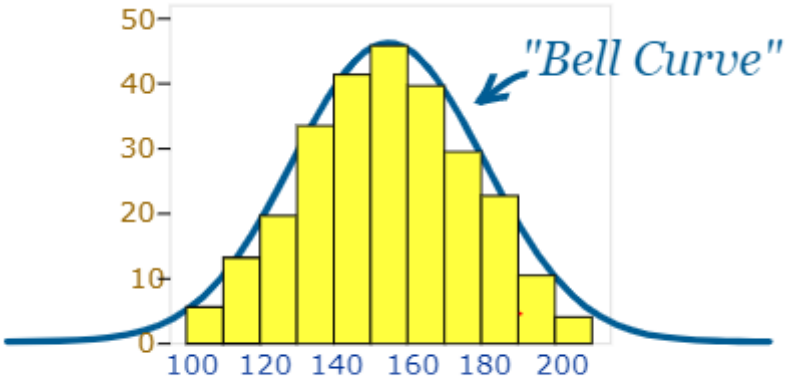
Mathematic Statistics

Normal Distribution / Gaussian Distribution

分布的意思是描述一组数，有多少数是大，多少数是小，这些大数和小数又在总体中的比例是多少。 Normal Distribution 的意思就是正常状态下的分布，It refers to the equation or graph which are bell-shaped。Normal Distribution 可以用来解释生活中很多现象，比如说一个国家成年男性身高分布，一个健康人一天的血压变滑。



A Normal Distribution

$$Normal\ Distribution = P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

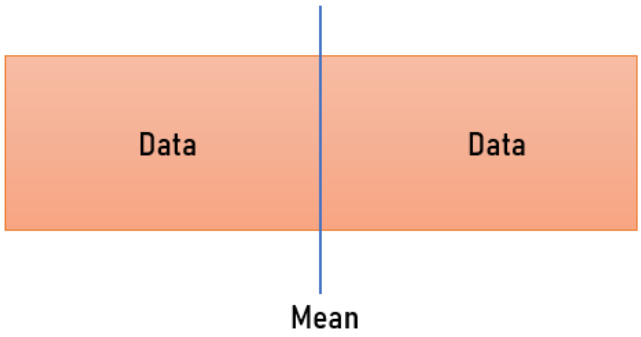
$$\mu \rightarrow Mean$$

$$\sigma \rightarrow Standard\ Distribution$$

$$x \rightarrow Normal\ Random\ Variable$$

Note – If mean μ is 0 and standard deviation σ is 1, then the distribution is known to be normal distribution.

What is Variance?



Variance 是用来算出每个 Data Point 跟 Mean 之间相差了多少。我想要知道这组数据集中和分散。A set of data with low variance – 这组数据都很集中，很靠近 Mean。A set of data with high variance – 这组数据很分散，偏离 Mean。

For each Data Point –

$$Deviation\ from\ mean = X_i - \bar{X}$$

$$Averaging\ All\ the\ points\ from\ Deviation = \frac{\sum(X_i - \bar{X})}{N}$$

会出现的问题是当 Data Point 小于 Mean，用上面的 Formula 来计算就会导致有 Negative Value。解决这个问题就是使用下面的 Formula 来计算。

$$Population\ Variance\ (\sigma^2) = \frac{\sum(X_i - \mu)^2}{N}$$

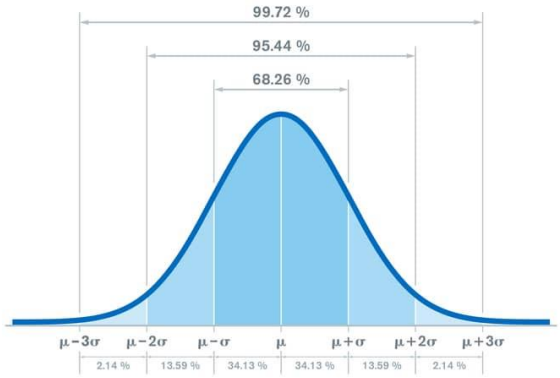
$$Sample\ Variance\ (S^2) = \frac{\sum(X_i - \bar{X})^2}{N - 1}$$

Population – 想要研究的所有对象组成的集合

Sample – 总体的子集

What is Standard Deviation σ?

