Computational Methods in Stochastics

Lecture I

Intro
Probability Review
Random variables
Common Distributions
Random Number Generators

Stochastic Modelling

The word "stochastic" derives from the Greek (στοχαζεσθαι to aim, to guess) and means ″random″ or "chance".

A deterministic model predicts a single outcome from a given set of circumstances.

A stochastic model predicts a set of possible outcomes weighted by their likelihoods or probabilities.

Stochastic modelling can be applied also to deterministic states whose outcome is not known (e.g. a hidden tossed coin).

Notation for stochastic processes is really compact: Computing a deterministic equation $x = -A\log(u)$ means computing x from a single value of u, whereas the stochastic equation $X = -A\log(U)$ requires computing from the distribution U.

Stochastic Modelling

Scientific modelling has three components: (i) *a natural phenomenon* under study, (ii) *a logical system* for deducing implications about the phenomenon, and (iii) *a connection* linking the elements of the natural system under study to the logical system used to model it.

Example: (i) The earth with airports, (ii) mathematics of spherical geometry, and (iii) viewing the airports in the physical system as points in the logical system.

The law of large numbers: The relative fraction of times in which an event occurs in a sequence of independent similar experiments approaches, in the limit of an infinite sequence, the probability of the occurrence of the event on any single trial.

Stochastic Modelling

A stochastic process is a family of random variables X_t, where t is a parameter running over a suitable index set T. (Sometimes we write $X(t)$ instead of X_t.)

In a common situation the index t corresponds to discrete units of time, and the index set is $T = \{0, 1, 2, \dots\}$.

Stochastic processes for which $T = [0, \infty)$ are important. t often represents time. It may also represent e.g. distance from an arbitrary origin, and X_t may count the number of defects in the interval $(0, t]$ along a thread, or the number of cars in the interval $(0, t]$ along a highway.

Stochastic processes are distinguished by their state space, or the range of possible values for the random values X_t, by their index set T, and by the dependence relations among the random variables X_t.

Probability Review

For a clear review, see: [Intro of the online book]

Let A and B be **events.**

The event that at least one of A or B occurs: $A \cup B$ (union).

The event that both A and B occur: $A \cap B$, or AB (intersection).

This notation extends to *finite* and *countable* sets of events $A_1, A_2, \ldots$:

At least one event occurs: $A_1 \cup A_2 \cup \cdots = \cup_{i=1}^{\infty} A_i$.

All events occur: $A_1 \cap A_2 \cap \cdots = \cap_{i=1}^{\infty} A_i$.

The probability of an event A: $\Pr\{A\}$.

The *certain* event is denoted by Ω: $\Pr\{\Omega\} = 1$.

The *impossible* event is denoted by $\emptyset$: $\Pr\{\emptyset\} = 0$.

Probability Review

Disjoint events, $A \cap B = \emptyset$, cannot both occur.

The *addition law* for disjoint events: $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$; if events A_i and A_j are disjoint for $i \neq j$, then $\Pr\{\cup_{i=1}^{\infty}\} = \sum_{i=1}^{\infty} \Pr\{A_i\}$.

The law of total probability: Let $A_1, A_2, \ldots$ be disjoint events for which $\Omega = A_1 \cup A_2 \cup \ldots$ (exactly one event will occur). Then $\Pr\{B\} = \sum_{i=1}^{\infty} \Pr\{B \cap A_i\}$ for any event B.

Many important equations and principles are derived from the law of total probability.

Events are said to be *independent* if $\Pr\{A \cap B\} = \Pr\{A\} \times \Pr\{B\}$, or $\Pr\{A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}\} = \Pr\{A_{i_1}\}\Pr\{A_{i_2}\} \cdots \Pr\{A_{i_n}\}$ for every infinite set of distinct indices $i_1, i_2, \ldots, i_n$.

Random Variables

See [Introduction to Probability, Statistics and Random Processes, Chapter 3]()

A useful non-formal definition:
A *random variable* is one that takes on its values by chance.

Random variables are denoted by capital letters, e.g. X, Y, and Z.

Real numbers are denoted by lowercase letters, e.g. x, y, and z.

The expression $\{X \leq x\}$ is the event that the random variable X assumes a value that is less than or equal to the real number x.

The probability that this event occurs is $\Pr\{X \leq x\}$.

Allowing x to vary, this probability defines the *distribution function* (or cumulative distribution function, CDF) of the random variable X as

$$F(x) = \Pr\{X \leq x\}, \qquad -\infty < x < +\infty.$$

Random Variables

Subscripts are used, when several random variables appear in the same context, $F_X(\xi) = \Pr\{X \leq \xi\}$ and $F_Y(\xi) = \Pr\{Y \leq \xi\}$.

It's easy to see that, for example, $\Pr\{X > a\} = 1 - F(a), \Pr\{a < X \leq b\} = F(b) - F(a)$,
and $\Pr\{X = x\} = F(x) - \lim_{\epsilon \downarrow 0} F(x - \epsilon) = F(x) - F(x-)$.

Discrete random variable X: There is a finite or denumerable set of distinct values $x_1, x_2, \ldots$ such that $a_i = \Pr\{X = x_i\} > 0$ for $i = 1, 2, \ldots$ and $\sum_i a_i = 1$.

The *probability mass function* (PMF) for the random variable X:

$$p(x_i) = p_X(x_i) = a_i \qquad \text{for } i = 1, 2, \ldots$$

The relation between the probability mass and distribution function:

$$p(x_i) = F(x_i) - F(x_i-) \text{ and } F(x) = \sum_{x_i \leq x} p(x_i).$$

Random Variables

For a *continuous random variable* X: $\Pr\{X = x\} = 0 \quad \forall x$.

If there is a nonnegative function $f(x) = f_X(x)$ defined for $-\infty < x < \infty$ such that

$$\Pr\{a < X \leq b\} = \int_a^b f(x) \qquad \text{for } -\infty < a < b < \infty,$$

then $f(x) = f_X(x)$ is called the *probability density function* (PDF) for the random variable X.

Then the *cumulative distribution* (CDF) takes the form:

$$F(x) = F_X(x) = \Pr\{X \leq x\} = \int_{-\infty}^{x} f(z)dz\,.$$

$$\Rightarrow f_X(x) = \frac{d}{dx} F_X(x).$$

Random Variables

Then the *continuous distribution function*

$$F(x) = \int_{-\infty}^{\infty} f(\xi)d\xi, \qquad -\infty < x < \infty.$$

If $F(x)$ is differentiable in x, $f(x) = \frac{d}{dx}F(x) = F'(x), \quad -\infty < x < \infty$

$\Longrightarrow \Pr\{x < X \leq x + dx\} = F(x + dx) - F(x) = dF(x) = f(x)dx.$

More precisely, $\Pr\{x < X \leq x + \Delta x + o(\Delta x)\}, \Delta x \downarrow 0.$

$o(\Delta x)$ represents any term for which $\lim_{\Delta x \downarrow 0} o(\Delta x)/\Delta x = 0.$

Random Variables

The *m*th *moment* of a discrete random variable X:

$$E[X^m] = \sum_i x_i^m \Pr\{X = x_i\}$$

If the infinite sum diverges, the moment is said not to exist.

