monty (switching offered every time) and sometimes "bad" Monty (switching offered only if it is disadvantageous to you)? Suppose he tosses a coin to decide whether he will be "good" or "bad." What is now the best strategy? Switch or not?

Try all this out in the lab!

Gambler's ruin

Let's look at the simple situation where a gambler and an opponent (= casino?) play a game. In each move, either the gambler or the opponent wins \$1 from the other party; suppose both have a probability of $\frac{1}{2}$ of winning each time.

The game continues until one of them has lost all his/her money.

Suppose the gambler starts with N dollars.

The combined starting capital of the gambler and the opponent is T (= total amount of money in game, which does not change).

The sample space is now the collection of all sequences W L L W W L .... (W = gambler wins, L = gambler loses), with the convention that we stop the sequence if one of them is ruined.

P_N = probability that the gambler will be ruined, starting with N \$.

$$\begin{aligned} P_N &= P(\text{ruin when starting with } N) \\ &= \tfrac{1}{2}P(\text{ruin when starting with } N+1) + \tfrac{1}{2}P(\text{ruin when starting with } N-1) \\ &= \tfrac{1}{2}P_{N+1} + \tfrac{1}{2}P_{N-1} \end{aligned}$$

P_0 = 1 (this probability is 1 because at the start of the game the gambler is already ruined if he has no money!)

P_T = 0 (this probability is 0 because the gambler has already the total sum, so that his opponent is already ruined, and the game is over; the gambler cannot lose now.)

$$\begin{aligned} P_1 &= \tfrac{1}{2}P_2 + \tfrac{1}{2}P_0 = \tfrac{1}{2}P_2 + \tfrac{1}{2} \\ &\Rightarrow P_2 = 2P_1 - 1 \\ P_2 &= \tfrac{1}{2}P_3 + \tfrac{1}{2}P_1 \\ &\Rightarrow P_3 = 2P_2 - P_1 = 4P_1 - 2 - P_1 = 3P_1 - 2 \\ P_3 &= \tfrac{1}{2}P_4 + \tfrac{1}{2}P_2 \\ &\Rightarrow P_4 = 2P_3 - P_2 = 6P_1 - 4 - 2P_1 + 1 = 4P_1 - 3 \end{aligned}$$

Induction: assume that $P_n = nP_1 - (n-1)$ is okay for $n = 1, 2, \ldots, K$. What then for $n = K+1$?

$$\begin{aligned} P_K &= \tfrac{1}{2}P_{K+1} + \tfrac{1}{2}P_{K-1} \\ \Rightarrow P_{K+1} &= 2P_K - P_{K-1} \\ &= 2\left[KP_1 - (K-1)\right] - \left[(K-1)P_1 - (K-2)\right] \\ &= (K+1)P_1 - K \Rightarrow \text{ Okay again!} \end{aligned}$$

$\Rightarrow$ formula holds for all n.

In particular

$$\begin{array}{rcccl} 0 & = & P_T & = & TP_1 - (T-1) \\ & \Rightarrow & P_1 & = & \frac{T-1}{T} \\ & \Rightarrow & P_n & = & n\frac{T-1}{T} - (n-1) \\ & & & = & \frac{n(T-1)-(n-1)T}{T} = \frac{T-n}{T} = 1 - \frac{n}{T} \end{array}$$

So if $n = \frac{T}{2}$ (gambler and opponent start with same capital) $\Rightarrow P_n = \frac{1}{2}$: there is an even chance that the gambler or the opponent ends up ruined.

If T is much bigger than n, then the gambler's probability of ending up ruined is very close to 1.

What if the gambler doesn't aim to "ruin the casino," but just to make some money? Example: the gambler has \$1000 and wants to stop after (and if) he makes \$10. That is like playing against an opponent who only has \$10. $\Rightarrow T = 1010$ now. $\Rightarrow$ Chance of winning \$10 $= 1 -$ chance of ruin $= 1 - (1 - \frac{1000}{1010}) = \frac{1000}{1010} \simeq 0.990 \Rightarrow$ pretty good ....

But the total expected winnings are still zero! (In practice, the game is not fair, so the casino **always** wins on average.)

The Normal Distribution

Suppose you toss a coin 100 times, and you count the number of heads you obtained. This is really similar to drawing beads from a large collection of white and black beads (where we assume the beads to be evenly distributed: 50% white, 50% black, and they are well mixed.) (Looking at the number of heads after 100 coin tosses is the same as taking out 100 times a bead from the huge bag of beads and counting how many white ones we got, assuming we return the bead we picked, every time, after noting its color.)

