Probability, Stochastic Processes, Statistics and Applications

Gavin Conran

Introduction

This revision/background paper covers the fundamentals of Probability, Stochastic Processes and Statistics with practical applications of each. The paper begins with a discussion on the difference between the two main theoretical approaches to Probability; Frequentist and Bayesian, coupled with the main Rules of Probability. After an exploration of four key mathematical tools: Principle Component Analysis (PCA); the Normal (Gaussian) Distribution; the Weiner Process; and Bayes' Rule, the paper, by applying this mathematical machinery, develops, step by step, a fully Bayesian approach to curve fitting. This approach lies at the heart of Pattern Recognition algorithms and techniques. The paper continues with an overview of the two key limit theorems of probability, the strong Law of Large Numbers and the Central Limit Theorem, followed by a discussion on hypothesis testing and the concept of statistical significance. The paper finishes with a discussion on how the Heat Equation is used to solve the Black-Scholes equation for pricing European options.

Gaussian and Bayesian Probability

Probability is a measure of the likelihood that an event will occur. The study of Probability can be divided into two main camps:

Frequentist Probability - *Assign numbers to decide some objective or physical state of affairs*:

- Probability of a Random Event denotes the Relative Frequency of Occurance of an experiment's outcome when repeating the experiment
- Carl Friedrich Gauss (1777 – 1855) gave his name to the term, Gaussian, used to describe the normal distribution

Bayesian Probability - *Assign numbers per subjective probability, i.e. as a degree of belief*:

- Includes Expert Knowledge as well as Experimantal Data to produce probabilities
- Expert knowledge is represented by some (subjective) ***Prior Probability Distribution***
- Experimental Data are incorporated into a ***Likelihood Function***
- Product of Prior and Likelihood, ***normalised***, returns a ***Posterior Probability Distribution***
- Thomas Bayes (1701- 1761) derived Bayes' Theorem, a simple statement about conditional probabilities, but the Bayesian interpretation of probability was developed mainly by Pierre-Simon Laplace (1749-1847), Napolean's examiner at the Ecole Militaire in 1784.

Event	Probabaility	Comment
A	$P(A)\in[0,1]$	
Not A	$P(A^c)=1-P(A)$	A^c : complement of A
A or B	$P(A\cup B)=P(A)+P(B)-P(A\cap B)$ $P(A\cup B)=P(A)+P(B)$	 If A and B are mutually exclusive
A and B	$P(A\cap B)=P(A\|B)P(B)=P(B\|A)P(A)$ $P(A\cap B)=P(A)P(B)$	 If A and B are independent
A given B	$P(A\|B)=P\frac{(A\cap B)}{P(B)}=\frac{P(B\|A)P(A)}{(P(B))}$	Conditional Probability

Table 1: Probability Notation

Rules of Probability

Rule	Discrete	Continuous
Sum Rule:	$P(X)=\sum_{Y} P(X,Y)$	$p(x)=\int p(x,y)dy$
Product Rule:	$P(X,Y)=P(Y\|X)P(X)$	$p(x,y)=p(y\|x)p(x)$
Bayes' Theorem:	$P(Y/X)=\frac{P(X\|Y)P(Y)}{P(X)}$	$p(y/x)=\frac{p(x\|y)p(y)}{p(x)}$
Normalisation Constant:	$P(X)=\sum_{Y} P(X\|Y)P(Y)$	$p(x)=\int p(x,y)dy$
where P(X) is a distribution over the Random Variable X P(Y) is a distribution over the Random Variable Y P(X, Y) is a Joint Probability => Probability of X and Y P(Y|X) is a Conditional Probability => Probability of Y given X		

Table 2: Rules of Probability

Note: The Normalisation Constant ensures that the sum of the Conditional Probability on the L.H.S of Baye's Theorem over all values equals one.

Probability Densities

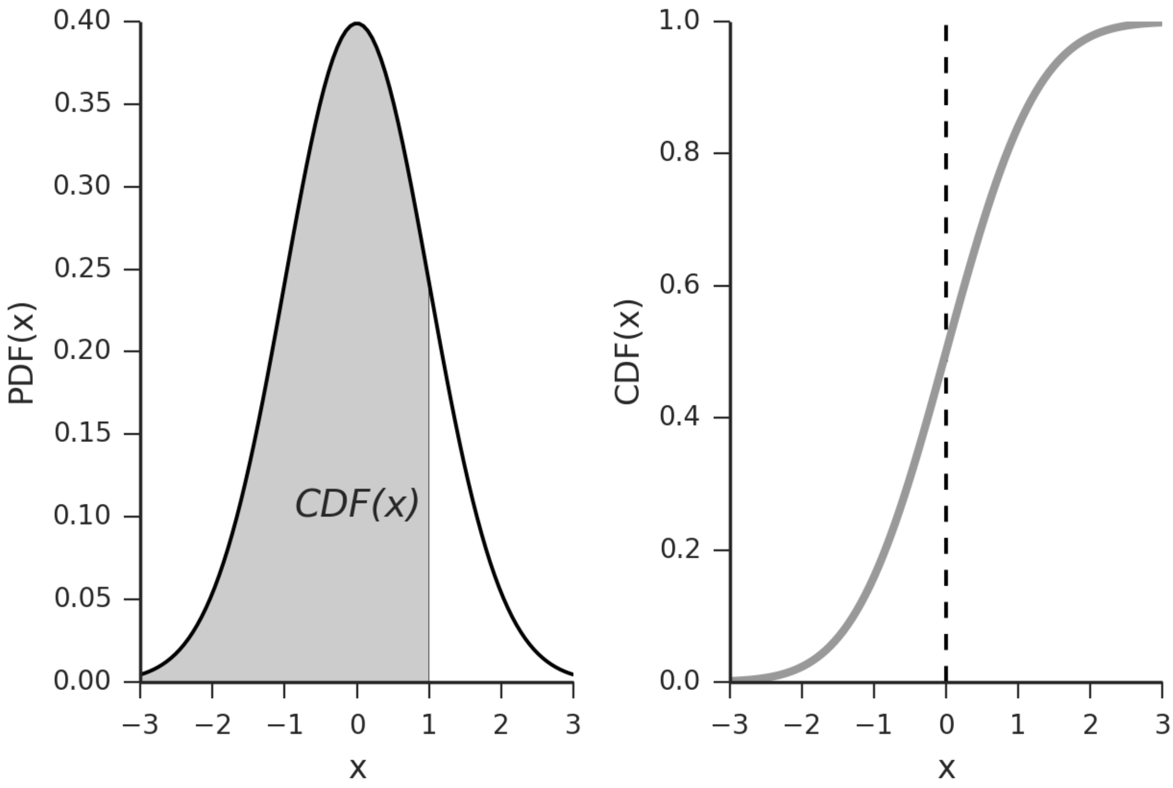


Illustration 1: PDF vs. CDF

- Concept of Probability for Discrete Variables can be extended to that of a ***Probability Density Function p(x), PDF(x),*** over a continuous variable x and is such that the probability of x lying on the interval (x, x+δx) is given by p(x)δx for δx→0.
- Probability Density, p(x), can be expressed as the derivative of the ***Cumulative Distribution Function P(x) - CDF(x)***.
- Probability that x will lie in an interval (a, b) is given by:

$$P(x \in (a,b)) = \int_a^b p(x)dx$$

- The probability that x lies in the interval (-∞, z) is given by the Cumultative Distribution Function is defined by:

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

which satisfies

$$CDF'(x) = PDF(x)$$

Expectation and Variance

Expectation (Weighted Averages of Functions):

The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the Expectation of $f(x)$ and is denoted as $\boldsymbol{E}[f]$.

Discrete Distribution:

$$\boldsymbol{E}[f] = \sum_x p(x)f(x)$$

Here the average is weighted by the relative probabilities of the different values of x.

Continuous Distribution:

$$\boldsymbol{E}[f] = \int p(x)f(x)dx$$

In either case, if we are given a finite number N of points drawn from the Probability Distribution or Probability Density then the Expectation can be approximated as a sum over these points:

Approximation:

$$\boldsymbol{E}[f] = \frac{1}{n}\sum_{n=1}^{N} f(x_n)$$

This approximation becomes exact in the Limit $N \rightarrow \infty$.

The ***Conditional Expectation*** wrt a Conditional Distribution:

$$\boldsymbol{E}_x[f|y] = \sum_x p(x|y)f(x)$$

Note: For a function of several variables we use a subscript to indicate which variable is being averaged over, e.g $\boldsymbol{E}_x[f(x,y)]$ denotes the average of the function $f(x,y)$ wrt the distribution of x.