The *m*th *moment* of a continuous random variable X:

$$E[X^m] = \int_{-\infty}^{\infty} x^m f(x) dx$$

(the integral must converge absolutely).

Random Variables

The first moment, $m = 1$, is called the *mean* or *expected value* of X, denoted by m_X or μ_X.

The mth *central moment* of X is defined as the mth moment of the random variable $X - \mu_X$.

The second central moment is called the *variance* of X and written σ_X^2 or Var$[X]$. $\mathrm{Var}[X] = E[(X-\mu)^2] = E[X^2] - \mu^2$.

σ_X is called the *standard deviation*.

The *median* of a random variable X is any value ν such that

$$\Pr\{X \geq \nu\} \geq \frac{1}{2} \text{ and } \Pr\{X \leq \nu\} \geq \frac{1}{2}.$$

Random Variables

$Y = g(X)$ is also a random variable. The expectation of $g(X)$:

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i)\Pr\{X = x_i\}.$$

For continuous X: $\quad E[g(X)] = \int g(x) f_X(x) dx.$

Generally, for both the discrete and continuous cases:

$$E[g(X)] = \int g(x) dF_X(x),$$

where F_X is the distribution function of the random variable X. (Lebesque-Stieltjes integral.)

Given a pair (X, Y) of random variables, their *joint distribution function* is given by

$$F_{XY}(x, y) = F(x, y) = \Pr\{X \leq x \text{ and } Y \leq y\}.$$

Random Variables

A joint distribution is said to possess a (joint) probability density if there exists a function f_{XY} of two real variables for which

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(\xi,\eta) d\eta d\xi \quad \forall x,y.$$

The *marginal distribution functions* of X and Y are $F_X(x) = \lim_{y\to\infty} F_{XY}(x,y)$ and $F_X(x) = \lim_{x\to\infty} F_{XY}(x,y)$, respectively.

The *marginal density functions* are $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(x) = \int_{-\infty}^{\infty} f(x,y)dx$.

If X and Y are jointly distributed, then $E[X+Y] = E[X] + E[Y]$.

Random Variables

The random variables X and Y are said to be *independent* if
$F(x, y) = F_X(x) \times F_Y(y) \quad \forall\, x, y.$

Then the joint density function $f(x, y) = f_X(x) f_Y(y) \;\; \forall\, x, y.$

Given that the jointly distributed X and Y have means μ_X and μ_X and finite variances, the *covariance* of X and Y is

$$\mathrm{Cov}[X, Y] = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

X and Y are said to be *uncorrelated* if $\sigma_{XY} = 0$.

Independent random variables having finite variances are uncorrelated, but the converse is not true; there are uncorrelated random variables that are not independent.

Correlation coefficient: $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$.

Random Variables

The *joint distribution function* of any finite collection $X_1, X_2, \dots, X_n$:

$$F(x_1, \dots, x_n) = F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \Pr\{X_1 \le x_1, \dots, X_n \le x_n)$$

Independence of $X_1, X_2, \dots, X_n$:

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n) \ \forall x_1, \dots, x_n$$

A joint distribution function is said to have a *probability density function* $f(\xi_1, \dots, \xi_n)$ if

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(\xi_1, \dots, \xi_n) d\xi_1, \dots, d\xi_n \ \forall x_1, \dots, x_n \,.$$

Random Variables

For jointly distributed $X_1, X_2, \ldots, X_n$ and arbitrary functions $h_1, \ldots, h_n$ of n variables each, the *expectation* is:

$$E\left[\sum_{j=1}^{m} h_j(X_1, \ldots, X_n)\right] = \sum_{j=1}^{m} E[h_j(X_1, \ldots, X_n)]\,.$$

Random Variables

If X and Y are independent random variables having distribution functions F_X and F_Y, respectively, then the distribution function of their sum $Z = X + Y$ is the *convolution* of F_X and F_Y:

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z-\xi)dF_Y(\xi) = \int_{-\infty}^{\infty} F_Y(z-\eta)dF_X(\eta)\,.$$

Respectively, for probability density functions:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-\xi)f_Y(\xi)d\xi = \int_{-\infty}^{\infty} f_Y(z-\eta)f_X(\eta)d\eta\,.$$

(for nonnegative random variables, replace the lower limit $-\infty$ by 0.)

The *variance*: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$

Random Variables

For all events A and B such that $\Pr\{B\} > 0$, the *conditional probability* of A given B is written

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \qquad \text{if } \Pr\{B\} > 0.$$

This is the *Bayes Theorem.*

If $\Omega = B_1 \cup B_2 \cup \cdots$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, then

$$\Pr\{A\} = \sum_{i=1}^{\infty} \Pr\{A \cap B_i\} = \sum_{i=1}^{\infty} \Pr\{A|B_i\} \Pr\{B_i\}.$$

This is the *law of total probability.*

Discrete Distributions

Bernoulli Distribution

Random variable X has two possible values 0 and 1.

The probability mass function (PMF) $p(1) = p$ and $p(0) = 1 - p$ where $0 < p < 1$.
The mean: $E[X] = p$.
The variance: $\mathrm{Var}[X] = p(1-p)$.

Bernoulli random variables occur frequently as indicators of events. The *indicator of an event* A is the random variable

$$\mathbf{1}(A) = \mathbf{1}_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

$\mathbf{1}_A$ is a Bernoulli random variable with parameter $p = E[\mathbf{1}_A] = \Pr\{A\}$.

Discrete Distributions

Binomial Distribution

Independent events $A_1, A_2, \dots, A_n$ all having the same probability $p = \Pr\{A_i\}$ of occurrence. Let Y count the total number of events among $A_1, A_2, \dots, A_n$ that occur. Then Y has a binomial distribution with parameters n and p.

