WEEK 5

PROBABILITY DISTRIBUTIONS:

CONTINOUS DISTRIBUTIONS

What would be the shape of the balls at the bottom of the line after releasing them?

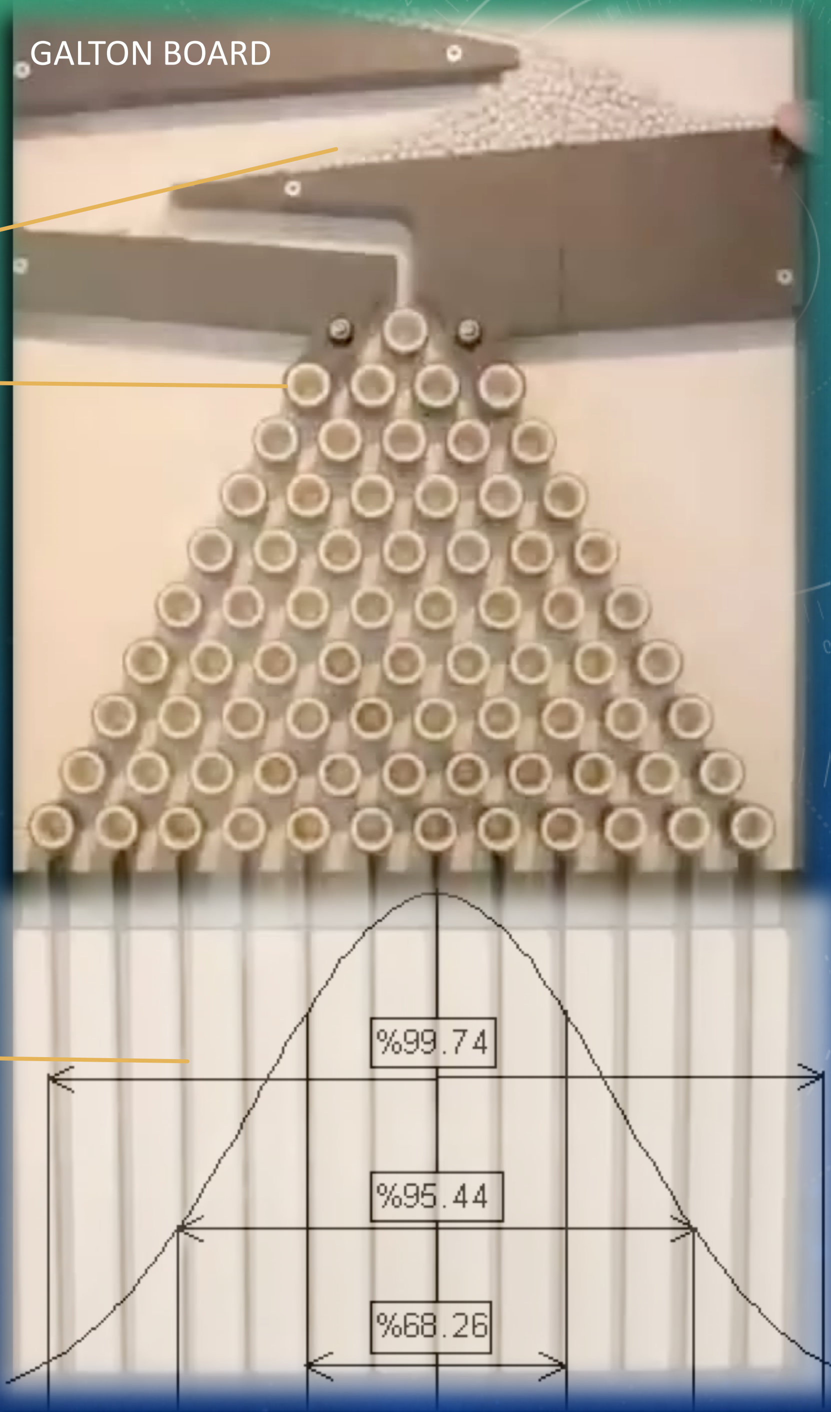
$$P(r) = \binom{n}{r} p^r q^{n-r}$$

- r is the bin position e.g. r=0 could be treated as the left-most bin, and r=n could be treated as the right-most bin.
- P is the probability of r
- p is the probability of bouncing right (if r=0 represents the left-most bin). (In an unbiased machine: p = 0.5.)
- N is the number of rows of pins i.e. the number of times a ball bounces.