Variance / Covariance

The **variance of** $\boldsymbol{f(x)}$ is defined by

$$\boldsymbol{var}[f]=\boldsymbol{E}[(f(x)-\boldsymbol{E}[f(x)])^2]$$

and provides a measure of how much variability there is in $f(x)$ around its mean value $\boldsymbol{E}[f(x)]$.

By expanding out the square, we see that the variance can also be written in terms of the Expectations of $f(x)$ and $f(x)^2$:

$$\boldsymbol{var}[f]=\boldsymbol{E}[f(x)^2]-\boldsymbol{E}[f(x)]^2$$

The **variance of the variable x** itself is given by:

$$\boldsymbol{var}[x]=\boldsymbol{E}[x^2]-\boldsymbol{E}[x]^2$$

The **Covariance of two Random Variables** x and y is defined by:

$$\begin{aligned}\boldsymbol{cov}[x,y]&=\boldsymbol{E}_{x,y}[\{x-\boldsymbol{E}[x]\}\{y-\boldsymbol{E}[y]\}]\\ &=\ \boldsymbol{E}_{x,y}[xy]-\boldsymbol{E}[x]\boldsymbol{E}[y]\end{aligned}$$

which expresses the extent to which x and y vary together.

If x and y are independent then their covariance vanishes.

Covariance Matrix

In the case of two vectors of Random Variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ the ***Covariance Matrix*** is given by:

$$\boldsymbol{cov}[\boldsymbol{X},\boldsymbol{Y}]=\boldsymbol{E}_{x,y}[\boldsymbol{X}\boldsymbol{Y}^{\boldsymbol{T}}]-\boldsymbol{E}[\boldsymbol{X}]\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{T}}]$$

and for ease of notation a vector $\boldsymbol{X}$ has the following Covariance Matrix:

$$\boldsymbol{cov}[\boldsymbol{X},\boldsymbol{X}]\equiv\boldsymbol{cov}[\boldsymbol{X}]$$

Principal Component Analysis (PCA)

Principle Component Analysis (PCA) is a technique used to emphasise ***variation*** and bring out strong patterns in a dataset. It is ofter used to make data easy to explore and visualise.

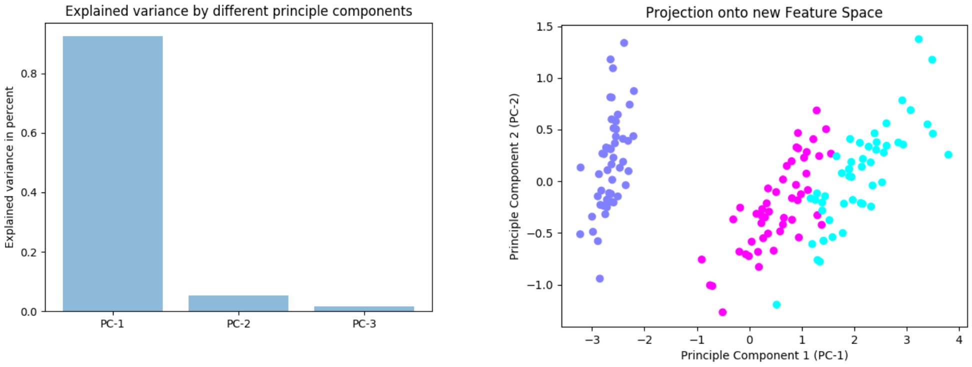


Illustration 2: PCA example with Iris Data-set

Covariance measures the statistical dependence / independence between two variables and strong statistically dependent variables can be considered as redundant observations of the system. This insight leads us to an easy way to identify redundant data by considering the covariance between data sets.

If we dig deeper into our covariance matrix we notice that it is a square, symmetric *m x m* matrix whose diagonal represents the variance between measurement types. Thus ***cov[X, Y]*** captures the correlations between all possible pairs of measurements, ***X*** and ***Y***. Redundancy is thus easily captured since if two data sets are identical, the off-diagonal term and diagonal term would be equal since $\sigma_{\boldsymbol{XY}}^2 = \sigma_{\boldsymbol{X}}^2 = \sigma_{\boldsymbol{Y}}^2$ if **X** = **Y**. Thus large diagonal terms correspond to redundancy while small diagonal terms suggest that the two measured quantities are close to statistically independent and have low redundancy.

It should also be noticed that large diagonal terms, or those with large variances, typically represent what we might consider the *dynamics of interest* since the large variance suggests strong fluctations in that variable. The insight given by the covariance matrix leads to our ultimate aim of:

1) removing redundancy
2) identifying those signals with maximum variance

Thus in a mathematical sense, we are simply asking to represent ***cov[X]*** so that the diagonals are ordered from the largest to smallest and the off-diagonals are zero i.e. our task is to ***diagonalise*** the covariance matrix.

The key idea behind diagonalisation is simply this: there exists an ***ideal basis*** in which ***cov[X]*** can be written (diagonalised) so that in this basis, all redundancies have been removed, and the largest variances of particular measurements are ordered, meaning that the system has been written in terms of its ***Principle Components***, as shown in Illustration 2 and thus ***reducing dimensionality***.

Gaussian (Normal) Distribution

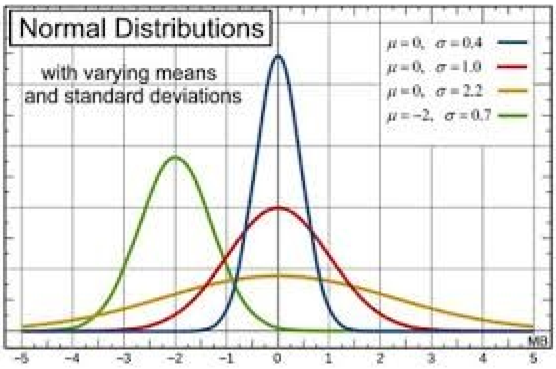


Illustration 3: Standard Normal Distributions with varying means and standard deviations

For the case of a single valued variable, x, a Gaussian Distribution, as shown in Illustration 3, is defined by:

$$N(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

which is governed by the two parameters:

μ : mean

σ^2 : variance; σ : standard deviation; $\frac{1}{\sigma^2} = \beta$: precision

Observations:

1) $N(x|\mu,\sigma^2)>0$
2) The Gaussian is normalised by the $\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}$ term
3) $\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\,dx=1$

Expectation, Variance and Likelihood Functions

Mean (1st order moment)

The average value of x under the distribution.

It also represents the mode, i.e. the maximum value of the distribution.

$$\boldsymbol{E}[x]=\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\,x\,dx=\mu$$

$$\boldsymbol{E}[x^2]=\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\,x^2 dx=\mu^2+\sigma^2$$

Variance (2nd order moment)

$$\boldsymbol{var}[x]=\boldsymbol{E}[x^2]-\boldsymbol{E}[x]^2=\sigma^2$$

The ***Likelihood Function*** of the Gaussian is given by:

$$p(\boldsymbol{X}|\mu,\sigma^2)=\prod_{n=1}^{N} N(x_n|\mu,\sigma^2)$$

For mathematical ease we use the ***Log Likelihood***

$$\ln p(\boldsymbol{X}|\mu,\sigma^2)=\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)$$

In the case of the Gaussian Distribution the solution for μ and σ^2 decouples so we can evaluate μ and then $\sigma^{2.}$.

Sample mean (maximising wrt μ)	Sample Variance (maximising wrt σ^2)
$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N} x_n$	$\sigma^2{}_{ML}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$

There is a significant limitaion to the Maximum Likelihood approach, i.e. ***Bias:***

- The maximum likelihood approach systemically under estimates the Variance of the distribution by a factor of $(\frac{N-1}{N})$
- Bias is due to computing the variance with the sample mean rather than the true mean when inferring μ and σ^2 from the Maximum Likelihood.
- Bias is also related to the problem of ***over-fitting*** encountered in polynomial curve fitting.
- A Bayesian approach removes the Bias.

Note: The 3rd order moment is refered to as ***Skewness*** and ***Kurtosis*** is the 4th order moment.

