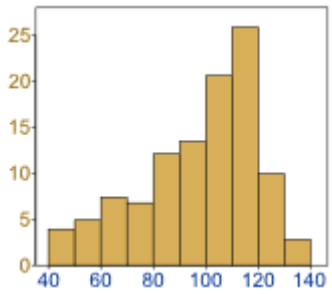
Normal Distribution

Def: The **normal distribution**, also known as the Gaussian distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean. Grouping takes place at values close to the mean and then tails off symmetrically away from the mean.

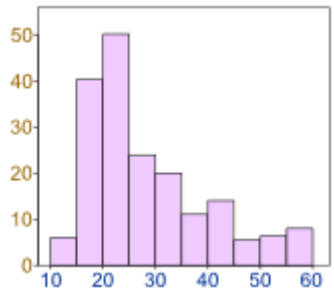
Need for normal Distribution:

Data can be "distributed" (spread out) in different ways.

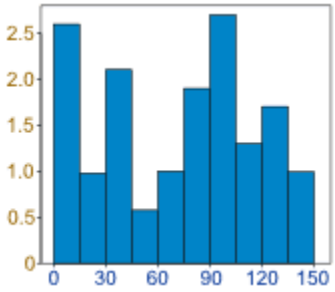
It can be spread out more on the left



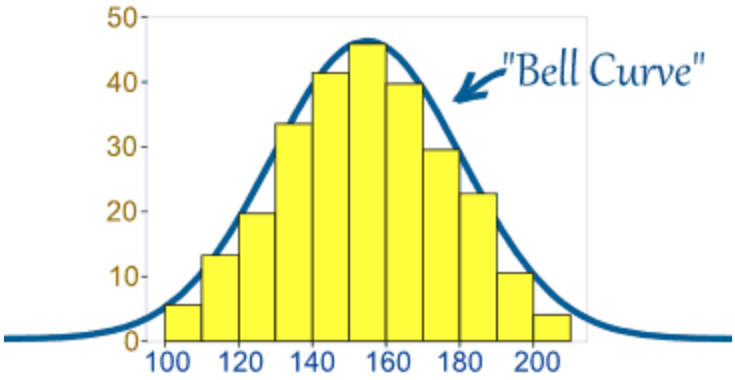
Or more on the right



Or it can be all jumbled up



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



A Normal Distribution

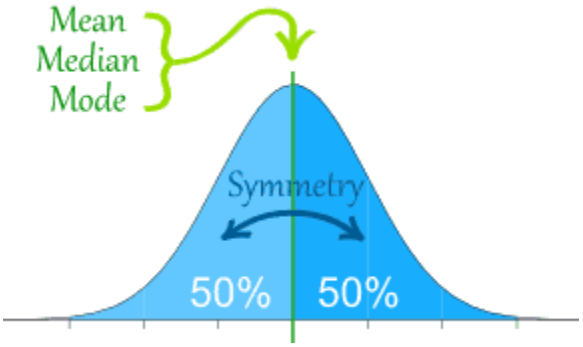
The "Bell Curve" is a Normal Distribution.
And the yellow histogram shows some data that
follows it closely, but not perfectly (which is usual).

It is often called a "Bell Curve"
because it looks like a bell.

Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the centre.
- 50% of values less than the mean
  and 50% greater than the mean

The "bell-shaped" curve of the Normal Distribution is called **Standard Normal Distribution** if it is a normal distribution with **mean 0 and standard deviation 1.**

Standard Deviation

The Standard Deviation is a measure of how spread out numbers are (or) **Standard Deviations** a measure of dispersion in a frequency distribution.

Its symbol is **σ** (the greek letter sigma)

The formula is easy: it is the **square root** of the **Variance.**

Variance

The Variance is defined as:

The average of the **squared** differences from the Mean.

To calculate the variance follow these steps:

- Work out the Mean(the simple average of the numbers)

- Then for each number: subtract the Mean and square the result (the *squared difference*).
- Then work out the average of those squared differences.