The probability mass function:

$$p_Y(k) = \Pr\{Y = k\} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \quad \text{for } k = 0,1 \dots, n.$$

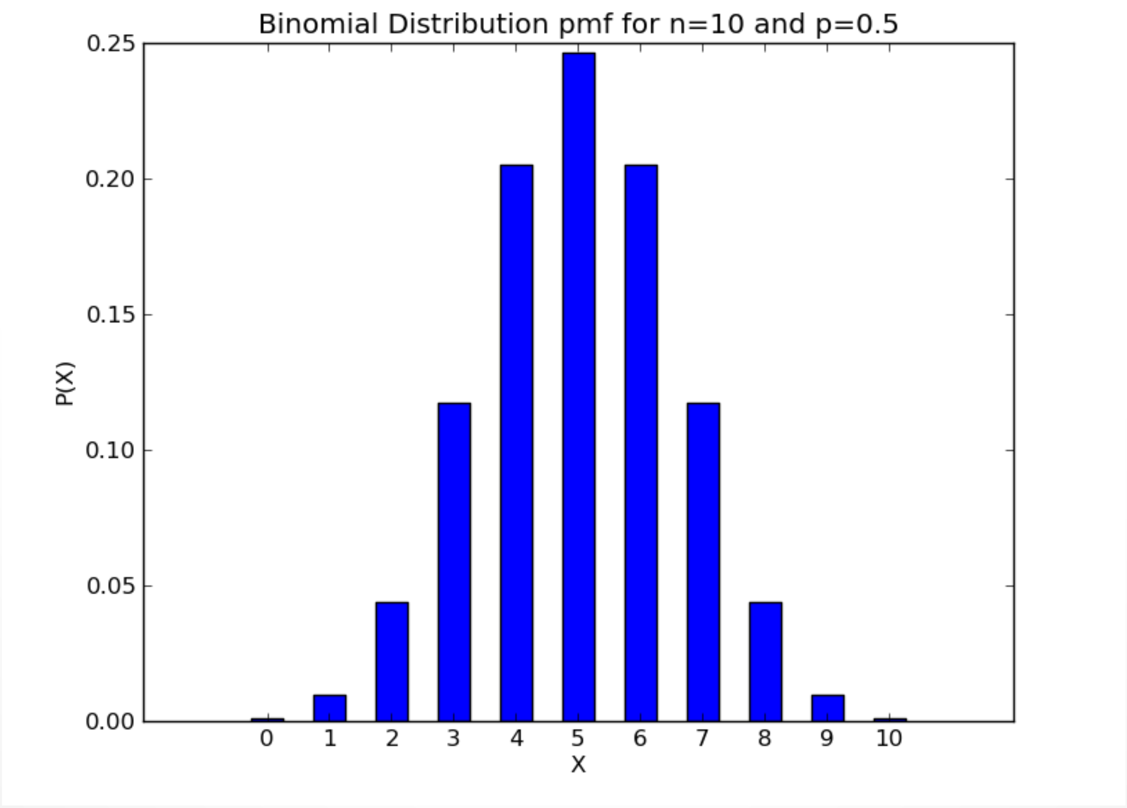
Trick worth remembering: write Y as a sum of indicators $Y = \mathbf{1}(A_1) + \cdots + \mathbf{1}(A_n)$ to determine the moments.

$$E[Y] = E[\mathbf{1}(A_1)] + \cdots + E[\mathbf{1}(A_n)] = np$$

$$\mathrm{Var}[Y] = \mathrm{Var}[\mathbf{1}(A_1)] + \cdots + \mathrm{Var}\,[\mathbf{1}(A_n)] = np(1-p)$$

Discrete Distributions

Notation: $X \sim Bin(n, p)$ means X is a binomial random quantity based on n independent trials, each occurring with probability p.



Discrete Distributions

The Poisson Distribution

The probability mass function of the Poisson distribution with parameter $\lambda > 0$

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \text{for } k = 0, 1, \ldots$$

In calculations related to the Poisson distribution the series expansion for the exponential comes in handy

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots$$

$$\sum_{k=0}^{\infty} kp(k) = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda, \ \sum_{k=0}^{\infty} k(k-1)p(k) = \lambda^2$$

Discrete Distributions

In terms of a Poisson distributed random variable $X \sim Po(\lambda)$:

Mean : $E[X] = \lambda$

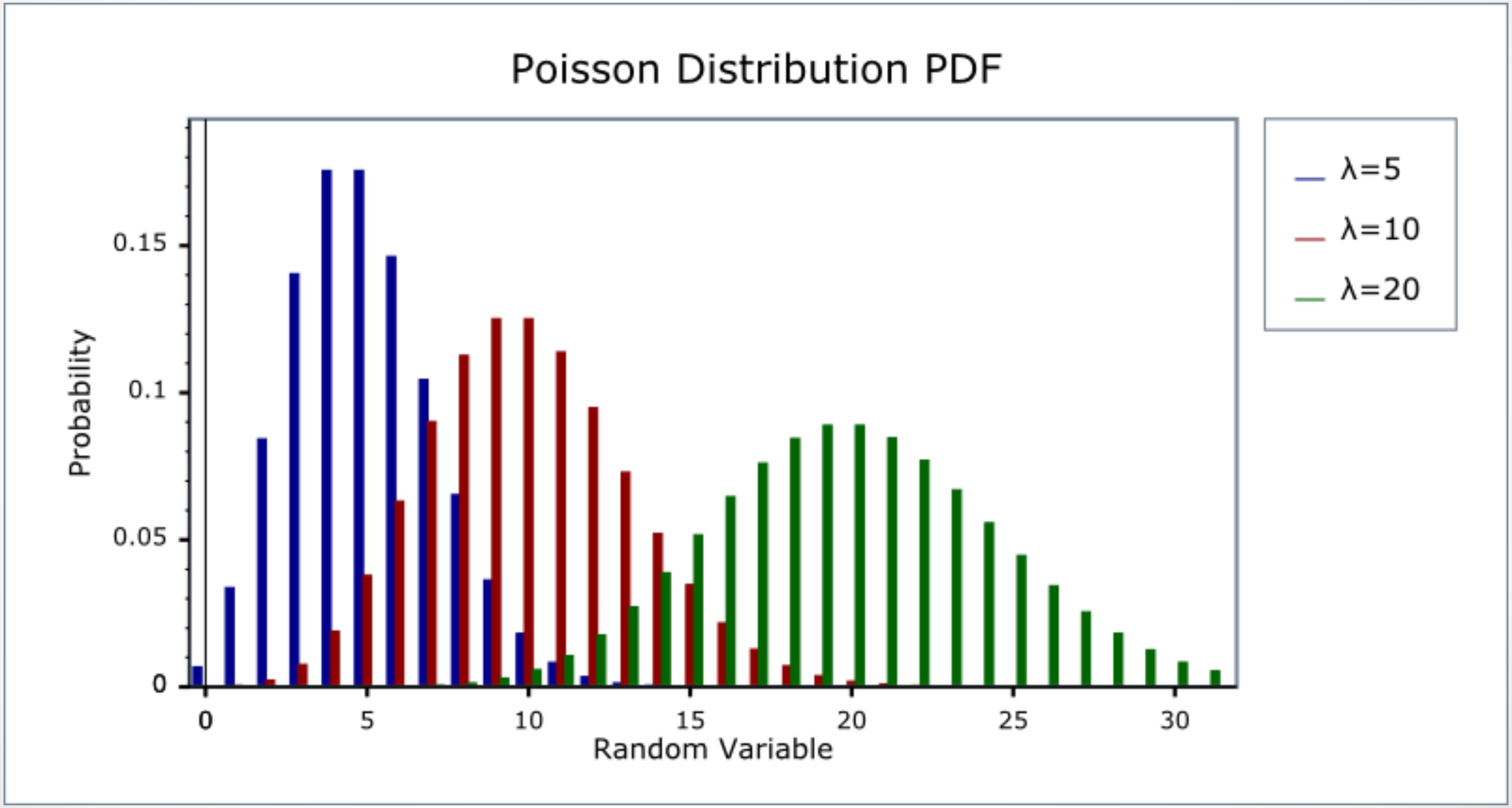
Variance: $\mathrm{Var}[X] = E[X^2] - \{E[X]\}^2 =$

$$= E[X(X-1)] + E[X] - \{E[X]\}^2 = \lambda$$

The simplest form of the *law of rare events:* The binomial distribution with parameter n and p converges to the Poisson distribution with parameter λ if $n \to \infty$ and $p \to 0$ so that $\lambda = np$ remains constant. (See 'An Introduction to Stochastic Modeling' for a proof.)

