The Normal (Gaussian) Distribution

A huge variety of phenomena will generate random numbers which can be modelled as being distributed by the **Normal** or ***Gaussian* Distribution**.

In other words, a *histogram* of a large quantity of measurements of random variable x, *normalized such that the area under the histogram is unity*, will tend towards the Normal Distribution if the number of samples is large enough.

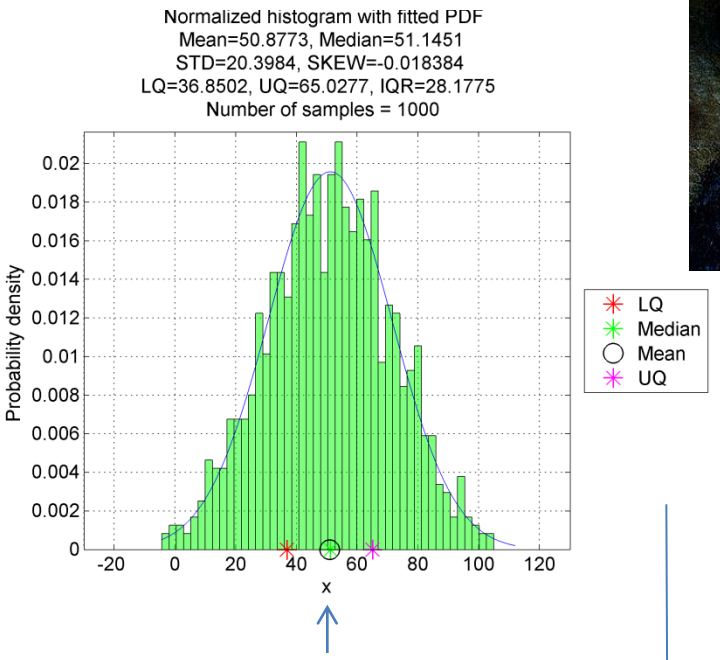
Indeed, the **Central Limit Theorem** states *that* ***"The distribution of the mean values of a set of independent random values tends towards a Gaussian Distribution if the number of samples is large enough."***
This is true even when the underlying probability distribution of the individual samples is unknown.

It also *means discrete distributions* such as *Binomial* and *Poisson* will tend towards a Normal Distribution when the number of independent trials becomes large. This allows large N Poisson and Binomial probabilities to be efficiently computed, as these distributions can be approximated by a Normal Distribution (with appropriate *continuity corrections*), using known values for the mean and variance appropriate to each distribution.

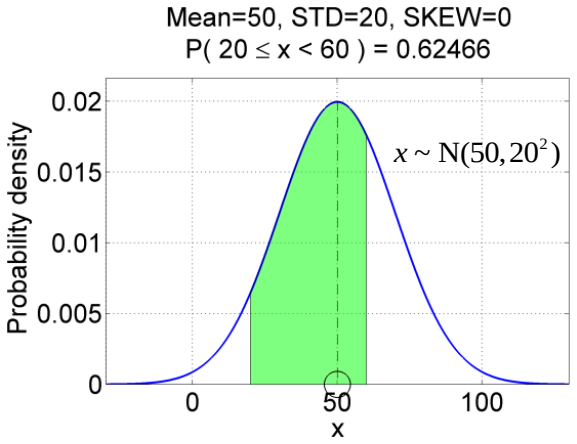
A **Normal Distribution** is a 'bell shaped' curve characterized by *two* parameters:

1. The **mean** μ, which describes the x value associated with the **peak** of the distribution
2. The **variance** σ^2 which characterizes the **width** of the bell curve. σ is called the **standard deviation**.

If random variable x is distributed by a Normal Distribution then we say: $x \sim \mathrm{N}(\mu, \sigma^2)$



In the above example, 1000 random numbers are generated from a Normal Distribution. The normalized histogram begins to take on the characteristic bell curve shape.



Johann Carl Friedrich Gauss
1777–1855

The Normal Distribution is often described as *Gaussian* in honour of Gauss, who developed much of the Mathematics associated with it.

