Probability distributions

Introduction

What is a probability?

If I perform *n* experiments and a particular event occurs on *r* occasions, the "relative frequency" of this event is simply $\frac{r}{n}$. This is an experimental observation that gives us *an estimate* of the probability – there will be some random variation above and below the actual probability.

- If I do a large number of experiments, the relative frequency gets closer to the probability.
- We *define* the probability as meaning the limit of the relative frequency as the number of experiments tends to infinity. Usually we can *calculate* the probability from theoretical considerations (number of beads in a bag, number of faces of a dice, tree diagram, any other kind of *statistical model*).

What is a random variable?

A random variable (r.v.) is the value that might be obtained from some kind of experiment or measurement process in which there is some random uncertainty.

- A discrete r.v. takes a finite number of possible values with distinct steps between them.
- A continuous r.v. takes an infinite number of values which vary smoothly. We talk not of the probability of getting a particular value but of the probability that a value lies between certain limits.
- Random variables are given names which start with a capital letter.
- An r.v. is a numerical value (eg. a head is not a value but the number of heads in 10 throws <u>is</u> an r.v.)

Eg the score when throwing a dice (discrete); the air temperature at a random time and date (continuous); the age of a cat chosen at random.

The binomial distribution (Statistics book p298)

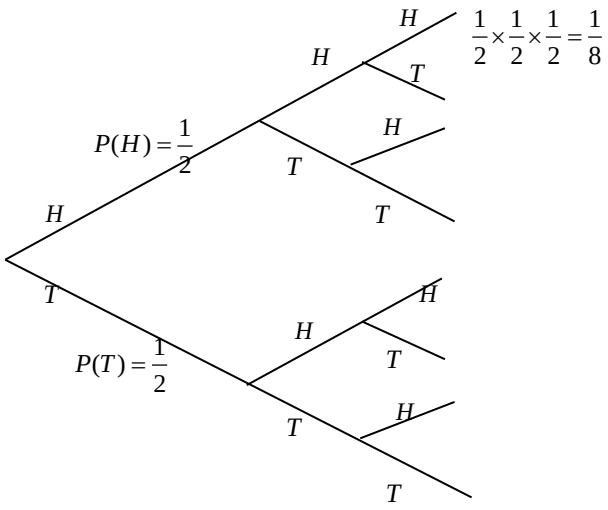
When we have a tree diagram, we can calculate the probability of each combination of outcomes (e.g. a head then a tail when throwing a coin twice).

Special case:

- If each fork has the same two choices
- The probabilities are the same at each fork

We can then use a formula to calculate the probabilities, without needing the tree diagram.

e.g. probability of getting a head and 2 tails in 3 throws of a coin:



There are three equivalent combinations of events (arrangements HTT, THT, TTH) so the probability is $P(\text{one head})=\frac{1}{8}\times 3=\frac{3}{8}$

In general, for any experiment in which the desired outcome can be obtained by a number of "equivalent routes" through the tree diagram that all have the same probability, we can define:

Probability = (probability for one route) ×(number of arrangements)

The binomial distribution describes the probability of getting *r* "successes" out of *n* trials in the kind of experiment with just two possible results ("success" or "failure") and in which successes are independent and have a constant probability *p* in all trials.

The formula for getting r "successes" out of n "trials" is

$$P(X=r)={}^{n}C_{r}\,p^{r}q^{n-r}$$

- p is the probability of each trial being a success,
- $q=1-p$ is the probability of each trial being a failure
- $p^{r}q^{n-r}$ is the probability of one path with *r* successes through a tree diagram

- "n choose r" ${}^{n}C_{r}$ is the number of possible paths through the tree. You can use the button on your calculator to find ${}^{n}C_{r}$.