Ref: http://www.statisticalconsultants.co.nz/blog/the-galton-box.html

Balls

Pins



Evenly spaced slots

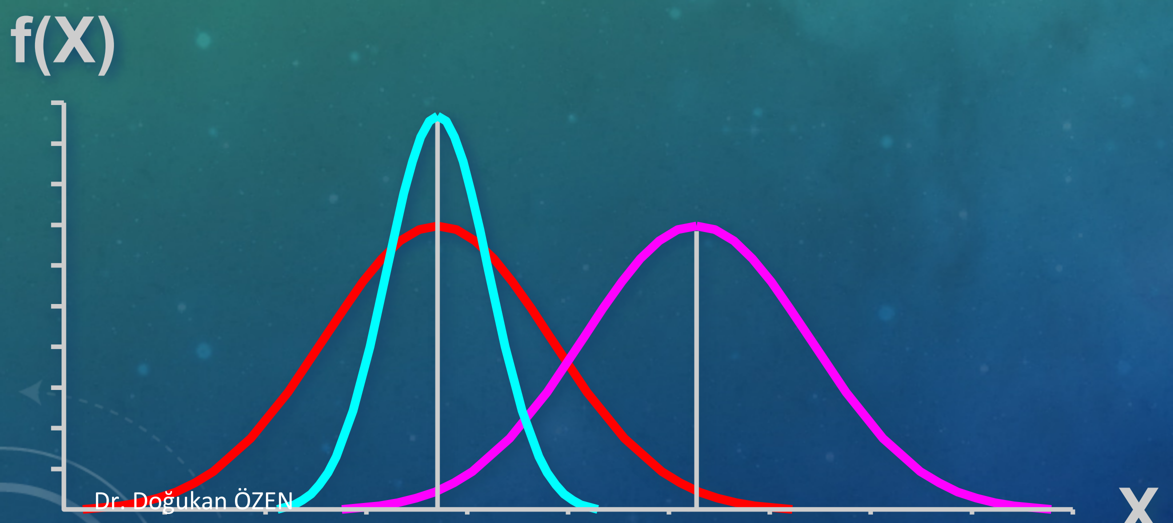
If the number of rows of pins is large enough, this would approximate a normal distribution due to the central limit theorem.

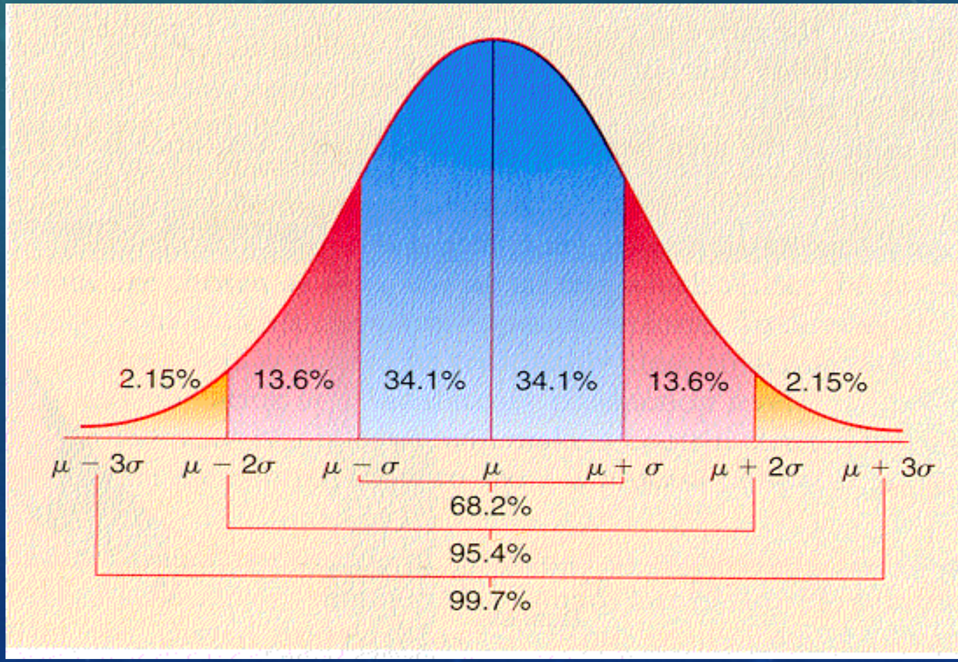
Central limit theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed

Ref: Galton, Sir Francis (1894). Natural Inheritance. Macmillan

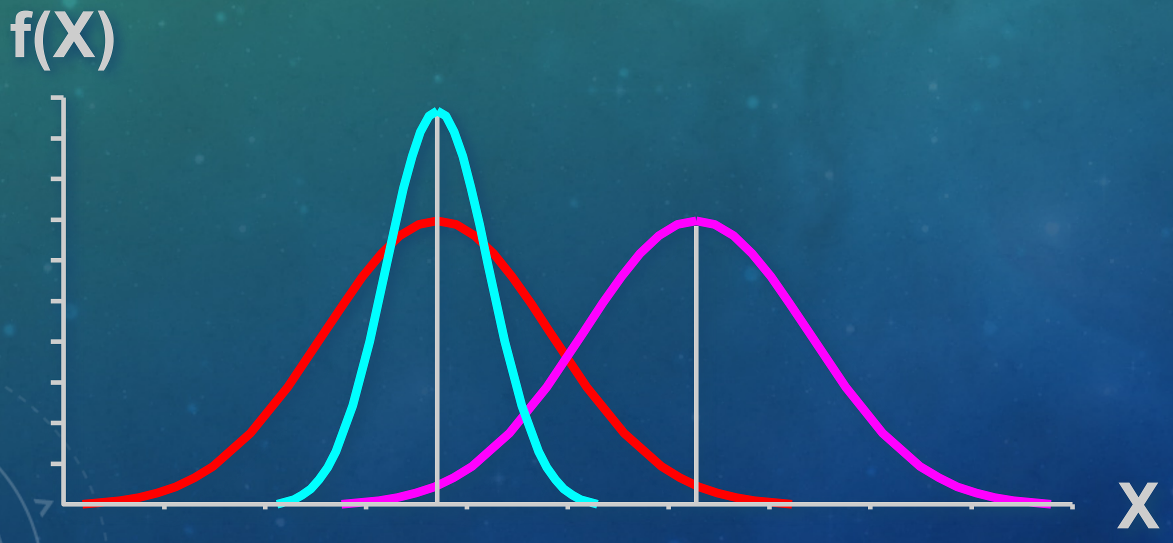
PROPERTIES OF NORMAL (OR GAUSSIAN) DISTRIBUTION

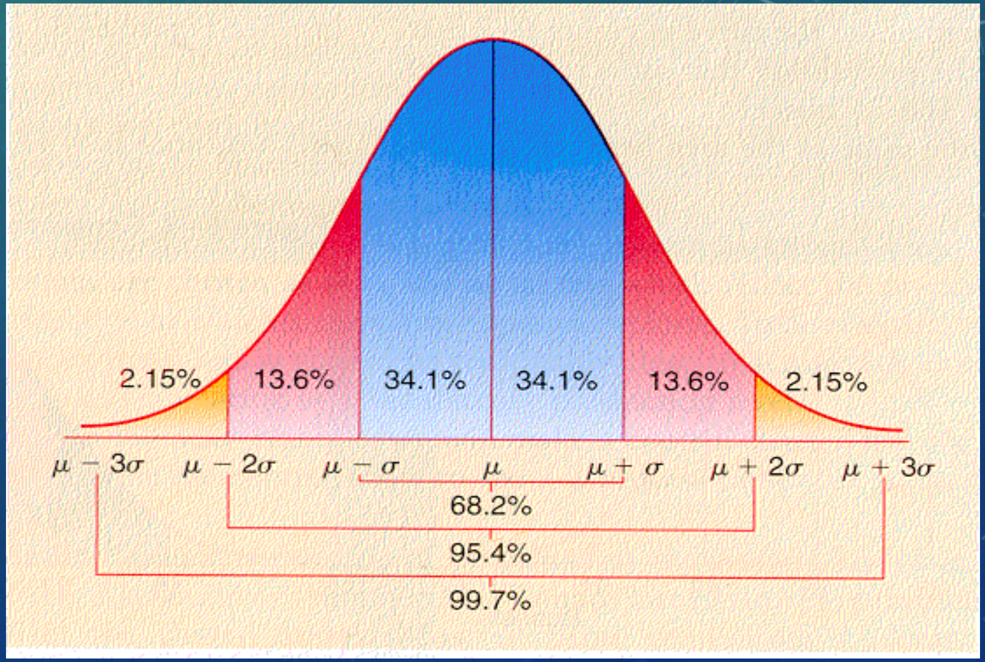
- The Normal distribution is completely described by two parameters: the mean and the standard deviation. These are usually denoted by the Greek letters μ and σ, respectively.
- It is unimodal
- It is symmetrical about its mean.
- Described as bell-shaped.
- Mean = median = mode