An important property: The sum of Poisson random quantities is also a Poisson random quantity.

Discrete Distributions



Discrete Distributions

The Multinomial Distribution

= joint distribution of r variables taking nonnegative values $0, \ldots, n$. The joint probability mass function

$$\Pr\{X_1 = k_1, \ldots, X_n = k_n\} = \begin{cases} \dfrac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r} & \text{if } \sum_{i=1}^{r} k_i = n \\ 0 & \text{otherwise} \end{cases}$$

Here, $p_i > 0$ for $i = 1, \ldots, r$ and $\sum_{i=1}^{r} p_i = 1$.

Mean: $E[X_i] = np_i$

Variance: $\mathrm{Var}[X_i] = np_i(1 - p_i)$

Covariance: $\mathrm{Cov}[X_i X_j] = -np_i p_j$

Multinomial is the generalisation of the binomial distribution.

Continuous Distributions

The Normal/Gaussian Distribution $\boldsymbol{N(\mu, \sigma^2)}$

The probability density function

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

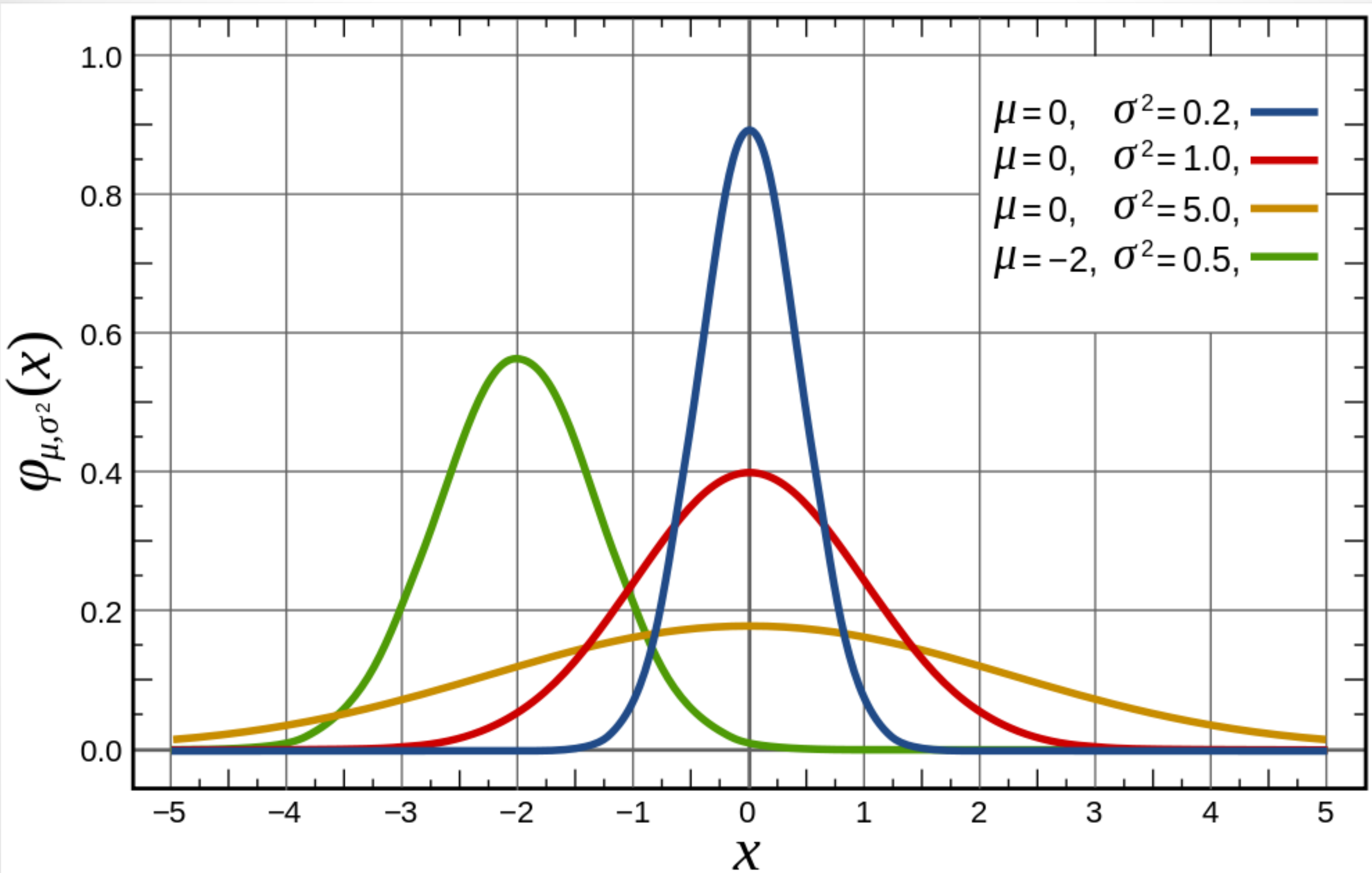
Mean: $E[X] = \mu$

Variance: $\mathrm{Var}[X] = \sigma^2$

Standard normal distribution when $\mu = 0$ and $\sigma = 1$; $N(0,1)$.

Continuous Distributions

The Normal Distribution



Continuous Distributions

[See the online book]

The *central limit theorem:* For partial sums $S_n = \xi_1 + \cdots + \xi_n$ of independent and identically distributed (i.i.d.) summands $\xi_1, \xi_2, \ldots$ having finite means $\mu = E[\xi_k]$ and finite variances $\sigma^2 = \mathrm{Var}[\xi_k]$,

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right\} = \Phi(x) \;\; \forall\, x.$$

Here, $$\Phi(x) = \int_{-\infty}^{x} \phi(\xi)d\xi = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi$$

(the standard normal distribution function)

So, the normal distribution results for numerous small additive, independent ξ, no matter how they are distributed.

Continuous Distributions

Equivalently, the *central limit theorem* for the sample mean $\overline{X}_n = \frac{1}{n} S_n$:

$$\lim_{n \to \infty} \Pr\left\{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right\} = \Phi(x) \ \forall\, x.$$

Continuous Distributions

The Lognormal Distribution

Here the natural logarithm of a nonnegative random variable V is normally distributed.