Stochastic Processes

Brownian Motion, B_t :

A stochastic process, $\{B_t : 0 \leqslant t \leqslant \infty\}$, is a standard Brownian motion if:

- $B_0 = 0$
- It has Continuous Sample Paths
- It has independent, normally-distributed increments

Wiener Process, W_t

The Weiner Process, W_t , is viewed as equivalent to a standard Brownian Motion, B_t and is characterised by the following characteristics:

- $W_0 = 0$
- W_t has ***ALMOST*** Continuous Sample Paths
- W_t has independent increments with distribution $W_t - W_s \sim N(0, t-s)$
- Mean (1st moment): $\boldsymbol{E}[W(t)] = 0$
- Variance (2nd moment): $\boldsymbol{E}[W(t)^2] = t$
- W_t is a martingale (and therefore a Markovian Process)

Ito Process

An adapted Stochastic Process, $dx = a\,dt + b\,dW(t)$, that can be expressed as the sum of an integral wrt Brownian Motion and integral wrt time.

$$X_t = X_0 + \int_0^t b\,dW + \int_0^t a\,ds$$

Where W is a Wiener Process
a, the Drift, is predictable and integratable
b, the Diffusion, is a predictable B-integrable process

Wiener Process with Drift (example of an Ito process)

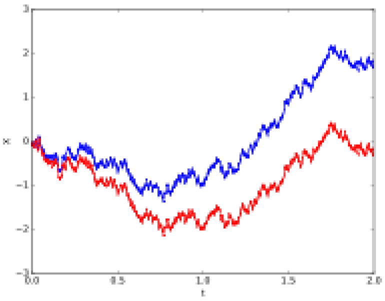


Illustration 4: Wiener processes with drift (blue) and without drift (red).

A Weiner process with drift, as shown in Illustration 4 above, is goverened by the same Differential Equation used to describe the, earlier described, Ito Process:

$$dx = a\,dt + b\,dW(t)$$

where a and b are constants

Ito‘s Lemma

Ito’s lemma is an identity in Ito calculus to find the differential of a time-dependent function of a Stochastic Process. It serves as the stochastic calculus counterpart of the ***chain rule*** (see Appendix II). It can be heuristically derived by forming the Taylor Series expansion of the function up to its second derivatives and retaining terms up to first order in time increment and second order in the Wiener process increment. The lemma is widely employed in mathematical finance, and its best known application is in the derivative of the ***Black-Scholes equation for option values***.

An informal derivation of Ito’s lemma follows:

Step 1: Assume X_t is an Ito drift-diffusion process that satisfies the Stochastic Differential Equation (SDE):

$$dX_t = \mu_t\,dt + \sigma_t\,dB_t$$

Where B_t is a Wiener Process

Step 2: If ***f(t, x)*** is a twice-differentiable scalar function, its expansion in a Taylor Series is

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dx + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}dx^2 + \ldots$$

Substuiting X_t for x and therefore $\mu_t dt + \sigma_t dB_t$ for dx gives:

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}(\mu_t\,dt + \sigma_t dB_t) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(\mu_t^2 dt^2 + 2\mu_t\sigma_t dt\,dB_t + \sigma_t^2 dB_t^2) + \ldots$$

Step 3: In the limit dt → 0, the terms dt^2 and $dt\,dB_t$ tend to zero faster than dB_t^2 which is of the order of complexity O(dt). Setting:

- dt^2 and $dt\,dB_t$ terms to zero,
- Substuiting dt for dB_t^2 (due to the variance of a Wiener Process)
- Collecting the dt and dB_t terms

we obtain, Ito’s lemma, as required:

$$df = \left(\frac{\partial f}{\partial t} + \mu_t\frac{\partial f}{\partial x} + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial x^2}\right)dt + \sigma_t\frac{\partial f}{\partial x}dB_t$$

Markov Process

- Only the present value of a variable is relevant for predicting the future
- The history of the variable and the way that the present has emerged from the past are irrelevant

Martingale Process

- All Martingales are Markovian
- A stochastic process where at any time, t, the expected value of the final outcome is the current value, i,e. $\boldsymbol{E}[X_T | X_t = x] = x$

Bayesian Probabilities

We want to quantify the uncertainty that surrounds the appropriate choice for the model parameters **W**. We can use the machinery of Probability Theory to describe the uncertainty in model parameters such as **W**. This is conducted using the following steps:

1) Capture assumptions made about **W**, before observing the data, in the form of a ***Prior Probability Disptribution***, p(**W**)
2) The effect of the observed data, $D=\{t_{1,}...,t_N\}$ is expressed through the ***Conditional Probability,*** p(D|**W**). This is also known as the ***Likelihood Function***.
3) Evaluate the uncertainty in **W** after we have observed D, using ***Bayes' Theorem***, in the form of the ***Posterior Probability***, p(**W**|D):

$$p(\boldsymbol{W}|D)=\frac{p(D|\boldsymbol{W})p(\boldsymbol{W})}{p(D)}$$

Posterior ~ Likelihood * Prior (The Denominator is the Normalisation Constant)

The ***Likelihood Function*** expresses how probable the observed data set is for different settings of the parameter vector **W**.

- Likelihood is NOT a probability distribution over **W**
- Its integral wrt **W** does not (necessarily) equal to 1

Integrating both sides of Bayes Theorem wrt W we can express the Denominator in terms of the Prior Distribution and the Likelihood Function:

$$p(D)=\int p(D|\boldsymbol{W})p(\boldsymbol{W})dW$$

Bayesian and Frequentist Probabilities interpret the Likelihood Function differently.

Frequentist: **W** is considered to be a ***fixed parameter, i.e. point estimate,*** whose value is determined by some form of ***Estimator***. ***Error Bars*** on this estimate are obtained by considering the distribution of possible data sets, D.

A popular Estimator is ***Maximum Likelihood***:

- **W** is set to the value that ***Maximises the Likelihood Function*** p(D|**W**) which corresponds to choosing the value of **W** for which the probability of the observed data is Maximised.
- In Machine Learning the negative Log Likelihood Function is called the Error Function
- Because the negaive log is a monotonically decreasing function

Maximising the Likelihood <=> Minimising the Error

A popular way to estimate Error Bars is the ***Bootstrap***:

- ***Multiple Data Sets*** are created by sampling the original Data Set, D, of size N with Replacement:
- Repeat L times to generate L Data Sets each of size N and and each obtained from the original data set, D.
- Statistical Accuracy of parameter Estimates can then be evaluated by looking at the ***variability of predictions*** between the different bootstrap data sets.

Bayesian: There is only a single Data Set, D (namely the one actually observed), and the ***uncertainty in the parameters*** is expressed through a probability distribution over **W**.

Application: Polynomial Curve Fitting

At first, the goal is to fit ‘observed’ data using a polynomial function of the form:

$$y(x,\boldsymbol{W})=w_0+w_1x+w_2x^2+\ldots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

where
M is the order of the polynomial
x^j denotes x raised to the power j
the polynomial coefficients, $w_{0,}\ldots,w_M$, are collectively denoted as the vector **W**

Once the vector **W** has been ***inferred*** the secondary goal is to use the polynomial function to ***predict*** $y(x,\boldsymbol{W})$ from new input data, x.

We will explore four approaches to this Curve Fitting problem:

- Plain vanilla Approach
- Introducing (Gaussian) Probability
- Prior Distribution leading to a more Bayesian Approach
- Fully Bayesian Approach

1) Plain vanilla approach

Error Function:

$$E(\boldsymbol{W})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{W})-t_n\}^2$$

Root Mean Square Error Function:

$$E_{RMS}=\sqrt{\left(\frac{2E(W^*)}{N}\right)}$$

We can solve the curve fitting problem by choosing the value of **W** for which E(**W**) is a small as possible. Because the error function is a quadratic function of the coefficient **W** its derivatives wrt the coefficients will be linear in the elements of **W** and the Minimisation of the Error Function has a unique solution, demoted by $\boldsymbol{W}^*$, which can be found in closed form.

The resulting polynomial is given by the function $y(x,\boldsymbol{W}^*)$.

Over-fitting

As shown in Illustration 5, over-fitting happens when, as the order of the polynomial increases, there is an excellent fit to the training data but the curve oscillates wildly.

This is due to the more flexible polynimials of large order becoming increasingly tuned to the Random Noise on the target values.

Increasing the size of the data set reduces the over-fitting problem but it is unsatisfactory to have to limit the number of parameters in a model according to the size of the available training set.