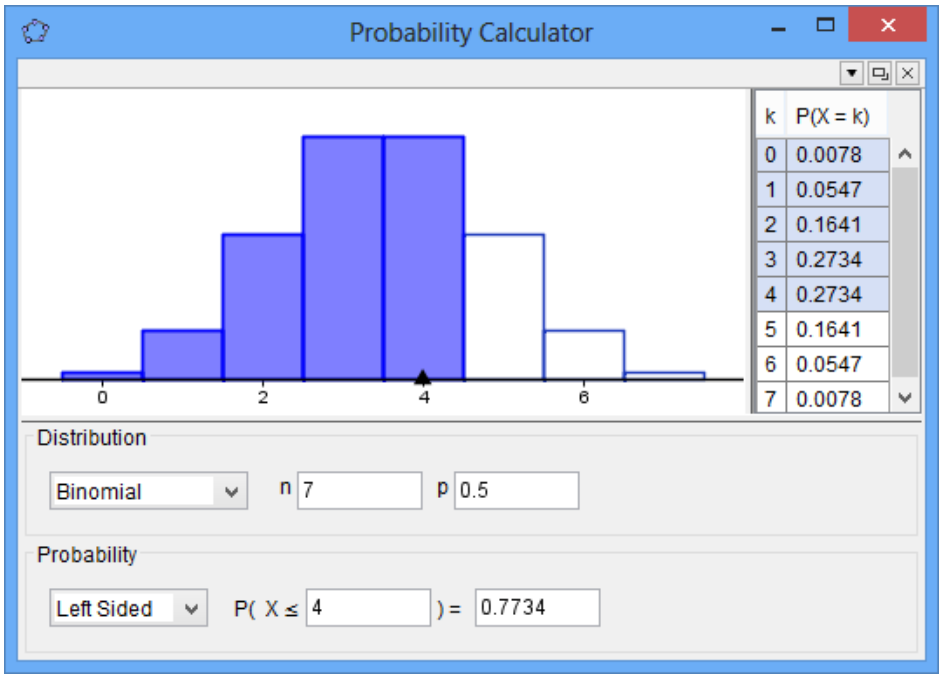
Example

A random variable "*X*" could be the number of heads in *n* throws of a coin.

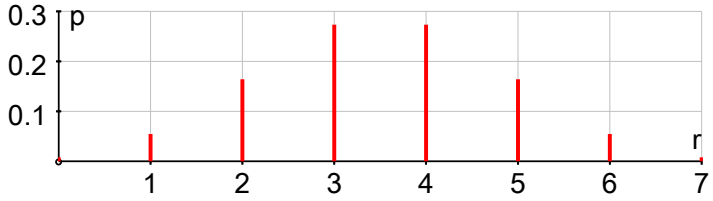
The probability of getting 3 heads in 7 throws of a coin is

$$P(X=3) = {}^{7}C_{3}0.5^{3}(1-0.5)^{7-3} = 0.2734$$

Geogebra calculates this nicely:



Autograph draws this as a bar graph:



- but won't tabulate the probabilities like Geogebra does.

Example.

The probability of a die landing on a "4" is $\frac{1}{6}$ so the probability of it not landing on 4 is

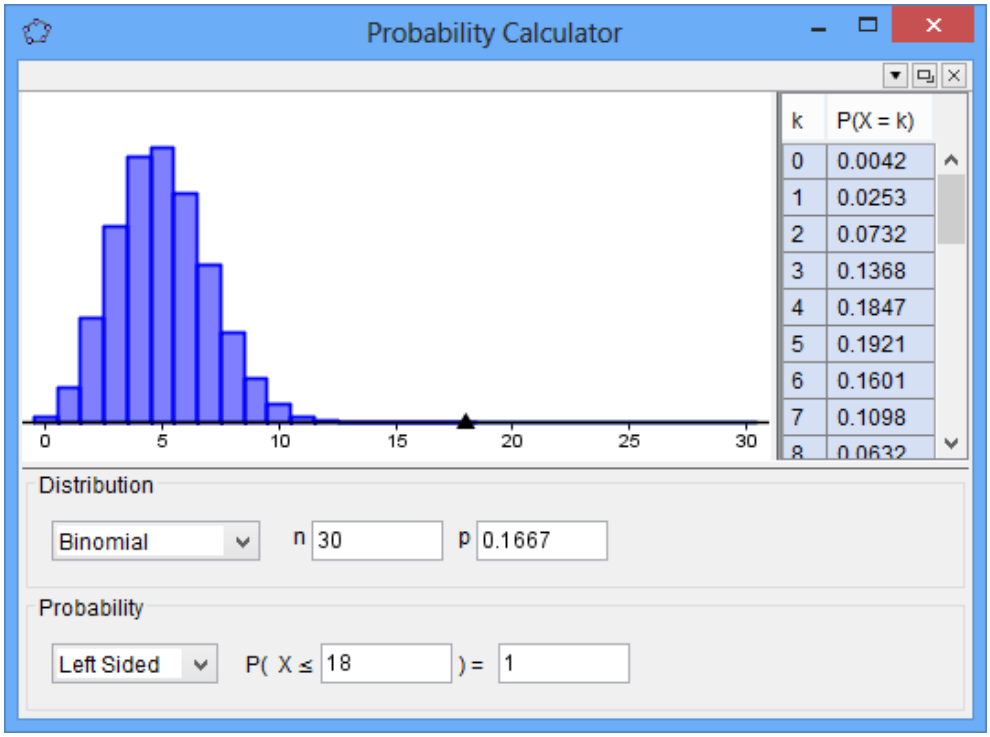
$$q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}.$$

The probability of getting 5 fours in 30 throws will be

$$P(X=5) = {}^{30}C_5 \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^{30-5} = 0.1921$$

(defining *X* as the number of sixes in 30 throws).