Standard Deviation σ is a measure of how spread out numbers are.



$$Population\ Standard\ Deviation\ \sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

Standard Deviation is just Square Root of Variance. 计算了 Variance 之后没办法 Plot Bell Curve Graph 因为所有数值都被 Squared 了当在计算 Variance 的时候。

Bayes' Theorem 贝叶斯定理

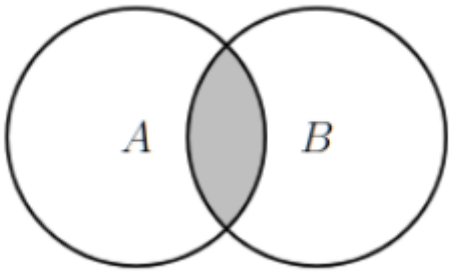
$$P(B|A) = \frac{P(B \cap A)}{P(A)} \rightarrow P(B \cap A) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) \rightarrow Probability\ of\ A\ given\ B\ is\ True$$

$$P(B|A) \rightarrow Probability\ of\ B\ given\ A\ is\ True$$

$$P(A), P(B) \rightarrow Independent\ Probabilities\ of\ A\ and\ B$$

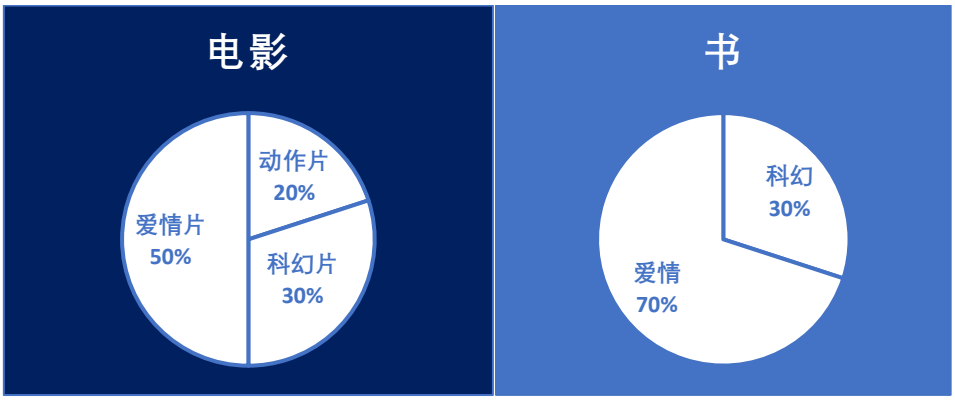


$P(B|A) \rightarrow A$发生的前提之下，B发生的机率 (上图灰色部分)

Example

100 个电影里，20 个动作片，30 个科幻片，50 个爱情片。

50 本书里，15 个科幻小说，35 个爱情小说。



$$P(电影)=\frac{100}{150},\ P(电子书)=\frac{50}{150},\ P(科幻)=\frac{45}{150},\ P(动作)=\frac{20}{150},\ P(爱情)=\frac{85}{150}$$

$$P(动作|电影)=\frac{20}{100},\ P(科幻|电影)=\frac{30}{100},\ P(爱情|电影)=50/100$$

$$P(科幻|书)=\frac{15}{50},\ P(爱情|书)=\frac{35}{50}$$

寻找已知是科幻题材，是电影的概率是多少，这时候就能够使用 Bayes' Theorem。

$$P(电影|科幻)=\frac{P(科幻|电影)P(电影)}{P(科幻)}$$

$P(电影|科幻)\rightarrow 验证后的概率\ (Posterior\ Probability)$ – 在观测到这个文件是科幻题材之后，我们知道了这个文件的部分信息，它是电影的概率改变了。因为这个概率 P(电影|科幻)是在观测之后才知道的，所以叫做后验概率。

$P(科幻|电影)\rightarrow Likelihood$ – 字典上意思是一件事发生的可能性或概率，在这个例子中它表示当文件是个电影时，它是科幻题材的概率。