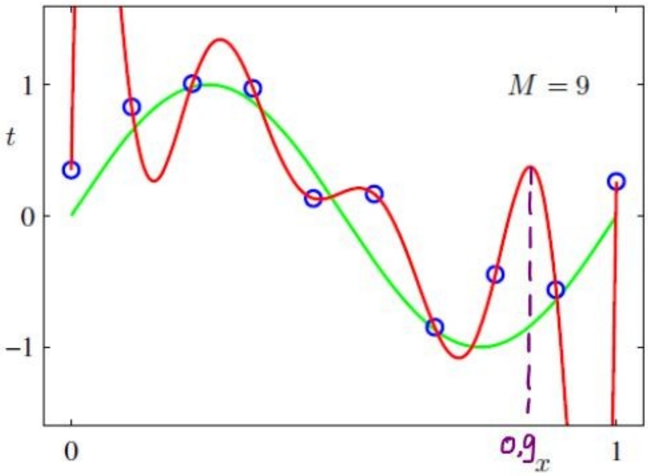


Illustration 5: Over-fitting. The green curve shows the function sin(2πx), the blue circles are the data points and the red line is the inferred polynomial with coefficients $\boldsymbol{W}^*$.

Regularisation

Regularisation controls the over-fitting phenomenon by adding a penalty term to the Error Function in order to discourage the coefficients from reaching large values.

Error Function with Regularisation:

$$\tilde{E}(\boldsymbol{W})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{W})-t_n\}^2+\frac{\lambda}{2}\|\boldsymbol{W}\|^2$$

where

$$\|\boldsymbol{W}\|^2=\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}=w_0{}^2+w_1{}^2+\ldots+w_m{}^2$$

and

λ Governs the relative importance of the regularisation term compared with the Sum-Of-Squares Error Term

2) Introducing (Gaussian) Probability

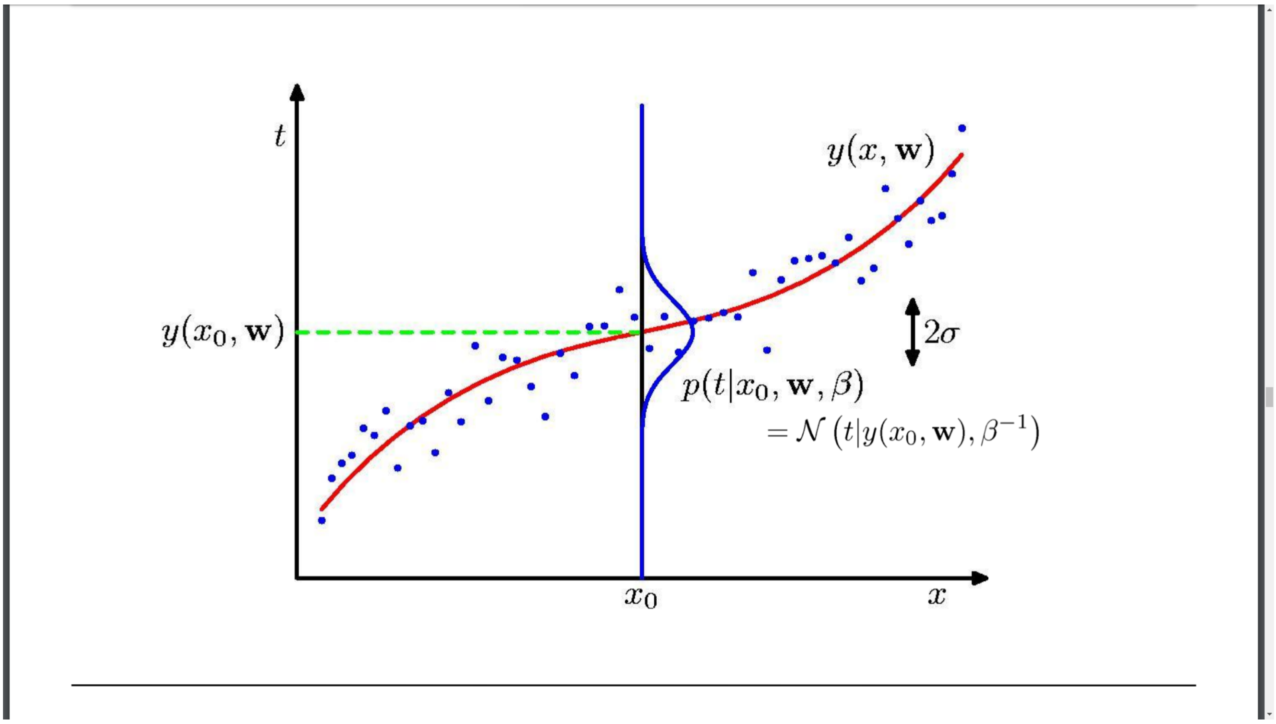

t
$y(x, \mathbf{w})$
$y(x_0, \mathbf{w})$
2σ
$p(t|x_0, \mathbf{w}, \beta)$
$= \mathcal{N}\left(t|y(x_0, \mathbf{w}), \beta^{-1}\right)$
x_0
x

Illustration 6: Gaussian Conditional Distribution for t given x in which the mean is given by the polynomial function y(x, W) and the precision is given by the parameter β which is related to the variance $\beta^{-1} = \sigma^2$.

We can express our uncertainty over the value of the target value, t, using a Probability Distribution as shown in Illustration 6. Assume that, given the value of x, the corresponding value of t has a Gaussian Distribution with the mean equal to $y(x, \boldsymbol{W})$. Thus we have:

$$p(t|x, \boldsymbol{W}, \beta) = N(t|y(x, \boldsymbol{W}), \beta^{-1})$$

where

β Is the ***Precision*** which corresponds to the inverse Variance of the Distribution

Likelihood Function

$$p(t|\boldsymbol{X}, \boldsymbol{W}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{W}), \beta^{-1})$$

Log Likelihood Function

$$\ln p(t|\boldsymbol{X}, \boldsymbol{W}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{W}), \beta^{-1}) + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

We now use the training data {**x**, **t**} to determine the values of the unknown parameters **W** and β by Maximum Likelihood. As before it is convenient to ***maximise*** the Logarithm of the Likelihood Function.

Maximising wrt **W** giving $\boldsymbol{W}_{\boldsymbol{ML}}$	Maximising wrt β giving β_{ML}
$E(\boldsymbol{W}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{W}) - t_n\}^2$	$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{W}_{\boldsymbol{ML}}) - t_n\}^2$

When considering the Maximum Likelihood solution to the polynomial coefficients $\boldsymbol{W}_{\boldsymbol{ML}}$:

- Determined by Maximising the Log Likelihhod Function of **W:**
  - Can omit last two terms of RHS because they do not depend on W
  - Scaling the Log Likelihhod by +ive constant coefficient does not alter the location of the maximum wrt **W** and so we can replace the coefficient $\beta/2$ with ½
- Instead of maximising the Log Likelihood we can equivalently Minimise the -ive Log Likelihood
- Therefore, we see that the ***Maximising Likelihood is equivalent***, so as far as determining W is concerned, to ***Minimising the Sum-Of-Squares Error Function***!!!

This is a similar process to determining $\boldsymbol{W}_{\boldsymbol{ML}}$ (governing the mean) and subsequently β_{ML} as with the case for the simple Gaussian Distribution.

Having determined the model parameters **W** & β we can now ***make predictions for new values of x***. Because we now have a ***Probablistic Model*** the predictions are expressed in terms of the Predictive Distribution that gives the Probability Distribution over t rather than simply a point estimate and is obtained by substituting the maximum Likelihood parameters into:

$$p(t|x,\boldsymbol{W},\beta)=N(t|y(x,\boldsymbol{W}),\beta^{-1})$$

To give the Predictive Distribution:

$$p(t|x,\boldsymbol{W}_{\boldsymbol{ML}},\beta_{ML})=N(t|y(x,\boldsymbol{W}_{\boldsymbol{ML}}),\beta_{ML}{}^{-1})$$

3) Prior Distribution leading to a more Bayesian Approach

Given the ***Prior Distribution***

$$p(\boldsymbol{W}|\alpha)=N(\boldsymbol{W}|0,\alpha^{-1}I)=(\frac{\alpha}{2\pi})^{\frac{(M+1)}{2}}\exp(-\frac{\alpha}{2}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W})$$

where

α is the Precision of the distribution
M + 1 is the total number of elements in vector **W** for a M^{th} order polynomial

Using ***Bayes' Theorem*** the ***Posterior Distribution*** is:

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{t},\alpha,\beta)\propto p(t|\boldsymbol{X},\boldsymbol{W},\beta)p(\boldsymbol{W}|\alpha)$$