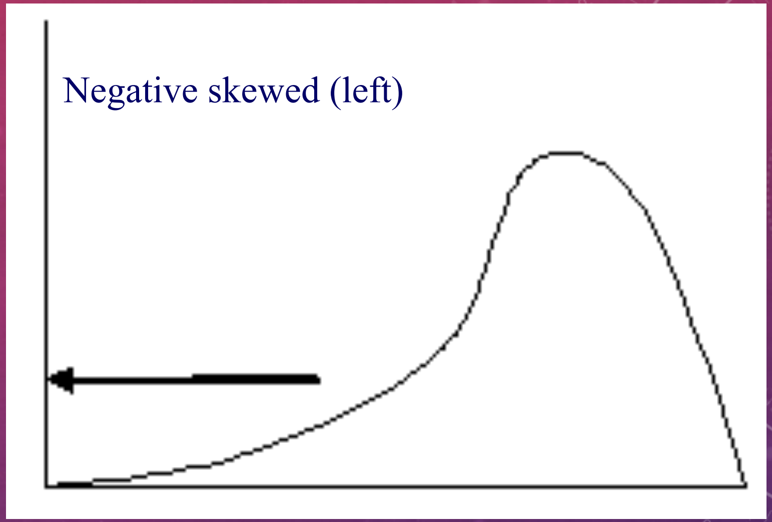
- If the standard deviation remains unchanged, increasing the value of the mean shifts the curve horizontally to the right. Conversely, decreasing the value of the mean shifts the curve horizontally to the left
- A decrease in the standard deviation of the curve makes the curve thinner, taller and more peaked. Con- conversely, an increase in the standard deviation makes the curve fatter, shorter and flatter
- The limits ($\mu - \sigma$) and ($\mu + \sigma$) contain 68.3% of the distribution
- The limits ($\mu - 2\sigma$) and ($\mu + 2\sigma$) contain 95% of the distribution
- The limits ($\mu - 3\sigma$) and ($\mu + 3\sigma$) contain 99% of the distribution



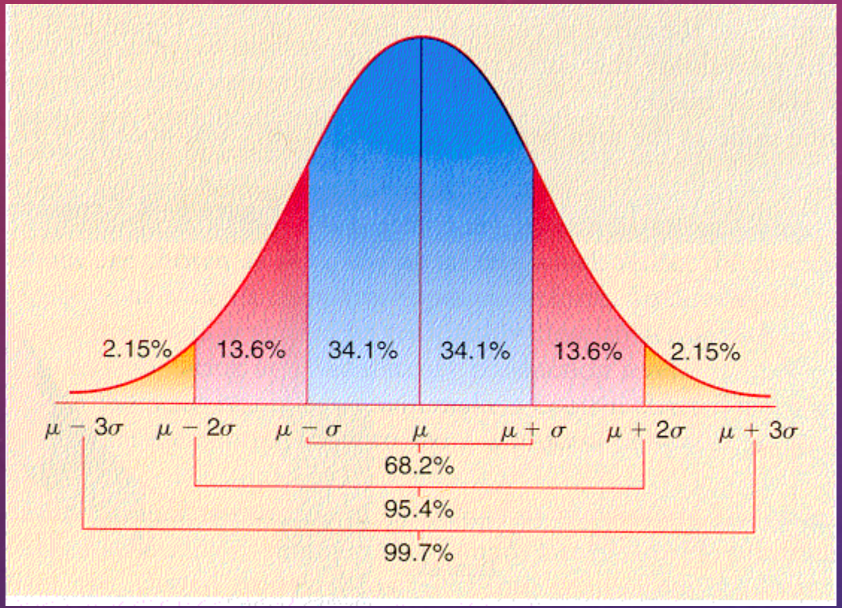


What would be the relationship between mean, median and mode when **the mass of the distribution is concentrated on the right or left side of the figure?**

