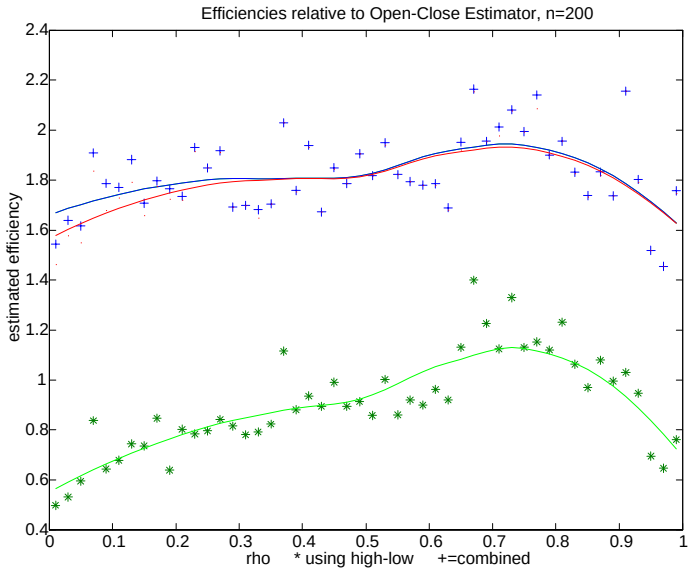

6.4 Estimation using the Term Structure of Interest rates

Largely for accounting reasons, and for the avoidance of arbitrage, financial analysts commonly estimate parameters of a model not from historical data but from some derivative. For example, the volatility parameter of an equity is often derived from the Black-Scholes price of options on the equity and the parameters of a diffusion model for interest rates from the term structure of interest rates. In general, of course, the decision whether to use implied values for parameters of efficient statistical estimators based on historical data is analogous to the choice between the real-world probability measure and the risk-neutral one. If one is interested in valuing derivatives employing a no-arbritrage principle, then the use of the risk-neutral measure is required. On the other hand, if one wishes to model the real-world behaviour of a given process, often statistical estimators, though complicated by the fact that parmaeters may change dynamically over time, provide a better fit to observed data and may be more useful in predicting the future.

It is common, for example, to assume a diffusion model for interest rates that permits time-varying coefficient;

$$dr_t = a(r_t, t)dt + \sigma(r_t, t)dW_t.$$

Consider a 0-coupon bond which, if invested today at time t returns 1\$ at time T. Then, if the current short rate is r_t, the value of this bond can be

written as a function

$$f(r_t, t) = E^Q[exp\{-\int_t^T r_s ds\}]$$

where E^Q denotes expectation under the risk-neutral measure. The *yield curve* describes the current expectations for average interest rates;

$$Yield(T-t) = -\frac{log(f(r_t, t))}{T-t}$$

For a given diffusion model, the function f can be determined by solving the PDE

$$\frac{\partial^2}{\partial t^2} f + a(x,t)\frac{\partial^2}{\partial t \partial x} f + \frac{1}{2}\sigma^2(x,t) f - xf = 0$$

subject to the boundary condition $f(x,T) = 1, \quad all \quad x \in \Re$. The more common models such as the Vasicek, the CIR and the Merton models for interest rate structure are such that the yield curve is *affine* or a linear function of the interest rate. In this case $f(x,t) = exp\{c(T-t)+d(T-t)x\}$ for some functions $c(.),\ d(.)$. Generally this linearity occurs provided that both the drift term and the square of the diffusion coefficient $\sigma^2(x,t)$ are linear in x.

One of the most popular current approaches to interest rates is the *HJM* or Heath-Jarrow-Morton model, which consists of modeling the *instantaneous forward rate* $f(t,u), \quad u \geq t$. Essentially, this function is assumed to follow a multidimensional diffusion. For details, see Chapter 8 of Duffie (1996).

Chapter 7

Miscellany

There are many othe models proposed for financial data that find support in some communities, and the debate about which are appropriate shows no sign of early resolution. The field of *artificial intelligence* offers *Neural Nets*, a locally simple model originally suggested as a design for the brain.

7.1 Neural Nets

A basic premise of much of modern research is that many otherwise extremely complex phenomena are much simpler when viewed locally. On this local scale, structures and organisation are substantially simpler. Complex societies of insects, for example, are organized with very simple interactions. Even differential equations like

$$\frac{dy}{dx} = ay$$

are used to describe the simple local structure of the more complicated exponential function.

Neural Nets are suggested as devices for processing information as it passes through a network. For example binary bits b_1, b_2, b_3 entering a given node j are processed with a very simple processor $g_j(b_1, b_2, b_3)$ which outputs a bit and then transmits it to another node. Thus a neural net consists of a description of the processors (usually simple functions weighted averages), an architecture describing the routing, and a procedure for estimating the parameters (for example the weights in the weighted average). They have the advantage of generality and flexibility- they can probably be modified to handle nearly any problem with some success. However, in specific models for which there are statistically motivated alternatives, they do not usually perform as well as a method designed for that model. Nevertheless, their generality makes them a popular research topic in finance.

7.2 Chaos, Long term dependence and non-linear Dynamics

Another topic, popularized in finance by books by Peters (...) and Gliek (1987), is *chaos*. Chaotic systems are generally purely deterministic systems that may resemble random or stochastic ones. For example if we define a sequence by a recursion of the form $x_t = f(x_{t-1})$ for some non-linear function f, the resulting system may have many of the apparent properties of a random sequence. Depending on the nature of the function f, the seqeunce may or may not appear "chaotic". Compare for example the bahaviour of the above recursion when $f(x) = ax(1-x), \quad 0 < x < 1, \quad , a < 4$ and a is small or a is near 4.

Similarly, the recursion

$$x_t = 1 - ax_{t-1}^2 + bx_{t-2}, \quad a = 1.4, \quad b = 0.3$$

describes a *bivariate* chaotic system, which, like an autoregressive process of order 2, requires two predecessors to define the current value. In general, a system might define x_t as a non-linear function of n predecessors. Then detecting chaos(or lack therof) is equivalent to determining whether the sequences $(x_t, x_{t+1}, \ldots x_{t+n}),\ t = 1, 2,$ fill $n+1$ dimensional space.

Tests designed to test whether a given sequence of stock returns are independent identically distributed generally result in rejecting this hypothesis but the most plausible explanation of this is not so clear. For example Hsieh (1991) tests for both chaotic behaviour and for arch-garch effects (predictable variance changes) and concludes that the latter is the most likely cause of apparent dependence in the data.

7.3 ARCH AND GARCH

One of the first noticable failures in the application of time series models to financial data such as a security price is the failure to adequately represent extended observed periods of high and low volatility. The innovations are supposed in the conventional ARMA models to be independent with 0 mean and constant variance σ^2 and the squared innovations should therefore be approximately independent (uncorrelated) variates.

The time series models discussed so far basically model the expected value of the series given the past observations, assuming that the conditional variance is constant. GARCH, or *Generalized Autoregressive Conditional Heteroscedasticity* takes this one step further, allowing this conditional variance to also be modeled by a time series. In particular, suppose that the innovations in an ARMA model are normally distributed given the past

$$a_t \sim N(0, h_t)$$

where the conditional variance h_t satisfies some ARMA relationship with

the squared innovations posing as the new innovations process.

$$\beta(B)h_t \;=\; \alpha_0 + \alpha(B)a_t^2$$

where $\beta(B) \;=\; 1 - \beta_1 B - \ldots \beta_r B^r$ and $\alpha(B) \;=\; \alpha_1 B + \ldots + \alpha_s B^s$.

The case $r \;=\; 0$ is the original ARCH *Autoregressive Conditional Heteroscedasticity* model, and the most common model takes $r \;=\; 0, s \;=\; 1$ so $h_t \;=\; \alpha_0 + \alpha_1 a_{t-1}^2$. For ARCH and GARCH models the parameters must be estimated using both the models for the conditional mean and the conditional variance and diagnostics apply to both models. The advantages of these models are that they provide both for some dependence among the observations through volatility rather than through the mean, and that they tend to have heavier tails. As a result, they provide larger estimated prices for deep out-of-the-money options, for example, which are heavily dependent on an accurate model for volatility.

7.3.1 ARCH(1)

The basic model investigated by Engle was the simplest case in which the process has zero conditional mean (it is reasonable to expect that the market has removed most or all of this) and but that the squares are significantly autocorrelated. Much financial data exibits this property to some degree. Engles ARCH(1) model is: $x_t \sim N(0, h_t)$ and

$$h_t \;=\; \alpha_0 + \alpha_1 x_{t-1}^2$$

whereas an ARCH regression model allows the conditional mean of x_t in (7.4) to depend on some observed predictors. The GARCH-IN-MEAN process fit by French et. al. allow the mean of x_t to be a function of its variance so that $x_t \sim N(a + bh_t^{p/2}, h_t)$. This would allow testing the hypotheses of relative risk aversion, for example. However, there is little evidence that b may be nonzero, and even less evidence to determine whether the linear relation should be between mean and standard deviation ($p = 1$) or between mean and variance ($p = 2$).

7.3.2 Estimating Parameters

The conditional log likelihood to be maximized with respect to the parameters $\alpha_i,\ \beta_j$ is:

$$log(L) \;=\; -\frac{1}{2}\sum_t [log\ h_t + \frac{\hat{a}_t^2}{h_t}]$$

Various modifications of the above GARCH model are possible and have been tried, but the spirit of the models as well as most of the methodology remains basically the same. There is also a system of Yule-Walker equations that can be

solved for the coefficients β_i in an ARCH model. If γ_i is the autocovariance function of the *innovations squared* a_i^2, then

$$\gamma_n = \sum_{i=1}^{s} \alpha_i \gamma_{n-i} + \sum_{i=1}^{r} \beta_i \gamma_{n-i}$$

for $n \geq r+1$. These provide the usual PAF for identification of the suitable order r of the autoregressive part.

7.3.3 Akaike's Information Criterion

Clearly, a model which leads to small estimated variances for the innovations is preferred, all else being equal, to one with large residual variation. In other words we are inclinded to minimize the estimated residual variance $\frac{1}{N-k}\sum \hat{a}_i^2$ (or equivalently its logarithm) in the selection of the model, where k is the number of autoregressive+moving average parameters in the model. However, such a criterion would encourage the addition of parameters for even a marginal improvement in residual variance, and so a better criterion penalizes against an increase in the number of parameters.

$$AIC = log[\frac{1}{N-k}\sum \hat{a}_i^2] + \frac{2k}{N}$$

The AIC criterion chooses that model which minimizes this quantity. It should be noted that the AIC put out by S is $-2log(L) + 2\times k$ and this is approximately N times the above value. The advantge in multiplying by N is that differences operate on a more natural scale. When nested models are compared (i.e. one model is a special case of the other), differences between values of the statistic $-2log(L)$ have a distribution which is Chi-squared with degrees of freedom the difference in the number of parameters in the two models under the null hypothesis that the simpler model holds.

7.3.4 Testing for ARCH effects

Most of the tests for the adequacy of a given time series model are inherited from regression, although in some cases the autocorrellation of the series induces a different limiting distribution. For example, if there is an ARCH effect, then there should be a significant regression of $\hat{a}_t^2$ on its predecessors $\hat{a}_{t-1}^2$, $\hat{a}_{t-2}^2$, $\hat{a}_{t-3}^2 \ldots$. Suppose we are able to obtain residuals $\hat{a}_l$, $\hat{a}_{l+1}, \ldots \hat{a}_N$ from an ARMA model for the original series. We test for ARCH effect by regressing the vector $(\hat{a}_{l+s}^2, \ldots \hat{a}_N^2)$ on a constant as well as the s "predictors"

$$(\hat{a}_{l+s-1}^2, \quad , \hat{a}_{N-1}^2),\ (\hat{a}_{l+s-2}^2, \quad , \hat{a}_{N-2}^2).\ \ldots (\hat{a}_l^2, \quad , \hat{a}_{N-s}^2)$$

and obtaining the usual coefficient of determination or *squared multiple correlation coefficient* R^2. Standardized, $(N-l)R^2$ has an approximate chi-squared

distribution with s degrees of freedom under the null hypothesis of homoscedasticity so values above the 95'th chi-squared percentile would lead to rejecting the homoscedasticity null hypothesis and concluding arch-like effects. The following table provides the approximate critical values for the chi-squared at the 5% level.

Chi-squared critical values
Conclude ARCH effects if $(N-l)R^2$ exceeds:

s	Critical Value
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

7.3.5 Example. Deutschmark Exchange

The original exchange rate series is given in Figure 7.1. There are approximately 4700 daily observations of the value of the Deutschmark priced in American dollars covering a period from May 16, 1972 to December 31, 1990. Fitting an

Figure 7.1: Deutschmark exchange rate
figure=dm.ps,height=3in,width=5in

AR model by AIC to the first differenced series of DM exchange rates led to an estimate $\hat{\sigma}^2 = 1.41141 \times 10^{-5}$ and an order AR(21) model with the following coefficients:

LAG	AR 21 coefficients	LAG	AR 21 coefficients
1	0.017195644	11	0.007974478
2	0.023371179	12	-0.021073932
3	0.017750192	13	0.026812630
4	-0.011688776	14	0.008615040
5	-0.002804700	15	0.037807230
6	0.023669761	16	-0.019937096
7	0.012745515	17	-0.022599909
8	0.040368907	18	-0.017522413
9	0.016654816	19	0.023648366
10	0.041129731	20	0.034871727

LAG	COEFF
21	-0.024458840

and the following values of the AIC as output from Splus corresponding to lags 1-37,

```
 14.3662109 14.4160156 13.6240234 14.1992188 15.7988281 17.7529297
 16.7841797 17.6533203 12.2275391 12.6582031  6.5322266  8.2451172
 8.5869141  6.9638672  8.5751953  4.5263672  4.5263672  4.3544922
 5.0292969  4.3935547  0.8105469  0.0000000  1.7773438  3.2275391
 4.7246094  5.5156250  5.4570312  7.4218750  9.3662109  6.0908203
 4.2158203  5.9472656  7.8769531  9.8632812 10.3320312 11.3339844
12.4189453
```

The minimum AIC has been subtracted from all values so it corresponds to the order 21 model, and after subtraction, gives tabulated value 0.0000000. These coefficients can be compared with twice the standard error of $1/N^{1/2}$ where $N = 4696$ and this is around 0.028. Since few of the coefficients exceed this value (except the coefficient for lag 10), there is little support for any non-trivial autoregressive effect. Fortunately, because the estimated coefficients are generally small as well, there will also be little difference in forecasting whether we use white noise or the suggested AR(21) model above, since the white noise model corresponds to putting all of the above (small) coefficients equal to 0.

However, if we save the residuals $\hat{a}_t$ from the above time series, and then study the series $\hat{a}_t^2$ we obtain the following:

```
    LAG  Autocorrelation Function  LAG  Autocorrelation Function

    [1,]  0.158922836               [6,]  0.046017088
    [2,]  0.101335056               [7,]  0.021516455
    [3,]  0.077508301               [8,]  0.009644603
    [4,]  0.008419966               [9,]  0.002976760
    [5,]  0.032856703               [10,]  0.105071306
                                    [11,] -0.030951777

  LAG  PARTIAL Autocorrelation  LAG  PARTIAL Autocorrelation

    [1,]  2.079889e-01          [6,]  5.511491e-02
    [2,]  1.328089e-01          [7,]  3.411203e-02
    [3,]  9.759974e-02          [8,]  2.005137e-02
    [4,]  3.005405e-02          [9,]  1.561776e-02
    [5,]  4.940604e-02          [10,]  1.002484e-01
                                [11,] -3.095178e-02
```

Since the AF seems to decay somehat more quickly than the PAF with 3 values significantly non-zero, we might try an ARCH(3) model to describe the process. The large coefficient at lag 10 gives some concern. Does it indicate some sort of biweekly seasonality that we may wish to remove (if it has any reasonable explanation)? If an explanation can be found, then the seasonality can be dealt with by using differences of the form $(1 - B^{10})$.

Diebold and Nerlove, in ARCH Models of Exchange Rate Fluctuations confirm the ARCH effect on the exchange rate for a number of different currencies. However, they observe substantially longer effects (e.g. at lag ≥ 12) although their data is weekly.

Chapter 8

Appendix A: Some Basic Theory of Probability

8.1 Probability Models.

Basic Definitions.

Probabilities are defined on sets or events, usually denoted with capital letters early in the alphabet such as A, B, C. These sets are subset of a *Sample Space or Probability Space* Ω, which one can think of as a space or set containing all possible outcomes of an experiment. We will say that an event $A \subset \Omega$ occurs if one of the outcomes in A (rather than one of the outcomes in Ω but outside of A) occurs. Not only should we be able to describe the probability of individual events, we should also be able to define probabilities of various combinations of them including

1. Union of sets or events $A \cup B =$ A or B (occurs whenever A occurs or B occurs or both A and B occur.)

2. Intersection of sets $A \cap B = A$ and B (occurs whenever A and B occur).

3. Complement : $A^c =$ not A (occurs when the outcome is not in A).

4. Set differences : $A \setminus B = A \cap B^c$ (occurs when A occurs but B does not)

5. Empty set : $\phi = \Omega^c$ (an impossible event-it never occurs since it contains no outcomes)

Recall *De Morgan's rules* of set theory: $(\cup_i A_i)^c = \cap_i A_i^c$ and $(\cap_i A_i)^c = \cup_i A_i^c$

Events are subsets of Ω . We will call $\mathcal{F}$ the class of all events (including ϕ and Ω).

Axioms of Probability

A probability measure is a set function $P : \mathcal{F} \rightarrow [0,1]$ such that

1. $P(\Omega) = 1$
2. If A_k is a disjoint sequence of events so $A_k \cap A_j = \phi, \quad k \neq j$, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Proposition

$P(\phi) = 0$.

Proposition

If $A_k, k = 1, \ldots N$ is a finite sequence of disjoint events so $A_k \cap A_j = \phi, \quad k \neq j$, then

$$P(\cup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i)$$

Proposition

$P(A^c) = 1 - P(A)$

Proposition

Suppose $A \subset B$. Then $P(A) \leq P(B)$.

Proposition

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proposition

(inclusion-exclusion) $P(\cup_k A_k) = \sum_k P(A_k) - \sum \sum_{i<j} P(A_i \cap A_j) + \sum \sum \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots$.

Proposition

$P(\cup_{i=1}^{\infty} A_i) \leq \sum_i P(A_i)$.

Proposition

Suppose $A_1 \subset A_2 \subset \ldots$. Then $P(\cup_{i=1}^{\infty} A_i) = lim_{i \rightarrow \infty} P(A_i)$.

Example.

A coin is tossed twice. List Ω and the class $\mathcal{F}$ of possible events. Define the probability of an event A to be

$$P(A) = \frac{number\ \ of\ \ points\ \ in\ \ A}{number\ \ of\ \ points\ \ in\ \ \Omega}$$

Would this be the correct definition of probability if we defined the sample space using the number of heads observed $\Omega = \{0, 1, 2\}$?

Counting Techniques

Permutations.

The number of ways of arranging n distinct objects in a row is $n! = n(n-1)\ldots 1$ and $0! = 1$. Define $n^{(r)} = n(n-1)\ldots(n-r+1)$ (called "n to r factors") for arbitrary n, and r a non-negative integer. Define $n^{(0)} = 1$.

Example.

How many distinct ways are there of rearranging the 15 letters

$$AAAAABBBBCCCDDE?$$

Example

There are ten students seated at a table of which 5 are Pure Math, and 5 are Impure Math. The organisers are concerned about (intellectual) disputes. How many arrangements are there such that no two pure math students sit together? If the students are seated at random, what is the probability no two pure math students are seated together?