Using ***Maximum Posterior (MAP)*** we can now determine **W** by finding the most probable value of **W** given the data. In other words ***Maximise the Posterior Distribution (MAP)***:

Taking the -ive Log of the Posterior Distribution and combining it with the Log Likelihood and the Prior Distribution

From earlier the Log Likelihood is:

$$\ln p(t|\boldsymbol{X},\boldsymbol{W},\beta)=-\frac{\beta}{2}\sum_{n=1}^{N}N(t_n|y(x_n,\boldsymbol{W}),\beta^{-1})+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$$

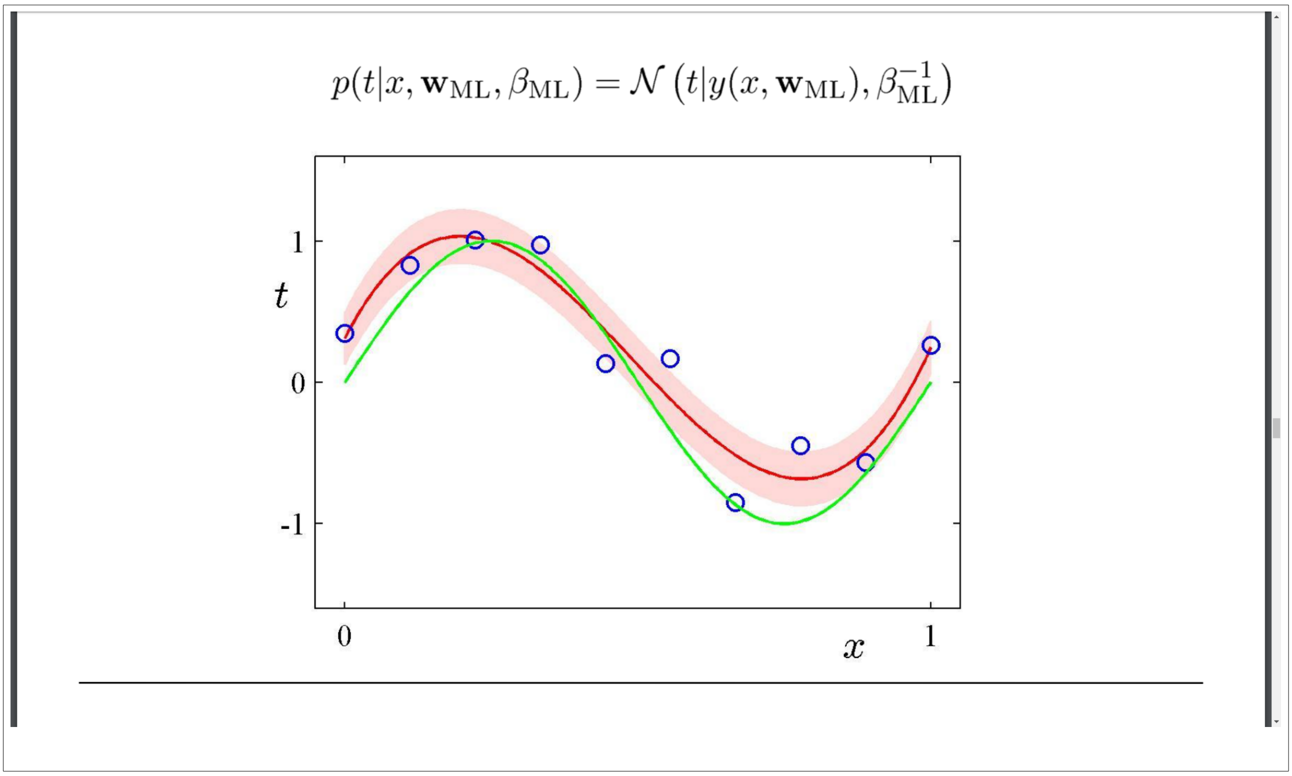
Therefore we find that the Maximing the Posterior is given by the Minimum of:

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{W})-t_n\}^2+\frac{\alpha}{2}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}$$

Which is equivalent to the ***Error Function with Regularisation***

$$\tilde{E}(\boldsymbol{W})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{W})-t_n\}^2+\frac{\lambda}{2}\|\boldsymbol{W}\|^2$$

Thus we see that maximising the Posterior Distribution is equivalent to Minimising the Regularised Sum-Of-Squares Error Function encountered earlier with a Regularisation parameter given by $\lambda=\frac{\alpha}{\beta}$.



$$p(t|x,\mathbf{w}_{\mathrm{ML}},\beta_{\mathrm{ML}})=\mathcal{N}\left(t|y(x,\mathbf{w}_{\mathrm{ML}}),\beta_{\mathrm{ML}}^{-1}\right)$$

*Illustration 7: Predictive Distribution with a point estimate of **W** and β*

We are still only making a ***point estimate of* W** , as shown in Illustration 7, and so does ***not*** amount to a ***full Bayesian*** approach and so we march on.

4) Fully Bayesian Approach

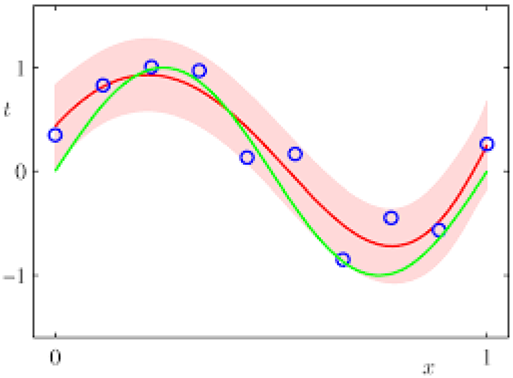


Illustration 8: Bayesian Curve Fitting. The red curve denotes the mean of the Predictive Distribution and the red region corresponds to +/- standard deviation around the mean. Both the mean and variance are dependent on x.

In a full Baysian approach, as shown in Illustration 8, we should consistently apply the Sum and Product Rules of Probability which requires we ***integrate over all values of* W**. Such ***Marginalisations*** lie at the heart of Bayesian Approaches for Pattern Recognition.

The goal is to evaluate the ***Predictive Distribution***

$$p(t|x,\boldsymbol{X},\boldsymbol{t})$$

Here we assume that α and β are fixed and known in advance.

A ***Bayesian approach*** simply corresponds to a consistent application of the Sum and Product Rules of Probability which allows the ***Predictive Distribution to be written as***:

$$p(t|x,\boldsymbol{X},\boldsymbol{t})=\int p(t|x,\boldsymbol{W})\,p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{t})\,dW$$

Where the (Gaussian) Predictive Distribution is given by:

$$p(t|x,\boldsymbol{W})=N(t|y(x,\boldsymbol{W}),\beta^{-1})$$

And the ***Posterior Distribution*** is given by

$$p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{t})$$

This integration can be performed analytically with the result that the ***Bayesian Predictive Distribution is given by the Gaussian***:

$$p(t|x,\boldsymbol{X},\boldsymbol{t})=N(t|m(x),S^2(x))$$

Where the mean is given by:

$$m(x)=\beta\phi(x)^T\boldsymbol{S}\sum_{n=1}^{N}\phi(x_n)t_n$$

And the variance is given by

$$S^2(x)=\beta^{-1}+\phi(x)^T\boldsymbol{S}\phi(x)$$

The matrix **S** is given by:

$$\boldsymbol{S}^{-1}=\alpha\boldsymbol{I}+\beta\sum_{n=1}^{N}\phi(x_n)\phi(x)^T$$

I is the Unit Matrix

And the vector $\phi(x)$ is defined with elements $\phi_i(x)=x^i\, for\; i=0,\ldots,m$

We see that the variance, as well as the mean, of the Predictive Distribution $p(t|x,\boldsymbol{X},\boldsymbol{t})$ is dependent on x.

Application: Hypothesis Testing and Statistical Significance

Two theorems of probability are important to understand.