Example

You and your friends have just measured the heights of your dogs (in millimetres):

The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

Answer:

Mean = *600 + 470 + 170 + 430 + 300***5** = *1970***5** = 394

so the mean (average) height is 394 mm. Let's plot this on the chart:

Now we calculate each dog's difference from the Mean:

To calculate the Variance, take each difference, square it, and then average the result:

$$\begin{aligned}\text{Variance: } \sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\ &= \frac{108,520}{5} = 21,704\end{aligned}$$

So the Variance is **21,704**

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation

σ = √21,704

= 147.32...

= **147** *(to the nearest mm)*

So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

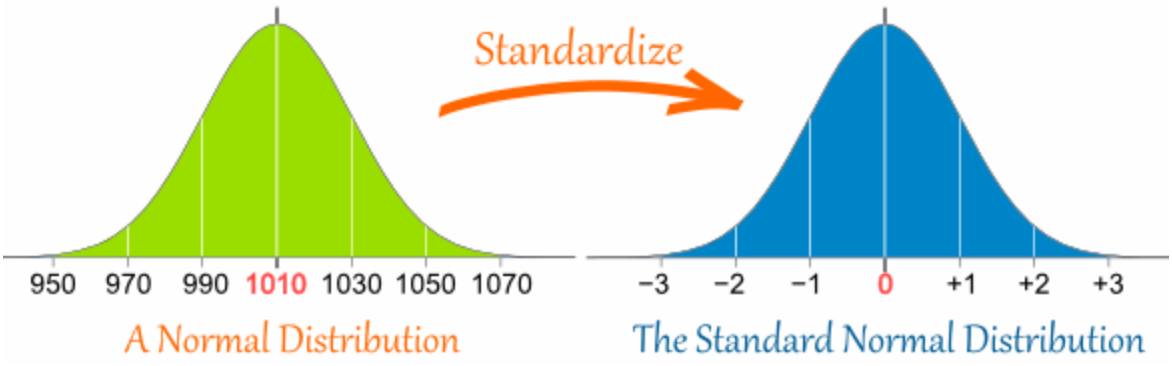
Standard Scores

The number of standard deviations from the mean is also called the "Standard Score", "sigma" or **"z-score".**

So to convert a value to a Standard Score ("z-score"):

- first subtract the mean,
- then divide by the Standard Deviation

And doing that is called "Standardizing":



We can take any Normal Distribution and convert it to The Standard Normal Distribution.

Here is the formula for z-score that we have been using: (Z table is attached in the last page)

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

- **z** is the "z-score" (Standard Score)
- **x** is the value to be standardized

- **μ** is the mean
- **σ** is the standard deviation

Problems on Normal Distribution

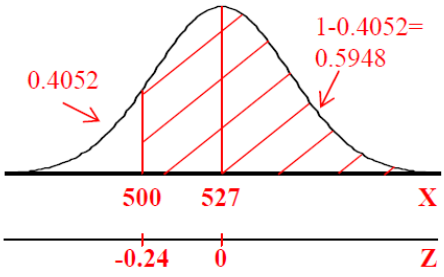
1. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?

   Normal Distribution $\quad Z = \dfrac{500 - 527}{112} = -0.24107$

   $\mu = 527$

   $\sigma = 112$

   $\Pr\{X > 500\} = \Pr\{Z > -0.24\} = 1 - 0.4052 =$ **0.5948**



Obtaining solution using R:

```
> pnorm(500, mean = 527, sd = 112, lower.tail = FALSE)
[1] 0.5952501
```