Combinations

Suppose the order of selection is not considered to be important. We wish, for example, to distinguish only different *sets* selected, without regard to the order in which they were selected. Then the number of distinct sets of r objects that can be constructed from n distinct objects is

$$\binom{n}{r} = \frac{n^{(r)}}{r!}$$

Note this is well defined for r a non-negative integer for any real number n .

8.2 Independence and Conditional Probabilities.

Independent Events.

Two events $A,\ B$ are said to be *independent* if

$$P(A \cap B) = P(A)P(B) \tag{2.1}$$

Compare this definition with that of *mutually exclusive or disjoint* events A, B . Events A, B are mutually exclusive if $A \cap B = \phi$.

Independent experiments are often built from *Cartesian Products* of sample spaces. For example if Ω_1 and Ω_2 are two sample spaces, and $A_1 \subset \Omega_1,\ A_2 \subset \Omega_2$ then an experiment consisting of *both of the above* would have sample space the Cartesian product

$$\Omega_1 \times \Omega_2 \ = \ \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

and probabilities of events such as $A_1 \times A_2$ are easily defined, in this case as $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$. Verify in this case that an event entirely determined by the first experiment such as $A = A_1 \times \Omega_2$ is independent of one determined by the second $B = \Omega_1 \times A_2$.

Definition.

A fine or countably infinite set of events $A_1, A_2, \ldots ..$ are said to be mutually independent if

$$P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}) \tag{2.2}$$

for any $k \geq 2$ and $i_1 < i_2 < \ldots i_k$.

Properties.

1. A,B independent implies $A,\ B^c$ independent.
2. Any A_{i_j} can be replaced by $A_{i_j}^c$ in equation (2.2).

Why not simply require that every pair of events is independent?

Example:

Pairwise independence does not imply independence. Two fair coins are tossed. Let A = first coin is heads, B = second coin is heads, C = we obtain exactly one heads. Then A is independent of B and A is independent of C but $A,\ B,\ C$ are **not mutually independent**.

Example

Players A and B decide to play chess until one of them wins. The probability A wins a given game is .3, the probability B wins is .2 and the probability of a draw is .5. What is the probability A wins first?

Lim Sup of events

For a sequence of events $A_n, n = 1, 2, ...$ we define another event $[A_n \ i.o.] = \limsup A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$. Note that this is the set of all points x which lie in infinitely many of the events $A_1, A_2, ...$ The notation i.o. stands for "infinitely often".

Borel Cantelli Lemmas

Clearly if events are individually too small, then there little or no probability that their lim sup will occur, i.e. that they will occur infinitely often.

Lemma 1: For an arbitrary sequence of events A_n , if $\sum_n P(A_n) < \infty$ then $P[A_n \ i.o.] = 0$.

Lemma 2: For a sequence of *independent events* A_n , $\sum_n P(A_n) = \infty$ implies $P[A_n \ i.o] = 1$.

Conditional Probability.

Suppose we are interested in the probability of the event A but we are given some relevant information, namely that another related event B occurred. How do we revise the probabilities assigned to points of Ω in view of this information? If the information does not effect the relative probability of points in B then the new probabilities of points outside of B should be set to 0 and those within B simply rescaled to add to 1.

Definition: Conditional Probability:

For $B \in \mathcal{F}$ with $P(B) > 0$, define a new probability

$$Q(A) \ = \ P(A|B) \ = \ \frac{P(A \cap B)}{P(B)}. \tag{2.3}$$

This is also a probability measure on the same space $(\Omega, \mathcal{F})$, and satisfies the same properties. Note that $P(B|B) = 1, \ \ P(B^c|B) = 0.$

Theorem: Bayes Rule

If $P(\cup_n B_n) = 1$ for a *disjoint* finite or countable sequence of events B_n all with positive probability, then

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)} \tag{2.4}$$

Theorem: Multiplication rule.

If $A_1 \dots A_n$ are arbitrary events,

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 A_1) \dots P(A_n|A_1 A_2 \dots A_{n-1}) \tag{2.5}$$

Example. Diagnostic Testing.

Suppose a blood test for HIV tests positive for 95% of the people who are actually HIV positive and tests negative for 99% of the people who are HIV negative. Suppose the probability that a male is HIV positive is .0001 and the probability that a female is HIV positive is .00005. Assume equal proportions of males and females in the population.

1. Find the probability that a randomly selected person who tested positive on the diagnostic test is indeed HIV positive.
2. Find the probability that a randomly selected male who tested positive on the diagnostic test is indeed HIV positive.

Random Variables and Discrete Distributions

Random Variables

Properties of $\mathcal{F}$.

The class of events $\mathcal{F}$ (called a $\sigma - a\,\mathrm{lg}\,ebra$ or $\sigma - field$) should be such that the operations normally conducted on events, for example countable unions or intersections, or complements, keeps us within that class. In particular it is such that

(a) $Varphi \in \mathcal{F}$

(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(c) If $A_n \in \mathcal{F}$ for all $n = 1, 2, ...,$ then $\cup_{n=1}^{\infty} \in \mathcal{F}$.

It follows from these properties that $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is also closed under countable intersections, or countable intersections of unions, etc.

Definition

Let X be a function from a probability space Ω into the real numbers. We say that the function is *measurable* (in which case we call it a random variable) if for $x \in \Re$, the set $\{\omega; X(\omega) \leq x\} \in \mathcal{F}$. Since events in $\mathcal{F}$ are those to which we can attach a probability, this permits us to obtain probabilities for the event that the random variable X is less than or equal to any number x.

Definition: Indicator random variables

For an arbitrary set $A \in \mathcal{F}$ define $I_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. This is called an *indicator random variable.* (sometimes a *characteristic function* in measure theory, but not here).

Definition: Simple Random variables.

Consider events $A_i \in \mathcal{F}$ such that $\cup_i A_i = \Omega$. Define $X(\omega) = \sum_{i=1}^{n} c_i I_{A_i}(\omega)$ where $c_i \in \Re$. Then X is measurable and is consequently a random variable. We normally assume that the sets A_i are disjoint. Because this is a random variable which can take only finitely many different values, then it is called *simple*. Any random variable taking only finitely many possible values can be written in this form.

Example.

A coin is tossed 10 times. X is the number of heads. Describe $(\Omega, \mathcal{F})$ and the function $X(\omega)$.

Notation

We will often denote the event $\{\omega \in \Omega; X(\omega) \leq x\}$ more compactly by $[X \leq x]$.

Theorem.

If X_1, X_2 are random variables, so is

1. $X_1 + X_2$
2. $X_1 X_2$
3. $min(X_1, X_2)$.

Cumulative Distribution Functions.

Definition.

The *cumulative distribution function* (c.d.f.) of a *Random variable* X is defined to be the function $F(x) = P[X \leq x],\ x \in \Re$.

Properties of C. D. F.

1. A c.d.f. $F(x)$ is non-decreasing. i.e. $F(x) \geq F(y)$ whenever $x \geq y$.
2. $F(x) \to 0$, as $x \to -\infty$.
3. $F(x) \to 1,\ x \to \infty$.
4. $F(x)$ is right continuous. i.e. $F(x) = lim\ F(x + h)$ as h decreases to 0.

There are two types of distributions that we consider in this course, discrete distributions and continuous ones. Discrete distributions are those whose cumulative distribution function at any point x can be expressed as a finite or countable sum of values. For example

$$F(x) = \sum_{i \leq x} p_i$$

for some probabilities p_i which sum to one. In this case the cumulative distribution is piecewise constant, with jumps at the values that the random variable can assume. The values of those jumps are the individual probabilities. For example $P[X = x]$ is equal to the size of the jump in the graph of the c.d.f. at the point x. We refer to the function $f(x) = P[X = x]$ as the *probability function* of the distribution.

Some Special Discrete Distributions

The Discrete Uniform Distribution

Many of the distributions considered so far are such that each point is equally likely. For example, suppose the random variable X takes each of the points $a, a+1, \ldots b$ with the same probability $\frac{1}{b-a+1}$. Then the c.d.f. is

$$F(x) = \frac{x - a + 1}{b - a + 1}, \quad x = a, a+1, \ldots b$$

and the probability function is $f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, \ldots b$ and 0 otherwise.

The Hypergeometric Distribution

Suppose we have a collection (the *population*) of N objects which can be classified into two groups S or F where there are r of the former and $N-r$ of the latter. Suppose we take a random sample of n items without replacement from the population. What is the probability that we obtain exactly x S's?

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}, \; x = 0, 1, \ldots$$

What is the possible range of values of x ? Note that as long as N, R, n, x are integers, this formula gives 0 unless x is in this range. (Note: while attempting to avoid being too judgemental, S above stands for *success* and F for *Failure*)

The Binomial Distribution

The setup is identical to that in the last paragraph only now we sample *with replacement*. Thus, for each member of the sample, the probability of an S is

$p = r/N$. Then the probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,\ 1, \ldots n$$

With any distribution, the sum of *all* the probabilities should be 1. Check that this is the case for the binomial, i.e. that

$$\sum_{x=0}^{n} f(x) = 1.$$

The Hypergeometric distribution is often approximated by the binomial distribution in the case N large. Problem 6 below justifies this approximation. Note that in the case of the binomial distribution, the two *parameters* (constants that one needs to determine the distribution) n, p are fixed, and usually known. For fixed sample size n we have counted X the number of S's in n *trials* of a simple experiment (e.g. tossing a coin).

The Negative Binomial distribution

The binomial distribution was generated by assuming that we repeated trials a fixed number n of times and then counted the total number of successes X in those n trials. Suppose we decide in advance that we wish a fixed number (k) of successes instead, and sample repeatedly until we obtain exactly this number. Then the number of trials X is random.

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots$$

A special case of most interest is the case $k = 1$ called the *Geometric* distribution. Then

$$f(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

The Poisson Distribution.

Suppose that a disease strikes members of a large population (of n individuals) independently, but in each case it strikes with very small probability p . If we count X the number of cases of the disease in the population, then X has the binomial (n, p) distribution. For very large n and small p this distribution can be again approximated as follows:

Theorem. Suppose $f_n(x)$ is the probability function of a binomial distribution with $p = \lambda/n$ for some fixed λ. Then as $n \to \infty$,

$$f_n(x) \to f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for each $x = 0, 1, 2, \ldots$.

The function $f(x)$ above is the probability function of a *Poisson Distribution* named after a French mathematician. This distribution has a single parameter λ, which makes it easier to use than the binomial, since the binomial requires knowledge or estimation of two parameters. For example the size n of the population of individuals who are susceptible to the disease might be unknown but the "average" number of cases in a population of this type λ could be obtained.

Example.

Phone calls arrive at a switchboard at an average rate of one every two minutes. If the operator nips out for a quick drink (5 minutes) what is the probability that there are no calls in this interval? What is the probability that there are more than three calls (in which case the supervisor is alerted).

8.3 Expected Values, Mean, Variances

Expected Value

An indicator random variable I_A takes two values, the value 1 with probability $P(A)$ and the value 0 otherwise. Its expected value, or average over many (independent) trials would therefore be $0(1-P(A))+1P(A)=P(A)$. This is the simplest case of an integral or expectation.

Recall that a simple random variable is one which has only finitely many distinct values c_i on the sets A_i where these sets form a partition of the sample space (i.e. they are disjoint and their union is Ω).

Expectation of simple random Variables.

For a simple random variable $X = \sum_i c_i I_{A_i}$, define $E(X) = \sum_i c_i P(A_i)$. The form is standard:

$$E(X) = \sum (values\ of\ X) \times Probability\ of\ values$$

Thus, for example, if a random variable X has probability function $f(x) = P[X=x]$, then $E(X) = \sum_x x f(x)$.

Properties.

For simple random variables $X,\ Y$,

1. $X(\omega) \le Y(\omega)$ for all ω implies $E(X) \le E(Y)$.
2. For real numbers $\alpha,\ \beta,\ E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proof. Suppose $X = \sum_i c_i I_{A_i} \le \sum_j d_j I_{B_j}$ where A_i forms a disjoint partition of the space Ω (i.e. are disjoint sets with $\cup_i A_i = \Omega$) and B_j also

forms a disjoint partition of the space. Then $c_i \leq d_j$ whenever $A_iB_j \neq \phi$. Therefore

$$E(X) = \sum_i c_i P(A_i) = \sum_i c_i \sum_j P(A_iB_j) \leq \sum_i \sum_j d_j P(A_iB_j) = \sum_j d_j P(B_j) = E(Y)$$

For the second part, note that $\alpha X + \beta Y$ is also a simple random variable that can be written in the form $\sum_i \sum_j (\alpha c_i + \beta d_j) I_{A_iB_j}$ where the sets A_iB_j form a disjoint partition of the sample space Ω. Now take expectation to verify that this equals $\alpha \sum_i c_i P(A_i) + \beta \sum_j d_j P(B_j)$.

Example Find the Expected value of X , a random variable having the Binomial(n, p) distribution.

Expectation of non-negative measurable random variables.

Definition: Suppose X is a non-negative random variable so that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then we define

$$E(X) \;=\; sup\{E(Y);\; Y\; simple,\; Y \leq X\}.$$

Expected value: discrete case.

If a random variable X has probability function $f(x) = P[X = x]$, then the definition of expected value in the case of *finitely many* possible values of x is essentially $E(X) = \sum_x xf(x)$. This formula continues to hold even when X may take a countably infinite number of values provided that the series $\sum_x xf(x)$ is absolutely convergent.

Example.

Find the expected value of a random variable X having the geometric distribution.

Notation.

Note that by $\int_A XdP$ we mean $E(XI_A)$ where I_A is the indicator of the event A.

Properties of Expectation.

Assume $X,\ Y$ are non-negative random variables. Then ;

1. If $X = \sum_i c_i I_{A_i}$ simple, $E(X) = \sum_i c_i P(A_i)$.
2. If $X(\omega) \leq Y(\omega)$ for all ω , $E(X) \leq E(Y)$.

3. If X_n increasing to X, then $E(X_n)$ increases to $E(X)$ (this is usually called the *Monotone Convergence Theorem*).

4. For non-negative numbers $\alpha,\ \beta,\ E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

Proof of Properties.

(1) If $Z \leq X$ and Z is a simple function, then $E(Z) \leq E(X)$. It follows that since X is a simple function and we take the supremum over all simple functions Z, that this supremum is $E(X)$.

(2.) Suppose Z is a simple function $\leq X$. Then $Z \leq Y$. It follows that the set of Z satisfying $Z \leq X$ is a subset of the set satisfying $Z \leq Y$ and therefore the supremum of $E(Z)$ over the former cannot be greater.

(3.) Since $X_n \leq X$ if follows form property (2) that $E(X_n) \leq E(X)$. Similarly $E(X_n)$ is monotonically non-decreasing and it therefore converges. Thus it converges to a limit satisfying

$$limE(X_n) \leq E(X).$$

We will now show that $lim\ E(X_n) \geq E(X)$ and then conclude equality holds above. Suppose $\epsilon > 0$ is arbitrary and $Y = \sum_i c_i I_{A_i}$ where $Y \leq X$ is a simple random variable. Define $E_n = \{\omega; X_n(\omega) \geq (1-\epsilon)Y(\omega)\}$ Note that as $n \to \infty$, this sequence of sets increases to a set containing $\{\omega; X(\omega) \geq (1-\epsilon/2)Y(\omega)\}$ and since $X \geq Y$ the latter is the whole space Ω. Therefore,

$$E(X_n) \geq \int_{E_n} X_n dP \geq (1-\epsilon)\int_{E_n} Y dP.$$

But

$$\int_{E_n} Y dP = \sum_i c_i P(A_i E_n) \to \sum_i c_i P(A_i)$$

as $n \to \infty$. Therefore

$$lim\ \ E(X_n) \geq (1-\epsilon)E(Y)$$

whenever Y is a simple function satisfying $Y \leq X$. Note that the supremum of the right hand side over all such Y is $(1-\epsilon)E(X)$. We have now shown that for any $\epsilon > 0$, $limE(X_n) \geq (1-\epsilon)E(X)$ and it follows that this is true also as $\epsilon \to 0$.

(4) Take two sequences of simple random variables X_n increasing to X and Y_n increasing to Y. Assume α and β are non-negative. Then by property 2. of 4.1.2,

$$E(\alpha X_n + \beta Y_n) = \alpha E(X_n) + \beta E(Y_n)$$

By monotone convergence, the left side increases to the limit $E(\alpha X + \beta Y)$ while the right side increases to the limit $\alpha E(X) + \beta E(Y)$. We leave the more general case of a proof to later.

General Definition of Expected Value.

For an arbitrary random variable X, define $X^+ = max(X, 0)$, $X^- = max(0, -X)$. Note that $X = X^+ - X^-$. Then we define $E(X) = E(X^+) - E(X^-)$. This is well defined even if one of $E(X^+)$ or $E(X^-)$ are equal to ∞ as long as both or not infinite since the form $\infty - \infty$ is meaningless.

Definition.

If both $E(X^+) < \infty$ and $E(X^-) < \infty$ then we say X is *integrable*.

Example:

Define a random variable X such that $P[X = x] = \frac{1}{x(x+1)}, x = 1, 2, ...$.

Is this random variable integrable?

General Properties of Expectation.

In the general case, expectation satisfies 1-4 of 4.1.8 above plus the the additional properties:

1. If $P(A) = 0$, $\int_A X(\omega)dP = 0$
2. If $P[X = c] = 1$ for some constant c, then $E(X) = c$.
3. If $P[X \geq 0] = 1$ then $E(X) \geq 0$. .

Other interpretations of Expected Value

For a discrete distribution, the distribution is often represented graphically with a bar graph or histogram. If the values of the random variable are $x_1 < x_2 < x_3 < \ldots$ then rectangles are constructed around each value, x_i, with *area* equal to the probability $P[X = x_i]$. In the usual case that the x_i are equally spaced, the rectangle around x_i has as base $(\frac{x_{i-1}+x_i}{2}, \frac{x_i+x_{i+1}}{2})$. In this case, the expected value $E(X)$ is the x-coordinate of the center of gravity of the probability histogram.

We may also think of expected value as a long run average over many independent repetitions of the experiment. Thus, $f(x) = P[X = x]$ is approximately the long run proportion of occasions on which we observed the value $X = x$ so the *long run average* of many independent replications of X is $\sum_x xf(x) = E(X)$.

8.4 Discrete Bivariate and Multivariate Distributions

Definitions.

Example.

Suppose we throw 2 dice and define two random variables $X =$ maximum of the two numbers observed and $Y =$ minimum. We wish to record the probability of all possible combinations of values for both X and Y. We may do so through a formula for these joint probabilities

$$P[X = x, Y = y] = f(x, y) = \begin{cases} 2/36 & x > y \\ 1/36 & x = y \\ 0 & x < y \end{cases}$$

for $x, y = 1, 2, \ldots 6$.

Definitions.

The function $f(x, y) = P[X = x, Y = y]$ giving the probability of all combinations of values of the random variables is called the *joint probability function* of X and Y. The function $F(x, y) = P[X \leq x, Y \leq y]$ is called the *joint cumulative distribution function.* The joint probability function allows us to compute the probability functions of both X and Y. For example

$$P[X = x] = \sum_{all\ y} f(x, y).$$

We call this the *marginal* probability function of X, denoted by $f_X(x) = P[X = x] = \sum_{all\ y} f(x, y)$. Similarly, $f_Y(y)$ is obtained by adding the joint probability function over all values of x. Finally we are often interested in the conditional probabilities of the form

$$P[X = x|Y = y] = f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

This is called the *conditional probability function* of X given Y.

Example Determine all marginal probability functions and conditional probability functions in Example 5.1.1.