Theorem: *Strong law of large numbers*:

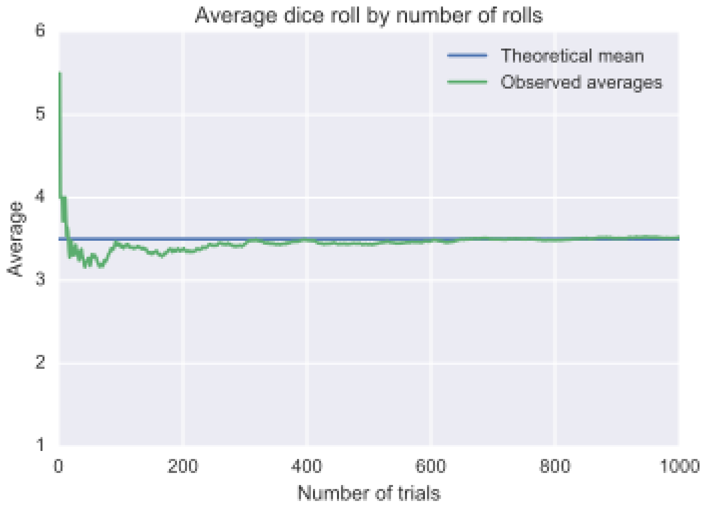


Illustration 9: Law of Large Numbers

Let $X_{1,} X_{2,} \ldots$ be a sequence of independent random variables having a common distribution, and let $E[X_j] = \mu$. Then, with probability one,

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \; as \; n \to \infty$$

Theorem: *Central Limit Theorem*:

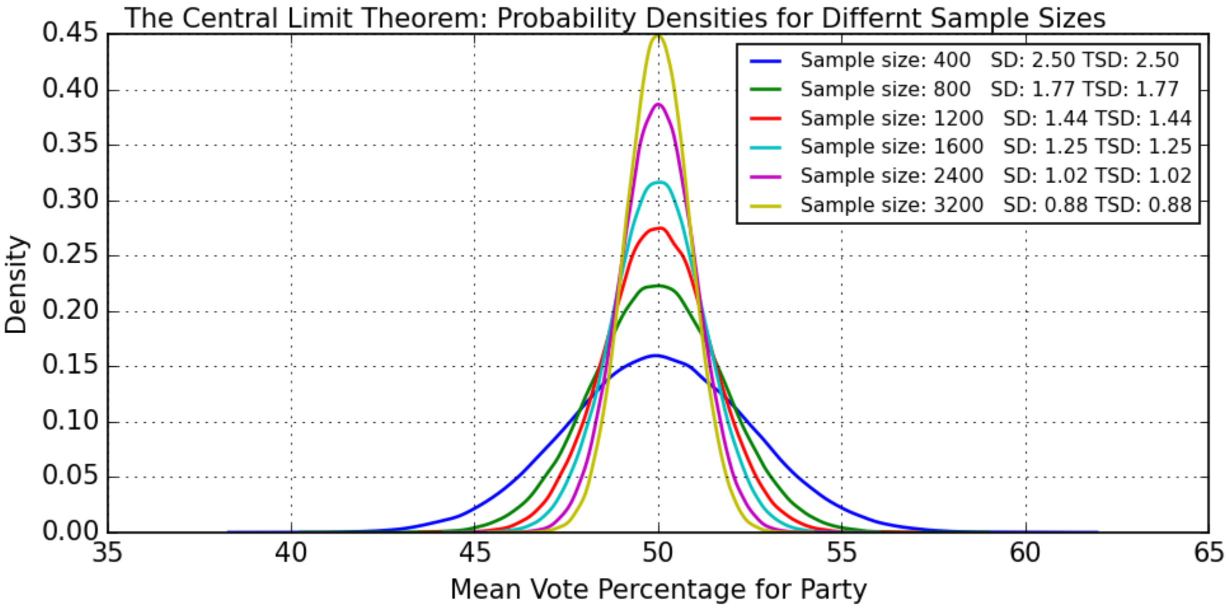


Illustration 10: Central Limit Theorem

Let $X_{1,} X_{2,} \ldots$ be a sequence of independent random variables each with mean μ and variance σ^2 . Then the ***distribution*** of the quantity

$$\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{(n)}}$$

tends to the ***standard normal distribution*** as $n\rightarrow\infty$ so that

$$P\{\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{(n)}}\leqslant a\}\rightarrow\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{a} e^{\frac{-x^2}{2}}dx \quad \text{as} \quad n\rightarrow\infty$$

The important part of this result:

It holds for any distribution of X_j !!!

Thus for a large enough sampling of data, i.e. a sequence of independent random variables that goes to infinity, one should observe the ubiquitous bell-shaped probability curve of the normal distribution.

This is a powerful result provided you have large enough sampling. As the normal distribution has some very beautiful mathematical properties it leads us to consider the Central Limit Theorem almost exclusively when testing for statistical significance.

Statistical Decisions

- The power of probability and statistics comes into play in making decisions, or drawing conclusions, based upon limited sampling of information of an entire ***Popuation*** (of data)
- Indeed, it is often impossible or impractical to examine the entire population
- Thus a sample of the population is considered instead. This is called the ***Sample***
- The goal is to infer facts or trends about the entire population with this limited sample size
- The process of obtaining samples is called ***Sampling***
- The process of drawing conclusions about this sample is called ***Statistical Inference***
  - An example is ***Polling*** data for elections.
- Using these ideas to make decisions about a population based upon the sample information is termed ***Statistical Decisions***
  - Example: deciding whether a new drug is effective in curing a disease or ailment
  - This will be the focus for the rest of this section

Statistical Decision Making

- All statistical tests are Hypothesis Driven. A ***Statistical Hypothesis*** is proposed and its validity investigated, usually a ***Null Hypothesis,*** H_0 , e.g. there is no difference between two procedures.
- Any Hypothesis which differs from a given hypothesis is called an ***Alternative Hypothesis ,*** H_1 **,** e.g. there is a difference between the two procedures.

Tests for Statistical Significance

- The ultimate goal in statistical decision making is to develop tests that are capable of allowing us to accept or reject a hypothesis with a certain, quantifiable ***confidence level.***
- Procedures that allow us to do this are called

    - tests of hypothesis,
    - tests of significance or
    - rules of decision.
- In making such statistical decisions, there is some probability that erroneous conclusion can be reached. Such errors are classified as follows:
    - ***Type I error****:* If a hypothesis is rejected when it should be accepted
    - ***Type II error***: If a hypothesis as accepted when it should be rejected
- What is imporatnt in making the statistical decision is the ***level of confidence*** of the test
- This probability is often denoted by α and in practice it is often taken to be 0.05 or 0.01 This corresponds to a 5% or 1% level of significance, or alternatively, we are 95% or 99% confident respectively that our decision is correct. Basically, in these tests, we will be wrong 5% or 1% of the time

Hyothesis Testing with the Normal Distribution

Figure 5.6: The normal distribution

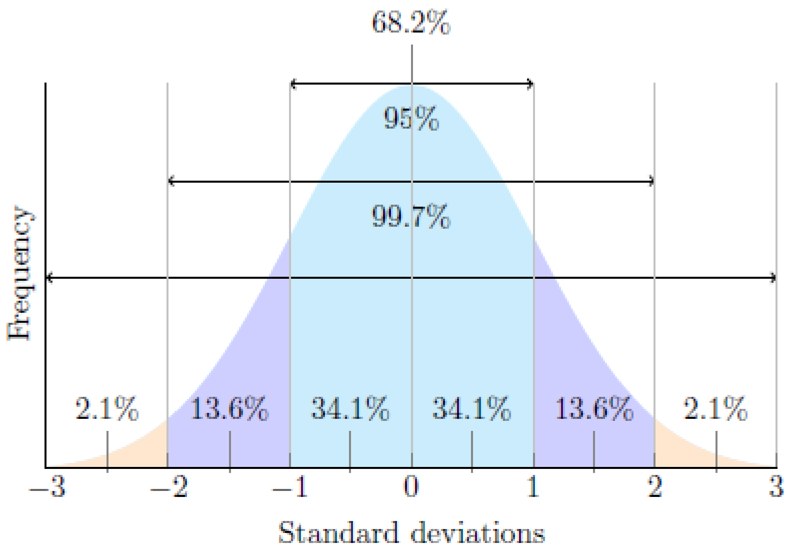


Illustration 11: Standard Normal Distribution Curve f(x) with the first, second and third standard deviation markers with the corresponding percentage of the population contained within each

- For large samples, the ***Central Limit Theorem*** asserts that the sample statistic has a normal distribution (or nearly so) with mean μ_s and standard deviation σ_s . Thus, as mentioned before, many statistical decision tests are specifically geared to the normal distribution.
- The standard normal density function has mean zero and unit variance as shown in Illustration 11. The first, second and third standard deviation markers indicate 68.2%, 95% and 99.7% of the population are contained within the respective standard deviation markers.
- The 95% standard deviation (or 2 * σ) markers indicates the ***region of acceptance of the hypothesis for*** $x \in [-1.96, 1.96]$
- Needless to say, the ***region of non-significance for*** $|x| > 1.96$ is the left and the right of the 95% standard deviation (or 2 * σ) markers