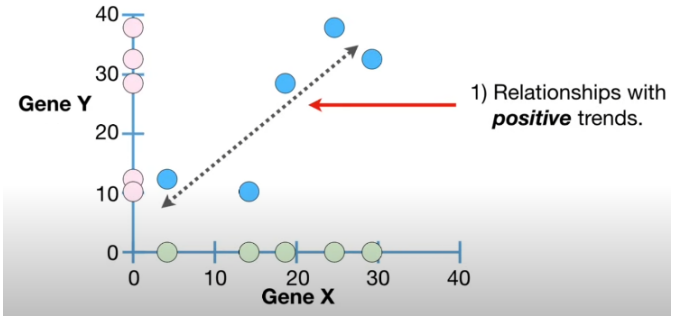
$P(电影)=先验证率\ (Prior\ Probability)$ – 在观测到这个文件是科幻题材之前，这个文件是未知的。我们的目标是算出它是电影的概率，而我们在观测之前已经知道了一个未知文件是电影的概率，因此 P(电影)叫做先验概率。

$P(科幻)=证据\ (Evidence)$ – 因为我们已经知道了它是科幻题材的，我们已经观察到了这个事实的发生，因此对我们来说它是一个证据，而我们观察到这个证据的概率 P(科幻)就叫证据。

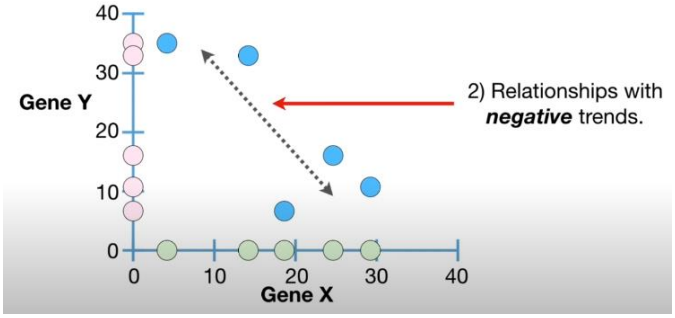
Covariance 协方差

Covariance 是用来知道两个 Parameters 之间是不是存在关系。但是 Covariance 的值不能告诉我们这两个 Parameters 之间的数据是 Close to Slope 还是 Far from Slop，个或者那个 Slope 有多 Steep。Covariance 可以分类成 3 种关系。

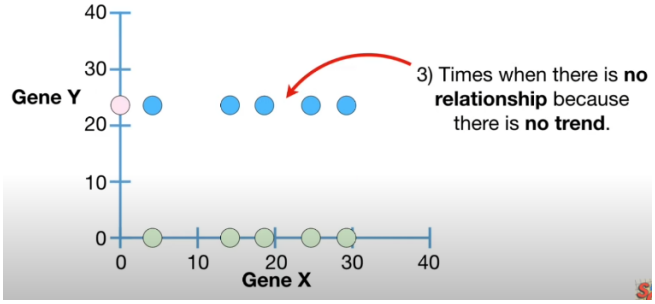
I. Relationships with **Positive Trends**

