



Statistics Assignment 2

HET551 – Design and Development Project 1

Michael Allwright - 5704634
Haddon O'Neill – 5703956
Monday, 13 June 2011



1 Simple Stochastic Processes

1.1 Mean, Variance and Covariance Derivation

The following sections show the derivation of various parameters of the data set Y_n . This data set is defined in Equation 1

$$y_n = \begin{cases} \sum_{i=0}^n X_i & n = 1, 2, \dots \\ 0 & n = 0 \end{cases} \quad (1)$$

1.1.1 Mean Derivation

$$m(n) = E[Y_n]$$

$$m(n) = E \left[\sum_{i=1}^n X_i \right]$$

$$m(n) = n E[X_i]$$

$$m(n) = n\mu$$

1.1.2 Variation Derivation

$$V(n) = \text{Var}(Y_n)$$

$$V(n) = \text{Var} \left(\sum_{i=1}^n X_i \right)$$

Since the variance of sum is the sum of variances for a random variable, this expression can be rewritten as:

$$V(n) = \sum_{i=1}^n \text{var}(X_i)$$

$$V(n) = n\sigma^2$$

1.1.3 Covariance Derivation

$$\text{cov}(Y_L, Y_K) = \text{Cov} \left(\sum_{i=1}^L X_i, \sum_{j=1}^K X_j \right)$$

$$\text{cov}(Y_L, Y_K) = \sum_{i=1}^L \sum_{j=1}^K \text{Cov}(X_i, X_j)$$

Representing this expression in a matrix it can be shown that across the diagonal we have the covariance of a single random variable with itself, and outside the diagonal we the covariance of two independent random variables.

$$\begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_K) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \dots & \text{Cov}(X_2, X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_L, X_1) & \text{Cov}(X_L, X_2) & \dots & \text{Cov}(X_L, X_K) \end{bmatrix} \begin{matrix} i \\ \downarrow \\ L \end{matrix}$$

$j \rightarrow K$

Using the basic rules of Covariance, $\text{Cov}(X, X) = \text{Var}(X)$, and $\text{Cov}(X, Y) = 0$ (for X and Y being independent random variables), the matrix can be simplified as follows

$$\begin{bmatrix} \text{Var}(X_1) & 0 & \dots & 0 \\ 0 & \text{Var}(X_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \text{Var}(X_{L|K}) \end{bmatrix}$$

Since $\text{Var}(X) = \sigma^2$ the $\text{cov}(Y_K, Y_L) = K\sigma^2$ given that $K < L$.

1.2 Covariance Simulation of the Binomial Distribution

1.2.1 Process Simulation for the case of X_i being binomially distributed

To simulate this process, the MATLAB script in Listing 1 was to generate a few trajectories. These trajectories are shown in Figure 1.

```
few = 5; %Few being the amount of trajectories plotted
p=0.25, n=100;
%Graph variables defined
colors = ['b' 'g' 'r' 'c' 'k'];
figure(1), clf(1), axis([0 100 0 40]), xlabel('n Trials'), ylabel('Trial success');
hold on;
for i=1:1:few
    Xi = [0 binornd(1,p,[1,n])];
    Y=[];
    Y=cumsum(Xi);
    stairs(Y,sprintf('%s',colors(i)));
end
hold off;
title(sprintf('Number of successes for Probability of %.3f over 0-%i trials',p,n));
saveas(gcf,sprintf('p=%.3f n=%i.pdf',p,n));
```

Listing 1: MATLAB Source Code to plot of several trajectories of the counting process

1.2.2 Estimation of the Covariance of two variables in the sequence Y_n through Simulation

To estimate the covariance through simulation, a MATLAB script which utilized 100,000 trials was executed. This script is shown in Listings 2, this script relies on the sample covariance function defined in Listings 3. The script returned the value 9.2037 for the values indexed at 50 and 80 in Y_n (arbitrarily chosen).