2. How high must an individual score on the GMAT in order to score in the highest 5%?

   Normal Distribution

   $\mu = 527$

   $\sigma = 112$

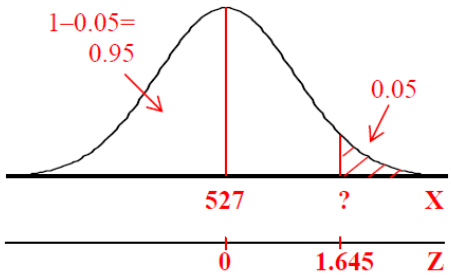
   $P(X > ?) = 0.05 \Rightarrow P(Z > ?) = 0.05$

   $P(Z < ?) = 1 - 0.05 = 0.95 \Rightarrow Z = 1.645$

   $X = 527 + 1.645(112)$

   $X = 527 + 184.24$

   $X =$ **711.24**



Obtaining solution using R:

```
> qnorm(0.05, mean = 527, sd = 112, lower.tail = FALSE)
```

[1] 711.2236

3. The length of human pregnancies from conception to birth approximates a normal distribution with a mean of 266 days and a standard deviation of 16 days. What proportion of all pregnancies will last between 240 and 270 days (roughly between 8 and 9 months)?

Normal Distribution $Z = \frac{240 - 266}{16} = -1.625$

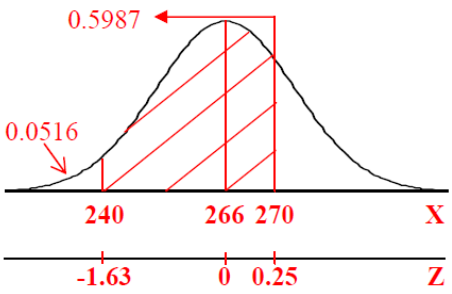
$\mu = 266$ $Z = \frac{270 - 266}{16} = 0.25$

$\sigma = 16$

$P(240 < X < 270) = P(-1.63 < Z < 0.25)$

$P(-1.63 < Z < 0.25) = P(Z < 0.25) - P(Z < -1.63)$

$P(-1.63 < Z < 0.25) = 0.5987 - 0.0516 =$ **0.5471**

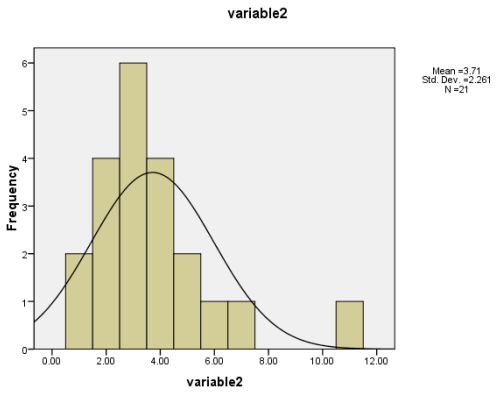


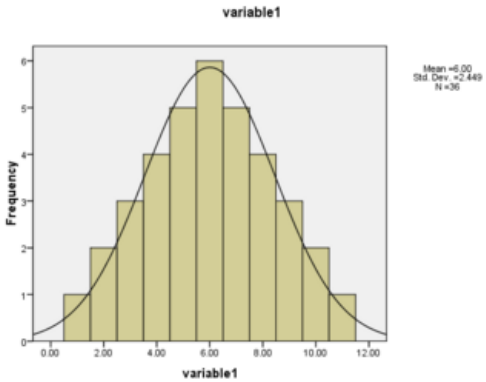
Obtaining solution using R:

```
> pnorm(270, mean = 266, sd = 16, lower.tail = TRUE)
[1] 0.5987063
> pnorm(240, mean = 266, sd = 16, lower.tail = TRUE)
[1] 0.05208128
> 0.5987063- 0.05208128
[1] 0.546625
```

How do I determine whether my data are normal?

1) **Use of Graphical re-presentations (Histograms) :** Look at a histogram with the normal curve superimposed. A histogram provides useful graphical representation of the data. - To provide a rough example of normality and non-normality, see the following histograms. The black line superimposed on the histograms represents the bell-shaped "normal" curve. Notice how the data for variable1 are normal, and the data for variable2 are non-normal. In this case, the non-normality is driven by the presence of an outlier.