Expected Values

For a single (discrete) random variable we determined the expected value of a function of X, say $h(X)$ by

$$E[h(X)] = \sum_{all\ x} (value\ of\ h) \times (Probability\ of\ value) = \sum_{x} h(x) f(x)$$

For two or more random variables we should use a similar approach. However, when we add over all cases, this requires adding over all values of x *and* y. Thus, if h is a function of both X and Y,

$$E[h(X,Y)] = \sum_{all\ x\ and\ y} h(x,y)f(x,y).$$

Definition: Independent Random Variables

Two random variables X, Y are said to be *independent* if the events $[X = x]$ and $[Y = y]$ are independent for all $x,\ y$, i.e. if

$$P[X = x, Y = y] = P[X = x]P[Y = y] \quad all \quad x, y$$

i.e. if

$$f(x,y) = f_X(x)f_Y(y) \quad all \quad x, y.$$

This definition extends in a natural way to more than two random variables. For example we say random variables $X_1, X_2, \ldots X_n$ are (mutually) independent if, for every choice of values $x_1, x_2, \ldots x_n$, the events $[X_1 = x_1], [X_2 = x_2], \ldots [X_n = x_n]$ are independent events. This holds if the joint probability function of all n random variables factors into the product of the n marginal probability functions.

Theorem

If X, Y are independent random variables, then

$$E(XY) = E(X)E(Y)$$

Example

Suppose X and Y are two independent random variables with the same distribution (i.e. same probability functions)

$$f_X(x) = (1-p)^x p, \quad x = 0, 1, \ldots$$

and

$$f_Y(y) = (1-p)^y p, \quad y = 0, 1, \ldots$$

where $0 < p < 1$. Find the probability function of $Z = X + Y$ and the conditional probability function $f_{X|Z}(x|z)$.

Definition: Variance

The variance of a random variable measures its variability about its own expected value. Thus if one random variable has larger variance than another, it *tends* to be farther from its own expectation. If we denote the expected value of X by $E(X) = \mu$, then

$$Var(X) = E[(X - \mu)^2].$$

Adding a constant to a random variable does not change its variance, but multiplying it by a constant does; it multiplies the original variance by the constant squared (see 5.1.13, property 2.)

Example

Suppose the random variable X has the binomial (n, p) distribution. Find $E(X)$ and $var(X)$.

Definition:Covariance.

Define the covariance between 2 random variables $X,\ Y$ as

$$cov(X, Y) = E[(X - EX)(Y - EY)]$$

Covariance measures the linear association between two random variables. Note that the covariance between two *independent random variables* is 0. If The covariance is large and positive, there is a tendency for large values of X to be associated with large values of Y . On the other hand, if large values of X are associated with small values of Y , the covariance will tend to be negative. There is an alternate form for covariance, generally easier for hand calculation but more subject to computer overflow problems: $cov(X, Y) = E(XY) - (EX)(EY)$.

Theorem.

For any two random variables $X,\ Y$

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

One special case is of fundamental importance: the case when X, Y are independent random variables and $var(X + Y) = var(X) + var(Y)$ since $cov(X, Y) = 0$.

Example

A population includes a proportion p of unemployed. An interviewer polls members of the population (you may assume with replacement since the population is large) at random until exactly k unemployed have been found and records $X =$ the total number polled. Find the variance of X when $k = 1$ and use this to determine $var(X)$ in general.

Properties of Variance and Covariance

For any random variables X_i and constants a_i

1. $Var(X_1) = cov(X_1, X_1)$.
2. $var(a_1X_1 + a_2) = a_1^2 var(X_1)$.
3. $cov(X_1, X_2) = cov(X_2, X_1)$.
4. $cov(X_1, X_2 + X_3) = cov(X_1, X_2) + cov(X_1, X_3)$.
5. $cov(a_1X_1, a_2X_2) = a_1a_2cov(X_1, X_2)$.

Correlation Coefficient

The covariance has an arbitrary scale factor because of property 5 above. This means that if we change the units in which something is measured, (for example a change from imperial to metric units of weight), the covariance will change. It is desirable to measure covariance in units free of the effect of scale. To this end, define the *standard deviation* of X by $SD(X) = \sqrt{var(X)}$. Then the *correlation coefficient* between X and Y is

$$\rho = \frac{cov(X, Y)}{SD(X)SD(Y)}$$

Theorem

For any pair of random variables X, Y, we have $-1 \leq \rho \leq 1$ with $\rho = \pm 1$ if and only if the points (X, Y) always lie on a line so $Y = aX + b$ for some constants a, b.

The Multinomial Distribution

Suppose an experiment is repeated n times (called "trials") where n is fixed in advance. On each "trial" of the experiment, we obtain an outcome in one of k different categories $A_1, A_2, \ldots A_k$ with the probability of outcome A_i given by p_i. Here $\sum_{i=1}^{k} p_i = 1$. At the end of the n trials of the experiment consider the count of X_i =number of outcomes in category i, $i = 1, 2, \ldots k$. Then the random variables $(X_1, X_2, \ldots X_k)$ have a joint *multinomial* distribution given by the joint probability function

$$P[X_1 = x_1, X_2 = x_2, \ldots X_k = x_k] = \binom{n}{x_1 \quad x_2 \quad \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

whenever $\sum_i x_i = n$ and otherwise this probability is 0. Note that the marginal distribution of each X_i is binomial (n, p_i) and so $E(X_i) = np_i$.

Example

In political poll of 1000 respondents, $X_1 = 400$ indicated that they would vote "Yes" in a referendum, $X_2 = 360$ indicated that they would vote "No" and the remainder were undecided. Write an expression for the probability of exactly this outcome assuming that $p_{no} = p_{yes} = .38$. Find the probability that $X_1 = x$ given that $X_1 + X_2 = y$.

Covariance of a linear transformation.

Suppose $X = (X_1, ..., X_n)'$ is a vector whosse components are possibly dependent random variables. We define the expected value of this random vector by

$$\mu = E(X) = \begin{pmatrix} EX_1 \\ . \\ . \\ . \\ . \\ EX_n \end{pmatrix}$$

and the covariance matrix by a matrix

$$V = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & . & . & cov(X_1, X_n) \\ & & & & \\ & & & & \\ cov(X_n, X_1) & . & . & . & var(X_n) \end{pmatrix}.$$

Then if A is a $q \times n$ matrix of constants, the random vector $Y = AX$ has mean $A\mu$ and covariance matrix AVA'. In particular if $q = 1$, the variance of AX is AVA'.

8.5 Continuous Distributions

Definitions

Suppose a random variable X can take any real number in an interval. Of course the number that we record is often rounded to some appropriate number of decimal places, so we don't actually observe X but $Y = X$ rounded to the nearest $\Delta/2$ units. So, for example, the probability that we record the number $Y = y$ is the probability that X falls in the interval $y - \Delta/2 < X \leq y + \Delta/2$. If $F(x)$ is the cumulative distribution function of X this probability is $P[Y = y] = F(y + \Delta/2) - F(y - \Delta/2)$. Suppose now that Δ is very small and that the cumulative distribution function is piecewise continuously differentiable with a derivative given in an interval by

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

Then $F(y+\Delta/2)-F(y-\Delta/2) \approx f(y)\Delta$ and so Y is a discrete random variable with probability function given (approximately) by $P[Y = y] \approx \Delta f(y)$. The derivative of the cumulative distribution function of X, provided it exists, is called the *probability density function* of the random variable X. Notice that an interval of small length Δ around the point y has approximate probability given by $length\ of\ interval \times f(y)$. Thus the probability of a (small) interval is approximately proportional to the probability density function in that interval, and this is the motivation behind the term *probability density*.

Example.

Suppose X is a random number chosen in the interval $[0,1]$. Any interval of length $\Delta \subset [0,1]$ is to have the same probability Δ regardless of where it is located. Then the cumulative distribution function is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

The probability density function is given by the derivative of the c.d.f. $f(x) = 1$ for $0 < x < 1$ and otherwise $f(x) = 0$. Notice that $F(y) = \int_{-\infty}^{y} f(x)dx$ for all y and the probability density function can be used to determine probabilities as follows;

$$P[a < X < b] = P[a \le X \le b] = \int_a^b f(x)dx.$$

In particular, notice that $F(b) = \int_{-\infty}^{b} f(x)dx$ for all b.

Example.

Is it always true that $F(b) = \int_{-\infty}^{b} F'(x)dx$*?* Let $F(x)$ be the binomial $(n, 1/2)$ cumulative distribution function. Notice that the derivative $F'(x)$ exists and is continuous except at finitely many points $x = 0,\ 1,\ 2,\ 3,\ 4$. Is it true that $F(b) = \int_{-\infty}^{b} F'(x)dx$?

Definition (cumulative distribution function)

Suppose the cumulative distribution function of a random variable $F(x)$ is such that its derivative $f(x) = F'(x)$ exists except at finitely many points. Suppose also that

$$F(b) = \int_{-\infty}^{b} f(x)dx \tag{6.1}$$

for all $b \in \Re$. Then the distribution is called *(absolutely) continuous* and the function $f(x)$ is called the *probability density function*.

Example.

Is it really necessary to impose the additional requirement (6.1) or this just a consequence of the fundamental theorem of calculus? Consider the case $F(x) = 0, x < 0,$ and $F(x) = 1,\ x \geq 0.$ This cumulative distribution function is piecewise differentiable (the only point where the derivative fails to exist is the point $x = 0$). But is the function the integral of its derivative?

For a continuous distribution, probabilities are determined by integrating the probability density function. Thus

$$P[a < X < b] = \int_a^b f(x)dx \tag{6.2}$$

A probability density function is not unique. For example we may change $f(x)$ at finitely many points and it will still satisfy (2) above and all probabilities, determined by integrating the function, remain unchanged. Whenever possible we will choose a continuous version of a probability density function, but at a finite number of discontinuity points, it does not matter how we define the function.

Properties of a Probability Density Function

1. $f(x) \geq 0$ for all $x \in \Re$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

The Uniform Distribution.

Consider a random variable X that takes values with a continuous uniform distribution on the interval $[a, b]$. Then the cumulative distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

and so the probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ and elsewhere the probability density function is 0. Again, notice that the definition of f at the points a and b does not matter.

Example

Let U have a continuous uniform distribution on the interval $[0, 1]$. Define the random variable $X = ln(1/U)$. Find the cumulative distribution function of X and determine whether it is absolutely continuous. If so, find its probability density function.

Expected Values for Continuous Distributions.

Suppose we were to approximate a continous random variable X having probability density function $f(x)$ by a a discrete random variable Y obtained by rounding X to the nearest Δ units. Then the probability function of Y is

$$P[Y = y] = P[y - \Delta/2 \leq X \leq y + \Delta/2] \approx \Delta f(y)$$

and its expected value is

$$E(Y) = \sum_y yP[y - \Delta/2 < X \leq y + \Delta/2] \approx \sum_y y\Delta f(y).$$

Note that as the interval length Δ approaches 0, this sum approaches the integral

$$\int xf(x)dx$$

and thus we define, for *continuous random variables*

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and for any function on the real numbers $h(x)$,

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

Example

Find the expected value and the variance of a random variable having probability density function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

The Exponential Distribution.

Consider a random variable X having probability density function

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x > 0$$

The cumulative distribution function is given by

$$F(x) = 1 - e^{-x/\mu}$$

and the moments are

$$E(X) = \mu, \quad var(X) = \mu^2$$

Such a random variable is called the *exponential distribution* and it is commonly used to model lifetimes of simple components such as fuses, transistors, etc that are not subject to wear and tear.

Example

If the lifetime of a transistor is exponentially distributed, find the probability that it will survive at least t more months given that it has already survived for x months. Compare this with the probability that a new transistor lives for at least t months.

Note: this property is called the memoryless property of the exponential distribution. A component with this distribution of liffetimes does not exhibit any evidence of aging.

Example

Show that for a uniform[a,b] random variable X, we have $E(X) = \frac{a+b}{2}$ and $var(X) = \frac{(b-a)^2}{12}$.

Generating Random variables with an Exponential Distribution.

Suppose that a computer has a built-in generator for the uniform[0,1] distribution (as is the case for nearly every higher-level computer language). How could I use a Uniform random variable U to generate an exponential random variable? Let $X = -\mu \ln(1-U)$. Find the cumulative distribution function of X.

Two methods for Computer Generation of Random Variables.

By far the simplest and most common method for generating non-uniform variates is based on the inverse cumulative distribution function. For arbitrary c.d.f. $F(x)$, define $F^{-1}(y) = min\ \{x; F(x) \geq y\}$. This defines a pseudo-inverse function which is a real inverse (i.e. $F(F^{-1}(y)) = F^{-1}(F(y)) = y$) only in the case that the c.d.f. is continuous and strictly increasing. However, in the general case of a possibly discontinuous non-decreasing c.d.f. the function continues to enjoy some of the properties of an inverse. In particular, in the general case, *If* F *is an arbitrary c.d.f. and* U *is uniform*$[0,1]$ *then* $X = F^{-1}(U)$ *has c.d.f.* $F(x)$.

Example: Generating a discrete random variable

Consider generating, using U a uniform$[0,1]$ random variable, a random variable X having the following probability function:

x	1	2	3	4	5	6
P[X=x]	0.1	0.3	0.2	0.1	0.1	0.2

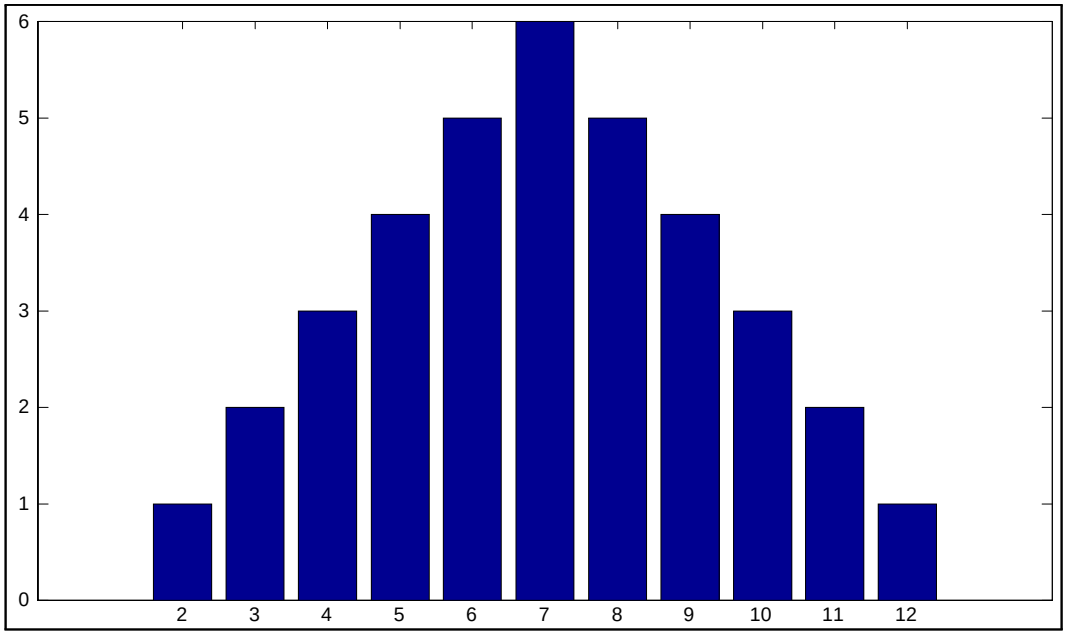

Figure 8.1:

Example: Generating a Geometric (p) random variable.

In this case, the c.d.f. is given by $F(x) = 1-(1-p)^{[x]}$, $x \geq 0$ where $[x]$ denotes the integer part of x. Then

$$X = 1 + [\frac{log(1-U)}{log(1-p)}] \quad or \quad 1 + [\frac{-E}{log(1-p)}]$$

where E is exponential(1) generates a geometric random variable. Compare the efficiency of this generator with one defined by $X = \min\{N; U_N < p\}$ where $U_1, U_2, ...$ are independent uniform[0,1] random variables.

The rejection method is useful if the density g is considerably simpler than f both to evaluate and to generate distributions from and if the constant c is close to 1. The number of iterations through the above loop until a point satisfies the condition has a geometric distribution with parameter $p = 1/c$ and mean c so when c is large, the rejection method is not very effective.

The Normal distribution

Normal Approximation to the Poisson distribution

Consider a random variable X which has the Poisson distribution with parameter μ . Recall that $E(X) = \mu$ and $var(X) = \mu$ so $SD(X) = \sqrt{\mu}$. We

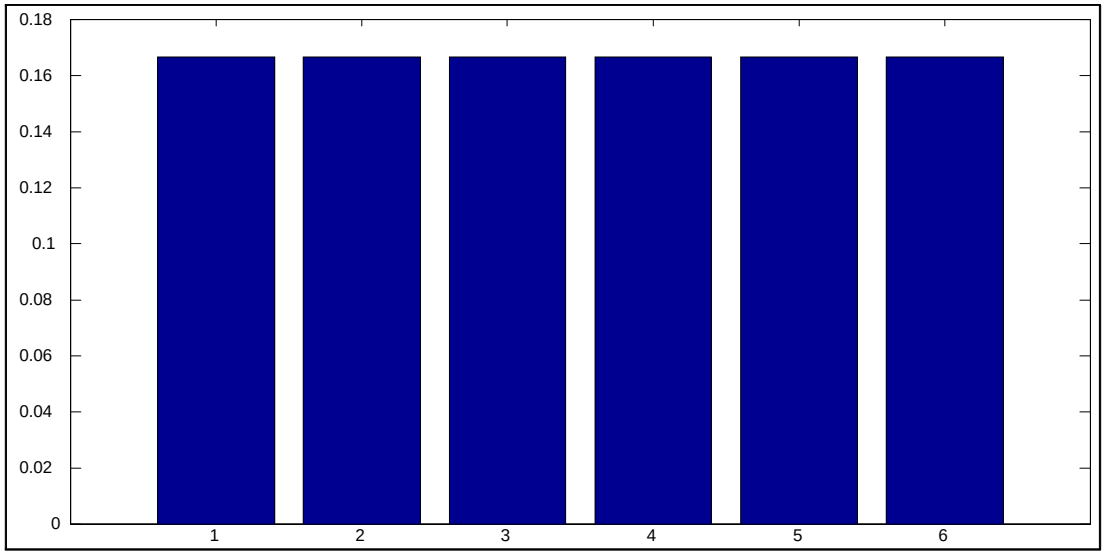

Figure 8.2:

wish to approximate the distribution of this random variable for large values of μ. In order to prevent the distribution from disappearing off to $+\infty$, consider the standardized random variable

$$Z = \frac{X - \mu}{\sqrt{\mu}}.$$

Then $P[Z = z] = P[X = \mu + z\sqrt{\mu}] = \frac{\mu^x}{x!}e^{-\mu}$ where $x = \mu + z\sqrt{\mu}$ is an integer. Using Stirling's approximation $x! \sim \sqrt{2\pi x}x^x e^{-x}$ and taking the limit of this as $\mu \to \infty$, we obtain

$$\frac{\mu^x}{x!}e^{-\mu} \sim \frac{1}{\sqrt{2\pi\mu}}e^{-z^2/2}$$

where the symbol $\sim$ is taken to mean that the ratio of the left to the right hand side approaches 1.