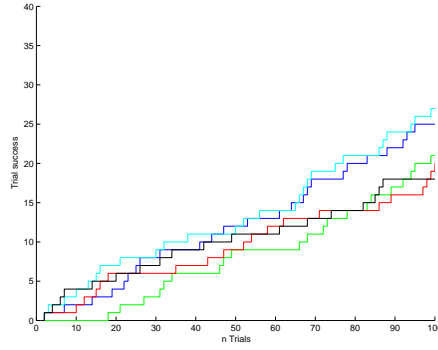


Figure 1: Number of successes for Probability of 0.25 over 0-100 trials

From Section 1.1.3 the analytical expression for Covariance is:

$$\text{cov}(Y_K, Y_L) = K\sigma^2 \text{ given that } K < L.$$

For the Bernoulli case is $\sigma^2 = p(1-p)$, substituting this into the expression yields

$$\text{cov}(Y_K, Y_L) = K\sigma^2 = Kp(1-p)$$

which for $K = 49$ and $L = 80$, results in the analytical calculation of 9.1875, this approximately matches the result from simulation.

1.2.3 Calculation of the Joint Probability of the Expression $P(X_{30} = 5, X_{40} = 7)$

The probability of $S_{n_2} = y_2$ occurring given that $S_{n_1} = y_1$ is given by

$$P(S_{n_1} = y_1, S_{n_2} = y_2) = \binom{n_2 - n_1}{y_2 - y_1} \binom{n_1}{y_1} p^{y_2} (1-p)^{n_2 - y_2}$$

For our chosen probability, that of $p = 0.25$

$$P(Y_{30} = 5, Y_{40} = 7) = \binom{40 - 30}{7 - 5} \binom{30}{5} 0.25^7 (1 - 0.25)^{40 - 7} \approx 0.0295$$

Using the MATLAB script in Listings 4, this analytical result was confirmed by simulation which gave the result $P(Y_{30} = 5, Y_{40} = 7) \simeq 0.0296$.

1.3 Stochastic Process of a Uniformly Distributed Variable

From Equation 1 we know that y_n is a cumulative summation of the randomly distributed variable X_i . To determine the function in Equation 2, we first analytically determine the mean and variance of $y(n)$.

$$\tilde{y}_t(n) = \frac{y(n) - m(n)}{\sqrt{v(n)}} \quad (2)$$

```

%Defining variables
p = 0.25, n = 100, trials = 100000;
Z = zeros(trials,(n+1));
acc = 0;

for i=1:1:trials
    Xi = [0 binornd(1,p,[1,n])];    %Inserts N(0)=0
    Y=[];
    Y=cumsum(Xi);                    %Returns
    if (Y(31)==5 && Y(41)==7)
        acc=acc+1;
    end
    Z(i,:)=Y;
end

%Using cov as Sample Variance
%Cov(Yk, Yl) where k < l
k=49;
l=80;
SampleCov(Z(:,k+1),Z(:,l+1))

%Comparing to Sample Variance
cov(Z(:,k+1),Z(:,l+1))

```

Listing 2: MATLAB Source Code to determine the covariance of two variables in the counting process

```

function [covariance] = SampleCov(x,y)
    acc=0;
    for i = 1:1:size(x,1)
        acc=acc+((x(i)-mean(x))*(y(i)-mean(y)));
    end
    covariance = acc/(size(x,1)-1);
end

```

Listing 3: MATLAB Source Code to Calculate the Sample Covariance

1.3.1 Mean Calculation for the Random Variable $X_i \sim Uniform(0, 2)$

The Probability Density Function (pdf) of a uniform random variable in the range $[0,2]$ is easily defined. It is simply a rectangular area which covers the range $[0,2]$ with a height of $\frac{1}{2}$, This ensures that the area under the pdf is unity.