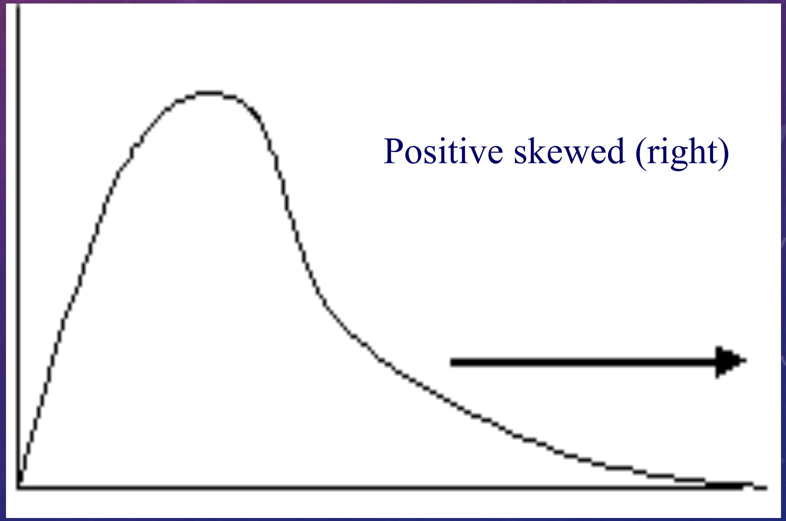
Hint: Remember what was told about the location of mean, median and mode (in week 3)



Arithmetic Mean < Median < Mode



Arithmetic Mean = Median = Mode



Arithmetic Mean > Median > Mode

WHAT TO DO WHEN YOUR DATA IS NOT NORMALLY DISTRIBUTED?