If $x \sim \mathrm{N}(\mu, \sigma^2)$ the probability of a random variable having value between x and $x + dx$ is given by $p(x)dx$, where:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

For simplicity we can define a **standardized Gaussian** using the substitution

$$z = \frac{x-\mu}{\sigma}$$

$$\therefore p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Note one can show $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = 1$

The probability of z being *less* than some value Z is therefore

$$P(z < Z) = \Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}z^2} dz$$

Note:

$$z \sim N(0,1)$$
$$P(z < Z) = \Phi(Z)$$
$$P(z > Z) = 1 - \Phi(Z)$$

The function $\Phi(Z)$ is called the *Cumulative (Normal) Distribution Function.* This *cannot* be evaluated analytically, but can be numerically determined using a table of results, or via a calculator or computer with the ***error function*** encoded.

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

This can be approximately evaluated by integrating terms of the *Maclaurin Expansion* of $\exp(-t^2)$

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \left(x - \tfrac{1}{3}x^3 + \tfrac{1}{10}x^5 - \tfrac{1}{42}x^7 + \tfrac{1}{216}x^9 - .... \right)$$

$\Phi(Z)$ can be written in terms of the error function by firstly noting the symmetry about $Z = 0$

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}z^2}\, dz$$

$$\Phi(Z) = \tfrac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^{Z} e^{-\frac{1}{2}z^2}\, dz$$

$$\tfrac{1}{2}z^2 = t^2 \quad \therefore z = t\sqrt{2} \quad \therefore dz = dt\sqrt{2}$$

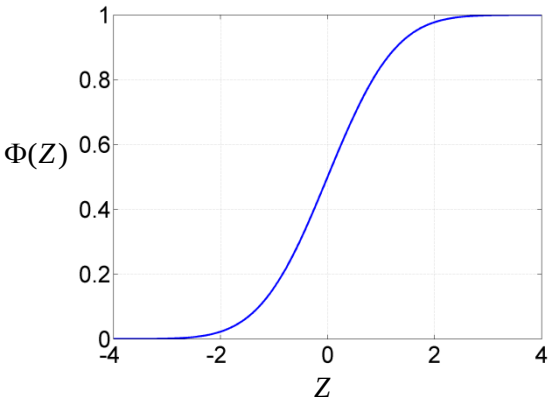
$$\therefore \frac{1}{\sqrt{2\pi}} \int_0^{Z} e^{-\frac{1}{2}z^2}\, dz = \frac{1}{\sqrt{\pi}} \int_0^{\frac{Z}{\sqrt{2}}} e^{-t^2}\, dt$$

$$\therefore \frac{1}{\sqrt{2\pi}} \int_0^{Z} e^{-\frac{1}{2}z^2}\, dz = \tfrac{1}{2} \operatorname{erf}\left(\tfrac{1}{\sqrt{2}} Z \right)$$

$$\therefore \Phi(Z) = \tfrac{1}{2} + \tfrac{1}{2} \operatorname{erf}\left(\tfrac{1}{\sqrt{2}} Z \right)$$

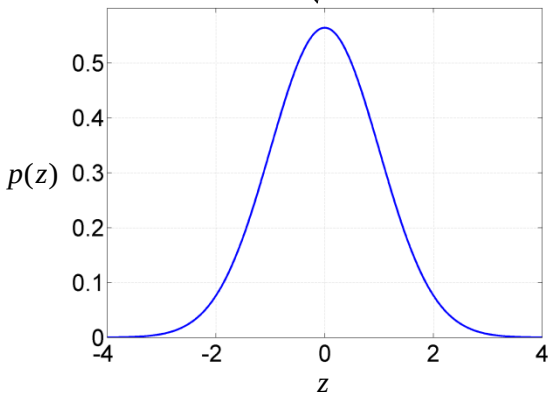
$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}z^2}\, dz$$

$$\Phi(Z) = \tfrac{1}{2} + \tfrac{1}{2} \operatorname{erf}\left(\tfrac{1}{\sqrt{2}} Z \right)$$



Cumulative Distribution Function
for the Standardized Normal Distribution

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$



Probability density function
for the Standardized Normal Distribution

$$z = \frac{x - \mu}{\sigma}$$

Example: The height x of a group of students is assumed to be normally distributed with mean 1.8m and standard deviation 0.2m.