2) **Skewness and Kurtosis:**

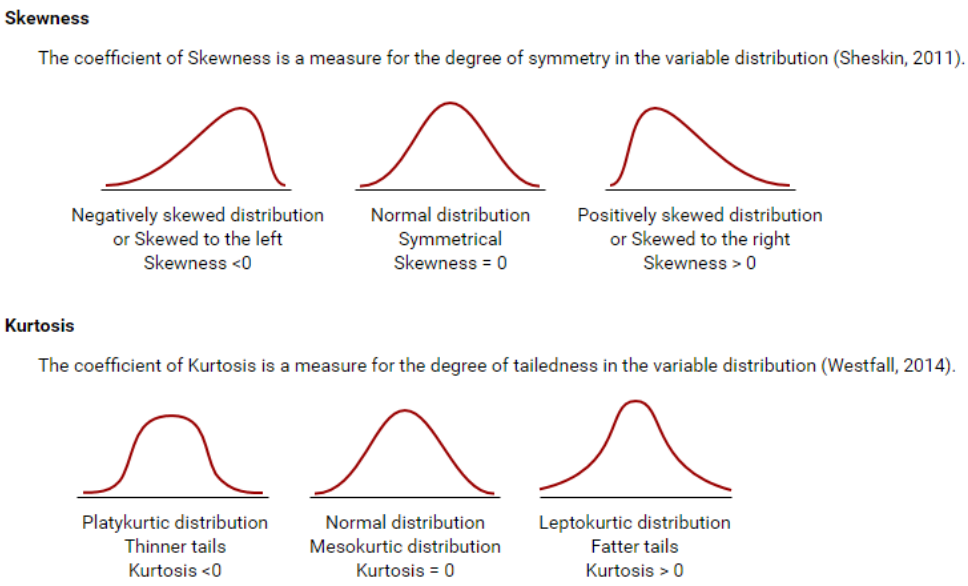
Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.A positively skewed distribution has scores clustered to the left, with the tail extending to the right. A negatively skewed distribution has scores clustered to the right, with the tail extending to the left. Skewness is 0 in a normal distribution, so the farther away from 0, the more non-normal the distribution.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Note: To perform kurtosis or skewness on data we have to install 'Moments' package in r.

Skewness=

Kurtosis=



3) **The Kolmogorov-Smirnov test (K-S) and Shapiro-Wilk (S-W) test** are designed to test normality by comparing your data to a normal distribution with the same mean and standard deviation of your sample. If the test is NOT significant, then the data are normal, so any value above .05 indicates normality. If the test is significant (less than .05), then the data are non-normal. See the data below which indicate variable1 is normal, and variable2 is non-normal. Also, keep in mind one limitation of the normality tests is that the larger the sample size, the more likely to get significant results. Thus, you may get significant results with only slight deviations from normality when sample sizes are large.