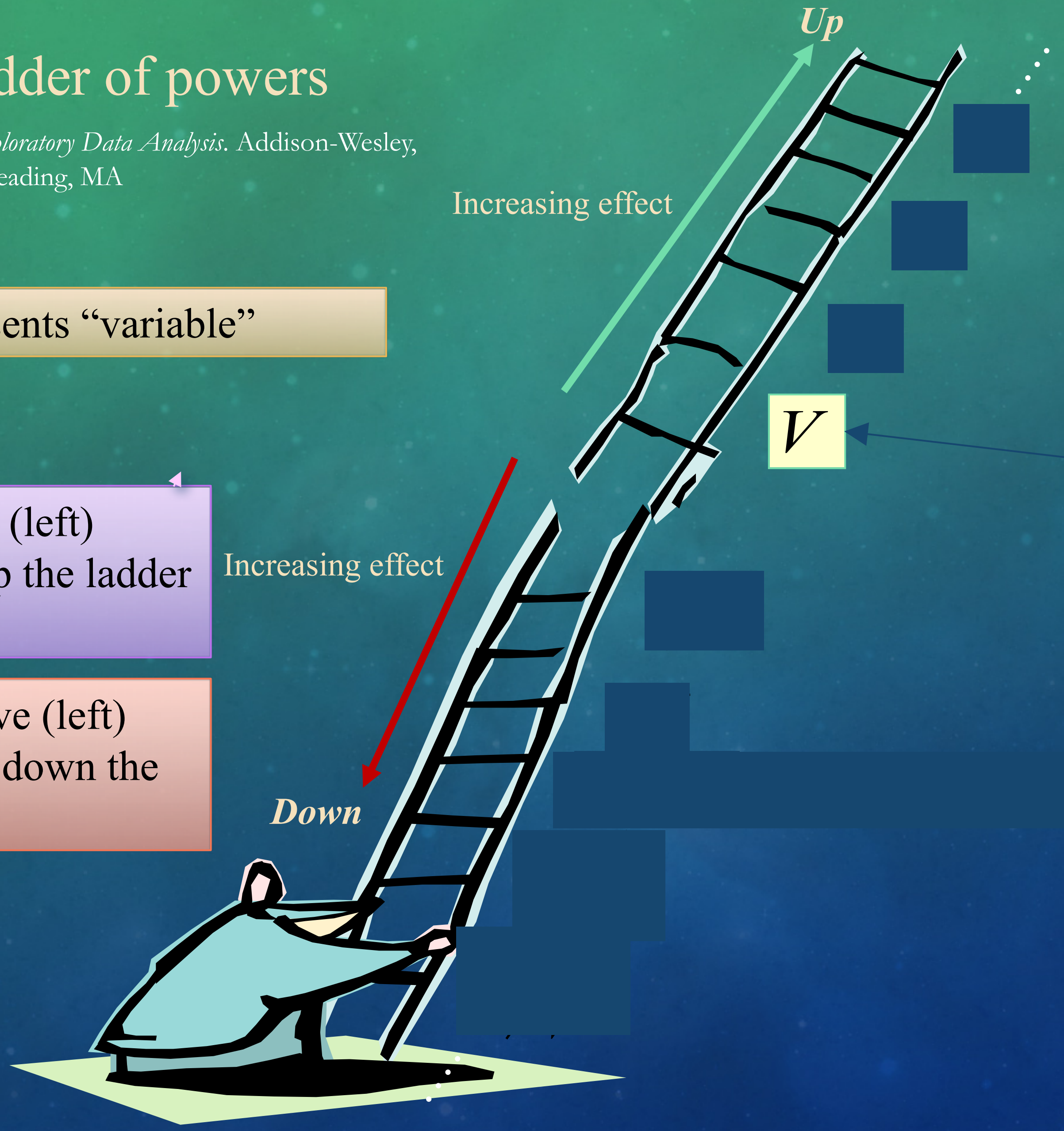
Tukey Ladder of powers

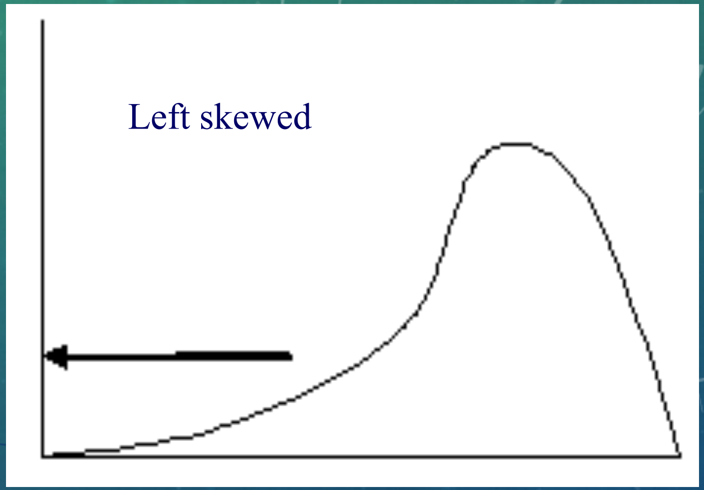
Ref: Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA

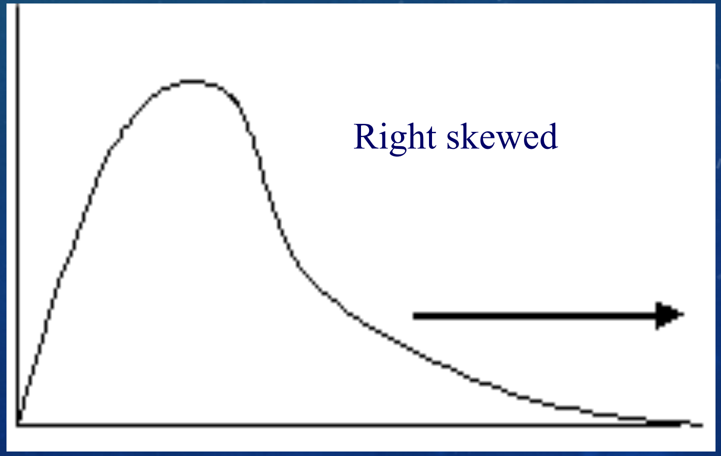
V, represents "variable"

To deal with negative (left) skewed data, climb up the ladder !

To deal with negative (left) skewed data, climb down the ladder !







STANDARD NORMAL DISTRIBUTION

- Every normal distribution have a mean (μ) and standard deviation (σ).

→ So, for each different mean and standard deviation we need to develop separate theoretical tables.

To overcome this problem, we can transform these distributions to a single standardized distribution and use only one theoretical table.

- To convert a normal distribution into a standard normal distribution with 0 mean and 1 standard deviation, we need a new variable, z, which is also named as standardized normal deviate.

Standard Normal Distribution

- In general terms, formula for "z" is given as;

$$z_i = \frac{x_i - \mu}{\sigma}$$

> The Standard Normal distribution is symmetrical around its mean of 0. Thus the tail area to the right of a value $z1$ is the same as the tail area to the left of $-z1$; equivalently, the probability that $z > z1$ is equal to the probability that $z < -z1$.