$$f(x) = \frac{1}{2}, 0 \leq x \leq 2 \text{ otherwise } 0$$

The expectation of X is then defined as:

$$E[X] = \int x f(x) dx$$

Substituting in the derived density function:

$$E[X] = \frac{1}{2} \int_0^2 x dx = \frac{1}{4} [x^2]_0^2 = 1 \quad (3)$$

```

%Finding P(X_n1 = y1, X_n2 = y2)
%P(X_n1 = y1, X_n2 = y2) via calculation
%where
n1=30, y1=5, n2=40, y2=7;

P = nchoosek((n2-n1),(y2-y1))*nchoosek(n1,y1)*p^y2*(1-p)^(n2-y2)

%Defining variables
p=0.25, n=100, trials = 100000, Z= zeros(trials,(n+1)), acc=0;
for i=1:trials
    Xi = [0 binornd(1,p,[1,n])];
    Y=[];
    Y=cumsum(Xi);
    if (Y(31)==5 && Y(41)==7)
        acc=acc+1;
    end
    Z(i,:)=Y;
end

%Results from Simulation,
ProbSim = acc/trials

```

Listing 4: MATLAB code to prove probability through simulation

1.3.2 Variance Calculation for the Random Variable $X_i \sim Uniform(0, 2)$

Using the result in Equation 3, the Variance can now be defined such that:

$$\begin{aligned}
 Var(X_i) &= \int_0^2 (x - E[X])^2 f(x) dx \\
 Var(X_i) &= \frac{1}{2} \int_0^2 (x^2 - 2x + 1) dx = \frac{1}{2} \left[\frac{1}{3}x^3 - x^2 + x \right]_0^2 = \frac{1}{2} \left(\frac{8}{3} - 4 + 2 \right) \\
 Var(X_i) &= \frac{1}{3} = 0.3\bar{3}
 \end{aligned}$$

1.3.3 Distribution of the Stochastic Process

The transformation of the random variable shown in Equation 2 can now be written numerically given $X_i \sim Uniform(0, 2)$.

$$\tilde{y}_t(n) = (y(n) - 1) \sqrt{3} \quad (4)$$

As shown in Figure 2, as n increases there is convergence towards a limiting distribution. From Figure 2, It appears that this distribution is Gaussian in shape, which is expected due to central limit theorem. The central limit theorem states that the summation of independent random variables leads to a Gaussian distribution.