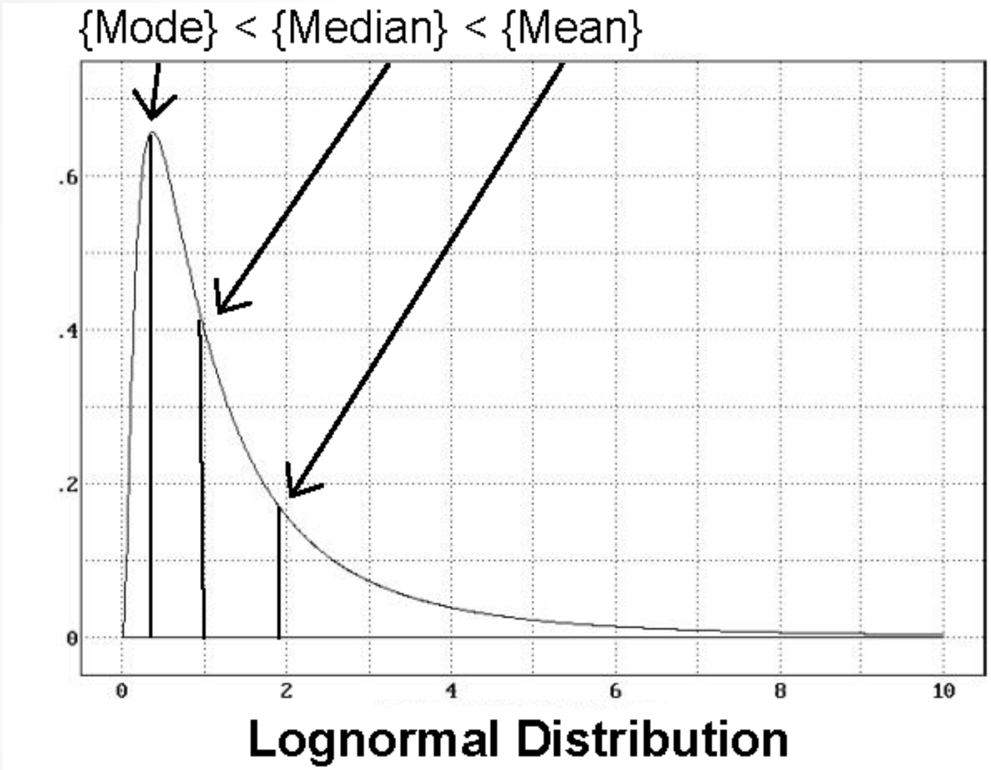
$$f_V(v) = \frac{1}{\sqrt{2\pi}\sigma v} \exp\left\{-\frac{1}{2}\left(\frac{\ln v - \mu}{\sigma}\right)^2\right\}, \qquad v \geq 0.$$

Mean: $$\mathrm{E[V]} = \exp\left\{\mu + \frac{1}{2}\sigma^2\right\}$$

Variance: $\mathrm{Var}[V] = \exp\left\{2\left(\mu + \frac{1}{2}\sigma^2\right)\right\}[\exp\{\sigma^2\} - 1].$

Lognormal distributions arise from *multiplicatively stochastic processes*. Here, the random process is described by a product (instead of a sum) of independent random variables: for a large number of variables the distribution $\ln f_V$ is normal (central limit theorem), so f_V is lognormal.

Continuous Distributions



The lognormal distribution is an example of a *fat-tailed* distribution. It is sometimes erroneously interpreted as a *logarithmic* distribution $p(x) \propto x^{-\alpha},\ x > 0$ and the constant $\alpha > 0$. Logarithmic binning is crucial for visualising and analysis of fat-tailed distributions (Lecture 2).

Continuous Distributions

The Exponential Distribution

The random variable T has an exponential distribution with parameter $\lambda > 0$ ($T \sim Exp(\lambda)$), if the probability density function is

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t \geq 0, \\ 0 & \text{for } t < 0. \end{cases}$$

The distribution function

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & \text{for } t \geq 0, \\ 0 & \text{for } t < 0. \end{cases}$$

Mean: $E[T] = \frac{1}{\lambda}$ Variance: $\mathrm{Var}[T] = \frac{1}{\lambda^2}$

There is an alternative definition for the parameter λ, so be sure to specify which one you are using.

Continuous Distributions

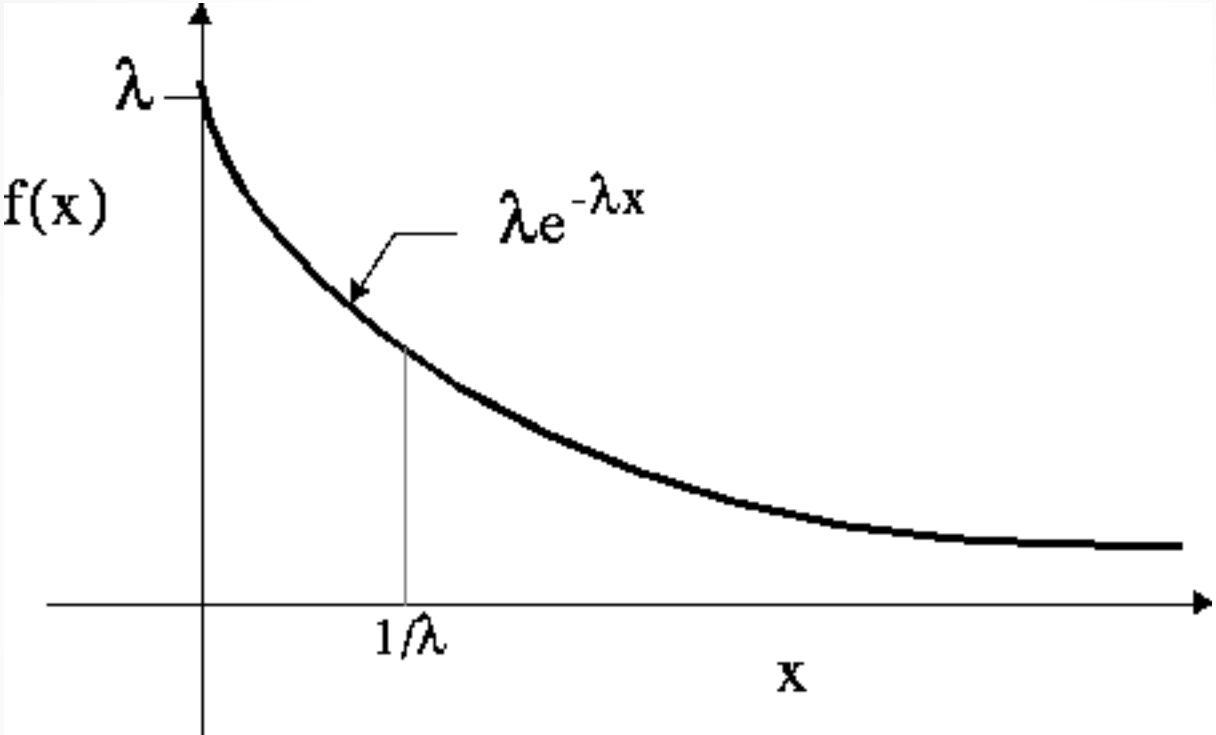


Figure 6. Exponential *pdf*

To see if a distribution is really exponential, plot it in the semilogarithmic coordinates (y-axis logarithmic); you should have a straight line. (For logarithmic distributions, plot both with coordinates logarithmic scales/binning.)