What is the probability that you get 61 heads or 61 white beads? To compute this probability, we

- identify the sample space and see how large it is; in this case there are 2^{100} possible sequences of 100 entries, each of which is "head" or "tail"; we can also represent these two possible outcomes by 1 and 0.
- count the number of "samples" corresponding to exactly 61 heads, i.e. the number of sequences with exactly 61 entries = 1, and 39 entries = 0. We did in fact see earlier

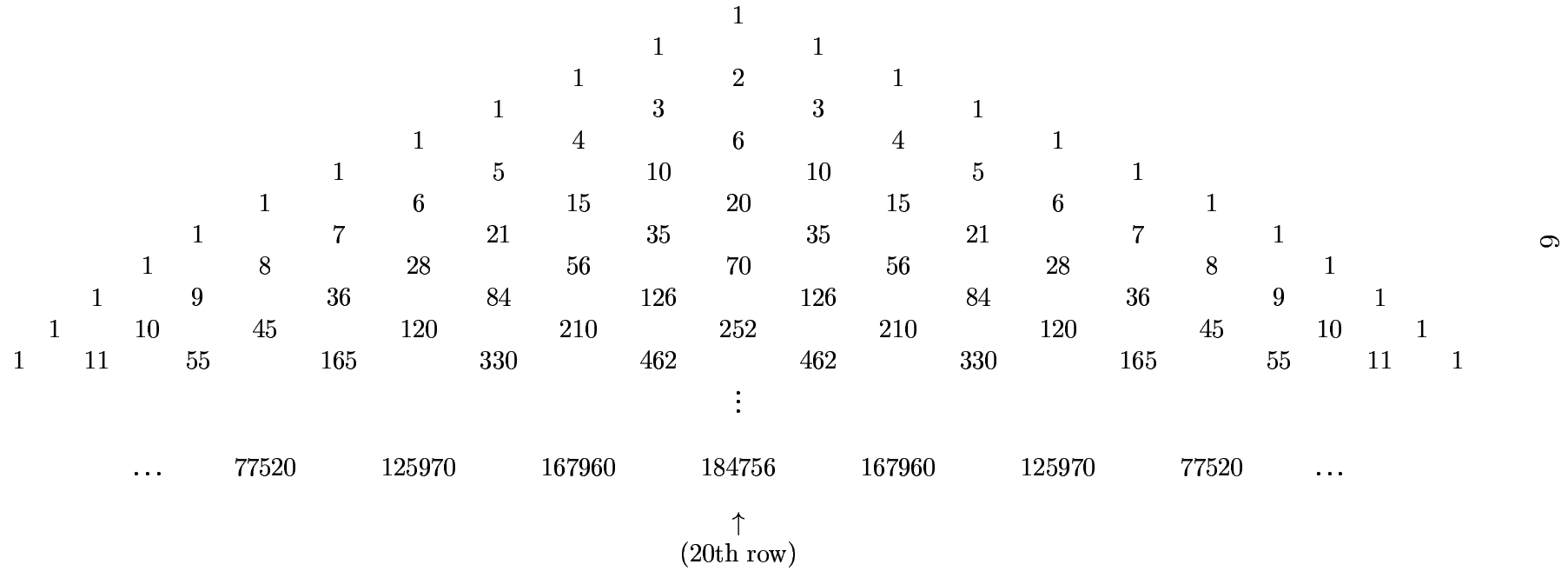
how to compute this number via Pascal's triangle (see next page). Remember:

$$\begin{aligned}(x+y)^n &= \underbrace{(x+y)(x+y)\cdots(x+y)}_{n \text{ factors}} \\ &= x^n + nx^{n-1}y + \cdots + \tfrac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}x^{n-k}y^k \\ &\quad + \cdots + nxy^{n-1} + y^n\end{aligned}$$

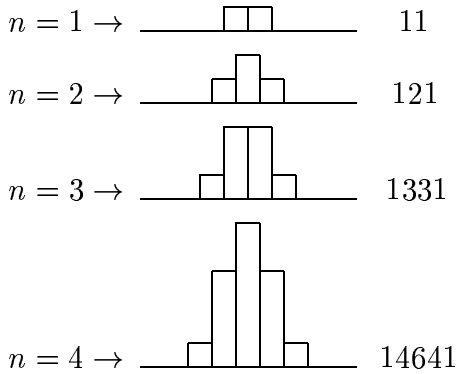
The coefficient of $x^{n-k}y^k$, namely $\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$, is exactly the number of different ways in which we can pick "y" k times when we work out the product of n factors, that is, it is the number of ways in which we can, given n possible picks, have k of these correspond to y (and the other $n-k$ picks correspond to x).