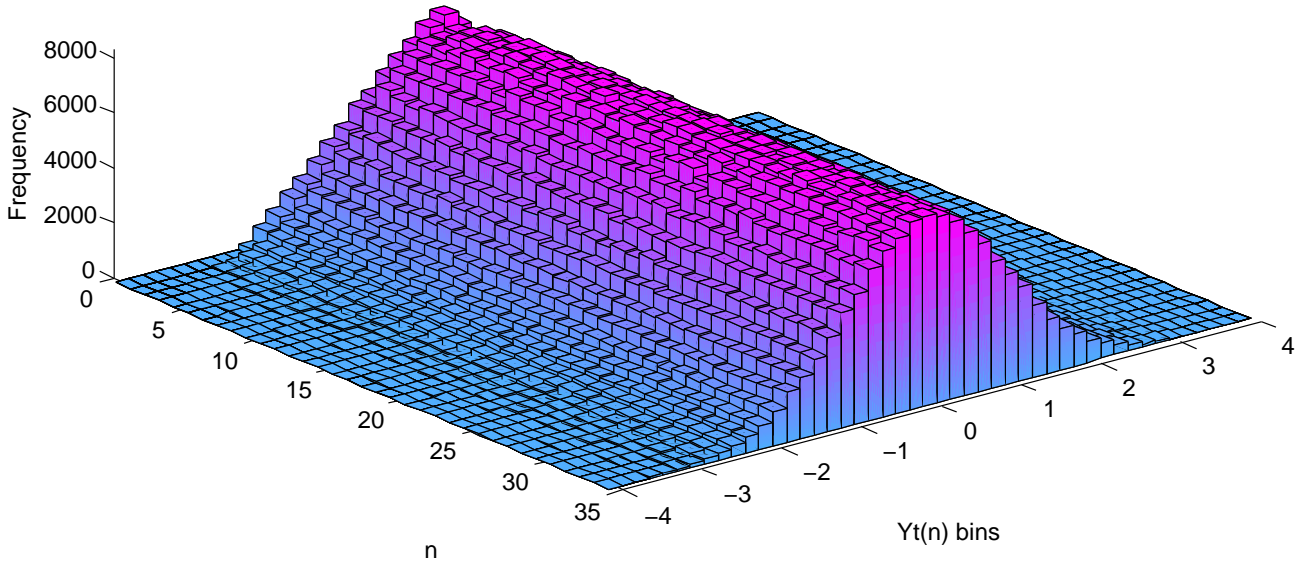


Figure 2: 3D Histogram showing change in distribution as n increases

2 The Multi-Dimensional Normal Distribution

2.1 Probability Density Function for the Bi-variate Normal Distribution

The Probability Density Function (pdf) for the Bi-variate Normal Distribution is as follows:

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_1^2} - \frac{2xy\rho}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right]} \quad (5)$$

2.2 Calculating the Correlation Coefficient

The expression for the covariance is shown in Equation 6. To determine the covariance of the pdf in Equation 5, The mean values μ_x and μ_y are set to zero and the pdf is substituted for the $f(x, y)$ term.

$$\text{Cov}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy \quad (6)$$

$$\text{Cov}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{xy}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} - \frac{2xy\rho}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right]} dx dy \quad (7)$$

To avoid any mistakes in evaluating this integral, this expression was evaluated and simplified using Wolfram's Mathematica Software Package.

To get a clean solution to this integral, some logical assumptions were made about the variables ρ , σ_x and σ_y these are as follows:

$$0 > \rho > 1$$

$$\sigma_x > 0$$

$$\sigma_y > 0$$

Given these assumptions, the covariance of the pdf in Equation 5 is defined by the expression in Equation 8.

$$\text{Cov}(x, y) = \rho \sigma_x \sigma_y \quad (8)$$

Comparing the result in Equation 8 with the relationship for the covariance and correlation coefficient in Equation 9, it shown that ρ is the correlation coefficient.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \quad (9)$$

2.3 Proof that the Bivariate Normal Distribution is a special case of the Multivariate Normal Distribution

The definition of the pdf for a multivariate joint Gaussian distribution is defined by:

$$f_X(x) \triangleq f_{X_1, X_2, \dots, X_n} = \frac{\exp\{-\frac{1}{2}(x - m)^T K^{-1}(x - m)\}}{(2\pi)^{n/2} |K|^{1/2}}$$

Where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad m = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix}, \quad K = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{bmatrix}$$

Substituting X and Y into the covariance matrix for $n = 2$,

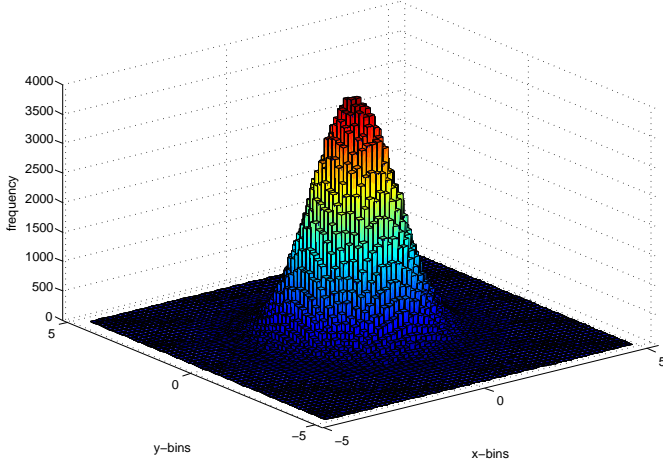
$$K = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{bmatrix}$$