You can see this in Geogebra (menu Tools/Special Object Tools/Probability Calculator).



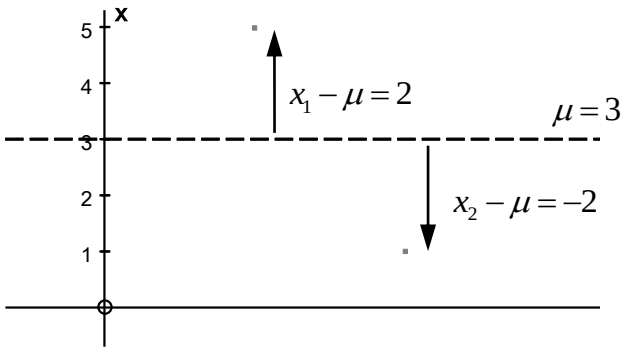
What is Standard Deviation? (Statistics book p145)

- By definition, in any list of numbers the mean of the difference between each value and the mean will be zero (some positive, some negative).
- To get a sensible measure of spread we square these differences (so all positive) and then average them to find the **variance**

This gives the basic definition $\text{variance} = \sigma^2 = \dfrac{\sum(x-\mu)^2}{n}$ where n is the number of data values, x is each value in the data and μ (pronounced "mu") is their mean.

- The formula can also be re-arranged as $\sigma^2 = \dfrac{\sum x^2}{n} - \mu^2$ (easier to use), often described as: **mean of (x^2) – (mean of x)^2**
- the variance (called σ^2) is the mean value of $(x-\mu)^2$
- the **standard deviation** σ is the square root of the variance (so its units are the same as the data, not data^2).
- Standard deviation is a measure of the "spread" of the data each side of the mean, just like range and inter-quartile range. In some ways it is better because it is (like the mean) based on every data value, not just the two at the 25% and 75% positions used for the IQR.

Simple example



Consider a data set of just two values, x = 5, 1

The mean $\mu = \dfrac{\sum x}{n} = \dfrac{5+1}{2} = 3$

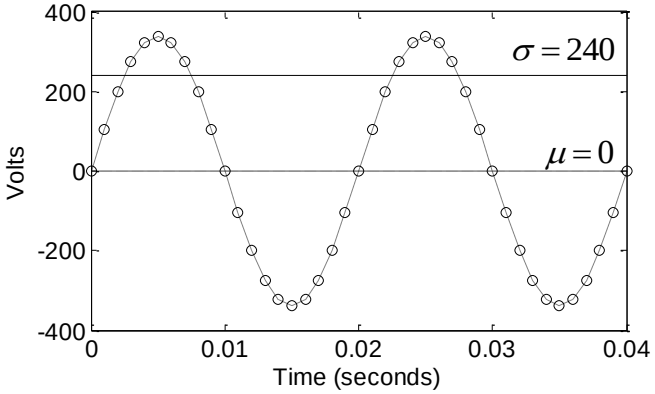
$\sigma^2 = \dfrac{\sum(x-\mu)^2}{n} = \dfrac{2^2+(-2)^2}{2} = \dfrac{8}{2} = 4$ (variance).

We now square root this to get the standard deviation $\sigma = \sqrt{4} = 2$

Everyday example

Electricity from power stations is *alternating current* (AC). The voltage is a sine wave that repeats 50 times per second. The voltage used in houses is 240 volts rms ("root mean square") meaning its standard deviation is 240 volts. The mean is 0 volts.

If we measured the mains voltage every $1/1000$ second and plotted the values it would look like this:



If the distribution of values follows a Normal distribution (below), we find for instance that a measurement of an item picked at random has a 95% chance of being within 1.96 standard deviations from the mean.

You can find standard deviation in many ways:

- Calculate it on paper or using a calculator
- In Excel, use a formula and specify a cell range e.g. =STDEV.P(A3:A42)
- In Autograph:

  (a) For 1-D data, having entered the data set, right-click on the data set name in the bottom bar and pick the "Show Statistics" menu; in the next pop-up window "Transfer to results box":

  Statistics for Raw Data 1

Number in sample, n:	40
Mean, x:	1100.83
Standard Deviation, x:	20.579
Range, x:	73
Lower Quartile:	1086.25
Median:	1096
Upper Quartile:	1124
Semi I.Q. Range:	18.875

  (b) For an x-y data set, with the data input window open click on "Show statistics".