What is the probability that a student is less than 1.7m tall?

$$x \sim \mathrm{N}(1.8, 0.2^2)$$

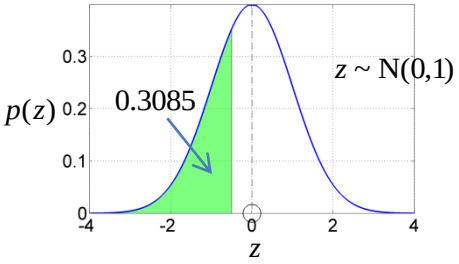
$$Z = \frac{1.7 - 1.8}{0.2} = -0.5$$

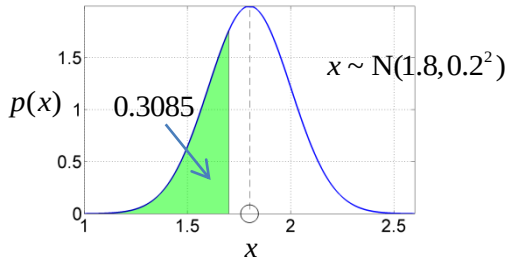
$$P(x < 1.7) = P(Z < -0.5)$$

$$P(Z < -0.5) = \Phi(-0.5)$$

$$\Phi(-0.5) = \tfrac{1}{2} + \tfrac{1}{2} \operatorname{erf}\left(\frac{-0.5}{\sqrt{2}} \right)$$

$$\therefore P(x < 1.7) = 0.3085$$





Tip: It is *always* a sensible idea to sketch the Standardized Normal Distribution function *p(z)* and sketch the area you want to calculate.

Example: The number of burpees x an athlete can perform in one minute is Normally distributed with mean 40 and standard deviation 10. What is the probability that an athlete does between 30 and 50 burpees in one minute?

$x \sim \mathrm{N}(40, 10^2)$

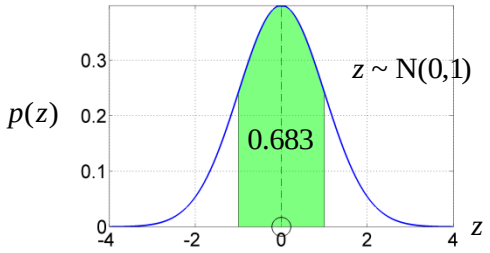
$$Z_+ = \frac{50-40}{10} = 1$$

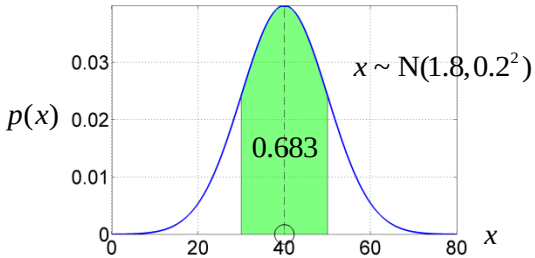
$$Z_- = \frac{30-40}{10} = -1$$

$$P(-30 \le x \le 50) = P(Z_- \le z \le Z_+) = P(-1 < Z < 1)$$

$$P(-1 < Z < 1) = \Phi(1) - \Phi(-1)$$

$$\therefore P(-30 \le x \le 50) = \boxed{0.683}$$





In many scenarios the *inverse problem* to those currently encountered might be posed. *Given a defined probability* for a normally distributed random variable x, what *range of* x does this correspond to?