- The vertical lines at x = -1.96 and x=1.96 indicate the 95% ***Confidence Interval***
- There are a number of tests for determing ***Confidence Intervals***:
  - Student’s t-test (involving measurements on the mean)
    - used when the sampling distribution is Gaussian but the sample size is small
  - chi-square test (involving tests on the variance)
    - used when the sampling distribution of the test statistic is a chi-squared distribution
  - F-test (involving ratios of variances)
    - used in the analysis of variance (ANOVA) between different models

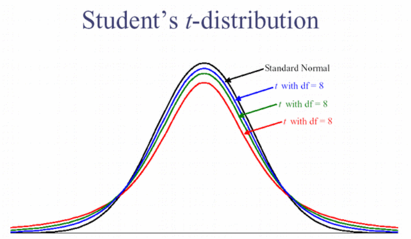
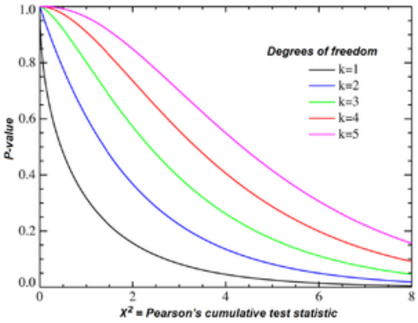
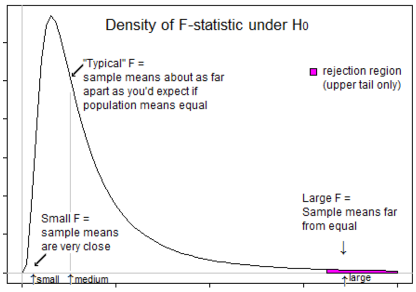
T-Test	Chi-Square Test	F-Test
		

Table 3: Graphical 'taste' of T-Test, Chi-Square test and F-Test

Application: Black-Scholes Equation

Stochastic Differential Equation (SDE) governing the price evolution of a European Call or European Put option under the Black-Scholes Model:

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0$$

where V: price of the option as a function of S
S: Stock price
t: time;
r: risk-free interest rate;
σ: Volitiity of the underlying Stock price

Key insight: one can ***perfectly hedge the option*** by buying and selling the underlying asset in just the right way and consequently ***eliminate risk***.

The hedge in turn implies there is only one right price for the option, as returned by the Black-Scholes formula.

Financial Interpretation

Equation can be rewritten as:

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}=rV-rS\frac{\partial V}{\partial S}$$

The equation can now be broken down into its separate components:

Left Hand Side: ***Time Decay term***

Theta, Θ, (1st Order Greek): measures the sensitivity to the passage of time and is ***loss*** in value due to having less time for exercising the option (-ive) $\Theta=-\frac{\partial V}{\partial t}$

Gamma, Γ,(2nd Order Greek): measures the rate of change in the delta, Δ, wrt the underlying price and is the ***gain*** in holding the option (+ive) $\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}$

Right hand Term: ***Riskless term***

Short (hedging) position rV

Riskless return of ***Long*** position $rS\frac{\partial V}{\partial S}$

Key insight: Loss from theta and gain from gamma offset each other and so the result is a return at the riskless rate.

Other key Greeks: Measures the sensitivity of the value of portfolio to a small change in a given underlying parameter

Delta, Δ, measures the sensitivity to the underlying asset's price $\Delta=\frac{\partial V}{\partial S}$

Vega, ν, measures the sensitivity to volatility $\nu=\frac{\partial V}{\partial \sigma}$

Rho, ρ, measures the sensitivity to the interest rate $\rho=\frac{\partial V}{\partial r}$

Derivation of the Black-Scholes Equation

Step 1: Price of underlying asset follows a geometric Brownian motion:

$$\frac{dS}{S}=\mu dt+\sigma dW$$

This states that the infinitisimal rate of return on the Stock has an expected value of μdt with a standard deviation of σdW .

Step 2: The payoff off of the option, $V(S,T)$, at maturity is known.
To find its ***value at an earlier time we need to know how V evolves as a function of S and t***.
By ***Ito's Lemma*** for two variables we know:

$$dV=(\mu S\frac{\partial V}{\partial S}+\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2})dt+\sigma S\frac{\partial V}{\partial S}dW$$

Step 3: Consider a certain portfolio, called the ***Delta-Hedge Portfolio*** consisting of

- One Short Option, V
- One Long share, S, at time t

The value of the holdings is:

$$\Pi=-V+\frac{\partial V}{\partial S}S$$

Over a time period, $[t,t+\Delta]$, the Profit/Loss from changes in the value of the holdings is:

$$\Delta\Pi=-\Delta V+\frac{\partial V}{\partial S}\Delta S$$

Step 4: Now we discretise the equations for dS/S and dV by replacing differentials with deltas:

$$\Delta S=\mu S\Delta t+\sigma S\Delta W \qquad \Delta V=(\mu S\frac{\partial V}{\partial S}+\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2})\Delta t+\sigma S\frac{\partial V}{\partial S}\Delta W$$

And appropriately sub them into the expression for $\Delta\Pi$:

$$\Delta\Pi=(-\frac{\partial V}{\partial t}-\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2})\Delta t$$

Note: ΔW term has vanished => the ***uncertainty has disappeared*** and the portfolio is now riskless.

Step 5: The rate of return on this porfolio must be equal to the rate of return on any other riskless instrument otherwise there would be an opportunity for arbitrage.
Now, assuming the risk-free rate of return, r, is r we must have over the time period $[t,t+\Delta]$

$$\Delta\Pi=r\Pi\Delta t$$

Step 6: If we now equate our two formulae for ΔΠ we obtain:

$$(-\frac{\partial V}{\partial t}-\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2})\Delta t=r(-V+\frac{\partial V}{\partial S}S)\Delta t$$

Step 7: Simplfying, we arrive at the Black-Scholes SDE:

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0$$

This 2nd order SDE holds for any type of option as long as its price, V, is twice differentiable wrt S and once wrt t. Different formulae for various options will arise from the choice of payoff function of expiry and appropriate boundary conditions.

Solving the Black-Scholes PDE via the Diffusion/Heat Equation

- The SDE is usually solved numerically by using a Finite Difference Method, e.g.:
  - Implicit Finite Difference Method (FDM)
  - Crank-Nicolson Method (2nd Order Implicit FDM)
  - Successive Over Relaxation (SOR) Method
- In some cases, it is possible to solve for an exact formula, such as the case of a European Call Option, as follows:

Step 1: To solve the Black-Scholes PDE for a Call Option recall that the PDE has Boundary Conditions:

- C(0, t) = 0 for all t
- C(S, t) → S as S→ ∞
- C(S, T) = max{S-K, 0} . This is the value of the option at maturity

where K is the option Strike Price.

Step 2: The solution to the SDE gives the value of the option at an earlier time:

$$\boldsymbol{E}[max\{S-K,0\}]$$

Step 3: To solve the SDE we recognise that it is a Cauchy-Euler equation which can be transformed into a ***Diffusion / Heat Equation*** by introducing the Change-Of-Variable transformation:

- $\tau=T-t$
- $u=C\,e^{r\tau}$
- $x=\ln(\frac{S}{K})+(r-\frac{1}{2}\sigma^2)\tau$

Then the Black-Scholes SDE becomes a Diffusion Equation

$$\frac{\partial u}{\partial\tau}=\frac{1}{2}\sigma^2\frac{\partial^2 u}{\partial x^2}$$

The terminal condition C(S, T) = max { S-K, 0} now becomes the Initial Condition:

$$u(x,0)=K(e^{max\{S-K,0\}}-1)=K(e^x-1)H(x)$$

where H(x) is the Heaverside Step Function

Step 4: using the standard Convolution Method for solving the Diffusion Equation given an initial function, u(x, 0), we have:

$$u(x,\tau)=\frac{1}{\sigma\sqrt{(2\pi\tau)}}\int_{-\infty}^{\infty}u_0(y)\exp[\frac{-(x-y)^2}{2\sigma^2\tau}]dy$$

which, after some manipulation, yields

$$u(x,\tau)=K\,e^{x+\frac{1}{2}\sigma^2\tau}N(d_1)-SN(d_2)$$

where N(·) is the standard normal Cumulative Distribution Function (CDF) and

$$d_1=\frac{1}{\sigma\sqrt{(\tau)}}[(x+\frac{1}{2}\sigma^2\tau)+\frac{1}{2}\sigma^2\tau] \qquad d_2=\frac{1}{\sigma\sqrt{(\tau)}}[(x+\frac{1}{2}\sigma^2\tau)-\frac{1}{2}\sigma^2\tau]$$