The standard normal distribution

Consider a continuous random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad -\infty < x < \infty$$

Such a distribution we call the *standard normal distribution* or the $N(0,1)$ distribution. The cumulative distribution function

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$$

is not obtainable in simple closed form, and requires either numerical approximation or a table of values. The probability density function $f(x)$ is symmetric about 0 and appears roughly as follows:

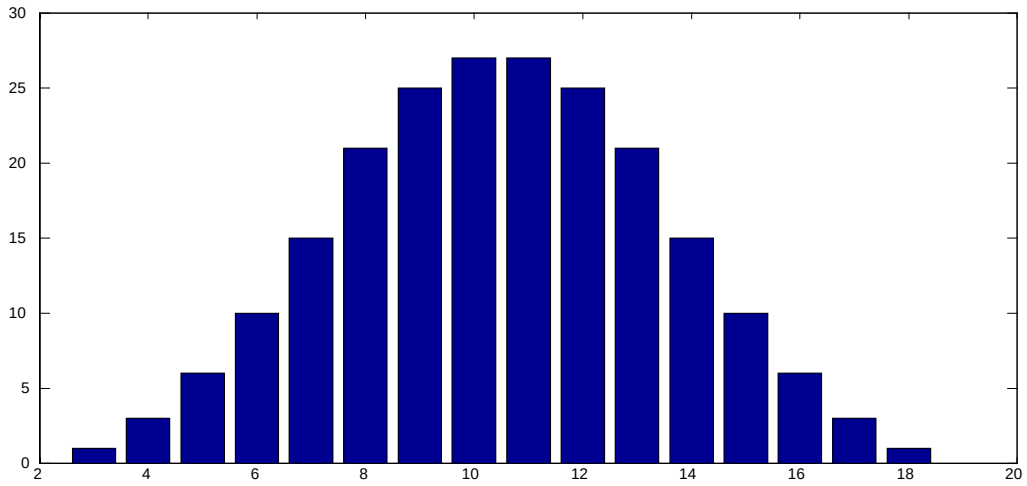

Figure 8.3: Standard Normal Probability Density Function

Example.

Prove that the integral of the standard normal probability density function is 1. The normal cumulative distribution function is as given below:

Note, for example that $F(-x) = 1 - F(x)$ for all x and if Z has a standard normal distribution

$$P[-1 < Z < 1] \approx .68 \quad \textit{and} \quad P[-2 < Z < 2] \approx .95.$$

Example.

If $Z \sim N(0,1)$ find $P[Z^2 \leq 3.84]$.

The General Normal Distribution.

If we introduce a shift in the location in the graph of the normal density as well as a change in scale, then the resulting random variable is of the form

$$X = \mu + \sigma Z, \quad Z \sim N(0,1)$$

for some constants $-\infty < \mu < \infty, \quad \sigma > 0.$

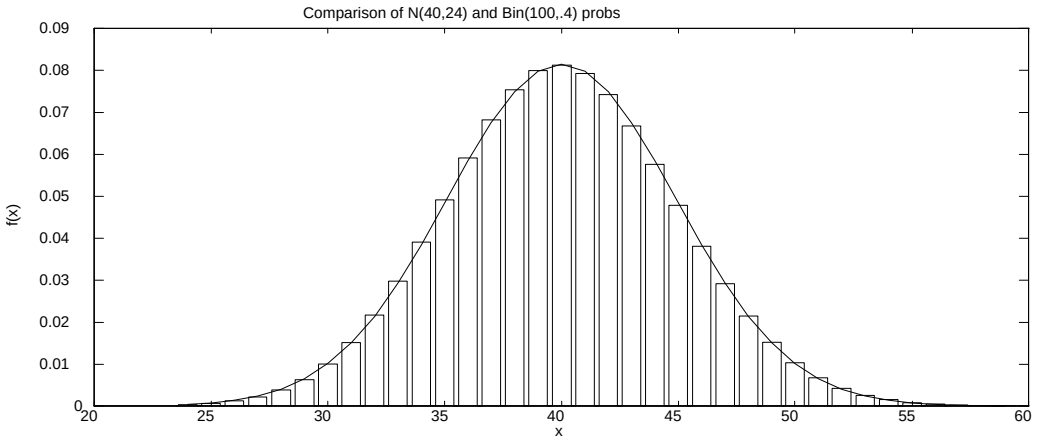

Figure 8.4:

Example.

Show that the probability density function of X is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X has the above normal distribution, we will denote this by $X \sim N(\mu, \sigma^2)$.

Moments

Show that the function $f(x;\mu,\sigma)$ integrates to 1 and is therefore a probability density function. Find the expected value and variance of a random variable having the probability density function $f(x;\mu,\sigma)$.

Linear Combinations.

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent random variables. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$.

Example.

Suppose $X_i \sim N(\mu, \sigma^2)$ are independent random variables. What is the distribution of the sample mean

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}?$$

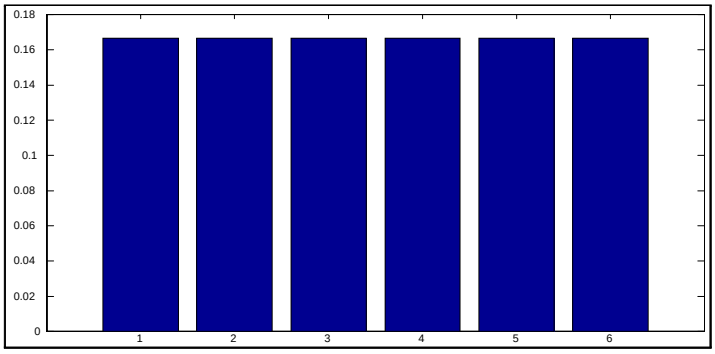

Figure 8.5:

Assume $\sigma = 1$ and find the probability $P[|\bar{X}_n - \mu| > 0.1]$ for various values of n. What happens to this probability as $n \to \infty$?

The Central Limit Theorem

The major reason that the normal distribution is the single most commonly used distibution is the fact that it tends to approximate the distribution of sums of random variables. For example, if we throw n dice and S_n is the sum of the outcomes, what is the distribution of S_n?. The tables below provide the number of ways in which a given value can be obtained. The corresponding probabability is obtained by dividing by 6^n.

$n = 1$,

1	2	3	4	5	6
1	1	1	1	1	1

$n = 2$,

2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	5	6	5	4	3	2	1

$n = 3$

3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

$n = 4$

4	5	6	7	8	.	.	.
1	4	10	20	35	.	.	.

The distributions show a simple pattern. For $n = 1$, the probability function is a constant (polynomial degree 0). For $n = 2$, two linear functions spliced together. For $n = 3$ a spline consisting of three quadratic pieces (polynomials of degree $n-1$). In general the histogram for S_n consists of n piecewise polynomials of degree $n-1$ which approach very rapidly the shape of the normal probability density function.

Example

Let $X_i = 0$ or 1 when the i'th toss of a biased coin is Tails or Heads respectively. What is the distribution of $S_n = \sum_{i=1}^{n} X_i$? Consider the standardized

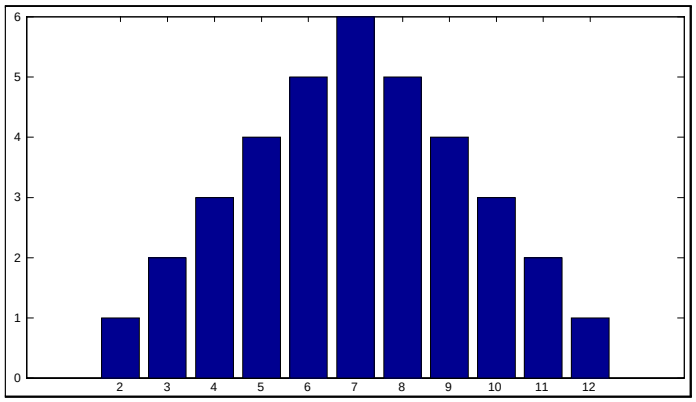

Figure 8.6:

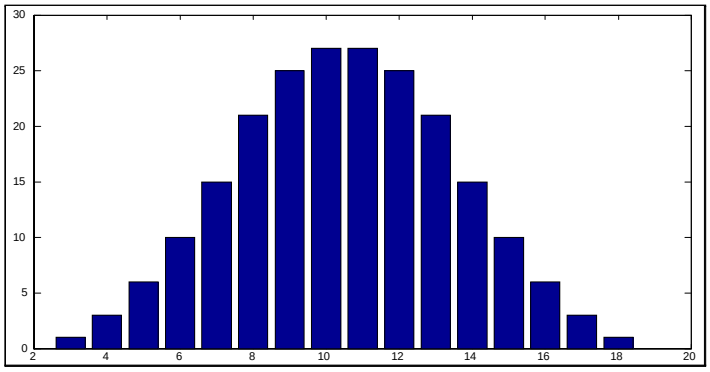

Figure 8.7:

random variable

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$$

Approximate the distribution of S_n^* for large values of n.

First let an integer $x \sim np + z\sqrt{np(1-p)}$ for fixed z. Then as $n \to \infty$, $x/n \to p$, $0 < p < 1$, Stirling's approximation implies that

$$\binom{n}{x} \sim \frac{\sqrt{2\pi}n^{n+1/2}e^{-n}}{2\pi x^{x+1/2}(n-x)^{n-x+1/2}} \sim \frac{1}{\sqrt{2\pi np(1-p)}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}.$$

Also using the series expansion $\ln(1+x) = x - \frac{1}{2}x^2 + O\left(x^3\right)$, putting $\sigma = \sqrt{\frac{p(1-p)}{n}}$, and noting $\sigma \to 0$ as $n \to \infty$,

$$\begin{aligned}
\ln\{\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}}\} &= x\ln(\frac{p}{p+z\sigma}) + (n-x)\ln(\frac{1-p}{1-p-z\sigma}) \\
&= -x\ln(1+\frac{z\sigma}{p}) - (n-x)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\ln(1+\frac{z\sigma}{p}) - n(1-p-z\sigma)\ln(1-\frac{z\sigma}{1-p}) \\
&= -n(p+z\sigma)\{(\frac{z\sigma}{p}) - \frac{1}{2}(\frac{z\sigma}{p})^2 + O(\frac{z\sigma}{p})^3\} \\
&\quad -n(1-p-z\sigma)\{-(\frac{z\sigma}{1-p}) - \frac{1}{2}(\frac{z\sigma}{1-p})^2 + O(\frac{z\sigma}{1-p})^3\} \\
&= -n\{z\sigma + \frac{z^2\sigma^2}{p} - \frac{1}{2}\frac{z^2\sigma^2}{p} - z\sigma + \frac{z^2\sigma^2}{1-p} - \frac{1}{2}\frac{z^2\sigma^2}{1-p} + O(\sigma^3)\} \\
&= -\frac{1}{2}z^2\sigma^2(\frac{n}{p} + \frac{n}{1-p}) + O(n^{-1/2}) \\
&= -\frac{z^2}{2} + O(n^{-1/2})
\end{aligned}$$

Therefore,

$$\begin{aligned}
P[S_n &= x] = P[S_n^* = z] = \binom{n}{x}p^x(1-p)^{n-x} \\
&\sim \binom{n}{x}(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}\frac{p^x(1-p)^{n-x}}{(\frac{x}{n})^x(1-\frac{x}{n})^{n-x}} \\
&\sim \frac{1}{\sqrt{np(1-p)}}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}.
\end{aligned}$$

This is the standard normal probability density function multiplied by the distance between consecutive values of S_n^*. In other words, this result says that

the area under the probabability histogram for S_n^* for the bar around the point z can be approximated by the area under the normal curve between the same two points $(z \pm \frac{1}{2\sqrt{np(1-p)}})$.

Theorem.

Let $X_i, \quad i = 1, \ldots n$ be independent random variables all with the same distribution, and with mean μ and variance σ^2 . Then the cumulative distribution function of

$$S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges to the cumulative distribution function of a standard normal random variable.

The proof of this result we will defer after the discussion of moment generating functions.

Consider, for example, the case where the X_i are indpendent each with a Bernoulli (p) distribution. Then the sum $\sum_{i=1}^n X_i$ has a binomial distribution with parameters n, p and the above theorem asserts that if we subtract the mean and we divide by the standard deviation of a binomial random variable, then the result is approximately standard normal. In other words, for large values of n a binomial random variable is approximately normal $(np, np(1-p))$. To veriify this fact, we plot both the binomial$(100, 0.4)$ histogram as well as the normal probability density function below.

Example.

Use the central limit theorem and the normal approximation to a probability histrogram to estimate the probability that the sum of the numbers on 6 dice is 20. What is the exact probability?

The Distribution of a Function of a Random Variable.

We have seen that if X has a normal distribution, then a linear function of X, say $aX + b$ also has a normal distribution. The parameters are easily determined since $E(aX + b) = aE(X) + b$ and $var(aX + b) = a^2 var(X)$. Is this true of arbitrary functions and general distributions? For example is X^2 normally distributed? The answer in general is NO. For example, the distribution of X^2 must be concentrated entirely on the positive values of x, whereas the normal distributions are all supported on the whole real line (i.e. the probability density functiopn $f(x) > 0, \quad$ all $x \in \mathcal{R}$. In general, the safest method for finding the distribution of the function of a random variable in the continuous case is to first find the cumulative distribution of the function and then differentiate to obtain the probability density function.

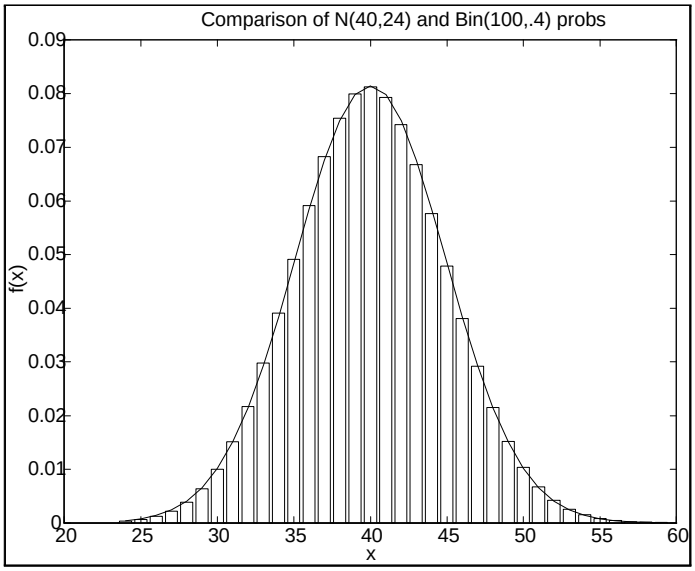

Figure 8.8:

Example.

Find the probability density function of $X = Z^2$ where Z has a standard normal distribution.

Theorem

Suppose a continuous random variable X has probability density function $f_X(x)$. Show that the probability density function of $Y = h(X)$ where $h(.)$ is a continuous monotone increasing function with inverse function $h^{-1}(y)$ is

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y)$$

Moment Generating Functions

Consider a random variable X. We have seen several ways of describing its distribution, using either a cumulative distribution function, a probability density function (continuous case) or probabililty function or a probability histogram or table (discrete case). We may also use some transform of the probability density or probability function. For example, consider the function defined by

$$M_X(t) = Ee^{tX}$$

defined for all values of t such that this expectations exists and is finite. This function is called the moment generating function of the (distribution of the)

random variable X. It is a powerful tool for detemining the distribution of sums of independent random variables and for proving the central limit theorem. In the discrete case we can write $M_X(t) = \sum_x e^{xt} P[X = x]$ and in the continuous case $M_X(t) = \int_{-\infty}^{\infty} e^{xt} f(x)dx$.

Properties of the Moment Generating Function

For these properties we assume that the moment generating function exists at least in some neighbourhood of the value $t = 0$, say for $-Varepsilon < t < Varepsilon$ for some $Varepsilon > 0$. We also assume that $\frac{d}{dt}E[X^n e^{tX}\} = E[\frac{d}{dt}X^n e^{tX}]$ for each value of $n = 0, 1, 2,$ for $-Varepsilon < t < Varepsilon$. The ability to differentiate under an integral or infinite sum is justified under general conditions involving the rate at which the integral or series converges.

1. $M'(0) = E(X)$
2. $M^{(n)}(0) = E(X^n), n = 1, 2, ...$
3. A moment generating function uniquely determines a distribution. In other words if $M_X(t) = M_Y(t)$ for all $-Varepsilon < t < Varepsilon$, then X and Y have the same distribution.
4. $M_{aX+b}(t) = e^{bt} M_X(at)$ for constants a, b.
5. If X and Y are independent random variables, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Example

Let X have a Binomial (n, p) distribution. Then the moment generating function of X is

$$M_X(t) = (pe^t + 1 - p)^n.$$

Example

Let X have a Poisson(λ) distribution. Then the moment generating function of X is

$$M_X(t) = \exp\{\lambda(e^t - 1)\}.$$

Example

Let X have an exponential distribution with mean μ. Then the moment generating function of X is

$$M_X(t) = \frac{1}{1 - \mu t} \quad \text{for } t < 1/\mu.$$

Example

Let X have a Normal (μ, σ^2) distribution. Then the moment generating function of X is

$$M_X(t) = \exp\{\mu t + \sigma^2 t^2/2\}.$$

Use this to show that the sum of independent normal random variables is also normally distributed.

Moment generating functions are useful for showing that a sequence of cumulative distribution functions converge because of the following result, stated without proof. The result implies that convergence of the moment generating functions can be used to show convergence of the cumulative distribution functions (i.e. convergence of the distributions).

Theorem

Suppose Z_n s a sequence of random variables with moment generating functions $M_n(t)$. Let Z be a random variable Z having moment generating function $M(t)$. If $M_n(t) \to M(t)$ for all t in a neighbourhood of 0, then

$$P[Z_n \leq z] \to P[Z \leq z]$$

as $n \to \infty$ for all values of z at which the function $F_Z(z)$ is continuous.

Proof of the Central Limit Theorem

We now use the properties of the moment generating function to prove the central limit theorem; i.e. that the cumulative distribution function of $S_n^* = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$ converges to the c.d.f. of the standard normal distribution as $n \to \infty$. Note that $S_n^* = \frac{1}{\sqrt{n}} X_i^*$ where $X_i^* = (X_i - \mu)/\sigma$ and so it is sufficient to prove this result for standardized random variables with mean 0 and variance 1. In this case, by the above theorem, it is sufficient to show that the moment generating function of S_n^* converges to the moment generating function of the standard normal, i..e to $m(t) = e^{t^2/2}$. Now let $L_n(t)$ be the logarithm of the moment generating function

$$L_n(t) = \ln[M_n(t)] = \ln[Ee^{S_n^* t}]$$

and

$$L(t) = \ln[M(t)] = \ln[Ee^{X^* t}]$$

Note that

$$L_n(t) = nL(t/\sqrt{n})$$

and that

1. $L(0) = 0$
2. $L'(0) = 0$
3. $L''(0) = 1$

Then

$$\begin{aligned}
\lim_{n\to\infty} L_n(t) &= \lim_{n\to\infty} L(t/\sqrt{n})/(n^{-1}) \\
&= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \text{ by L'Hospital's rule} \\
&= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-L^{''}(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \text{ by L'Hospital's rule} \\
&= \lim_{n\to\infty} L^{''}(t/\sqrt{n})\frac{t^2}{2} \\
&= \frac{t^2}{2}
\end{aligned}$$

It follows on exponentiating that $M_n(t)$ converges to $e^{t^2/2}$ which is the $N(0,1)$ moment generating function and therefore the cumulative distribution function of S_n^* converges to the normal cumulative distribution function pointwise (since the latter c.d.f. is continuous everywhere).

8.6 Stochastic Processes

A Stochastic process is an indexed family of random variables X_t for t ranging over some index set T such as the integers or an interval of the real line. For example a sequence of independent random variables is a stochastic process, as is a Markov chain. For an example of a continuous time stochastic process, define X_t to be the price of a stock at time t (assuming trading occurs continously over time).