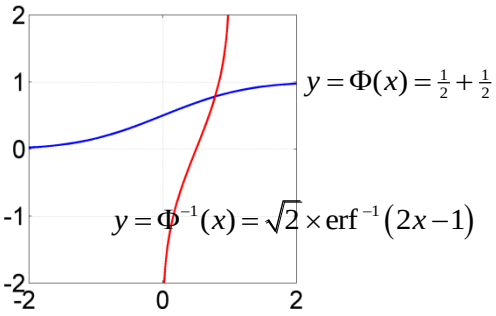
$$\Phi(x) = \tfrac{1}{2} + \tfrac{1}{2}\,\mathrm{erf}\left(\tfrac{1}{\sqrt{2}}x\right)$$

$$2\Phi - 1 = \mathrm{erf}\left(\tfrac{1}{\sqrt{2}}x\right)$$

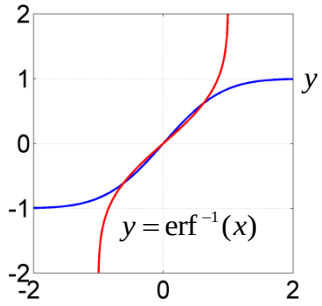
$$\mathrm{erf}^{-1}(2\Phi - 1) = \tfrac{1}{\sqrt{2}}x$$

$$x = \sqrt{2} \times \mathrm{erf}^{-1}(2\Phi - 1)$$

$$\boxed{\Phi^{-1}(x) = \sqrt{2} \times \mathrm{erf}^{-1}(2x-1)}$$



Note both $\Phi(x)$ & $\Phi^{-1}(x)$ are typically functions encoded in modern calculators or computer software such as MATLAB. If you don't have access to these, printed tables are regularly available.

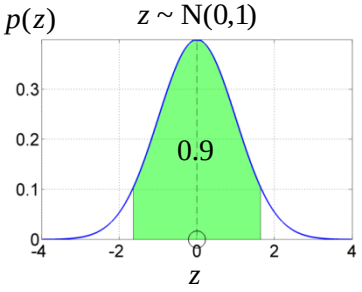


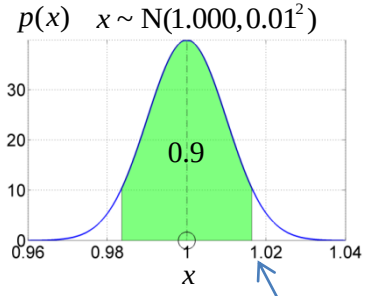
Example: ACME Corporation makes widgets of mass x which are Normally Distributed with mean 1.0000 kg and standard deviation 0.0100 kg. How can we express x in the form

$x = (1.0000 \pm \Delta x)$ kg

to "90% confidence" ?

This is called a **confidence interval** problem





$$x \sim N(1.0000, 0.01^2)$$

$$Z = \frac{\Delta x}{0.01}$$

$$\Phi(Z) = 0.95 \quad \leftarrow \quad P(x > 1 + \Delta x) = 0.05,\ P(x < 1 - \Delta x) = 0.05$$

$$Z = \Phi^{-1}(0.95) = 1.64$$

$$\therefore \Delta x = 0.0164$$

$$\therefore \boxed{x = (1.0000 \pm 0.0164)\ \mathrm{kg}}$$

$$1 + \Delta x = 1.0164$$

i.e. we expect 90% of widgets made to be within this range

A polynomial expansion is used to evaluate the inverse error function

$$\mathrm{erf}^{-1}(x) = \tfrac{1}{2}\sqrt{\pi}\left(x + \frac{\pi}{12}x^3 + \frac{7\pi^2}{480}x^5 + \frac{127\pi^3}{40,320}x^5 + \frac{4,369\pi^4}{5,806,080}x^9 + \frac{34,807\pi^5}{182,476,800}x^{11} + \ldots\right)$$

Approximating a Binomial Distribution by a Normal Distribution

$$P(x \mid N, p) = \frac{N!}{(N-x)!x!} p^x (1-p)^{N-x}$$

$$\mu = E[x] = Np$$

$$\sigma^2 = V[x] = Np(1-p)$$

If $N \gg 1$

$$x \sim \mathrm{B}(N, p) \approx \mathrm{N}(Np, Np(1-p))$$

Example:
One thousand people each roll a fair dice. What is the probability that up to 15% of them roll a six?