Step 5: Reverting u, x, τ to the original set of variables yields the follwing solution to the Black-Scholes equation:

$$C(S,t)=N(d_1)S-N(d_2)Ke^{-r(T-t)}$$

$$d_1=\frac{1}{\sigma\sqrt{(T-t)}}[\ln(\frac{S}{K})+(r+\frac{\sigma^2}{2})(T-t)] \qquad d_2=\frac{1}{\sigma\sqrt{(T-t)}}[\ln(\frac{S}{K})+(r-\frac{\sigma^2}{2})(T-t)]$$

Implied Volatility

There are three main assumptions that go into the Black Scholes formula that must be understood:

- Black-Scholes assumes a ***constant volatility*** through the life of the option
- BS assumes no early exercise and is specifically designed for European style options
- The volatility is the standard deviation of log returns and is derived from historical prices of a given security

As can be seen, the price of a European Call option is equal to a probability of an event occurring times the stock price minus a probability of an event occuring times the present value of the exercise price. With this in mind, the only thing that traders need to be aware of is the volatility term, σ^2, in d_1 and d_2. It is worth noting that, as the volitility increases, both d_1 and d_2 increase, thus increasing the price of the option.

This, essentially, is how volatility affects the Black Scholes Model and provides the main practical use of Black-Scholes, especially for volatility traders, computing the ***Implied Volatility***.

Notice that we can obtain the fair value price of the European Call with the 5 Black-Scholes inputs known to us:

- Stock Price (S): can be found on any financial web site
- Strike Price (K): Found on the strike chain for various maturity dates
- Risk free rate (r): Typically the yield on the 3-month treasury bill
- Time (T-t): Known by the expiration date we choose
- ***Volaitility*** (σ^2): calculated using the standard deviation of daily log returns applied to historical stock prices

If we know the market price of the option we can use the Black-Scholes model to compute the ***implied volatility*** of the underlying stock by using techniques, such as Newton's Method for Root Finding. In doing so, we have ***calibrated the model*** and we can use this 'calibrated' model to compute the price of other options based on the underlying stock, S.

It is worth noting:

- I have a particular interest in the Black-Scholes model as I included it in my MBA thesis. Allegedly, it was the main mathematical tool investors were using to price options based on the underlying Russian Ruble which, as it turned out, was the root cause of the contagion from the Russian financial crisis of the 1990s.
- As mentioned, the ***Black-Scholes*** model assumes ***constant volatility***, which, as proven by the mis-adventures in Russia, is incorrect. In fact, the volitility is arbitrary.
- The ***Heston model***, a ***stochastic volitility*** model (***volitility smile*** model), is a closed form solution for pricing models that seeks to overcome some of the short comings presented in the Black-Scholes option pricing model. Other stochastic volitility models include the SABR model, the Chen model and the GARCH model. Studying such models brings us back to Fourier and signal processing techniques as a mathematical tool for analysis.
- Never say never, but at the moment, a fuller discussion of the Heston model, or other stochastic volitility models, is beyond the scope of this document.

Applications

- Finance
  - Option Pricing and make trading decisions
  - Portfolio Management
  - Gambling: Casinos and Betting
- Government
  - Environmental Regulation
  - Entitlement Analysis
  - Financial Regulation
- Politics
  - Surveys
  - Polling
- Image Processing
  - Medical / Biology image processing, e.g. interpretation of X-ray / MRI images
  - Automatic Character Recognition, e.g. zip code
  - Finger print / face / iris recognition
  - Remote sensing / Reconnaissance: aerial and satellite image interpretations
- Science
  - Biology e.g. disease spread
  - Ecology e.g. biological Punnett Squares
  - Statistical Mechanics
  - Quantum Mechanics
- Machine Learning
  - Supervised Learning, e.g. Neural Networks
  - Unsupervised Learning, e.g. Clustering
- Data Analytics
  - Business Analytics
  - Sports Analytics
  - Marketing Analytics
  - Operations Analysis
  - HR Analytics
  - Fraud Detection

References

Papers / Books

Although not directly referenced in the paper, the following texts were used extensively:

Kutz, N. (2013). Data-Driven Modeling & Scientific Computation: Methods for Complex Systems & Big Data

Bishop, M. (2006). Pattern Recognition and Machine Learning

Hastie, T., Tibshirani, R. and Friedman, J. (2008). The Elements of Statistical Learning

Cvitanic, J. and Zapatero, F. (2004). Introduction to the Economics and Mathematics of Financial Markets

Zivot, E. (2016). Computational Finance and Financial Econometrics

Jacquier, A. (2015). Numerical Methods in Finance

Krauth, W. (2006). Statistical Mechanics: Algorithms and Computation

Aspuru-Guzik, A. (2018). The Quantum World

Wikipedia (multiple links)

https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html

Appendix I: Multi-Variate Functions

For reference I have collected here multi-variate functions and constructs which correspond to single variable functions/constructs discussed earlier in this document.

Probability Function

If we have several continuous variables, $x_{1,}\ldots,x_D$, denoted collectively by the vector X then we can define a Joint Probability Function:

$$P(X)=p(x_{1,}\ldots,x_D)$$

Such that the probability of X falling in an infinitesimal volume δX containing the point X is given by:

$$P(X)\delta x$$

This multi-variate density must satisfy:

$$P(X)\geq 0$$

$$\int P(X)dX=1$$

This integral is taken over the whole of X space

Gaussian Distribution

$$N(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}^2)=\frac{1}{(2\pi)}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}))$$

where:

$\boldsymbol{\mu}$: Mean => D-Dimensional Vector

$\boldsymbol{\Sigma}$: Covariance => D x D matrix

$|\boldsymbol{\Sigma}|$: determinant of $\boldsymbol{\Sigma}$ => Scalar

Covariance

In the case of two vectors of Random Variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ the ***Covariance Matrix*** is given by:

$$\boldsymbol{cov}[\boldsymbol{X},\boldsymbol{Y}]=\boldsymbol{E}_{x,y}[\boldsymbol{X}\boldsymbol{Y}^{\boldsymbol{T}}]-\boldsymbol{E}[\boldsymbol{X}]\boldsymbol{E}[\boldsymbol{Y}^{\boldsymbol{T}}]$$

and for ease of notation a vector $\boldsymbol{X}$ has the following Covariance Matrix:

$$\boldsymbol{cov}[\boldsymbol{X},\boldsymbol{X}]\equiv\boldsymbol{cov}[\boldsymbol{X}]$$

Brownian Motion

An n-dimensional process $W_t\{W_t^{(1)},\ldots W_t^{(n)}\}$ is a standard n-dimensional Brownian Motion if each $W_t^{(i)}$ is a Brownian Motion and the $W_t^{(i)}{}'s$ are independent of each other.

Appendix II: The Chain Rule

In calculus, the chain rule is a formula for computing the derivative of the composition of two or more functions. That is, if *f* and *g* are functions, then the chain rule expresses the derivative of their composition *f* ° *g* (the function which maps *x* to *f(g(x))*) in terms of the derivatives of *f* and *g* and the product of functions as follows:

$$(f\circ g)'=(f'\circ g)\cdot g'$$

This may equivalently be expressed in terms of the variable. Let *f* ° *g*, or equivalently, $F(x)=f(g(x))$ for all x. Then one can also write

$$F'(x)=f'(g(x))g'(x)$$

The chain rule may be written in Leibniz's notation in the following way.

If a variable *z* depends on the variable *y*, which itself depends on the variable *x*, so that *y* and *z* are therefore dependent variables, then *z*, via the intermediate variable of *y*, depends on *x* as well.

The chain rule then states:

$$\frac{dz}{dx} = \frac{dz}{dy}\cdot\frac{dy}{dx}$$

The two versions of the chain rule are related; if z = f(y) and y = g(x) then

$$\frac{dz}{dx} = \frac{dz}{dy}\cdot\frac{dy}{dx} = f'(y)g'(x) = f'(g(x))g'(x)$$

In integration, the counterpart to the chain rule is the substitution rule.