Markov Chains

Consider a sequence of (discrete) random variables $X_1, X_2, \ldots$ each of which takes integer values $1, 2, \ldots N$ (called *states*). We assume that for a certain matrix P (called the *transition probability matrix*), the conditional probabilities are given by corresponding elements of the matrix; i.e.

$$P[X_{n+1} = j | X_n = i] = P_{ij}, \quad i = 1, \ldots N, \quad j = 1, \ldots N$$

and furthermore that the chain only cares about the last state occupied in determining its future; i.e. that

$$P[X_{n+1} = j | X_n = i,\ X_{n-1} = i_1, X_{n-2} = i_2 ... X_{n-l} = i_l] = P[X_{n+1} = j | X_n = i] = P_{ij}$$

for all $j, i, i_1, i_2, \ldots$. Then the sequence of random variables X_n is called a *Markov Chain*. Markov Chain models are the most common simple models for dependent variables, including weather (precipitation, temperature), movements of security prices etc.

Properties of the Transition Matrix P

Note that $P_{ij} \geq 0$ for all i, j and $\sum_j P_{ij} = 1$ for all i. This last property implies that the $N \times N$ matrix $P - I$ (where I is the identity matrix) has rank at most $N - 1$ because the sum of the N columns of $P - I$ is identically 0.

Example. Rain-No rain

Suppose that the probability that tomorrow is rainy given that today is not is α and the probability that tomorrow is dry given that today is rainy is β.

Example. Gambler's Ruin

A gambler at each play of a game either wins \$1 or loses \$1 ith probabilities p, 1-p respectively. The gambler quits playing when his fortune reaches either 0 or M. Then the total fortune of the gambler at time t follows a Markov chain. What is the transition probability matrix?

The distribution of X_t

Show that if the chain is started by randomly choosing a state for X_0 with distribution $P[X_0 = i] = q_i,\ i = 1, 2, \ldots N$, then the distribution of X_1 is the vector $q'P$ where $\underline{q}$ is the column vector of values q_i. Similarly the distribution of X_t is the vector $\underline{q}'P^t$ where P^t is the product of the matrix P with itself t times. Under very general conditions, it can be shown that these probabilities converge and in many such cases, the limit does not depend on the initial distribution q.

Definition

A *limiting distribution* of a Markov chain is a vector ($\underline{\pi}$ say) of long run probabilities of the individual states so

$$\pi_i = lim_{t \to \infty} P[X_t = i].$$

Definition

A *stationary distribution* of a Markov chain is the column vector ($\underline{\pi}$ say) of probabilities of the individual states such that

$\underline{\pi}'P = \underline{\pi}'$.

Theorem

Any limiting distribution of a Markov Chain must be a stationary distribution.

Proof.

Note that $\pi' = \lim_{n\to\infty} q'P^n = \lim_{n\to\infty}(q'P^n)P = (\lim_{n\to\infty} q'P^n)P = \pi'P$.

Example

Consider a Markov chain with transition probability matrix

$$P = \begin{pmatrix} .9 & .1 \\ .2 & .8 \end{pmatrix}$$

Find $lim_{t\to\infty}P^t$ and the limiting distribution of the Markov chain. Show that in general for a 2×2 transition matrix, the stationary distribution is proportional to (P_{21}, P_{12}).

Example: Binary information:

Suppose that $X_1, X_2, \ldots$ is a sequence of binary information (Bernoulli random variables) taking values either 0 or 1. Suppose that the probability that a 0 is followed by a 1 is p and the probability that a 1 is followed by a 0 is given by q where $0 < p,\ q < 1$. Find the transition matrix for the Markov chain and the long run proportion of zeros in the sequence.

When is the limiting distribution of a Markov chain unique and independent of the initial state of the chain?

Definition: irreducible, aperiodic

We say that a Markov chain is *irreducible* if every state can be reached from every other state. In other words for every pair i, j there is some m such that $P^m_{i,j} > 0$. We say that the chain is *aperiodic* if $gcd\{N; P^{(N)}_{ii} > 0\} = 1$. For a *periodic chain* (i.e. one with period>1) returns to a state can occur only at multiples of the period $gcd\{N; P^{(N)}_{ii} > 0\}$.

Theorem

If a Markov chain is irreducible and aperiodic, then there exists a *unique* limiting distribution $\underline{\pi}$. In this case $P^n \to \pi'\underline{1}$ the matrix whose rows are all identically π' as $n \to \infty$.

Generating Functions.

Definition: Generating function

Let $a_0, a_1, a_2, \ldots$ be a finite or infinite sequence of real numbers. Suppose the power series

$$\mathcal{A}(t) = \sum_{i=0}^{\infty} a_i t^i$$

converges for all $-\epsilon < t < \epsilon$ for some value of $\epsilon > 0$. Then we say that the sequence has a *generating function* $\mathcal{A}(t)$.

Note. Every bounded seqence has a generating function since the series $\sum_{i=0}^{\infty} t^i$ converges whenever $|t| < 1$. Thus, discrete probability functions have generating functions. The generating function of a random variable X or its associated probability function $f_X(x) = P[X = x]$ is given by

$$\mathcal{F}_X(t) = \sum_x f_X(x) t^x = E(t^X).$$

Note that if the random variable has finite expected value, then this converges on the interval $t \in [-1, 1]$.

The *joy of generating functions* is that they provide a transform of the original distribution to a space where many operations are made much easier. We will give examples of this later. The most important single property is that they are in one-one correspondence with distributions such that the series converges; for each distribution there is a unique generating function and for each generating function there is a unique distribution.

As a consequence of this representation and the following theorem we can use generating functions to determine distributions that would otherwise be difficult to identify.

Theorem

Suppose a random variable X has generating function $\mathcal{F}_X(t)$ and Y has generating function $\mathcal{F}_Y(t)$. Suppose that X and Y are independent. Then the generating function of the random variable $W = X+Y$ is $\mathcal{F}_W(t) = \mathcal{F}_X(t)\mathcal{F}_Y(t)$.

Example

Find the distributions that corresponds to the following generating functions:

(a) $\mathcal{F}(t) = \frac{t}{3-2t}$

(b) $\mathcal{F}(t) = e^{\lambda(t-1)}$

Example

Find the generating function of the Binomial (n, p) distribution. Suppose X_1 and X_2 are independent random variables, both with this binomial distribution.

Find the distribution of $W = X_1 + X_2$. Notice that whenever a Moment generating function exists, we can recover the generating function from it by replacing e^t by t.

Example.

One of six defferent varieties of coupons is placed in each box of cereal. Find the distribution of the number of cereal boxes you need to buy to obtain all six coupons. (Answer: the (probability) generating function of the number is

$$\frac{5!t^6}{(6-t)(6-2t)(6-3t)(6-4t)(6-5t)}$$

$= \frac{5}{324}t^6 + \frac{25}{648}t^7 + \frac{175}{2916}t^8 + \frac{875}{11\,664}t^9 + \frac{11\,585}{139\,968}t^{10} + \frac{875}{10\,368}t^{11} + O\left(t^{12}\right)$
and this expansion as a power series provides the probabilities)

The Poisson Process.

One of the simplest continuous time stochastic processes is the *Poisson Process*. Suppose N_t denotes the total number of arrivals into a system (such as the number of customers arriving at a queue) untill time t. Note that the number of arrivals in time interval $(a, b]$ is then $N_b - N_a$. Assume the following properties;

(a) The probability of exactly one arrival in a small interval of length Δt is $\lambda\Delta t + o(\Delta t)$. (Note that the probability does not depend on where the interval is, only on its length).

(b) The probability of two or more arrivals in an interval of length Δt is $o(\Delta t)$ where by definition of the o notation, $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$.

(c) For disjoint intervals $I_i = (a_i, b_i]$ (so $I_i \cap I_j = \phi, \quad i \neq j$), the number of arrivals in these intervals $Y_i = N_{b_i} - N_{a_i}$ are mutually independent random variables.

Theorem.

Under the above conditions, (a)-(c), the distribution of the process $N_t, \quad t \in T$ is that of a *Poisson process*. This means that the number of arrivals $N_b - N_a$ in an interval $(a, b]$ has a Poisson distribution with parameter $\lambda(b-a) = \lambda\times$*the length of the interval*, and the number of arrivals in disjoint time intervals are independent random variables. The parameter λ specifies the *rate* of the Poisson process.

Example.

1. (a) (a) Show that if $N(t)$ is a Poisson process and $T_1, T_2, ...$ are the times of the first event, and the time between the first and second events, etc. then T_1, $T_2, ...$ are independent random variables, each with an exponential distribution with expected value $1/\lambda$.

(b) Show that if $T_1, T_2, \ldots T_n$ are independent random variables each with an exponential (1) distribution, then the sum $\sum_{i=1}^{n} T_i$ has a (gamma) probability density function

$$f(x) = \frac{1}{(n-1)!} x^{n-1} e^{-x}, x > 0.$$

Example.

Suppose emergency calls to 911 follow a Poisson process with an average of 10 calls per hour. What is the probability that there are no calls in a five minute period? What is the probability that there are more than 100 calls in an 8-hour shift? Given that there are 5 calls in the first hour, what is the probability that the first call occurred in the first x minutes?

Poisson Process in space.

In an analogous way we may define a Poisson process in space as a distribution governing the occurrence of random points with the properties indicated above; The number of points in a given set S has a Poisson distribution with parameter $\lambda \times |S|$ where $|S|$ is the area or volume of the set, and if $Y_1, Y_2, \ldots$ are the number of points occurring in disjoint sets $S_1, S_2, \ldots,$ they are mutually independent random variables.

Example

Bacteria are immersed in contaminated water at a rate of λ per nanolitre (note: 1 nanolitre $= 10^{-9}$ litres). What is the probability that there are no bacteria in a sample of 1 ml. if $\lambda = 10^{-3}$? What is the probability of more than 1200 bacteria in 1 ml if $\lambda = 10^{-3}$?

8.7 Conditional Expectation and Martingales

8.7.1 Conditional Expectation.

Theorem.

Let $\mathcal{G} \subset \mathcal{F}$ be sigma-algebras and X a random variable on $(\Omega, \mathcal{F}, P)$. Assume $E(X^2) < \infty$. Then there exists an almost surely unique $\mathcal{G}$-measurable Y such that

$$E[(X-Y)^2] \;=\; inf_Z E(X-Z)^2 \quad (6.1)$$

where the infimum is over all $\mathcal{G}$-measurable random variables. *Note.* We denote the minimizing Y by $E(X|\mathcal{G})$.

For two such minimizing Y_1, Y_2, i.e. random variables Y which satisfy (6.1), we have $P[Y_1 = Y_2] = 1$. This implies that conditional expectation is almost surely unique.

Example.

Suppose $\mathcal{G} = \{\varphi, \Omega\}$. What is $E(X|\mathcal{G})$?

Example.

Suppose $\mathcal{G} = \{\varphi, A, A^c, \Omega\}$ for some event A. What is $E(X|\mathcal{G})$? Consider the special case: $X = I_B$.

Example.

Suppose $\Omega = (0, 1]$ and the function $X(\omega)$ is Borel measurable. Assume that $\mathcal{G}$ is generated by the intervals $(\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, 2, ..., n$. What is $E(X|\mathcal{G})$?

Properties of Conditional Expectation.

(a) If a random variable X is $\mathcal{G}$-measurable, $E(X|\mathcal{G}) = X$.

(b) If a random variable X independent of a sigma-algebra $\mathcal{G}$, then $E(X|\mathcal{G}) = E(X)$.

(c) For any square integrable $\mathcal{G}$-measurable Z, $E(ZX) = E[ZE(X|\mathcal{G})]$.

(d) (special case of (c)): $\int_A XdP = \int_A E(X|\mathcal{G})dP$ for all $A \in \mathcal{G}$.

(e) $E(X) = E[E(X|\mathcal{G})]$.

(f) If a $\mathcal{G}$-measurable random variable Z satisfies $E[(X - Z)Y] = 0$ for all other $\mathcal{G}$-measurable random variables Y , then $Z = E(X|\mathcal{G})$.

(g) If Y_1, Y_2 are distinct $\mathcal{G}-$measurable random variables both minimizing $E(X - Y)^2$, then $P(Y_1 = Y_2) = 1$.

(h) *Additive* $E(X + Y|\mathcal{G}) = E(X|\mathcal{G}) + E(Y|\mathcal{G})$.
Linearity $E(cX + d|\mathcal{G}) = cE(X|\mathcal{G}) + d$.

(i) If Z is $\mathcal{G}-$measurable, $E(ZX|\mathcal{G}) = ZE(X|\mathcal{G})$ a.s.

(j) If $\mathcal{H} \subset \mathcal{G}$ are sigma-algebras, $E[E(X|\mathcal{G})|\mathcal{H}] = E(X|\mathcal{H})$.

(k) If $X \leq Y$, $E(X|\mathcal{G}) \leq E(Y|\mathcal{G})$ a.s.

(l) *Conditional Lebesgue Dominated Convergence.* If $X_n \to X$ a.s. and $|X_n| \leq Y$ for some integrable random variable Y , then $E(X_n|\mathcal{G}) \to E(X|\mathcal{G})$ *in distribution*

Notes. In general, we define $E(X|Z) = E(X|\sigma(Z))$ and conditional variance $var(X|\mathcal{G}) = E\{(X - E(X|\mathcal{G}))^2|\mathcal{G}\}$. For results connected with property (l) above providing conditions under which the conditional expectations converge, see Convergence in distribution of conditional expectations, (1994) E.M. Goggin, *Ann. Prob* 22, 2. 1097-1114.

Conditional Expectation for integrable random variables.

For non-negative integrable X choose simple random variables $X_n \uparrow X$. Then $E(X_n|\mathcal{G}) \uparrow$ and so it converges. Define $E(X|\mathcal{G})$ to be the limit. In general, for random variables taking positive and negative values, we define $E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G})$.

8.7.2 Martingales.

Intuitively, a martingale is the total fortune of an individual participating in a "fair game". In order to be fair, the expected value of one's future fortune given the history of the process up to and including the present should be equal to one's present wealth. Suppose the fortune at time s is denoted X_s. The current process and any other related processes up to time s generate a sigma-algebra $\mathcal{F}_s$. Then the assertion that the game is fair implies $E(X_t|\mathcal{F}_s) = X_s$ for $t > s$.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *martingale* if

(a) $\mathcal{F}_t$ is increasing (in t) family of sigma-algebras

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_t|\mathcal{F}_s) = X_s$ a.s.

Example.

Suppose Z_t are independent random variables with expectation 0. Define $\mathcal{F}_t = \sigma(Z_1, Z_2, \ldots Z_t)$ and $S_t = \sum_{i=1}^{t} Z_i$. Then $\{(S_t, \mathcal{F}_t),\ t = 1, 2, ...\}$ is a martingale.

Example.

Let X be any integrable random variable, and $\mathcal{F}_t$ an increasing family of sigma-algebras. Put $X_t = E(X|\mathcal{F}_t)$. Then $(X_t, \mathcal{F}_t)$ is a martingale.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *reverse martingale* if

(a) $\mathcal{F}_t$ is decreasing (in t) family of sigma-algebras.

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_s|\mathcal{F}_t) = X_t$ a.s.

Example.

Let X be any integrable random variable, $\mathcal{F}_t$ be any decreasing family of sigma-algebras. Put $X_t = E(X|\mathcal{F}_t)$. Then $(X_t, \mathcal{F}_t)$ is a reverse martingale.

Definition.

$\{(X_t, \mathcal{F}_t); t \in T\}$ is a *sub (super) martingale* if

(a) $\mathcal{F}_t$ is increasing (in t) family of sigma-algebras.

(b) Each X_t is $\mathcal{F}_t-$ measurable and $E|X_t| < \infty$.

(c) For each $s < t$, $s, t \in T$, $E(X_t|\mathcal{F}_s) \geq (\leq)\ X_s$ a.s.

Example.

Let Y_i be independent identically distributed, $\mathcal{F}_n = \sigma(Y_{(1)}, ..., Y_{(n)}, Y_{n+1}, Y_{n+2}, ...)$, where $(Y_{(1)}, ..., Y_{(n)})$ denote the order statistics. Then $\mathcal{F}_n$ is a decreasing family of sigma fields and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = E(Y_1|\mathcal{F}_n)$ is a reverse martingale.

Definition.

A random variable τ is a (optional) *stopping time* for a martingale $(X_t, \mathcal{F}_t)$ if for each t, $[\tau \leq t] \in \mathcal{F}_t$.

Definition.

For an optional stopping time τ define

$$\mathcal{F}_\tau = \{A \in \mathcal{F}; A \cap [\tau \leq t] \in \mathcal{F}_t, \quad for\ all\ t\}.$$

Then this is a sigma-algebra.

Theorem.

If $(X_t, \mathcal{F}_t)\ t = 1, 2, \ldots n$ is a (sub) martingale and α, β are stopping times with values in $\{1, .., n\}$, such that $\alpha \leq \beta$, then

$$E(X_\beta|\mathcal{F}_\alpha)\ (\geq) = X_\alpha$$

(Sub)martingale Convergence Theorem.

Let $(X_n, \mathcal{F}_n);\ n = 1, 2, \ldots$ be a submartingale such that $lim\ sup_{n\to\infty} E|X_n| < \infty$. Then there is a (finite) random variable X such that $X_n \to X$ a.s.

Reverse martingale convergence Theorem.

If $(X_n, \mathcal{F}_n);\ n = 1, 2, \ldots$ is a reverse martingale,

$$X_n \to E(X_1 | \cap_{n=1}^{\infty} \mathcal{F}_n) \quad a.s.$$

8.7.3 Martingales and Finance

Let $S(t)$ denote the price of a security at the beginning of period $t = 0, 1, 2, ...T$. We assume that the security pays no dividends. Define the *(cumulative) returns process* associated with this security by R_S where

$$\Delta R_S(t) = R_S(t) - R_S(t-1) = \frac{\Delta S(t)}{S(t-1)} = \frac{S(t) - S(t-1)}{S(t-1)}, \quad R_S(0) = 0.$$

Then $100\Delta R_S(t)\%$ is the percentage return in an investment in the stock in the $t-1'$st period. The returns process is a more natural characterisation of stock prices than the original stock price process since it is invariant under artificial scale changes such as stock splits etc. Note that we can write the stock price in terms of the returns process;

$$S(t) = S(0) \prod_{i=1}^{t} (1 + \Delta R_S(i)).$$

Now consider another security, a *riskless discount bond* which pays no coupons. Assume that the price of this bond at time t is $B(t), \quad B(0) = 1$ and $R_B(t)$ is the return process associated with this bond. Then $\Delta R_B(t) = r(t)$ is the interest rate paid over the $t-1$'st period. It is usual that the interest paid over the $t-1$st period should be declared in advance, i.e. at time $t-1$ so that if $S(t)$ is adapted to a filtration $\mathcal{F}_t$, then $r(t)$ is *predictable*, i.e. is $\mathcal{F}_{t-1}-$measurable. The *discounted stock price process* is the process given by

$$S^*(t) = S(t)/B(t).$$

Consider a *trading strategy* of the form $(\beta(t), \alpha(t))$ representing the total number of shares of bonds and stocks respectively held at the beginning of the period $(t-1, t)$. Since our investment strategy must be determined by using only the present and the past values of this and related processes, both $\beta(t)$ and $\alpha(t)$ are predictable processes. Then the value of our investment at time $t-1$ is $V_{t-1} = \beta(t)B(t-1) + \alpha(t)S(t-1)$ and at the end of this period, this changes to $\beta(t)B(t) + \alpha(t)S(t)$ with the difference $\beta(t)\Delta B(t) + \alpha(t)\Delta S(t)$ representing the *gain* over this period. An investment strategy is *self-financing* if the value after rebalancing the portfolio is the value before- i.e. if all investments are paid for by the above gains. In other words if $V_t = \beta(t)B(t) + \alpha(t)S(t)$ for all t. An *arbitrage opportunity* is a trading strategy that makes money with no initial investment; i.e. one such that $V_0 = 0, \quad V_t \geq 0$ for all $t = 1, \ldots T$ and $E(V_T) > 0$. The basic theorem of no-arbitrage pricing is the following:

Theorem

There are no arbitrage opportunities in the above economy if and only if there is a measure Q equivalent to the underlying measure P i.e. $P << Q$ and $Q << P$ such that under Q the discounted process is a martingale; i.e. $E_Q(S^*(t)|\mathcal{F}_{t-1}] = S^*(t-1)$ a.s. for all $t \leq T$.