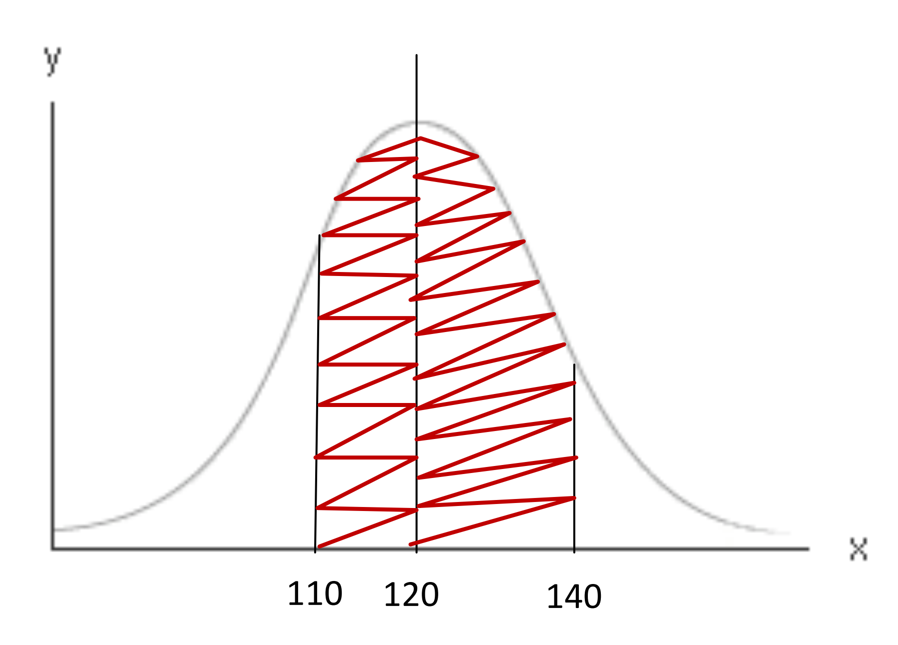
The values of standard normal variable z tell us by how many standard deviations the values of y deviate from the mean

> The values of z are sometimes called **critical values** or **percentage points**, as each defines a percentage of the total area under the probability density function.

AN EXAMPLE...

- Suppose that mean systolic blood pressure of a population is 120 mmHg with a standard deviation of 25 mmHg. What percentage of people with systolic blood pressure are between 110-140 mmHg?

$$z_i = \frac{x_i - \mu}{\sigma}$$



Step 1:

$$z_1 = \frac{110 - 120}{25} = -0.4$$

$$z_2 = \frac{140 - 120}{25} = 0.8$$

Step 2: find the corresponding "frequencies" for the related "z" values.

For z_1 = -0.4 ➡ 0.1554

For z_2 = 0.8 ➡ 0.2881

Step 3:

0.1554 + 0.2881

= **0.4435**

44.35% of people in the community have a systolic blood pressure of 110-140 mmHg.

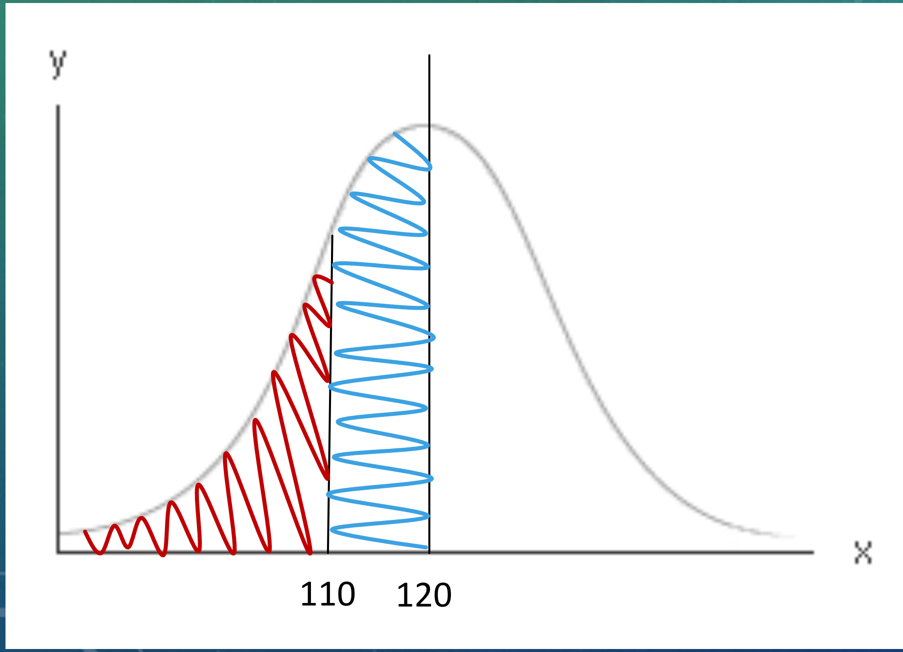
EXAMPLE 2:

SUPPOSE THAT MEAN SYSTOLIC BLOOD PRESSURE OF A POPULATION IS 120 MMHG WITH A STANDARD DEVIATION OF 25 MMHG. WHAT PERCENTAGE OF PEOPLE WITH SYSTOLIC BLOOD PRESSURE ARE < 110 MMHG?

$$z_i = \frac{x_i - \mu}{\sigma}$$

Step 1:

$$z_1 = \frac{110 - 120}{25} = -0.4$$



Step 2: find the corresponding "frequencies" for the related "z" values.