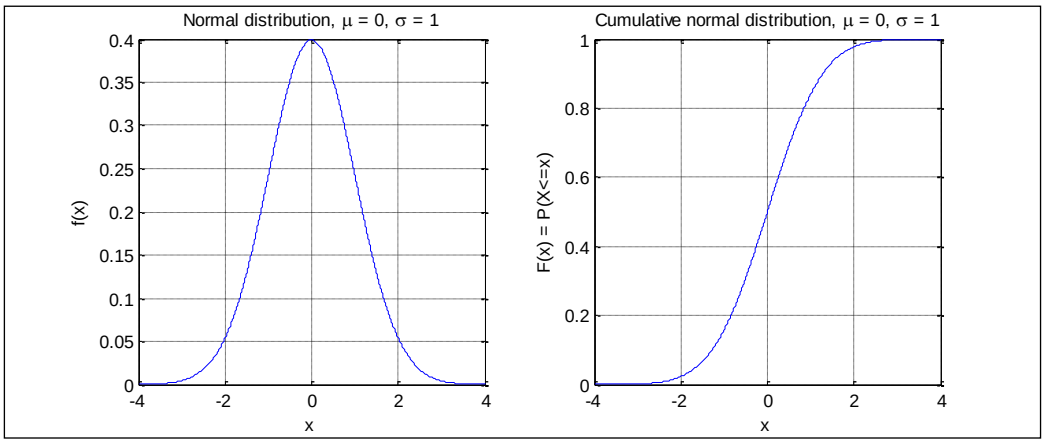
Number of points, n:	40
Mean, x:	1101
Mean, y:	3548
Standard Deviation, x:	20.58
Standard Deviation, y:	24.01
Correlation Coeff, r:	-0.8706
Spearman's Ranking Coeff:	-0.9411
y-on-x Regression Line:	y=-1.016x+4666
x-on-y Regression Line:	x=-0.7462y+3748

This gives you mean and standard deviation (but for 2D data, not the IQR) for both the x and y-data which you can paste into Word.

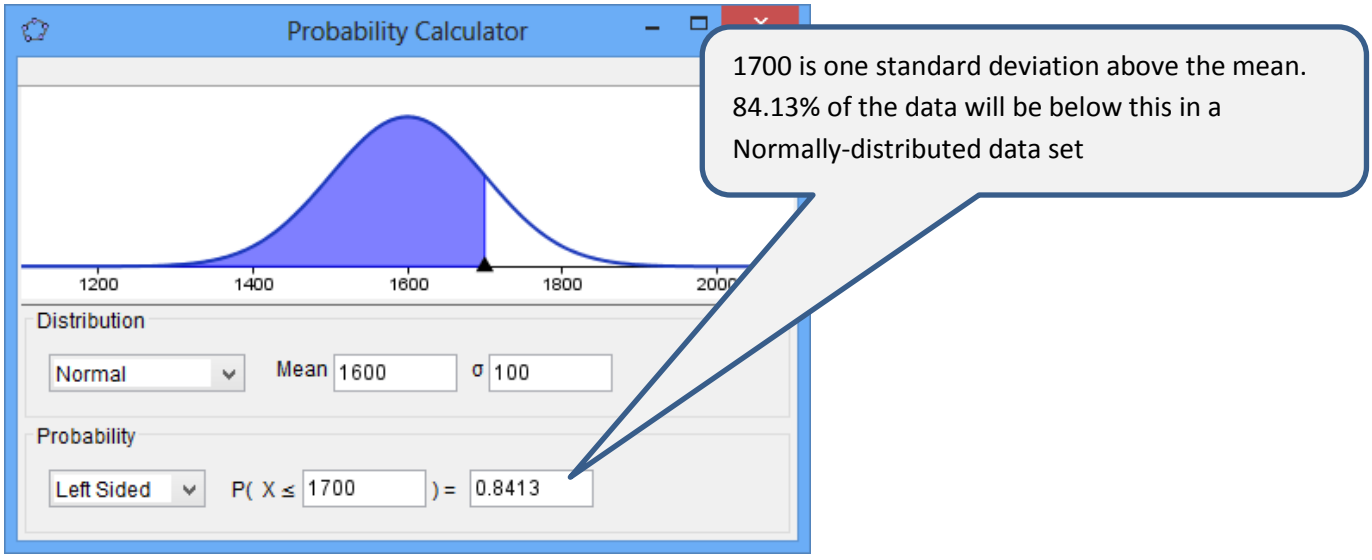
The Normal distribution (Statistics book p302)

The Normal distribution is a curve that often models the histogram shape for continuous variables. It is defined such that the area under the curve between any two x-values, gives you the probability that the random variable will lie in this range. The total area under the curve = 1.