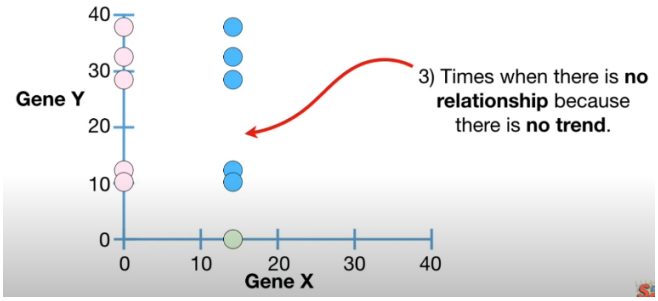


II. Relationships with **Negative Trends**



III. Times when there is **No Relationship** because there is **No Trend**

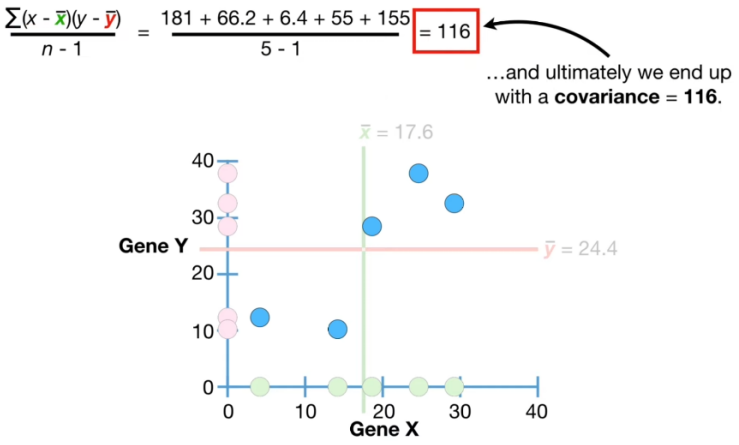




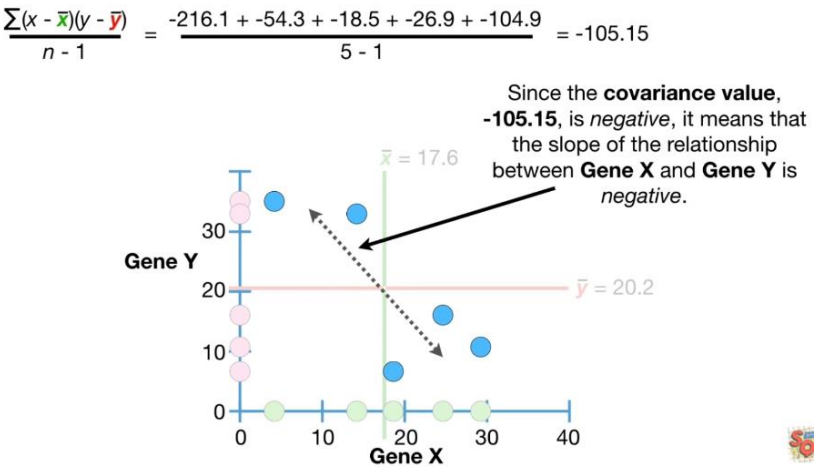
Covariance 也是一个计算的垫脚石 to something that is interesting, like **Correlation**。

$$Covariance = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

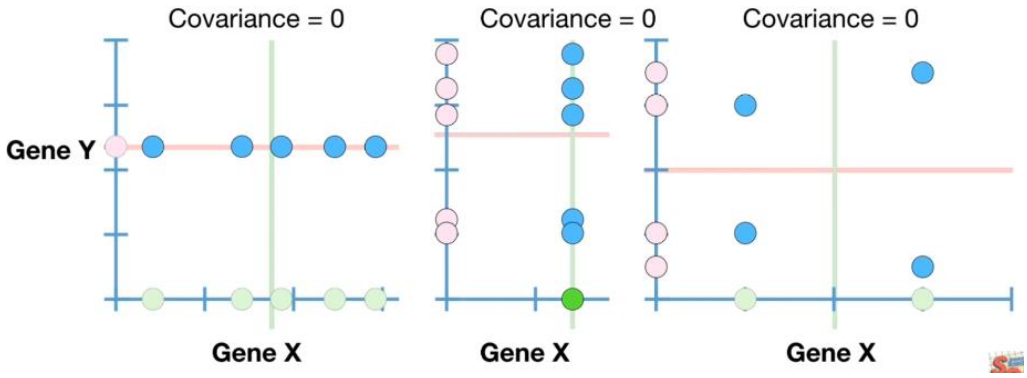
当计算出 Covariance 的值是 Positive，代表着这两个 Parameters 之间的关系是 Positive Slope。但是不能从 Covariance 的值知道 Slope 有多 Steep。



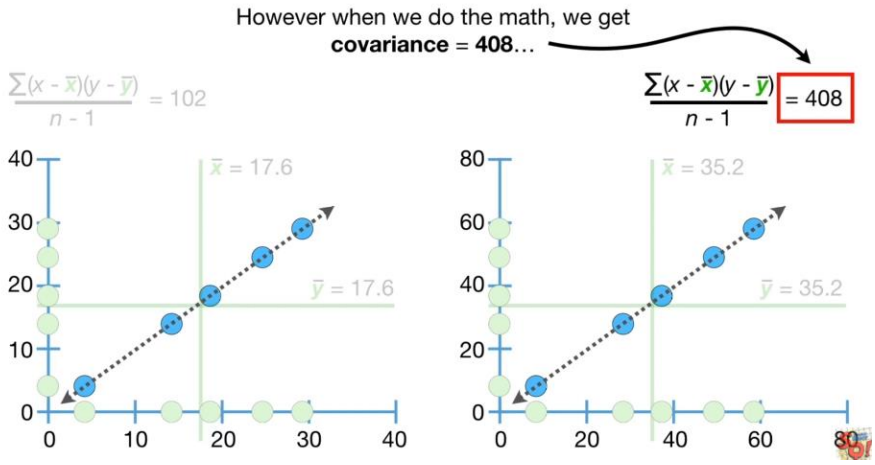
当计算出 Covariance 的值是 Negative，代表着这两个 Parameters 之间的关系是 Negative Slope。