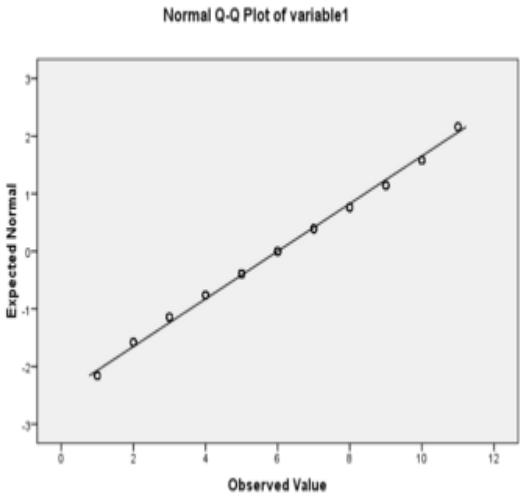
Tests of Normality

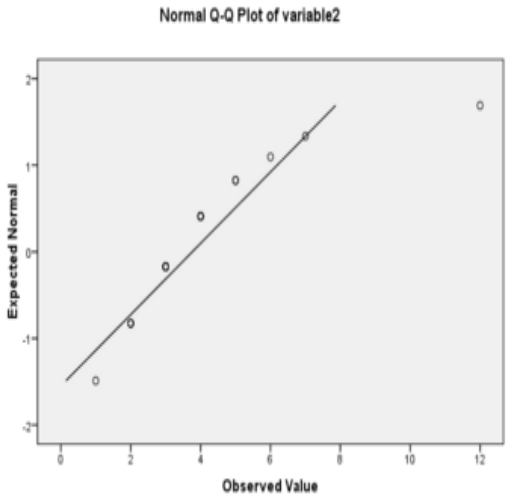
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
variable1	.083	36	.200*	.981	36	.782
variable2	.223	21	.008	.805	21	.001

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

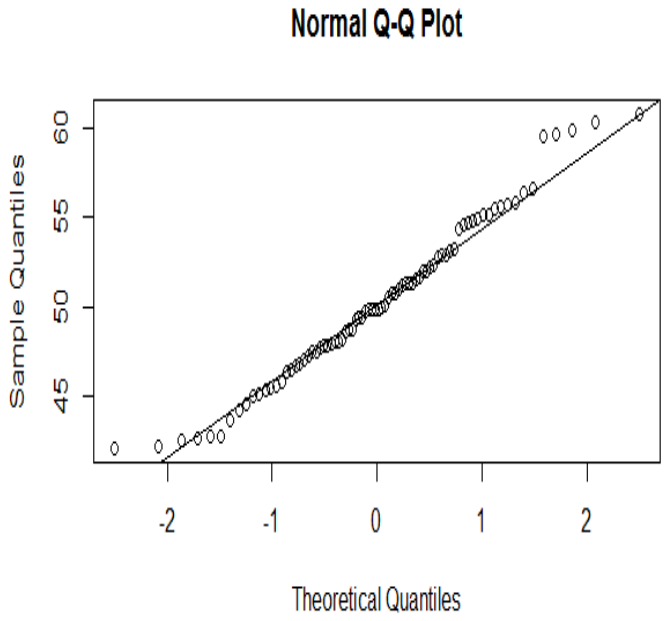
4) Look at normality plots of the data. "Normal **Q-Q Plot**" provides a graphical way to determine the level of normality. The black line indicates the values your sample should adhere to if the distribution was normal. The dots are your actual data. If the dots fall exactly on the black line, then your data are normal. If they deviate from the black line, your data are non-normal. - Notice how the data for variable1 fall along the line, whereas the data for variable2 deviate from the line.





QQ-plot examples to chech normality in R

```
> set.seed(183)
> x <- rnorm(80, mean=50, sd=5)
> qqnorm(x)
> qqline(x)
```



From the above qq-plot we can say that given contents are highly normally distributed because most of the points are located on the black line.

Normal distribution functions in R:

There are four functions that can be used to generate the values associated with the normal distribution. You can get a full list of them and their options using the help command:

```
> help(Normal)
```

	PURPOSE	SYNTAX	EXAMPLE
RNORM	Generates random numbers from normal distribution	rnorm(n, mean, sd)	rnorm(1000, 3, .25) Generates 1000 numbers from a normal with mean 3 and sd=.25
DNORM	Probability Density Function (PDF)	dnorm(x, mean, sd)	dnorm(0, 0, .5) Gives the density (height of the PDF) of the normal with mean=0 and sd=.5.
PNORM	Cumulative Distribution Function (CDF)	pnorm(q, mean, sd)	pnorm(1.96, 0, 1) Gives the area under the standard normal curve to the left of 1.96, i.e. ~0.975
QNORM	Quantile Function – inverse of pnorm	qnorm(p, mean, sd)	qnorm(0.975, 0, 1) Gives the value at which the CDF of the standard normal is .975, i.e. ~1.96