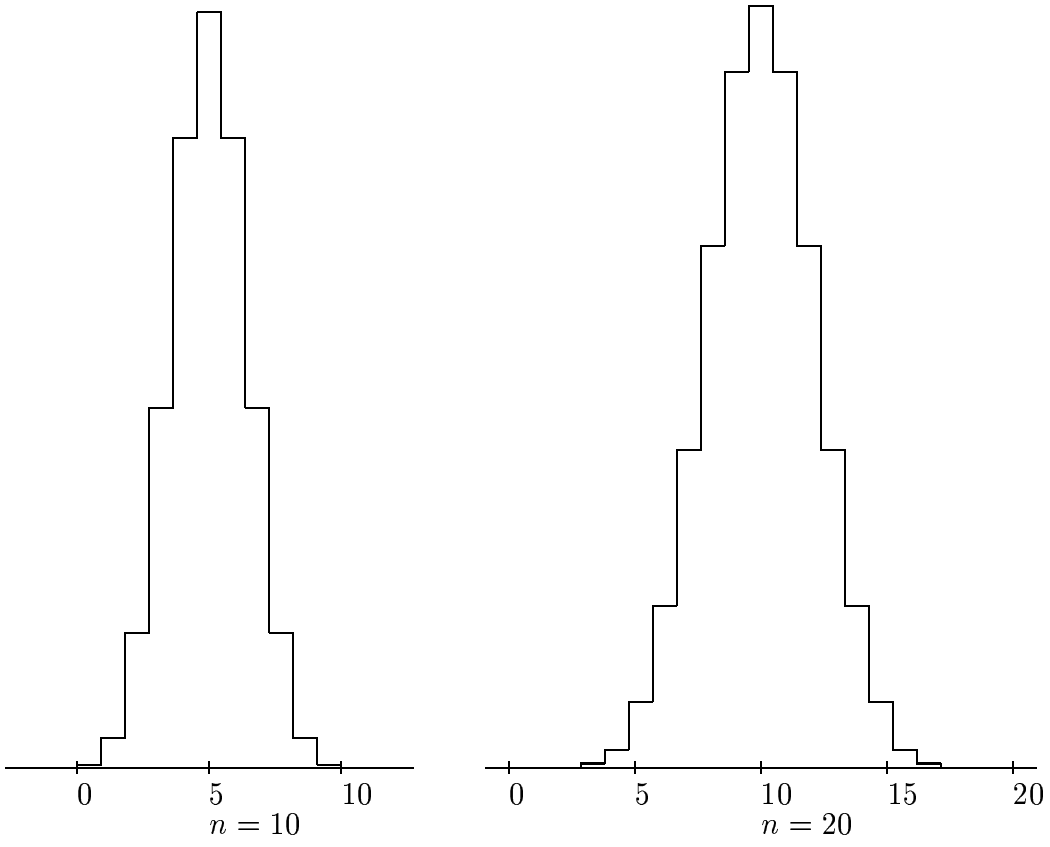
The probability that you have exactly 61 "heads" in 100 coin tosses is thus $\frac{\frac{100\times 99\times\cdots\times 40}{61\times 60\times\cdots\times 2\times 1}}{2^{100}}$, which is about .07 . (Note: you can compute this number on your calculator, especially if it allows you to program. But you have to be careful: if you start by computing $100\times 99\times\ldots\times 40$, everything will overflow. A better way is to write it as a product of many fractions that are each not too large or too small. To do this, let's first cancel out factors; we can rewrite the whole expression as $\frac{100\times 99\times\ldots\times 62}{39\times 38\times\ldots\times 2\times 1}\times\frac{1}{2^{100}}$, and now break it up $\frac{100}{2^2\times 39}\times\frac{99}{2^2\times 38}\times\ldots\times\frac{62}{2^2\times 1}\times\frac{1}{2^{22}}$. This should all remain tractable on your calculator.)

Pascal's triangle



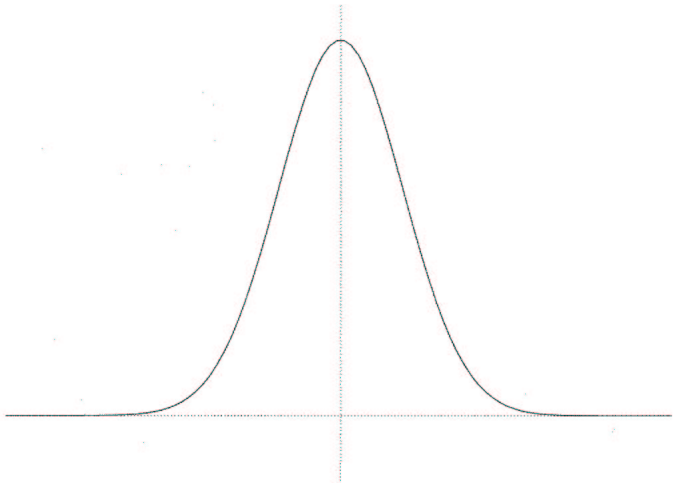
We can make little histograms for these values in Pascal's triangle for the first few values of n. Here they are for $n = 1$, 2, 3, 4, 10 and 20:





If we continue to make histograms like this, and arrange our graphs so that the center falls in the middle of the horizontal axis, and we adjust the "ticks " on the horizontal axis so that

1 inch always corresponds (more or less) to $\sqrt{N}$ (i.e. we adjust the scale of the graph as we change N), then these graphs start to look very similar as N grows; they look more and more like the "normal" curve shown below.



This is also borne out by mathematical analysis. If we look at the entries in the Nth row in Pascal's triangle, and divide them by 2^N, then we get the values of the probability that the number, x, of heads tossed equals k, which we denote by Prob $[x = k]$; for $|k - \frac{N}{2}| \leq \sqrt{N}$ this can be shown to be almost the same as $\frac{\sqrt{2}}{\sqrt{\pi N}} e^{\frac{-2(k-\frac{N}{2})^2}{N}}$. (The larger N is, the better this approximation is.)

This is the formula for a normal distribution! The general formula for such a normal distribution is given by $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$, where m stands for the **mean** and σ stands for the **standard deviation**.

In summary: repeating an experiment N times (here coin tossing), and then looking at the distribution for the outcome $S_N = x_1 + x_2 + \cdots + x_N$ (total number of heads) leads to a normal distribution for "large" N. (Mathematically speaking, we are never "quite" there, but the statement gets more accurate as N becomes larger and larger, "N going to infinity." Practically speaking, the approximation is already pretty good for pretty reasonable N.)

Normal distributions are completely characterized by their center and their width.

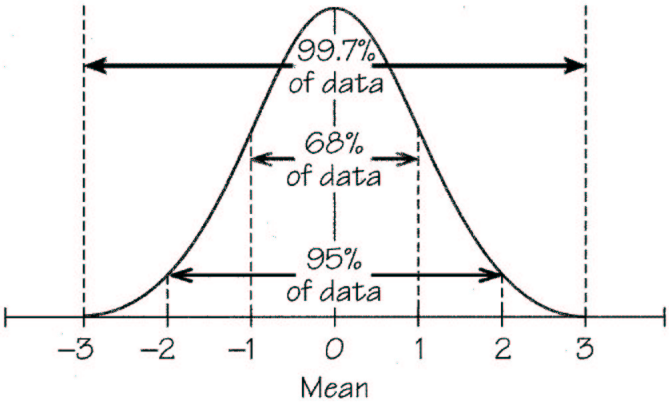
$\rightarrow$ **standard deviation** and **mean** characterize the distribution.

If the distribution of data is close to normal, then the corresponding m and σ are given by

$$m = \frac{(x_1+x_2+\cdots+x_N)}{N} \quad , \qquad \sigma = \sqrt{\frac{(x_1-m)^2+(x_2-m)^2+\cdots+(x_N-m)^2}{N}} \quad .$$

These formulas are again true "in the limit as N goes to infinity"; in practice this means that they are pretty good for large N.

The role of σ is explained in the following figure. For data that are distributed according to a normal curve, 68 % of the data lie within a distance σ of the center; 95 % lie within 2σ of the center, and 99.7 % within 3σ of the center.

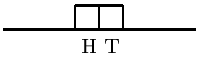


The 68–95–99.7 rule for normal distributions.

The formulas to compute m and σ are also used, sometimes, to get an idea of how much data are spread or concentrated, even if the data are not normally distributed, but then the meaning of σ is much less clear. In particular, the 68-95-99.7 % rule is no longer true then!

The Central Limit Theorem

Note that the probability distribution for the average number of heads approached a bell curve, even though the probability distribution for each of the individual x_j ($x_j = 0$ if toss j gives Heads, $x_j = 1$ if toss j gives Tails) was not normal at all! In our coin tossing case, we have:

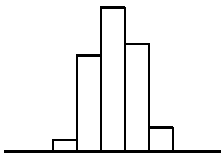


Probability distribution for one toss

This "tending to a normal distribution" for $x_1 + \cdots + x_N$, or for $\frac{x_1+\cdots+x_N}{N}$ is a general fact,

true for any reasonable probability distribution of the individual x_j. (It would be true, for instance, for a bent coin as well.)