Continuous Distributions

The exponential distribution is encountered in memoryless processes, which is why it is relevant for (continuous) Markov chains. (Markov property: the next state is determined only by the present state.)

T is a lifetime. The unit has survived up to time t. What is the conditional distribution of the remaining life $T - t$?

$$\Pr\{T - t > x | T > t\} = \frac{\Pr\{T > t + x, T > t\}}{\Pr\{T > t\}}$$

$$= \frac{\Pr\{T > t + x\}}{\Pr\{T > t\}} \quad (x > 0)$$

$$= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}$$

Continuous Distributions

The Uniform Distribution

The probability density function for a random variable U distributed uniformly over the interval $[a, b]$, where $a < b$:

$$f_X(u) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le u \le b, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution function

$$F_X(x) = \begin{cases} 0 & \text{for } u \le a, \\ \dfrac{x-a}{b-a} & \text{for } a < x \le b, \\ 1 & \text{for x} > b. \end{cases}$$

Continuous Distributions

Notation: $X \sim U(a, b)$

Mean: $E[X] = \frac{1}{2}(a + b)$ Variance: $\mathrm{Var}[X] = \frac{(b-a)^2}{12}$

The *standard* uniform distribution on the unit interval $[0, 1]$ has $a = 0$ and $b = 1$. A random variable having this distribution is usually denoted by $U \sim U(0,1)$.

$\mathrm{E}(U) = 1/2, \mathrm{Var}(U) = 1/12.$

Standard (pseudo)random number generators implement a random variable uniformly distributed over the interval $(0, 1]$.

Continuous Distributions

The Gamma Distribution

The random variable X has a gamma distribution with parameters $\alpha, \beta > 0$, written $X \sim Ga(\alpha, \beta)$, if it has PDF

$$f(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ is the *gamma function*.

$Ga(1, \lambda) = \text{Exp}(\lambda)$, so the gamma distribution is a generalisation of the exponential distribution.

The mean and variance:

$$E[X] = \frac{\alpha}{\beta}, \qquad \text{Var}[X] = \frac{\alpha}{\beta^2}.$$

Continuous Distributions

A useful property that can be utilised in sampling variates from the gamma distribution is that if $Y = X_1 + X_2$, $X_1 \sim Ga(\alpha_1, \beta)$, and $X_2 \sim Ga(\alpha_2, \beta)$ are independent and $Y \sim Ga(\alpha_1 + \alpha_2, \beta)$, then

$$Y \sim Ga(\alpha_1 + \alpha_2, \beta).$$

Continuous Distributions

Frequently CDF-based quantities are used to characterise a distribution:

Median m: $P(X \le m) = \frac{1}{2}$, or equivalently, $F_X(m) = 0.5$.

Lower quartile l: $F_X(l) = 0.25$

Upper quartile u: $F_X(u) = 0.75$

(The *cumulative distribution* (CDF):

$$F(x) = F_X(x) = \Pr\{X \le x\} = \int_{-\infty}^{x} f(z)dz\,.)$$

Quantifying noise

In stochastic modelling or when analysing data that is "noisy" or stochastic, one needs a measure of how noisy some random variable X is.

One can use the *variance* $\sigma^2 = \mathrm{Var}[X]$, or, for having the noise in the same units as X, the *standard deviation* $\mathrm{SD}[X] = \sigma$.

A consistent way is to give the noise magnitude relative to the random quantity X:

1. The *coefficient of variation* is defined as

$$\mathrm{CV[X]} = \frac{\mathrm{SD}[X]}{|\mathrm{E}[X]|} = \frac{\sigma}{|\mu|}$$

2. *Signal-to-noise ratio* $\mathrm{SNR} = {}^{1}/_{\mathrm{CV}[X]}$ is more commonly used in engineering.

3. The *dispersion index*, or *variance-to-mean ratio*:

$$\mathrm{VMR}[X] = {}^{\mathrm{Var}[X]}/_{\mathrm{E}[X]} = {}^{\sigma^2}/_{\mu}.$$

Random Number Generators

RNGs

Most computational methods use random numbers, for example Monte Carlo, molecular dynamics, stochastic optimization and cryptography. And, of course, all machine learning rests upon massive amounts of random numbers.

Deterministic algorithms. (Pseudorandom numbers.) Pseudorandom number generators (RNG's): Deterministic algorithms that mimic randomness ➔ Generated numbers are only "pseudo-random" but approximate real random numbers reasonably well.

In what follows, the term "random number" means pseudorandom number.

RNGs

Most RNG algorithms produce pseudorandom uniformly distributed numbers $x_i \in (0,1], i = 1, 2, \dots, n$. Roughly, $X \sim U(0, 1)$

These uniformly distributed random numbers can be used further to produce different distributions of random numbers, in other words to *simulate* different distributions.

RNGs

A good RNG has the following *properties*:

The numbers have a *correct distribution*:

- in simulations, the sequence of random numbers must be uncorrelated
- in numerical integration, it is important that the distribution is flat (uniform)

The sequence must have a *long period*:

- all RNGs repeat the same sequence of numbers eventually, but the sequence must be sufficiently long

The sequences should be *reproducible*:

- for testing of simulation parameters
- for stopping a simulation and continuing later

Starting from the same seed number gives you the same sequence → store the seed, when testing.

The RNG must be *fast*. Simulations need loads of them.

Linear congruential generators (LCG)

LCGs are based on the integer recursion relation

$$x_{i+1} = (ax_i + b) \bmod m$$ AND to scale to (0,1): x_{i+1}/m.

where integers a, b, and m are constants.

LCG generates a sequence x_1, x_2, … of random integers that are distributed within the intervals

- $[0, m-1]$ (if $b > 0$) or
- $[1, m-1]$ (if $b = 0$). (This means you need to start with $x_0 \neq 0$ in order to get $x_i \neq 0$.)

Scaling: divide by m for the interval

- $(0,1]$ ($b > 0$) or
- $(0,1)$ ($b = 0$).

Classification:

- mixed ($b > 0$): LCG(a, b, m)
- multiplicative ($b = 0$): MLCG(a, m)

Parameter m determines the *period* P of the LCG (it is usually close to the largest integer of the computer).

A seed x_0 is needed as input.

LCGs

Two standard LCGs:

GGL

MLCG(16807, 2^{31}-1): $x_{i+1} = (16807 x_i) \bmod (2^{31} - 1)$

Available in some numerical software packages such as subroutine RAND in Matlab.