$$K = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}$$

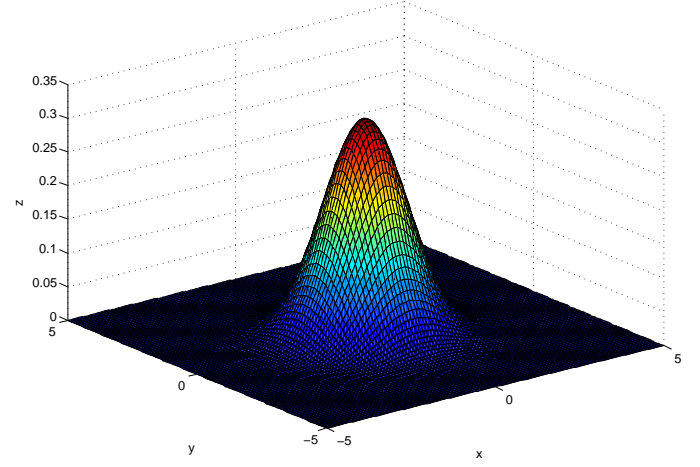
The inverse of the covariance matrix is defined as follows

$$K^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix}$$

Substituting this back into the joint Gaussian multivariate pdf, where the term φ is the exponent, in the term e^φ , such that:



(a) 3D Histogram of the Rayleigh Distribution



(b) pdf of the Bivariate Normal Distribution

Figure 3: Comparison of the Gaussian and Rayleigh Distributions

$$\begin{aligned}\varphi &= \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} (x - m_1, y - m_2) \begin{bmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix} \begin{bmatrix} x - m_1 \\ y - m_2 \end{bmatrix} \\ \varphi &= \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} (x - m_1, y - m_2) \begin{bmatrix} \sigma_2^2 (x - m_1) - \rho \sigma_1 \sigma_2 (y - m_2) \\ -\rho \sigma_1 \sigma_2 (x - m_1) + \sigma_1^2 (y - m_2) \end{bmatrix} \\ \varphi &= \frac{\left(\frac{x - m_1}{\sigma_1}\right)^2 - 2\rho \frac{(x - m_1)}{\sigma_1} \frac{(y - m_2)}{\sigma_2} + \left(\frac{y - m_2}{\sigma_2}\right)^2}{(1 - \rho^2)}\end{aligned}\quad (10)$$

Then m_1 and m_2 are the centering points for the X , Y and $n = 2$ normal distribution, setting the centering points m_1 and m_2 to zero standardizes the expression in Equation 10 giving the standard bivariate normal distribution's exponent from the pdf.

$$\varphi = \frac{\left(\frac{x}{\sigma_1}\right)^2 - \frac{2\rho xy}{\sigma_1 \sigma_2} + \left(\frac{y}{\sigma_2}\right)^2}{(1 - \rho^2)}$$

2.4 Rayleigh Distribution

The Rayleigh method generates independent pairs of normally distributed variables. The MATLAB code in Listing 5, generates these values and plots a 3D histogram showing the distribution for the random process.

For comparison, the code also generates a surface plot of the Probability Density Function for the Bivariate Normal Distribution. The plots for the Rayleigh Histogram and Bivariate Normal pdf surface are shown in Figures 3a and 3b respectively.

As can be from the plots in Figures 3a and 3b, Both of these plots appear to share the Gaussian shape in both dimensions, centered around the same mean.

```

function myrayleigh( sigma, numvals )
    nqz = 1.0000e-015; % not quite zero constant

    angle = random('unif',0,2 * pi, 1, numvals);
    radius = sigma * sqrt(-2 * log(random('unif',nqz,1,1,numvals)));

    [x y] = pol2cart(angle,radius);

    figure(1), clf(1), hist3([x' y'], [60 60]);
    xlabel('x-bins'), ylabel('y-bins'), zlabel('frequency');

    set(gcf,'renderer','opengl');
    set(get(gca,'child'),'FaceColor','interp','CDataMode','auto');

    pause;

    mx=[0 0]'; Cx=[1 0; 0 1]; x=-5:0.1:5;
    for i=1:length(x),
        for j=1:length(x),
            f(i,j)=(1/(2*pi*det(Cx)^1/2))*exp((-1/2)*
                ([x(i) x(j)]-mx')*inv(Cx)*([x(i);x(j)]-mx));
        end
    end
    figure(1), clf(1), surf(x,x,f);
    xlabel('x'), ylabel('y'), zlabel('z');
end

```

Listing 5: MATLAB Source Code to generate a histogram of the Rayleigh Distribution and Standard Bivariate Normal Plane for Comparison.