Normal distributions are also observed a lot in practice. If you try to measure a quantity really carefully, say the mass of the electron, then you can't trust one single measurement as being accurate: your measurement will also contain some unavoidable error (because, however carefully you calibrate all your apparatus, it will never be "infinitely" precise). So then you repeat the measurement several times, and you plot the observed values. Typically, they cluster around some central value, and when you make a histogram around this central value, indicating how many of the observations fell within a central bin, or in the next bin, or even further, ... you get something that looks close to a normal distribution, such as:



How can we explain this? After all, we are not averaging anything here? The error itself, however, is a conglomerate of many different things (none of which we can control, otherwise we would get rid of it!—With technological advances, we can and do indeed reduce the error when we repeat those classical measurements). The total error can therefore be viewed as the sum total of all these different influences. On this sum the central limit theorem plays its role $\rightarrow$ normal distribution.

What about for other types of measurements? For instance, the height of young men between 25 and 30 in the US? Or the number of calories in people's diets? Or how well they see small printed characters?

Here again, in a population where there are no obvious inhomogeneities (an example of inhomogeneities that could spoil this assumption: ethnic origin of first-generation immigrants for the height question, or also for the calorific intake), one usually observes normal distributions around a central value, again because (it is believed that) the deviation from the average is caused by many different independent factors. (Do you really believe this? Isn't height, for instance, largely determined by genetics? How would you then still explain observing a normal distribution?)

There is one difference between these normal distributions and the one we saw at the start, for the averages of N coin tosses.

In the latter case, the distribution became more narrow and peaked as N increased. That

is because we were truly dealing with N repeats of identical experiments. Then

$E(x_1 + \cdots + x_N) = E(x_1) + E(x_2) + \cdots + E(x_N) = NE(x)$
$\uparrow$
"expected value"

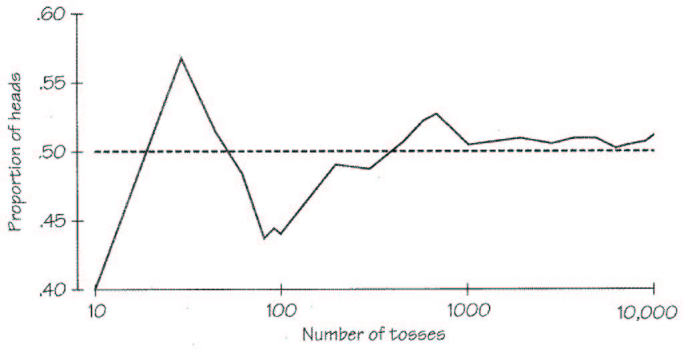
$\Rightarrow E\left(\frac{x_1+\cdots+x_N}{N}\right) = E(x)$

$V(x_1 + \cdots + x_N) = E((x_1 - \overline{x})^2) + \cdots + E((x_n - \overline{x})^2) = V(x_1) + \cdots + V(x_N) = NV(x)$
$\uparrow$
"variance"

$\Rightarrow V\left(\frac{x_1+\cdots+x_N}{N}\right) = \frac{1}{N^2}V(x_1 + \cdots + x_N) = \frac{1}{N}V(x)$

$\sigma\left(\frac{x_1+\cdots+x_N}{N}\right) = \left[V\left(\frac{x_1+\cdots+x_N}{N}\right)\right]^{\frac{1}{2}} = \frac{1}{\sqrt{N}}\sigma(x)$
$\uparrow$
"standard deviation"

The curve in the figure below, showing the number $\frac{x_1+\cdots+x_N}{N}$ as N increases from 1 to 10,000 for the 10,000 actual coin tosses done by John Kerrich in WWII, illustrates that the deviation from the average gets smaller as N increases.



Percent of heads versus number of tosses in Kerrich's coin-tossing experiment. (David Freedman et al., *Statistics*, Norton, New York, 1978.)

When you do a physics experiment, and you are as careful as you possibly can be, you will see this effect as well (to some extent), if you are measuring something like the mass of the electron, which has (we believe) a fixed, definite value. But when trying to ascertain the average height of young men between 25 and 30 in the US, we are finding the normal distribution that really truly exists in that population (although the fact that it is normal can be explained tentatively by the central limit theorem as well), and you cannot hope to spirit its spread away by doing sufficiently many experiments!

The following figure gives a whimsical summary of the importance of the normal curve.

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BRAODEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ⧫ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCINECES AND
IN MEDECINE AGRICULTURE AND ENGINEERING ⧫
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

(W.J. Youden)