Proof; See Pliska (3.19)) page 94.

Note: The measure Q is called the equivalent martingale measure and is used to price derivative securities. For any attainable contingent claim X; (a for any random variable X which can be written as a linear function of the available investments), the arbitrage-free price at time t is given by the conditional expected value under Q of the discounted return X given $\mathcal{F}_\tau$.

Chapter 9

Appendix B: Stochastic Integration and Continuous Time Models

The single most important continuous time process in the construction of financial models is the Brownian motion process. A Brownian motion is the oldest continuous time model used in finance and goes back to Bachelier around the turn of the last century. It is also the most common building block for more sophisticated continuous time models called diffusion processes.

The Brownian motion process is a random continuous time process $W(t)$ defined for $t \geq 0$ such that $W(0)$ takes some predetermined value, usually 0, and for each $0 \leq s < t$, $W(t) - W(s)$ has a normal distribution with mean $\mu(t-s)$ and variance $\sigma^2(t-s)$. The parameters μ and σ are the drift and the diffusion parameters of the Brownian motion and the special case $\mu = 0, \sigma = 1$, $W(t)$ is often referred to as a standard Brownian motion or a Wiener process. Further properties of the Brownian motion process that are important are:

A Brownian motion process exists such that the sample paths are each continuous functions (with probability one)

The joint distribution of any finite number of increments $W(t_2) - W(t_1), W(t_4) - W(t_3), ...W(t_k) - W(t_{k-1})$ are independent normal random variables for $0 \leq t_1 \leq t_2 ... \leq t_k$.

Further properties can be derived from these. For example suppose we consider a sum of squared increments of the form $\sum_{i=1}^{k-1}(W(t_{i+1}) - W(t_i))^2$ for for $0 = t_1 \leq t_2 ... \leq t_k = t$ If we allow the number of increments k to go to infinity and the mesh size $\max(t_{i+1} - t_i)$ to go to zero, then it is easy to show that the limit of this random sum of squares is a totally non-random value t,

this limit taken in probability. A sum of squares process defined in this way is

The Stochastic Exponential and logarithm.

It is natural to define the exponential of a process

$$\mathcal{E}(X) = \prod(1 + dX) = \exp\{X - \frac{1}{2} < X^c >\} \prod(1 + \Delta X)e^{-\Delta X}$$

Then if $Y = \mathcal{E}(X)$, we have

$$dY = Y_- dX$$

and therefore we define the stochastic logarithm

$$\mathcal{L}(Y) = \int \frac{1}{Y_-} dY$$

9.1 Ordinary Differential Equations

Consider a stochastic differential equation of the form **??** in the special case that the drift term is linear $a(X_t) = \alpha + \beta X_t$. Most of the standard models for interest rates, for example, take this form, including the CIR, Vasicek, Geometric Brownian motion.

Suppose we wish to determine the expected value of the process. We have seen that if we denote the expected value by $m(s) = E(X_s|X_0)$ then it satisfies the ordinary differential equation $m'(t) = \alpha + \beta m(t)$. This is an example of the simplest form of ordinary differential equation, one in which the derivatives are of order at most 1 and the coefficients are constant. Let us consider the general first order differential equation

$$\frac{dy}{dt} + p(t)y = g(t).$$

These are solved by introducing an integrating factor $\mu(t)$ satisfying

$$\mu(t)[\frac{dy}{dt} + p(t)y] = [\mu(t)y]'.$$

In order that μ satisfy this, we require $\mu p = \mu'$ or

$$\mu(t) = exp\{\int^t p(s)ds\}.$$

Then $(\mu y)' = \mu g$ resulting in the general solution

$$y(t) = \frac{1}{\mu(t)}[\int^t \mu(s)g(s)ds + c].$$

MAPLE can be used to solve differential equations such as this one. For example we would first define the differential equation (here we substituted a, b for α, β) and then request a solution; (:> is the MAPLE prompt)

```
> deq := diff(y(x),x) - a - b*y(x) = 0;
> dsolve(deq, y(x));
```

We might have wished to specify an initial condition in the above equation. Suppose we wish to specify $y(0) = c$

```
> dsolve(deq,y(0)=c,y(x));
```

and we have the solution

$$y(x) \ = \ -a/b \ + \ exp(bx)[\ \frac{a}{b} \ + \ c] \ . \tag{1.7}$$

Consider as an example the Cox-Ingersoll-Ross model, here written with a slightly different specification of parameters. If r_t denotes the spot interest rate at time t ,

$$dr_t \ = \ (\ \alpha + \ \beta r_t)\ dt \ + \ \sigma r_t^{1/2}\ dW_t \tag{1.8}$$

If we let $m(t)$ denote the mean, taking expectations on both sides gives

$$m'(t) = (\alpha + \beta m(t))$$

and therefore the solution is given by (1.7) with $a,\ b$ replaced by α, β.

Non-linear ordinary differential equations are usually somewhat more difficult to solve. These are equations of the form

$$\frac{dy}{dx} = f(x, y)$$

where $f(x, y)$ is not a linear function of y. For example if the function $f(x, y)$ is a quadratic function of y, the equation is called *Ricatti's equation.* Suppose this equation is homogeneous, and can be written in the form

$$M(x, y)dx + N(x, y)dy = 0. \tag{9.1}$$

There is sometimes a function $\mu(x, y)$ called an *integrating factor* satisfying

$$\frac{\partial}{\partial y}(\mu M) = \frac{\partial}{\partial x}(\mu M)$$

and in this case the homogeneous equation can be solved explicitly. Note that after multiplication by μ, the equation

$$\mu M dx + \mu N dy = 0$$

is exact in the sense that it is obtained as the differential of an equation of the form $\psi(x, y) = c$. When no explicit solution to a differential equation can be obtained, we may either approximate the solution numerically or obtain a power series expansion of the solution, to as many terms as are wished.

For example, consider the differential equation

$$\frac{d^2}{dt^2}y(t) + 5\frac{d}{dt}y(t) + 6y(t) = 0$$

This is solved in MAPLE with

```
> de1 := diff(y(t),t$2) + 5*diff(y(t),t) + 6*y(t) = 0;
> dsolve(de1, y(t));
```

yielding:

$$y(t) = C1exp(-3t) + C2exp(-2t)$$

and although in this case an analytic solution is available, we might for a more difficult differential equation wish a series expansion, as obtained by

```
> dsolve(de1, y(t), series);
```

Suppose the initial condition is $y'(0) = c$.

This yields the series expansion of the solution

$$y(t) = y(0) + ct + \left(-3\,y(0) - \frac{5\,c}{2}\right)\,t^2 + \left(5\,y(0) + \frac{19\,c}{6}\right)t^3$$

$$+ \left(-\frac{19\,y(0)}{4} - \frac{65\,c}{24}\right)t^4 + \left(\frac{13\,y(0)}{4} + \frac{211\,c}{120}\right)t^5 + O\left(t^6\right)$$

Problem:

Solve the ordinary differential equation

$$3t^2y''(t) - 2ty'(t) + 2y(t) = 0$$

and sketch the possible solutions.

Problem: (Finding the stationary distribution of an Ito process).

Consider an Ito process of the form:

$$dX_t = a(X_t)dt + \sigma(X_t)dW_t. \tag{9.2}$$

Suppose there is a stationary density $\pi(x)$ satisfying

$$\int p(s,z,t,x)\pi(z)dz = \pi(x) \text{ for all } s < t, \text{ and all } x.$$

Multiply Kolmogorov's forward differential equation (1.3) by $\pi(z)$ and integrate over z. Thus show that the stationary distribution $\pi(x)$ must satisfy a differential equation of the form

$$\frac{d^2}{dx^2}(\sigma^2(x)\pi(x)) = 2\frac{d}{dx}(a(x)\pi(x))$$

Solve this differential equation to obtain the form of the stationary distribution assuming that the diffusion models a process (like interest rate, or asset price) that is positive, so that $\pi(0) = \pi'(0) = 0$.

9.2 Systems of Ordinary Differential Equations.

Consider now a system of ordinary differential equations of the form

$$\frac{d}{dt}\mathbf{y}(t) = -\mathbf{p}(t)\mathbf{y}(t) + \mathbf{g}(t) \tag{1.9}$$

where $\mathbf{y}(\mathbf{t})$ is an n-dimensional column vector of functions $y_i(t)$and $\mathbf{p}(t)$ is a $n \times n$ matrix of functions and $\mathbf{g}$ is an n-dimensional column vector of functions. Then the solution is exactly analogous to the one-dimensional case. First, for a square matrix A we define the exponential

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} \tag{1.10}$$

Then the integrating factor μ is defined by integrating componentwise: μ(t) = e $^{\int_{t_0}^{t} \mathbf{p}(s)\ ds}$.

Note that this is a matrix. The solution is then given by $y(t) = [\,\mu(t)]^{-1}\,[\int_{t_0}^{t} \mu(s)\, g(s)\, ds\,]$. Maple also permits the solution of systems of differential equations. For example consider the second order differential equation $y''(x) = y(x)$. This is equivalent to the system $y'(x) = z(x), \quad z'(x) = y(x)$. Solve as follows:

```
> sys := diff(y(x),x)=z(x), diff(z(x),x)=y(x)
> fcns := y(x), z(x)
> dsolve(sys,y(0)=0,z(0)=1,fcns);y(x)=1/2 exp(x -1/2 exp(-x);z(x)
=1/2 exp(x) + 1/2 exp(- x)
```

Problem:

Find the general solution to the following system of first order differential equations:

$$\begin{aligned} g'(x) &= -\sqrt{f(x)}, \\ h''(x) &= \frac{g(x)}{f(x)} \\ f'(x) &= \exp\{-f(x)\} \end{aligned}$$

9.3 Partial Differential Equations

Many of the pricing formulae encountered in finance can be derived as solutions to one or more partial differential equations, including the most important, the Black-Scholes formula. In general, this is because the most common models for the underlying asset are diffusion models. Maple provides some facility for the solution of simple partial differential equations. For example:

:>PDE := x*diff(f(x,y),y)-y*diff(f(x,y),x) = 0;

:>pdsolve(PDE);

provides the solution $f(x, y) = F1(x + y)$ for arbitrary function $F1$.

Before discussing methods of solution in general, we develop the Black-Scholes equation in a general context. Suppose that a security price satisfies

$$dS_t \;=\; a(S_t,\; t\;)\; dt \;+\; \sigma(S_t,\; t)\; dW_t \tag{1.11}$$

Our assumed market allows investment in the stock as well as in discount bonds, whose price at time t is β_t. There are various other assumptions as well; for example partial shares may be purchased, there are no dividends paid and no commissions, and no possibility of default for the bonds. Since bonds are assumed risk-free, they satisfy an equation

$$d\beta_t = r_t \beta_t dt$$

where r_t is the risk-free (spot) interest rate at time t.

We wish to determine $V(S_t, t)$, the value of an option on this security when the security price is S_t, at time t. Suppose the option has expiry date T and a general payoff function which depends only on S_T, the process at time T.

A quick reminder of one of the most important single results of the twentieth century in finance and in science. This single mathematical result underlies the research leading to 1997 Nobel Prize to Merton and Black for their work on hedging in financial models.

Ito's lemma.

Suppose S_t is a diffusion process satisfying

$$dS_t \;=\; a(S_t,\; t\;)\; dt \;+\; \sigma(S_t,\; t)\; dW_t$$

and suppose $V(S_t,t)$ *is a smooth function of both arguments. Then* $V(S_t,t)$ *also satisfies a diffusion equation of the form*

$$dV = [a(S_t,\ t\ )\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t,\ t)}{2}\ \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]dt + \ \sigma(S_t,\ t)\frac{\partial V}{\partial S}dW_t. \qquad (1.12)$$

The proof of this result is technical but the ideas behind it are simple. Suppose we expand an increment of the process $V(S_t,t)$.

$$V(S_{t+h},t+h) \approx V(S_t,t) + \frac{\partial V}{\partial S}(S_{t+h}-S_t) + \frac{1}{2}\frac{\partial^2 V}{\partial S^2}(S_{t+h}-S_t)^2 + \frac{\partial V}{\partial t}h \qquad (1.13)$$

where we have ignored remainder terms that are $o(h)$. Note that substituting from (1.11) into (1.13), the increment $(S_{t+h}-S_t)$ is approximately normal with mean $a(S_t,\ t\ )\ h$ and variance $\sigma^2(S_t,\ t\ )\ h$. Consider the term $(S_{t+h}-S_t)^2$. Note that it is the square of the above normal random variable and has expected value $\sigma^2(S_t,t)h + a^2(S_t,t)h^2$. The variance of this random variable is $O(h^2)$ so if we ignore all terms of order $o(h)$ the increment $V(S_{t+h},t+h) - V(S_t,t)$ is approximately normally distributed with mean

$$[a(S_t,\ t\ )\frac{\partial V}{\partial S} + \frac{\sigma^2(S_t,\ t)}{2}\ \frac{\partial^2 V}{\partial S^2} + \frac{\partial V}{\partial t}]h$$

and standard deviation $\sigma(S_t,\ t)\frac{\partial V}{\partial S}\sqrt{h}$ justifying (but not proving!) the relation **??**.

By Ito's lemma, provided V is smooth, it also satisfies a diffusion equation of the form **??**. We should note that when V represents the price of an option, some lack of smoothness in the function V is inevitable. For example for a European call option with exercise price K, $V(S_T,T) = \max(S_T-K,0)$ does not have a derivative at the exercise price. Fortunately, such exceptional points can be worked around in the argument. For hedging purposes, is it possible to find a self-financing portfolio consisting only of the security and the bond which exactly replicates the option price process $V(S_t,t)$? Suppose such a linear combination is $u_tS_t + w_t\beta_t$ where the predictable functions $u_t,\ w_t$ represent the number of shares of stock and bonds respectively owned at time t. The portfolio is assumed to be self-financing and this requires that all returns obtain from the changes in the value of the securities and bonds held, i.e. it is assumed that $dV = u_tdS_t + w_td_t.\ Substituting from (1.11),$

$$dV = u_tdS_t + w_td\beta_t = [u_ta(S_t,t) + w_tr_t\beta_t]dt + u_t\sigma(S_t,t)dW_t \qquad (1.14)$$

It follows on comparing the coefficients of dt and dW_t in **??** and (1.14), that $u_t = \frac{\partial V}{\partial S}$ called the *delta* corresponding to *delta hedging*. Consequently,

$$V = \frac{\partial V}{\partial S} S_t + w_t \beta_t$$

and solving for w_t we obtain:

$$w_t = \frac{1}{\beta_t}[V - \frac{\partial V}{\partial S} S_t].$$

The conclusion is that it is possible to dynamically choose a trading strategy, i.e. the weights w_t, u_t so that our portfolio of stocks and bonds **perfectly replicates** the value of the option. If we own the option, then by shorting Delta units of stock, we are **perfectly** hedged in the sense that our portfolio replicates a risk-free bond. Surprisingly, in this ideal word of continuous processes and continuous time trading commission-free trading, the perfect hedge (said to exist only in a Japanese garden), is possible. The equation we obtained by equating both coefficients in **??** and (1.14) is of the form;

$$-r_t V + r_t S_t \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} + \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{1.15}$$

The negative of the first two terms $r_t(V - S_t \frac{\partial V}{\partial S})$ represents the amount made by the portion of our portfolio devoted to risk-free bonds. The last two terms represents the return on a hedged portfolio long one option and short delta stocks. This fundamental equation is evidently satisfied by any option price process where the underlying security satisfies a diffusion equation and the option value at expiry depends only on the value of the security at that time. The type of option determines the terminal conditions and usually uniquely determines the solution. It is extraordinary that this equation in no way depends on the drift coefficient $a(S_t, t)$. This is the remarkable feature of the arbitrage-free theory. Essentially, no matter what the drift term for the particular security is, in order to avoid arbitrage, all securities are priced as if they had drift the spot interest rate. This PDE governs most derivative products, European call options, puts, futures or forwards. However, the boundary conditions and hence the solution depends on the particular derivative. The solution to such an equation is possible analytically in a few cases, while in many others, numerical techniques are necessary. One special case of this equation deserves particular attention. In the case of geometric Brownian motion, $a(S_t, t) = \mu S_t$ and $\sigma(S_t, t) = \sigma S_t$ for constants μ, σ Assume that the spot interest rate is a constant r and that a constant rate of dividends D_0 is paid on the stock. In this case, the equation specializes to

$$-rV + \frac{\partial V}{\partial t} + (r - D_0) S \frac{\partial V}{\partial S} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 V}{\partial S^2} = 0. \tag{1.16}$$

Note that we have not used *any* of the properties of the particular derivative product yet, nor does this differential equation involve the drift coefficient μ.

We should also note that the assumption that there are no transaction costs is essential to this analysis, as we have assumed that the portfolio is continually rebalanced.