The first function we look at it is ***dnorm***. Given a set of values it returns the height of the probability distribution at each point. If you only give the points it assumes you want to use a mean of zero and standard deviation of one. There are options to use different values for the mean and standard deviation, though:

```
> dnorm(0)
[1] 0.3989423
> dnorm(0)*sqrt(2*pi)
[1] 1
> dnorm(0,mean=4)
[1] 0.0001338302
> dnorm(0,mean=4,sd=10)
[1] 0.03682701
>v <- c(0,1,2)
> dnorm(v)
[1] 0.39894228 0.24197072 0.05399097
> x <- seq(-20,20,by=.1)
> y <- dnorm(x)
```

```
> plot(x,y)
> y <- dnorm(x,mean=2.5,sd=0.1)
> plot(x,y)
```

The second function we examine is **pnorm**. Given a number or a list it computes the probability that a normally distributed random number will be less than that number. This function also goes by the rather ominous title of the "Cumulative Distribution Function." It accepts the same options as dnorm:

```
> pnorm(0)
[1] 0.5
> pnorm(1)
[1] 0.8413447
> pnorm(0,mean=2)
[1] 0.02275013
> pnorm(0,mean=2,sd=3)
[1] 0.2524925
> v <- c(0,1,2)
> pnorm(v)
[1] 0.5000000 0.8413447 0.9772499
> x <- seq(-20,20,by=.1)
> y <- pnorm(x)
> plot(x,y)
> y <- pnorm(x,mean=3,sd=4)
> plot(x,y)
```

If you wish to find the probability that a number is larger than the given number you can use the *lower.tail* option:

```
> pnorm(0,lower.tail=FALSE)
[1] 0.5
> pnorm(1,lower.tail=FALSE)
[1] 0.1586553
> pnorm(0,mean=2,lower.tail=FALSE)
[1] 0.9772499
```

The next function we look at is ***qnorm*** which is the inverse of pnorm. The idea behind *qnorm* is that you give it a probability, and it returns the number whose cumulative distribution matches the probability. For example, if you have a normally distributed random variable with mean zero and standard deviation one, then if you give the function a probability it returns the associated Z-score:

```
> qnorm(0.5)
[1] 0
> qnorm(0.5,mean=1)
[1] 1
> qnorm(0.5,mean=1,sd=2)
[1] 1
> qnorm(0.5,mean=2,sd=2)
[1] 2
> qnorm(0.5,mean=2,sd=4)
[1] 2
```

```
> qnorm(0.25,mean=2,sd=2)
[1] 0.6510205
> qnorm(0.333)
[1] -0.4316442
> qnorm(0.333,sd=3)
[1] -1.294933
> qnorm(0.75,mean=5,sd=2)
[1] 6.34898
> v = c(0.1,0.3,0.75)
> qnorm(v)
[1] -1.2815516 -0.5244005  0.6744898
> x <- seq(0,1,by=.05)
> y <- qnorm(x)
> plot(x,y)
> y <- qnorm(x,mean=3,sd=2)
> plot(x,y)
> y <- qnorm(x,mean=3,sd=0.1)
> plot(x,y)
```

The last function we examine is the ***rnorm*** function which can generate random numbers whose distribution is normal. The argument that you give it is the number of random numbers that you want, and it has optional arguments to specify the mean and standard deviation:

```
> rnorm(4)
[1]  1.2387271 -0.2323259 -1.2003081 -1.6718483
> rnorm(4,mean=3)
[1] 2.633080 3.617486 2.038861 2.601933
> rnorm(4,mean=3,sd=3)
[1] 4.580556 2.974903 4.756097 6.395894
> rnorm(4,mean=3,sd=3)
[1]  3.000852  3.714180 10.032021  3.295667
> y <- rnorm(200)
> hist(y)
> y <- rnorm(200,mean=-2)
> hist(y)
> y <- rnorm(200,mean=-2,sd=4)
> hist(y)
> qqnorm(y)
> qqline(y)
```