Simple and fast, but suffers from a short period of $2^{31} - 1 \approx 2 \times 10^9$ steps.

Problems due to correlation.

LCGs

RAND

LCG(69069, 1, 2^{32}): $x_{i+1} = (69069 x_i + 1) \bmod (2^{32})$

Also has problems due to correlations.

Visual Test

LCGs have a serious drawback of correlations between consecutive numbers $x_{i+1}, x_{i+2}, \ldots, x_{i+d}$ in the sequence.

In *d*-dimensional space, the points given by these *d* numbers order on parallel *hyperplanes*. The average distance between these planes, whose dimension is $d - 1$, is constant. The smaller the number of these planes, the less uniform is the distribution. In Assignment 1: hyperplanes = lines.

Lagged Fibonacci Generators

These are generalisations of LCGs: The period of a LCG can be increased by the form

$$x_i = (a_1 x_{i-1} + a_2 x_{i-2} + \ldots + a_p x_{i-p}) \bmod m$$

where $p > 1$ and $a_p \neq 0$.
An LF generator requires an initial set of elements x_1, x_2, …, x_r and then uses the integer recursion

$$x_i = (x_{i-r} \otimes x_{i-s}) \bmod m$$

where r and s are two integer lags satisfying $r > s$ and $\otimes$ is one of the following binary operations: +, -, ×, or $\oplus$ (exclusive-or). (To clarify: binary operations, x_i's are integers.)

LFGs

The corresponding generators are termed LF(r, s,$\otimes$).

The initialization requires a set of q random numbers that can be generated for example by using another RNG.

The properties of LF-generators are not very well known but a definite plus is long period. Some evidence suggests that the exclusive-or operation should not be used.

LFGs

RAN3

LF(55,24,-): $x_i = (x_{i-55} - x_{i-24}) \bmod m$

$m = 2^{32}$

- also called a subtractive method
- period $2^{55} - 1$
- initialisation requires 55 numbers
- does not suffer from similar correlations as LCGs

(Assignment 1.)

Shift Register Generators

These can be viewed as the special case $m = 2$ of LF generators.

Feedback shift register algorithms are based on the theory of primitive trinomials

$$P(x;p,q) = x^p + x^q + 1 \qquad (x = 0,1)$$

Given such a primitive trinomial and p initial binary digits, a sequence of bits $b = \{b_i\}$ ($i = 0, 1, 2, \ldots$) can be generated using the following recursion formula:

$$b_i = b_{i-p} \oplus b_{i-p+q}$$

where $p > q$.

Shift Register Generators

Using the recursion formula, random words W_i of size i can be formed by

$$W_i = b_i b_{i+d} b_{i+2d} \cdots b_{i+(l-1)d}$$

where d is a chosen *delay*.

The resulting binary vectors are treated as random numbers.

It can be shown that if p is a *Mersenne prime*, which means that $2^p - 1$ is also a prime, then the sequence of random numbers has a maximal possible period of $2^p - 1$.

If interested, see https://en.wikipedia.org/wiki/Linear-feedback_shift_register

Shift Register Generators

In **generalized feedback shift register** (GFSR) generators, *i*-bit words are formed by a recursion where two bit sequences are combined using the binary operation $\oplus$:

$$W_i = W_{i-p} \oplus W_{i-q}$$

The best choices for q and p are Mersenne primes, which satisfy the condition $p^2 + q^2 + 1 = \text{prime}$.

Examples of pairs that satisfy this condition:

$$\begin{aligned} p &= 98 & q &= 27 \\ p &= 250 & q &= 103 \\ p &= 1279 & q &= 216, 418 \\ p &= 9689 & q &= 84, 471, 1836, 2444, 418 \end{aligned}$$

Generalized feedback shift register generators are denoted by GFSR(p, q, $\oplus$).

Shift Register Generators

R250

R250 for which p = 250 and q = 103 has been the most commonly used generator of this class.

The 32-bit integers (32-bit words)are generated by

$$x_i = x_{i-250} \bigoplus x_{i-103}$$

250 uncorrelated seeds (random integers) are needed to initialize R250.

The latest 250 random numbers must be stored in memory.

The period length is 2^{250} - 1.

Shift Register Generators

R250 does not exhibit similar pair correlations as the LCG generators.

However, R250 has strong triple correlations:

$$\langle x_i \; x_{i-250} \; x_{i-103} \rangle \neq 0$$

In addition, R250 fails in some important physical applications such as random walks and simulations of the Ising model.

An efficient way of reducing correlations is to decimate the sequence by taking only every *k*th number ($k = 3, 5, 7, \ldots$).

Combination Generators

It seems natural that shuffling a sequence or combining two separate sequences might help in reducing correlations.

The combination sequence z_i is defined by $z_i = x_i \otimes y_i$

where x_i and y_i are from some (good) generators and $\otimes$ denotes a binary operation (+, -, ×, $\oplus$).

Combination Generators

RANMAR

- the best known and tested combination generator

The first RNG is a lagged Fibonacci generator

$$x_i = \begin{cases} x_{i-97} - x_{i-33} & \text{if } x_{i-97} \geq x_{i-33} \\ x_{i-97} - x_{i-33} + 1 & \text{otherwise} \end{cases}$$

Only 24 most significant bits are used for single precision reals.
The second part of the generator is a simple arithmetic sequence for the prime modulus 2^{24} - 3 = 16777213.

The sequence is defined as

$$y_i = \begin{cases} y_i - c & \text{if } y_i \geq c \\ y_i - c + d & \text{otherwise} \end{cases}$$

c=7654321/16777216,
d=16777213/16777216

Combination Generators

The final random number z_i is produced by combining x_i and y_i:

$$z_i = \begin{cases} x_i - y_i & \text{if } x_i \geq y_i \\ x_i - y_i + 1 & \text{otherwise} \end{cases}$$

The total period of RANMAR is about 2^{144}.

The code is available in the lectures in MyCourses.

RANMAR is first initialized by the call (in C)
crmarin(seed);
Here, seed is an integer seed.
The call
cranmar(crn,len);
fills the vector crn of length len with uniformly distributed random numbers.

Combination Generators

RANMAR is a very fast generator.

RANMAR has also performed well in several tests, and should thus be suitable for most applications.

RNG Tests

No single test can prove that a RNG is suitable for all applications

It is always possible to construct a test where a given RNG fails (since the numbers are not truly random but generated by a deterministic algorithm).

Classification of test methods

1. Theoretical tests

- based on theoretical properties of algorithms
- exact but often very difficult to perform
- only *asymptotic*: important correlations between consecutive number sets are not measured

RNG Tests

2. Empirical tests

- based on testing algorithms and their implementations in practice
- can be tailored to measure particular correlations
- suitable for all algorithms
- often difficult to say how much testing is sufficient
- further division into *standard tests* (statistical tests) and *application specific tests* (physical quantities)

3. Visual tests

- can be used to locate global or local deviations from randomness
- e.g. pairs of random numbers can be used to plot points in a unit square

Mersenne Twister RNG

In large simulations currently the best and computationally heaviest RNG is the **Mersenne Twister**, see: https://en.wikipedia.org/wiki/Mersenne_Twister

Python uses the Mersenne Twister as the core generator. It produces 53-bit precision floats and has a period of 2**19937-1. The underlying implementation is in C.