Let x be the number of people that roll a six.
Since $x \sim \mathrm{B}(1000, 1/6)$

$$P(x \le 150) = \sum_{n=0}^{150} \binom{1000}{n} \left(\tfrac{1}{6}\right)^n \left(\tfrac{5}{6}\right)^{1000-n}$$

This is going to be hard to compute. Since N is large we can *approximate* our solution by using the Normal Distribution

$$B(1000, \tfrac{1}{6}) \approx N\left(\tfrac{1000}{6}, \tfrac{1000}{6} \times \tfrac{5}{6}\right)$$

$$Z = \frac{150.5 - \frac{1000}{6}}{\sqrt{\frac{1000}{6} \times \frac{5}{6}}} = -1.3718...$$

$$P(x \le 150) \approx \Phi(Z) = 0.0850$$

Continuity correction applied – since the Normal Distribution is *continuous* whereas the Binomial distribution is *discrete*

We would expect a much higher value for the probability that up to 17% throw a six, since 1/6 = 0.166666...

$$Z = \frac{170.5 - \frac{1000}{6}}{\sqrt{\frac{1000}{6} \times \frac{5}{6}}} = 0.3253...$$

$$P(x \le 150) \approx \Phi(Z) = 0.6275$$

Approximating a Poisson Distribution by a Normal Distribution

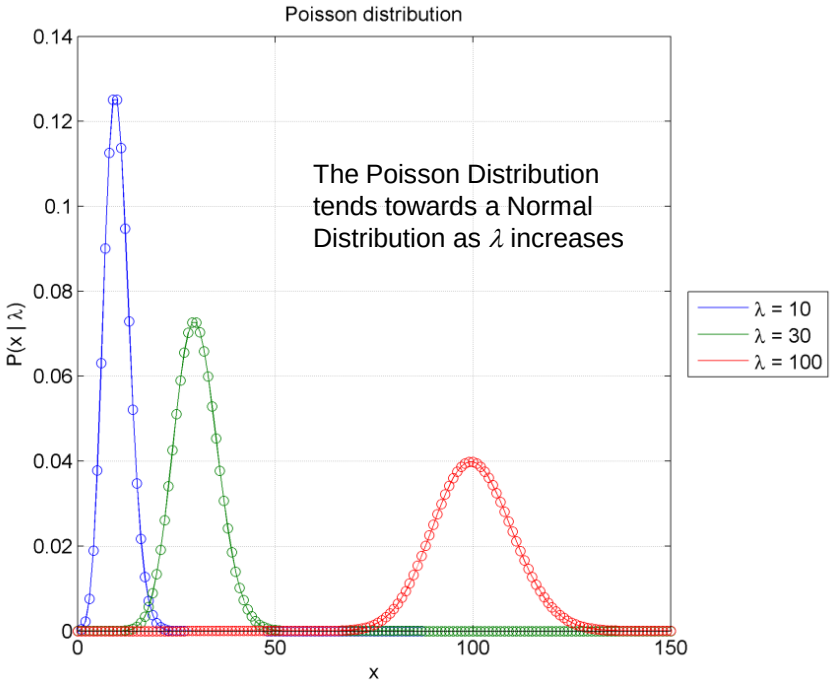
$$P(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\mu = E[x] = \lambda$$

$$\sigma^2 = V[x] = \lambda$$

If $\lambda \gg 1$

$$x \sim \mathrm{Po}(\lambda) \approx \mathrm{N}(\lambda, \lambda)$$



Example:
The number of decays x per minute of a radioactive isotope is modelled using a Poisson distribution with mean rate $\lambda = 100$.

Calculate the probability that:

(i) more than 110 decays per minute are observed
(ii) exactly 99 decays per minute are observed

$$x \sim \mathrm{Po}(100)$$

$$Z = \frac{110.5 - 100}{\sqrt{100}} = 15.827$$

$$\mathrm{P}(x > 110) \approx 1 - \Phi(Z) = 0.1469$$

$$Z_- = \frac{98.5 - 100}{\sqrt{100}} = -0.15$$

$$Z_+ = \frac{99.5 - 100}{\sqrt{100}} = -0.05$$

$$\mathrm{P}(x = 99) \approx \Phi(Z_+) - \Phi(Z_-) = 0.03968$$

Continuity correction applied – since the Normal Distribution is *continuous* whereas the Poisson Distribution is *discrete*