We have now seen two derivations of parabolic partial differential equations, so-called because like the equation of a parabola, they are first order (derivatives) in one variable (t) and second order in the other (x). Usually the solution of such an equation requires reducing it to one of the most common partial differential equations, the heat or diffusion equation, which models the diffusion of heat along a rod. This equation takes the form

$$\frac{\partial}{\partial t}u = k\frac{\partial^2}{\partial x^2}u \tag{1.17}$$

A solution of **??** with appropriate boundary conditions can sometime be found by the separation of variables. We will later discuss in more detail the solution of parabolic equations, both by analytic and numerical means. First, however, when can we hope to find a solution of **??** of the form $u(x,t) = g(x/\sqrt{t})$. By differentiating and substituting above, we obtain an ordinary differential equation of the form

$$g''(\omega) + \frac{1}{2k}\omega g'(\omega) = 0, \omega = x/\sqrt{t} \tag{1.18}$$

Let us solve this using MAPLE.

```
eqn := diff(g(w),w,w)+(w/(2*k))*diff(g(w),w)=0;
dsolve(eqn,g(w));
```

and because the derivative of the solution is slightly easier (for a statistician) to identify than the solution itself,

```
> diff(%,w);
```

giving

$$\frac{\partial}{\partial w}g(\omega) = C_2 \exp\{-w^2/4k\} = C_2 \exp\{-x^2/4kt\} \tag{1.19}$$

showing that a constant plus a constant multiple of the Normal $(0, 2kt)$ cumulative distribution function or

$$u(x,t) = C_1 + C_2\frac{1}{2\sqrt{\pi kt}}\int_{-\infty}^{x} \exp\{-z^2/4kt\}dz \tag{1.20}$$

is a solution of this, the heat equation for $t > 0$. The role of the two constants is simple. Clearly if a solution to **??** is found, then we may add a constant and/or multiply by a constant to obtain another solution. The constant in general is determined by initial and boundary conditions. Similarly the integral can be removed with a change in the initial condition for if u solves **??** then so does $\frac{\partial u}{\partial x}$. For example if we wish a solution for the half real $x > 0$ with initial condition $u(x,0) = 0, u(0,t) = 1$ all $t > 1$, we may use

$$u(x,t) = 2P(N(0,2kt) > x) = \frac{1}{\sqrt{\pi kt}}\int_{x}^{\infty} \exp\{-z^2/4kt\}dz, t > 0, x \geq 0.$$

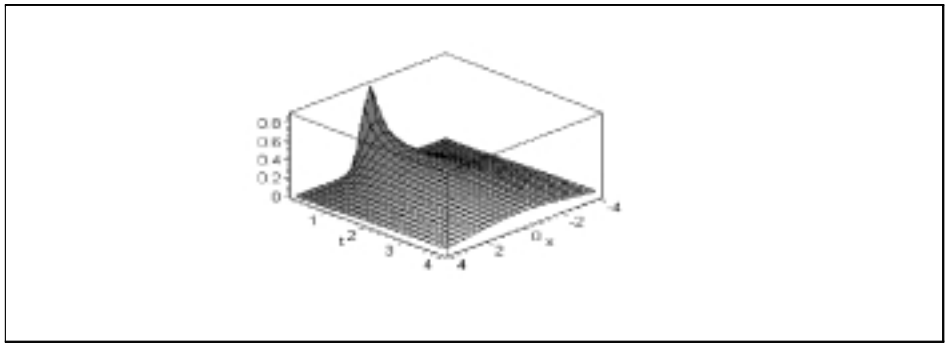

Figure 9.1:

Let us consider a basic solution to **??**:

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-x^2/4kt\} \tag{1.21}$$

This connection between the heat equation and the normal distributions is fundamental and the wealth of solutions depending on the initial and boundary conditions is considerable. We plot a fundamental solution of the equation as follows:

```
> u(x,t) := (.5/sqrt(Pi*t))*exp(-x^2/(4*t));
> plot3d(u(x,t),x=-4..4,t=.02..4,axes=boxed);
```

FIGURE 1.1: u(x,t)

As $t \to 0$, the function approaches a spike at $x = 0$, usually referred to as the "Dirac delta function" (although it is no function at all) and symbolically representing the derivative of the "Heaviside function". The Heaviside function is defined as $H(x) = 1, x \geq 0$ and is otherwise 0 and is the cumulative distribution function of a point mass at 0. Suppose we are given an initial condition of the form $u(x,0) = u_0(x)$. To this end, it is helpful to look at the solution $u(x,t)$ and the initial condition $u_0(x)$ as a distribution or measure (in this case described by a density) over the space variable x. For example the density $u(x,t)$ corresponds to a measure for fixed t of the form $\nu_t(A) = \int_A u(x,t)dx$. Note that the initial condition compatible with the above solution (1.20) can be described somewhat clumsily as "$u(x,0)$ corresponds to a measure placing all mass at $x = x_0 = 0$ ".In fact as $t \to 0$, we have in some sense the following convergence $u(x,t) \to \delta(x) = dH(x)$, the dirac delta function. We could just as easily construct solve the heat equation with a more general initial condition of the form $u(x,0) = dH(x - x_0)$ for arbitrary x_0 and the solution takes the form

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \exp\{-(x - x_0)^2/4kt\}. \tag{1.22}$$

Indeed sums of such solutions over different values of x_0, or weighted sums, or their limits, integrals will continue to be solutions to **??**. In order to achieve the initial condition $u_0(x)$ we need only pick a suitable weight function. Note that

$$u_0(x) = \int u_0(z) dH(z-x)$$

Note that the function

$$u(x,t) = \frac{1}{2\sqrt{\pi kt}} \int_{-\infty}^{\infty} \exp\{-(z-x)^2/4kt\} u_0(z) dz \tag{1.22}$$

solves **??** subject to the required boundary condition.

Problem

Use separation of variables to solve the heat equation **??** on $0 < x < 1, t > 0$ subject to initial condition $u(x,0) = u_0(x)$ and $u(0,t) = u(1,t) = 0, t > 0$.

We may also solve the heat equation with given initial/boundary conditions using the *Laplace Transforms.* The problem is to solve **??** with $k = 1$ subject to

$$u(x,0) = u_0(x), u(x,t) \text{ bounded.}$$

Define the Laplace transform with respect to the variable t to be $U(x,s) = \int_0^{\infty} u(x,t) \exp\{-st) dt$. We suppose that we may differentiate twice under the integral sign so that **??** implies

$$sU(x,s) - u_0(x) = \frac{\partial^2}{\partial x^2} U(x,s)$$

which can be solved as an ordinary differential equation for fixed s. The solution is

$$U(x,s) = \frac{1}{2\sqrt{s}} \int_{-\infty}^{\infty} \exp\{-\sqrt{s}|x-y|\} u_0(y) dy$$

The solution (1.22) to the heat equation can no be found by inverting this Laplace transform

Chapter 10

Appendix: Numerical Solutions of DE's and PDE's

10.0.1 Difference and Differential Operators and solving ODE's.

Normally, differential and partial differential equations are solved numerically by replacing the derivatives by differences. Recall that a Taylor series approximation to a function f takes the form

$$f(x+h) = f(x) + ah + \frac{b}{2}h^2 + O(h^3)$$

where $a = f'(x)$ and $b = f''(x)$ are the first and second derivative of the function at x respectively.

If we wish to estimate the first derivative a we might use the forward (first) difference approximation $\frac{f(x+h)-f(x)}{h} = a+\frac{b}{2}h+O(h^2)$ or the backward analogy $\frac{f(x)-f(x-h)}{h} = a - \frac{b}{2}h + O(h^2)$ but note that the approximation based on symmetric first differences appears better;

$$\frac{f(x+h)-f(x-h)}{2h} = a + O(h^2).$$

Similarly, in approximating the second derivative b we can use the second difference

$$\frac{f(x+h)-2f(x)+f(x-h)}{h^2} = b + O(h)$$

These approximations, generally based on central differences rather than forward or backward differences form the basis of the more precise numerical solutions to differential and partial differential equations.

Many of these approximations are easily obtained using difference and differential operator notation, designed to permit easy access to various formulae for approximating derivatives. To begin with, suppose we are interested in approximating the derivative of a function $f(x)$ of a single variable. Denote by Df the derivative function f', so

$$D^0 f(x) = f(x),\ (Df)(x) = f'(x), \quad (D^2 f)(x) = f'' ,$$

etc. Then for h small, by a Maclaurin's series expansion,

$$f(x+h) = \sum_{i=0}^{\infty} \frac{h^i D^i}{i!} f(x) = e^{hD} f(x)$$

where the second equality is really the definition of the operator e^{hD}. Thus the forward difference $\Delta f(x) = f(x+h) - f(x)$ can be written in operator notation $\Delta f(x) = e^{hD} f(x) - f(x)$ or symbolically $\Delta = e^{hD} - 1$ and

$$D = \frac{1}{h} log(1+\Delta) = \frac{1}{h}(\Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \dots)$$

Similarly

$$D^2 = \frac{1}{h^2}(\Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \dots).$$

The first term in these expansions provides a simple estimator of the first and second derivative respectively. Thus, for example, the estimator $\frac{\Delta^2}{h^2}$ provides a simple approximation to the second derivative. The next term in the series provides an indication of the order of the error. We have seen that central differences tend to provide more accurate estimates of derivatives. Introducing the notation for central differences $\delta f(x) = f(x+h/2) - f(x-h/2)$, note that

$$\delta = (1+\Delta)^{1/2} - (1+\Delta)^{-1/2} \ = (e^{hD/2} - e^{-hD/2}) = 2sinh(hD)$$

Expanding the inverse of the hyperbolic sine in a series,

$$D = \frac{1}{h} sinh^{-1}(\delta/2) = \frac{1}{h}(\delta - \frac{1}{24}\delta^3 + \dots)$$

and similarly

$$D^2 = \frac{1}{h^2}(\delta^2 - \frac{1}{12}\delta^4 + \dots).$$

The more rapid convergence of the estimators using central differences rather than forward or backward differences is apparent. Thus, the estimator of first derivative $\frac{\delta}{h}$ has error $O(h^2)$ and the estimator of second derivative

$$\frac{\delta^2}{h^2} f = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

has error $O(h^2)$.

Let us now see how these difference approximations to derivatives can be used to numerically solve an ordinary differential equation. The first and simplest method is the Euler method, but there is a hugh number of more sophisticated alternatives including the trapezoidal rule, or the improved Euler method, the modified Euler method, and the Runge-Kutta method. Consider the simplest, the forward Euler method and the simple differential equation of second order

$$\frac{d^2}{dt^2}y(t) = 100e^{-10t} + 100e^{10t} \tag{2.1}$$

with initial conditions $y(0) = 2, y'(0) = 0$. Note that if we replace the second derivative by a second difference, the equation becomes

$$\frac{1}{h^2}\Delta^2 y(t) = 100e^{-10t} + 100e^{10t}, t = 2h, 3h, ... \tag{2.2}$$

allowing us to approximate values of the function on the lattice of points of the form $kh, k = 2, 3, ...$ using the initial values. The values corresponding to $k = 0, 1$ result from the initial conditions.

Problem.

Solve the equation (2.1) with initial conditions $y(0) = 2, y'(0) = 0$ numerically using (2.2) recursively and $h = 0.1$.

Use the simplest Euler approximation for the initial conditions by replacing $y'(0) = 0$ by the condition $\Delta y(0) = 0$. Compare with the exact solution.

Problem.

Solve the equation (2.1) with initial conditions $y(0) = 2, y'(0) = 0$ numerically using the recursion

$$\frac{1}{h^2}\delta^2 y(t) = 100e^{-10t} + 100e^{10t}, t = 2h, 3h, ... \tag{2.3}$$

and $h = 0.1$. Compare with the exact solution.

The improved Euler method is a simple modification of the Euler method for solving an equation of the form $y'(t) = f(t, y)$. The problem with Euler's method is it uses as slope the derivative at one end only of the interval and an improved approximation of the derivative using the average slope of at the two endpoints is usually more accurate. In this case, we make a preliminary estimate of $y((k+1)h)$ denoted by $y^* = hf(kh, y(kh)) + y(kh)$ and then solve for $y((k+1)h)$ the equations

$$y((k+1)h) = \frac{f(kh, y(kh)) + f((k+1)h, y^*)}{2}h + f(kh, y(kh)), k = 1, 2, ...$$

The Runge-Kutta method differs from the above two methods only in the way in which the slope of the line is estimated. For example we may take an average of the slope $f(t,y)$ at points $t = kh, (k+1/2)h, (k+1)h$ and *various* corresponding values of y in the interval. In fact it is possible to select 4 combinations of values (t,y) so that the approximation is perfect when $y(t)$ is a polynomial of degree 4. In general. we may select k points so that the approximation is perfect for a polynomial of degree k. This is a general description of the Runge-Kutta method.

$y((k+1)h) =$ (weighted average of values of $f(t,y)$ in the interval)$h + f(kh, y(kh)), k = 1, 2, ...$

Problem:

Use Euler's method and the improved Euler's method to solve the equation $y'(t) = ty(t)$ using step size $h = 0.1$ and 0.05 and initial condition $y(0) = 1$. Compare the solution with the exact solution on the interval $0 < t < 1$. How does the error change as we (a) increase h, (b) let t get farther away from the initial condition $t = 0$.

Example: Numerical Methods for ODE's in MAPLE:

Consider the second order differential equation

$$\frac{d^2}{dt^2}y(t) = ty(t)$$

with initial conditions $y(0) = 0, y'(0) = 1$.. We wish a numerical solution. There are a number of methods available in MAPLE including classical, lsode, mgear, rfk45, taylorseries,dverk78. Each has a number of options. The simplest method is classical[foreuler] for forward euler.

> deq := diff(y(t), t$2) = y(t)*t;

> ans:=dsolve(deq,y(t),numeric,method=classical[foreuler],
initial=array([0,1]), start=0);

We may then plot the result for $0 < t < 2$ as follows;

> with(plots);

> odeplot(ans,[t,y(t)],0..2);

Compare with an alternate more accurate method;

> ans2 := dsolve(deq3, y(t), numeric, method=mgear[msteppart], initial=array([2,0]), start=0);

> odeplot(ans2,[t,y(t)],0..2, axes=boxed);

> ans(2);

[t = 2, y(t) = 3.588338680829067, y'(t) = 4.643239714056109]

> ans2(2);

[t = 2, y(t) = 3.611076600977132, y'(t) = 4.676276660728110]

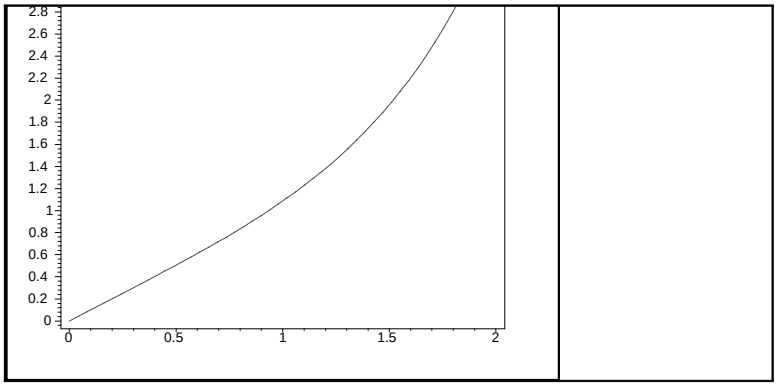

Figure 10.1:

Stability:

Instability is the possible accumulation of relatively small errors over time, and results in errors that get larger as $t \to \infty$. A stable numerical method is forgiving of small errors over time, while an unstable method may exaggerate these errors until they are the dominant part of the function. For an example of instability, consider the ordinary differential equation

$$\frac{dy}{dt} = y - t$$

with solution given by $y = (y_0 - 1)e^t + (t + 1)$ where the initial value is $y(0) = y_0$. Suppose that true value of the y_0 is 1 but it has been approximated numerically by $1 + \epsilon$ with ϵ small. The error might, for example, be the error due to expressing the initial condition with floating point arithmetic. For partial differential equations, often the initial condition is also approximated using a lattice of values, and this results in initial error. Note that even if the computer were able to obtain the exact analytic solution to the differential equation with the erroneous initial value, the difference between the true solution $t + 1$ and the approximation is ϵe^t (assuming no subsequent errors). This grows without bound as $t \to \infty$.

10.0.2 Numerical Methods for P.D.E.'s. Explicit Finite Difference Method.

We now return to the heat conduction or diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{2.4}$$

subject to the initial condition $u(x, 0) = u_0(x)$. A simple approach to solving this equation numerically is to approximate the first derivative on the left and

the second derivative on the right using finite differences and solve the resulting system of equations for the approximate value of the function on grid points. Suppose, for example we replace the derivative on the left side by a forward difference and the second derivative on the right side by a central second difference, This allows us to solve for the values of $u(x,t+\Delta t)$ in terms of the values of $u(x,t)$for all $x = n\Delta x$ and $t = m\Delta t,\ m = 0,1,2,\ldots.$ Denoting $U_{i,j} = u(i\Delta x, j\Delta t)$the equation becomes

$$U_{i,j+1} - U_{i,j} = \frac{\Delta t}{(\Delta x)^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})$$

or

$$U_{i,j+1} = rU_{i+1,j} + (1-2r)U_{i,j} + rU_{i-1,j} \tag{2.5}$$

where $r = \frac{\Delta t}{(\Delta x)^2}$. It turns out that this strategy of solving for $U_{i,j+1}$ in terms of its predecessors is not stable, i.e. the cumulative error does not converge, unless $r < 1/2$. This imposes a lower limit on the possible size of Δx, viz. $(\Delta x)^2 > 2\Delta t$. This method is called the *explicit finite difference method*, not because it is uncensored (although this is also true) but because we are able to solve explicitly for the values at time step $j+1$ using the values at time step j.

Consider the stability of an equation of the form (2.5). Let $U^*_{i,j}$ be the solution to (2.5) beginning with the exact initial conditions and $U_{i,j}$ the values obtained solving (2.5) if we begin with the (slightly) erroneous values $U_{i,0}$. Then the error due to the error in the initial condition is $E_{i,j} = U^*_{i,j} - U_{i,j}$. It is easy to see that this, since it is a linear combination of terms satisfying (2.5), also satisfies the equation (2.5). Suppose, for example, $E_{i,j} = \epsilon\lambda^j sin(i\omega)$ for each $i,\ j$ and for some real λ and frequency ω and for (small) ϵ. In general it is possible to show that the general solution of (2.5) is a linear combination of such terms with different values of λ, ω. Substituting in (2.5) and solving,

$$\lambda = 1 + r\frac{sin((i+1)\omega) - 2sin(i\omega) + sin((i-1)\omega)}{sin(i\omega)}.$$

Since $sin(A+B) = sin(A)cos(B) + cos(A)sin(B)$ this results in

$$\lambda = 1 + 2r[cos\omega - 1] = 1 - 4rsin^2(\omega/2)$$

Note that if $r > 1/2$, it is possible that for some frequencies ω, the corresponding value of $\lambda < -1$. This results in λ^j blowing up in magnitude as $j \to \infty$, i.e. the absolute value of at least some of the errors will go to infinity as $j \to \infty$. We have a similar instability if $r < 0$. However, if $0 < r < 1/2$ the errors will all converge to 0 as $j \to \infty$ and the solution to (2.5) is stable. For the explicit method to be stable, we need $\Delta t < \frac{1}{2}(\Delta x)^2$.

Problem:

What is the general form for the solution of the difference equation $y_{k+2} - 3y_{k+1} + 2y_k = (1/2)^k, k = 1,2,...$?

(Hint: solve as you would a differential equation: first, find the roots of the characteristic equation and hence all solutions to the homogeneous difference equation. Then find a particular solution to the homogeneous equation.)

Problem:

Consider solving the partial difference equation (2.5) by separation of variables. i.e. assume a solution exists of the form $U_{i,j} = d_i\phi_j$ for sequences d_i, ϕ_j. Show that the general solution is of the form $\phi_j = \epsilon\lambda^j$ for some λ and $d_i = C_1 \sin(i\omega) + C_2 \cos(i\omega)$ for some C_1, C_2, ω.

The next problem shows that if we were to replace the left side of (2.4) by a central difference that should be more accurate, we nevertheless obtain a method which is unstable for all values of r.

Problem:

We have already seen that *central differences* are generally more precise estimates of derivatives than are forward or backward differences. Suppose we replace (2.2) using a central difference estimator of $\frac{\partial u}{\partial t}$. Show that we obtain the partial difference equation

$$U_{i,j+1} - U_{i,j-1} = \frac{2\Delta t}{\Delta x^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) \tag{2.6}$$

The *Fourier* method approaches the stability of this difference equation by mapping the changes in the two directions into the complex plane. For example, suppose we consider a solution which is a constant multiple of $U_{i,j} = z^i q^j$ for some complex $z = e^{\beta\Delta x\sqrt{-1}}$ and real $q = e^{\alpha\Delta t}$, where α, β are arbitrary real numbers. General solutions could be obtained by taking linear combinations of such terms for different values of α and β and then taking the real or imaginary part of the linear combination. Substituting in (2.6), obtain the equation

$$q - \frac{1}{q} = 4r[cos(\beta\Delta x) - 1] = -8rsin^2(\beta\Delta x/2).$$

Show that even for r close to zero but positive, there is a solution for q with $|q| > 1$ leading to an exploding solution to the difference equation as $j \to \infty$. This indicates instability in the difference equation (2.6) making it undesirable for any value of $r = \frac{\Delta t}{\Delta x^2}$.