In Python you invoke the Mersenne-Twister RNG included in module random (**import random**) by random.random().
See the link: https://github.com/james727/MTP

Using Random Numbers

Example of a simple test

The *moment test* is a simple procedure to check that your RNG implementation works as it should.

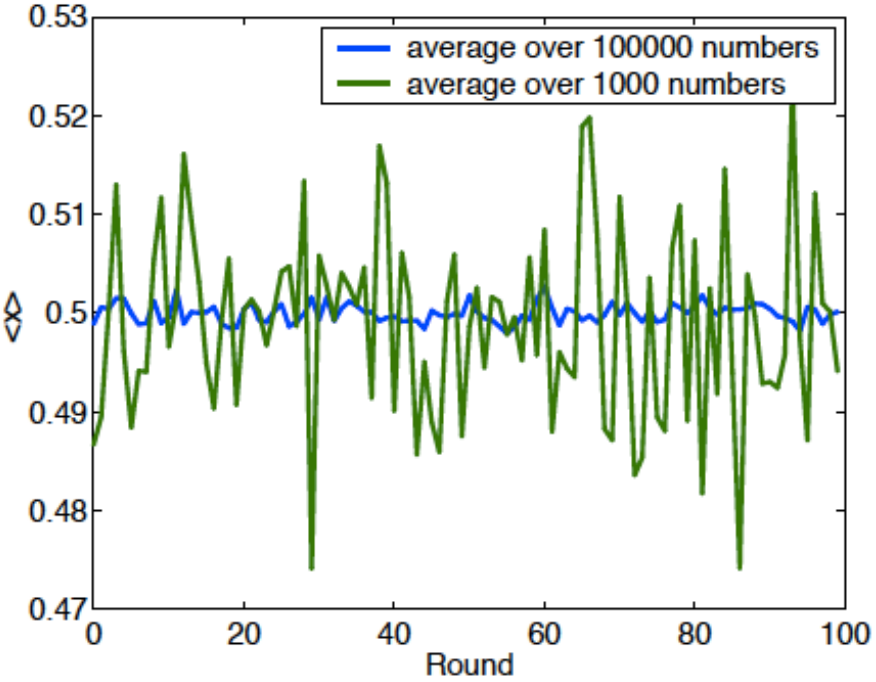
The moments of the uniform distribution are known

$$\langle x^k \rangle = \frac{1}{k+1}$$

If we generate random numbers that should be uniformly distributed, the moments calculated from these numbers should be approximately equal to the analytical values *within statistical fluctuations*.

Using Random Numbers

Example: the mean value of random numbers for 100 independent measurements over $N = 100$ and $N = 1000$ random numbers using RANMAR.



The 'measured' values fluctuate around the correct value 0.5. Error goes as $1/\sqrt{N}$ for uncorrelated random numbers (idealisation).

Using Random Numbers

Central limit theorem

For any independently measured values $M_1, M_2, \ldots, M_m$ that come from the same (sufficiently short-ranged) distribution $p(x)$, the average

$$\langle M \rangle = \frac{1}{m}\sum_{i=1}^{m} M_i$$

will asymptotically follow a Gaussian distribution (normal distribution), whose mean is $< M >$ (equal to the mean of the parent distribution $p(x)$) and standard deviation is $1/\sqrt{N}$ times the standard deviation of $p(x)$.

We can use this result to analyse the errors in the calculated values of any of the moments.

Using Random Numbers

Example

Denote the *second moment* $M = \langle x^2 \rangle$

The errors should follow the normal distribution and the width of this distribution should behave as $1/\sqrt{N}$

Let's take a set of m independent 'measurements' of the second moment, each consisting of an average obtained from N random numbers.

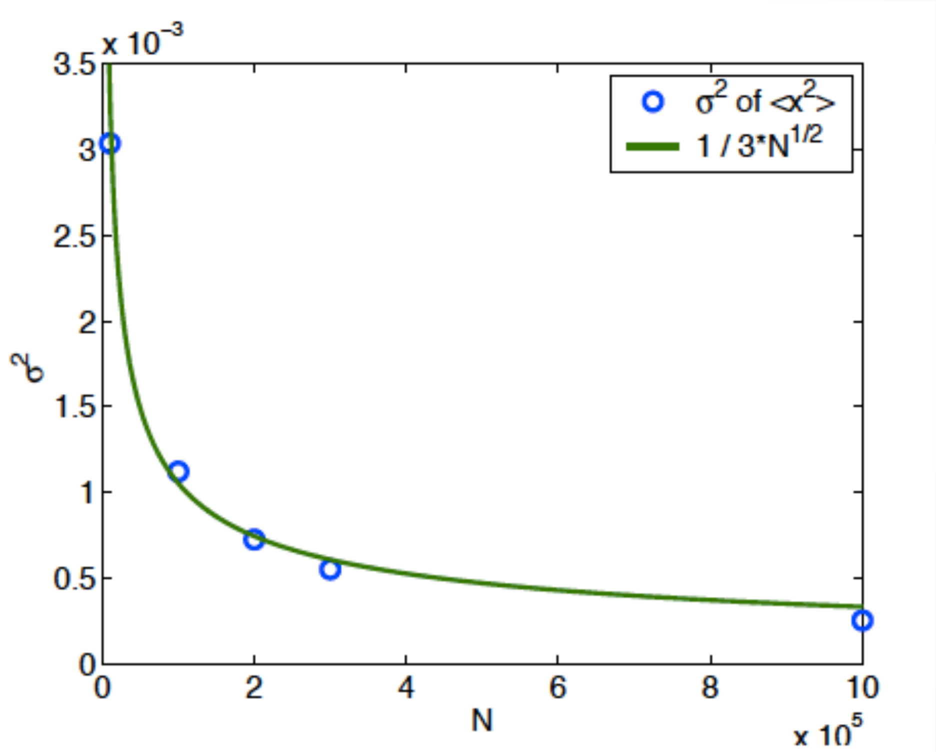
From each measurement we obtain a single value M_α.

The average of all m measurements is $\langle M \rangle = \frac{1}{m} \sum_{\alpha=1}^{m} M_\alpha$
and the variance is given by

$$\sigma^2 = \langle M^2 \rangle - \langle M \rangle^2$$

Using Random Numbers

Here we have the variance of $M = < x^2 >$ obtained from $m = 1000$ measurements, each consisting of an average from N random numbers.



The variance behaves like $1/\sqrt{N}$, meaning that the second moment obeys the scaling of the central limit theorem. Uniformly distributed random numbers from RANMAR.