For z_1 = -0.4 ⇨ 0.1554 BLUE AREA

Step 3: Remember the properties of normal distribution.
Distribution is symmetrical around mean.

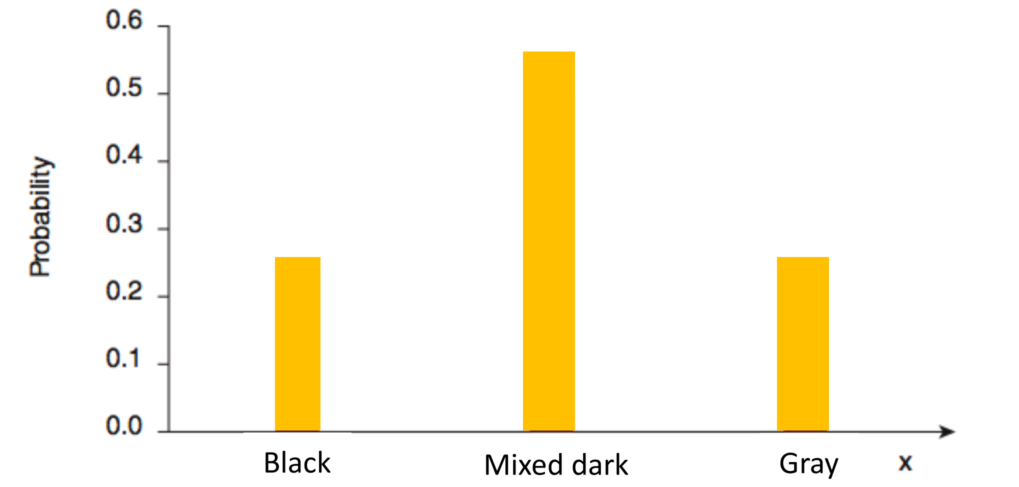
0.50 - 0.1554 = **0.3446 (RED AREA)**

34.46% of people in the community have a systolic blood pressure of < 110 mmHg.

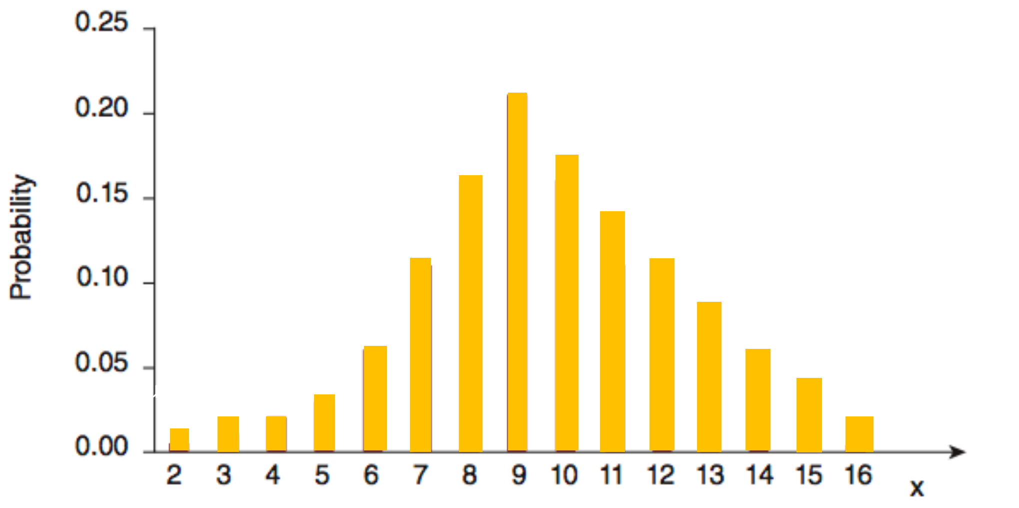
RELATIONSHIPS BETWEEN DISTRIBUTIONS

- The Binomial and Poisson distributions are skewed when sample sizes are small, although they become more symmetrical as sample sizes increase.
- Each distribution approaches Normality for large enough sample sizes when a smooth curve is drawn joining the discrete probability values.

i) Categorical random variable with 3 values



ii) Discrete random variable with 15 values



iii) Probability density function of a continuous random variable