- Normal distributions occur naturally in situations where the data value is the sum of mean of many parts, each having its own random variation.
- It is a symmetrical distribution with mean = median
- It models a continuous variable that can take values from –infinity to +infinity (but in practice, beyond 3 standard deviations from the mean the probability is so small that it can be ignored).

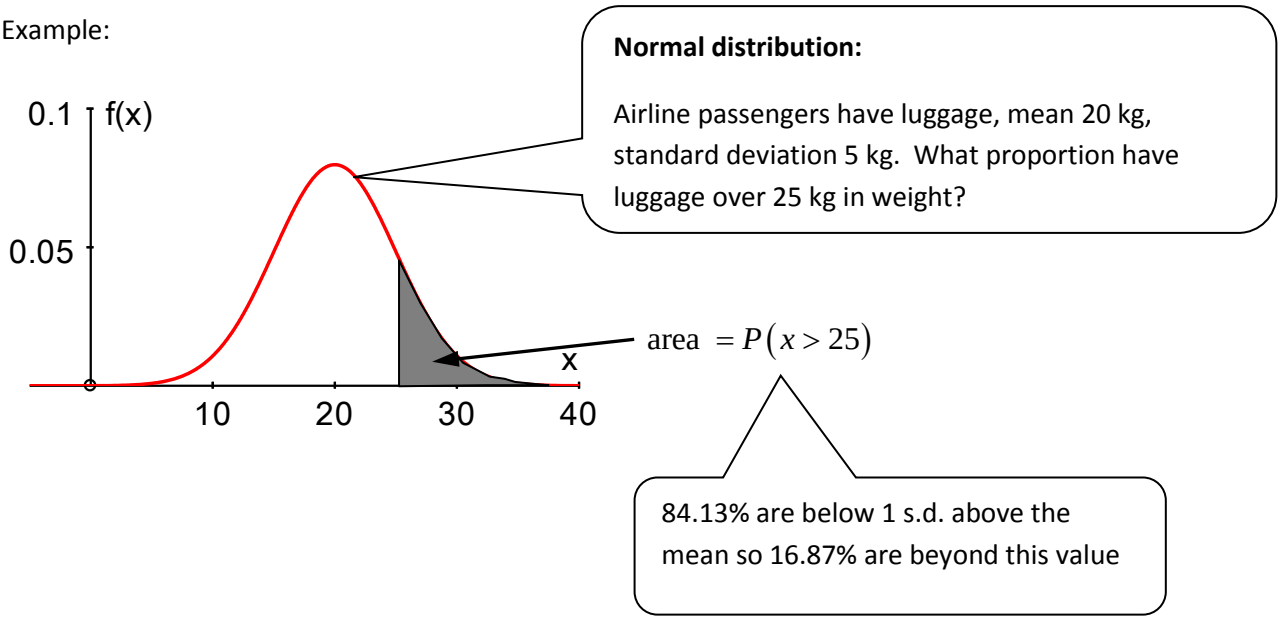


You can look up probabilities in Geogebra or via a formula in Excel.



(equivalent to Excel's formula =NORM.DIST(1700, 1600, 100, TRUE)).

Example:



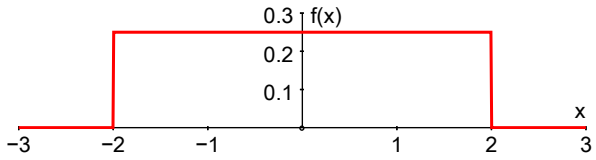
For data with a normal distribution, the:

- first quartile is 0.675 standard deviations below the mean
- third quartile is 0.675 standard deviations above the mean
- Hence the IQR = 1.35 standard deviations, standard deviation = 0.74×IQR
- 95% of the data is within 2 standard deviations (above and below) the mean
- 99.7% of the data is within 3 standard deviations (above and below) the mean

---

Not all symmetrical distributions are Normal!

A distribution could be more like a rectangle:



(IQR = 1.73×standard deviation, standard deviation = 0.58×IQR, all the data within 1.73 standard deviations of the mean)

Or more like a triangle:

(IQR = 1.43×standard deviation, standard deviation = 0.7×IQR, all the data within 2.45 standard deviations of the mean)

Both these lack the long “tails” of the Normal distribution so their s.d. is smaller, as a fraction of IQR.