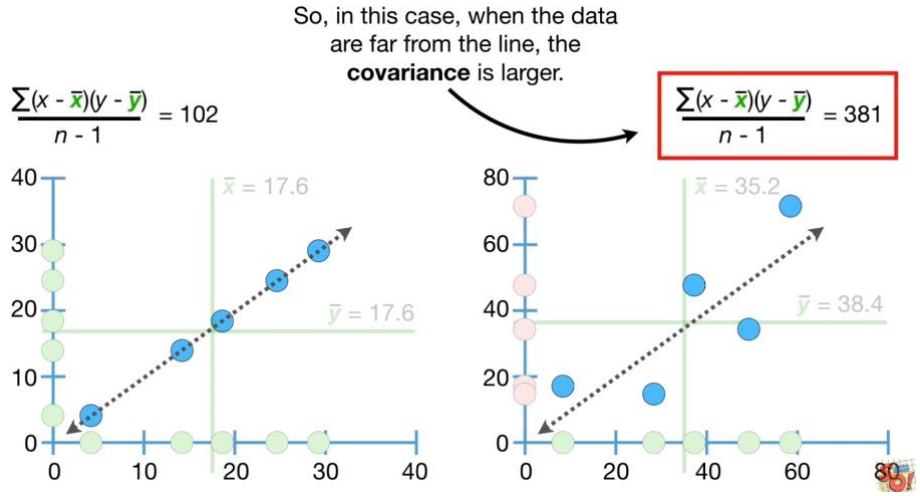
当 Covariance 算出来是 0，代表着这两个 Parameters 之间没有任何关系。

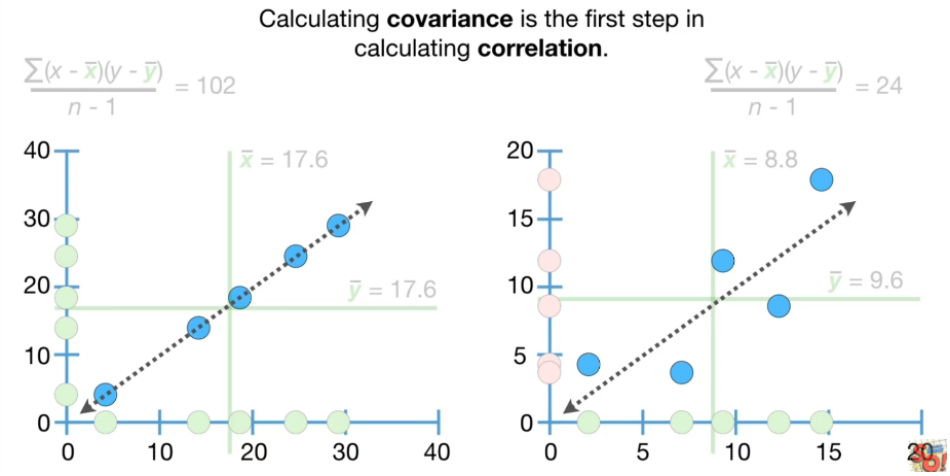


除此之外，当这两个 Parameters 之间的关系不变但是 Scale 变了，Covariance 的值也会改变。Scale 变大，Covariance 的值就会变大，Scale 变小，Covariance 的值就会变小。



当数据离 Line 远但是 Scale 是大的，这时候算出来的 Covariance 也是会比当数据就在 Line 上面但是 Scale 是小的，还要大。





从上面的图，可以看出 Covariance 只能用来知道这两个 Parameters 的 Data 的关系是 Positive, Negative 还是没有关系。

Correlation