For the explicit finite-difference method it can be proven that if we Δt and Δx tend to zero in such a way that the ratio $r = \frac{\Delta t}{(\Delta x)^2}$ remains between 0 and $\frac{1}{2}$ then the finite difference approximation converges to the actual solution. To see this we can consider the difference between the exact solution and the finite-difference approximation:

$$D_{n,m} = U_{n,m} - V_{n,m},$$

where $V_{n.m}$ is the solution of the explicit finite-difference method. From the Taylor's theorem applied to the forward and the central difference approximation we can find that

$$D_{n.m} = (1-2r)D_{n,m} + r(D_{n+1,m} + D_{n-1,m}) + \Delta t(R_1\Delta t + R_2(\Delta x)^2),$$

where R_1 and R_2 are two bounded in absolute value functions. For $\hat{D}^m = \max_n |D_{n,m}|$ and $\hat{D}^{m+1} = \max_n |D_{n,m+1}|$, the largest errors at time-step m and $m+1$ respectively, we get

$$\hat{D}^{m+1} \leq (|1-2r| + 2|r|)\hat{D}^m + \Delta t(R_1\Delta t + R_2(\Delta x)^2).$$

Provided that $0 \leq r \leq \frac{1}{2}$, we have $|1-2r| + 2|r| = 1$, and hence

$$\hat{D}^{m+1} \leq \hat{D}^m + \Delta t(R_1\Delta t + R_2(\Delta x)^2).$$

By induction it follows

$$\hat{D}^{m+1} \leq \hat{D}^0 + (m+1)\Delta t(R_1\Delta t + R_2(\Delta x)^2),$$

so if we assume zero error at time-step $m = 0$, which we can do since $V_{n,0} = U_{n,0}$ from the initial condition, we see that

$$\hat{D}^{m+1} \leq (m+1)\Delta t(R_1\Delta t + R_2(\Delta x)^2) \to 0 \quad \text{as} \quad \Delta t \to 0,$$

which proves that the method gives an approximation which converges to the actual solution. A modification to this argument shows that if $r > \frac{1}{2}$ the error actually grows without bound as we let $\Delta t \to 0$. Therefore, for explicit finite-difference method for the diffusion equation the stability and the convergence problems are equivalent.

The moral of this story is that one should be careful with discrete approximations to continuous differential equations. The stability of the method and the convergence of the numerical solution are among the many concerns. Various methods have been proposed that solve some of the stability problems encountered above, and these are implemented in many packages. We discuss these methods further in the following sections.

10.0.3 Implicit Finite Difference Solutions to the Diffusion Equation.

The implicit finite difference scheme for solving (2.4) is similar to the explicit method in that we use first differences for the derivative on the left hand side of (2.4) but we use the backward difference instead. This results in a system of linear equations in $u(x,t)$ in terms of the values of $u(x, t-\Delta t)$for all $x = n\Delta x$ and $t = m\Delta t$, $m = 0, 1, 2, \ldots$, and these equations fortunately take a simple form.

$$U_{i,j} - U_{i,j-1} = r(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}) \tag{2.7}$$

or

$$-rU_{i+1,j} + (1+2r)U_{i,j} - rU_{i-1,j} = U_{i,j-1}$$

where $r = \Delta t/(\Delta x)^2$. Boundary conditions determine the value of $U_{i,j}$on the upper and lower boundary of a rectangle, e.g. for $i = \pm N$, and all j, while initial conditions prescribe the value of the function at time 0.

The implicit method requires solving this system of equations at time j in terms of the solution at time $j-1$. There are three ways in which solutions to systems such as this can be approached. The first is matrix inversion. For large matrices this is very difficult. The second method uses an LU decomposition of the matrix into two components, L , a lower triangular matrix, and U an upper triangular matrix. The advantage of triangular matrices is that systems of equations of the form $Lx = y$, for example, can be easily solved by simple substitution. A third method, called successive over relaxation (SOR) is a method for solving the equation

$$MU_{*,j} = b_j$$

for a matrix M and vector b_j. Here $U_{*,j}$ denotes the column vector $(U_{-N,j},\ U_{-N+1,j}, \ldots ,\ U_{N,j})$, b depends on $U_{*,j-1}$ and the boundary conditions and

$$M = \begin{pmatrix} 1+2r & -r & 0 & . & . & 0 \\ -r & 1+2r & -r & 0 & . & 0 \\ 0 & -r & 1+2r & -r & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -r & 1+2r \end{pmatrix}.$$

This is then rewritten

$$U_{i,j} = \frac{1}{1+2r}[b_{i,j} + r(U_{i-1,j} + U_{i+1,j})] \tag{2.8}$$

and then starting with an initial estimate of the solution (always for fixed j), substitute in the left to obtain an updated estimate on the right. This continues until there is little change in the components of $U_{*,j}$. There is one simple modification of this that is normally applied in practice, motivated by the fact that sequences often converge in such a way that the errors decrease like a geometric series. When this is the case, convergence may be accelerated by over-relaxation (extrapolating the solution). This is most easily described with an example. Consider the sequence $x_n = 0.1 + 0.9x_{n-1} = 0.1000, 0.1900, 0.2710, 0.3439, 0.4095, 0.4686, 0.5217, \ldots$ if we begin with $x_1 = 0.1$. Clearly this sequence converges rather slowly to 1. Consider the sequence defined by $y_n = 0.1 + 0.9x_{n-1}$, and $x_n = x_{n-1} + \omega(y_n - x_{n-1})$ for some over-relaxation parameter $\omega > 1$. How much faster does the sequence converge if $\omega = 5$? The first few terms are$0.1, 0.595, 0.7975, 0.8987$ which is clearly a substantial improvement. Of course $\omega \approx 9$ provides even faster convergence.

Problem (SOR):

Consider the sequence

$$x_n = \frac{x_{n-1}}{2} + \frac{1}{x_{n-1}}$$

$x_1 = 10$. Investigate the speed of convergence of this sequence and one obtained by successive over-relaxation. For what value of the parameter ω does the speed of convergence seem to be greatest?

The SOR method consists of first solving

$$Y_{i,j}^{k+1} = \frac{1}{1+2r}[b_{i,j} + r(U_{i-1,j}^{k+1} + U_{i+1,j}^{k})]$$

and then putting

$$U_{i,j}^{k+1} = U_{i,j}^{k} + \omega(Y_{i,j}^{k+1} - U_{i,j}^{k}), k = 1, 2, 3, ...$$

until convergence.

10.0.4 The Crank-Nicolson Method

This method differs from both the forward and backward approach of the explicit and fully implicit methods primarily in that the first forward or backward difference is now replaced by a symmetric first difference, improving the accuracy to $O(\Delta t)^2$. Note that

$$\frac{\partial u}{\partial t}(x, t + \Delta t/2) \ = \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} + O(\Delta t)^2.$$

Similarly, we can obtain an approximation to the right hand side accurate to the same order by averaging the symmetric second difference at points t and $t + \Delta t$, i.e. using

$$\frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{2(\Delta x)^2}$$

$$+\frac{u(x + \Delta x, t + \Delta t) - 2u(x, t + \Delta t) + u(x - \Delta x, t + \Delta t)}{2(\Delta x)^2}$$

Setting these equal results in a system of linear equations in the values of $u(x, t + \Delta t)$ solvable in terms of the values of $u(x, t)$ for all x. These equations take the form;

$$U_{i,j+1} - U_{i,j} = \frac{r}{2}\{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1} + U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\}$$

The system of equations that result from the Crank-Nicolson approach is similar to the form shown of the implicit method

$$MU_{*,j+1} = BU_{*,j} + b_{*,j}$$

where

$$M = \begin{pmatrix} 1+r & -r/2 & 0 & . & . & 0 \\ -r/2 & 1+r & -r/2 & 0 & . & 0 \\ 0 & -r/2 & 1+r & -r/2 & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & -r/2 & 1+r \end{pmatrix},$$

$$B = \begin{pmatrix} 1-r & r/2 & 0 & . & . & 0 \\ r/2 & 1-r & r/2 & 0 & . & 0 \\ 0 & r/2 & 1-r & r/2 & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & r/2 & 1-r \end{pmatrix},$$

and $b_{*,j}$ is a vector of zeros in the center, with non-zero elements at the boundary which depend on the boundary conditions.

As we will show in the next section, the Crank-Nicolson method has convergence and stability properties similar to that of the implicit method, the method is stable if $r > 0$.

10.0.5 Stability and Consistency of the Crank-Nicolson.

We now consider the stability of the Crank-Nicolson method using a matrix approach. Consider once again the numerical solution to the heat equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u$$

on $0 < x < 1$ subject to the initial condition $u(x,0) = f(x)$ and some boundary conditions $u(1,t) = g_1(t),\ \ u(0,t) = g_2(t)$ for all t . We have seen that methods such as the Crank-Nicolson involve discretizing in both the time and space directions, and then writing a recursive formula of the form

$$MU_{*,j+1} = BU_{*,j} + b_{*,j}$$

where the matrices M and B are defined in the last section . Note that

$$U_{*,j+1} = PU_{*,j} + M^{-1}b_{*,j} = \ldots$$

$$= P^{j+1}U_{*,0} + P^jM^{-1}b_{*,0} \ + \ P^{j-1}M^{-1}b_{*,1} + \ldots + M^{-1}b_{*,j} \qquad (2.9)$$

where P is the matrix $M^{-1}B$. Let the eigenvalues of the matrix P be $\lambda_1, \lambda_2, \ldots$ and the corresponding normalized eigenvectors $v_1, v_2, \ldots$. Recall from linear algebra that P^j has the same eigenvectors with eigenvalues λ_n^j. Then provided that these eigenvectors span the space, we can write $U_{*,0} = \sum_i \alpha_i v_i$ for some constant coefficients α_i, and then

$$P^j U_{*,0} = \sum_i \alpha_i \lambda_i^j v_i. \tag{2.10}$$

Similarly, each of the other terms on the right hand side of (2.9) can also be written in a similar form with coefficients involving powers of the eigenvalues λ_i^j. For stability with respect to small errors in the initial conditions, we would like to be assured that as j increases, the vector on the right hand side of (2.10) remains bounded. This is clearly the case if all of the eigenvalues λ_i are less than or equal to 1. A similar argument implies stability with respect to small errors in the initial conditions. Thus the condition for matrix stability is more generally, for a recursion of the form $MU_{*,j+1} = BU_{*,j} + b_{*,j}$, that the maximum eigenvalue of P , i.e. the maximum root of the equation $det(M - \lambda B) = 0$ is less than one in absolute value. In the case of the Crank Nicolson method, this results in the same condition as does the Fourier method, i.e. that $0 < r$.

Besides stability, which ensures that errors do not tend to grow without bound, we generally require some indication that the discrete approximation to a partial DE is close to the continuous solution. Let the difference operator F correspond to the discrete Crank-Nicolson method.

$$FU_{i,j} = \frac{U_{i,j+1} - U_{i,j}}{\Delta t} - \frac{\{U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1} + U_{i+1,j} - 2U_{i,j} + U_{i-1,j}\}}{2(\Delta x)^2}$$

Then the difference between the discrete Crank-Nicolson and the corresponding differential operator can be expanded, as was done in section 2.1, to obtain

$$FU_{i,j} - (\frac{\partial}{\partial t}u - \frac{\partial^2}{\partial x^2}u) = \frac{1}{2}(\Delta t)\frac{\partial}{\partial t}[\frac{\partial}{\partial t}u - \frac{\partial^2}{\partial x^2}u] - \frac{1}{12}(\Delta x)^2\frac{\partial^4 u}{\partial x^4} \tag{2.11}$$

$$+\frac{1}{6}(\Delta t)^2\left(\frac{\partial^3 u}{\partial t^3} - \frac{3}{2}\frac{\partial^4 u}{\partial x^2 \partial t^2}\right) + \ldots$$

Note that as long as both Δt and Δx approach 0, the error in the approximation on the right side of (2.11) also approaches 0 (assuming sufficient smoothness of the solution). Thus the numerical solution is consistent under these conditions.

10.0.6 The Method of Lines.

In general, the method of lines is a device which reduces a partial differential equation to a coupled system of ordinary differential equations. Consider for an example the non-linear equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u + (\frac{\partial}{\partial x}u)^2$$

Suppose we use symmetric first and second order differences for the derivatives on the right side of this equation, but continue to leave the left side as a derivative. Then at a given point (x,t) the equation takes the form

$$\frac{\partial}{\partial t}u(x,t) = \frac{1}{h^2}[u(x+h,t) - 2u(x,t) + u(x-h,t)] + \frac{1}{4h^2}[u(x+h,t) - u(x-h,t)]^2.$$

If we now denote $U_i(t) = u(ih,t), \quad i = 0,1,2,\ldots$, then this becomes a coupled system of first order differential equations

$$U_i'(t) = \frac{1}{h^2}[U_{i+1}(t) - 2U_i(t) + U_{i-1}(t)] + \frac{1}{4h^2}[U_{i+1}(t) - U_{i-1}(t)]^2, \quad i = 1,\ 2,\ldots$$

which, together with the appropriate initial or boundary conditions, may be solved, for example in MAPLE, to obtain an approximate solution to the PDE.

10.0.7 Finite Elements and the Galerkin Method.

Much of the theory of ordinary and partial differential equations parallels corresponding results in linear algebra. We begin with some elementary results concerning linear operators. Consider a vector space spanned by a complete set of vectors $v_1, v_2, \ldots$. By this we mean that any element of the vector space can be written as a limit of a linear combination of the spanning vectors $lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i v_i$. Suppose the vector space has an inner product denoted by (u,v). A complete inner product space is called a Hilbert Space. In the case of a finite-dimensional vector space, this inner product is usually the dot product between the two coefficient vectors. However, we will deal here with an infinite dimensional vector space in which the vectors are functions. Now let A be a positive linear operator on the vector space. This means that when v s a vector, so is Av and it is linear; $A(v_1+v_2) = Av_1 + Av_2$. Of course we usually represent linear operators in the finite dimensional case by matrices, but when the vectors consist of functions, then objects like derivatives of the function, first differences, etc. are also linear operators. A linear operator is positive definite if it is symmetric ($(Av,u) = (v,Au)$ for all u,v) and $(Av,v) \geq C||v||^2$ for all v and some $C > 0$.

We now consider a particular minimization problem, easiest to interpret and prove in the finite dimensional case, but useful in the infinite dimensional problems in differential equations as well. To motivate the result, pretend for the moment that the vectors f,u,v are finite-dimensional column vectors and A is a positive definite matrix. Suppose we wish to minimize the length of the residuals $Av - f$ over possible vectors v restricted to some linear subspace of Euclidean space. The notion of distance is adapted to the positive definite matrix A, so that the squared distance is given by $(Av-f)^T A^{-1}(Av-f) = (Av,v) - 2(f,v) + (f,f)$. Since we are minimizing over v, it is equivalent to minimize the quantity $(Av,v) - 2(f,v)$ and this minimization is equivalent to minimizing the "length" of the residuals $Av - f$.

Proposition

Suppose A is a linear operator on a given domain D_A in a Hilbert space. Suppose we wish to minimize the quadratic functional

$$(Av, v) - 2(f, v) \tag{2.12}$$

over all vectors v for a given function $f \in D_A$. Then the following are equivalent

(a) v^* is the desired minimum.

(b) v^* is the unique solution to the equation $Av^* = f$.

(c) Provided D_A is spanned by base vectors $v_1, \ldots v_n$ scaled to have length 1, then $v^* = \sum a_i v_i$ where $a_1, a_2, \ldots a_n$ solves the system of equations

$$\sum_i (Av_i, v_j)a_i = (f, v_j), \quad j = 1, 2, \ldots n \tag{2.13}$$

The last equation above can be written in simpler form and applies generally, even in circumstances in which the operator is not positive definite. Suppose we wish to solve $Av^* = f$. Suppose there are finitely many basis vectors $v_1, \ldots v_n$ that we think "almost" span the desired space. Then any proposed solution that is a linear combination of these vectors, say $v_n^* = \sum_{i=1}^n a_i v_i$ must be assessed through the size of the residuals $Av_n^* - f$. As in regression, we have achieved the projection of the solution onto this finite dimensional subspace only if these residuals are perpendicular to each of the basis vectors, i.e. if $(Av_n^* - f, v_j) = 0, \quad j = 1, 2, \ldots n$. This is the system of equations (2.13).

Now consider as a simple example the heat equation

$$\frac{\partial}{\partial t}u = \frac{\partial^2}{\partial x^2}u, \quad 0 < x < 1, \quad 0 < t$$

subject to initial and boundary conditions

$$u(x, 0) = 1, \quad 0 < x < 1$$

$$u(0, t) - \frac{\partial}{\partial x}u(0, t) = 0, \quad t > 0$$

$$\frac{\partial}{\partial x}u(1, t) = 0, \quad t > 0$$

We begin with a base of functions of x that satisfy the boundary conditions. For example, the polynomials

$$v_j(x) = (1 + x) - \frac{x^{j+1}}{j + 1}$$

all satisfy the boundary conditions

$$v_j(0) - v_j'(0) = 0, \qquad v_j'(1) = 0.$$

Now consider approximating a solution in the form

$$\bar{u}(x,t) = a_1(t)v_1(x) + a_2(t)v_2(x)$$

In this example, the space of functions should include (possibly infinite) linear combinations of twice differentiable functions of x which satisfy the boundary conditions, with coefficients that are differentiable functions of t. The inner product is $(u,v) = \int_0^1 u(x,t)v(x,t)dx$. The residual at time t is

$$A\bar{u} - f = \frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}, \quad t > 0$$

and

$$1 - \bar{u}(x,0), \quad t = 0.$$

Then the conditions (2.13) lead to the fact that the residuals are orthogonal to the two basis vectors, i.e. that

$$\int_0^1 (1 - \bar{u}(x,0))v_j(x)dx = 0, \quad j = 1,\ 2$$

$$\int_0^1 [\frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}]v_j(x)dx = 0, \quad j = 1,\ 2 \tag{10.1}$$

These equations reduce to

$$\frac{9}{5}a_1(0) + \frac{691}{360}a_2(0) = \frac{4}{3}$$

$$\frac{691}{360}a_1(0) + \frac{1291}{630}a_2(0) = \frac{17}{12}$$

and

$$\frac{9}{5}a_1'(t) + \frac{691}{360}a_2'(t) + \frac{4}{3}a_1(t) + \frac{17}{12}a_2(t) = 0$$

$$\frac{691}{360}a_1'(t) + \frac{1291}{630}a_2'(t) + \frac{17}{12}a_1(t) + \frac{23}{15}a_2(t) = 0$$

and the solution to this system of equations is given by

$$a_1(t) = 0.586e^{-.742t} + 2.448e^{-11.770t}, \quad a_2 = 0.144e^{-.742t} - 2.295e^{-11.770t}$$

and so the approximate solution to the initial-boundary value problem is

$$\bar{u}(x,t) = 0.586e^{-.742t} + 2.448e^{-11.770t}(1 + x - x^2/2)$$

$$+0.144e^{-.742t} - 2.295e^{-11.770t}(1 + x - x^3/3).$$

Of course, because the second exponent is so large and negative, it is clear that the first term in these coefficients dominates for moderate or large values of t.

Example.

The methods we consider here are generally applicable to numerically solving an equation of the form

$$\frac{\partial}{\partial t}u + A(u) = 0$$

where $A(u)$ is usually required to be a positive definite linear differential operator with respect to x. An example of a positive definite operator is one of the form $A(u) = -\frac{\partial}{\partial x}[a(x,t)\frac{\partial}{\partial x}u(x,t)]$ for a non-negative function a. It is easy to show, for example, that the operator $-\frac{\partial^2}{\partial x^2}$ is positive definite on a suitable subspace determined by the boundary conditions. We return to solving equation (2.13) on $0 < x < \pi$, $0 < t < 1$ under the boundary, initial conditions $u(x,0) = 1$, $0 < x < \pi$, and $u(0,t) = u(\pi,t) = 0$. The method of separation of variables leads to solutions of the form $u_n(x,t) = e^{-n^2t}sin(nx)$ and it is easy to see in this case that if we try a linear combination of the form $u(x,t) = \sum_n a_n u_n$ the equation and boundary conditions are satisfied for $a_n = \frac{4}{\pi n}$, $n = 1,3,5,\ldots$, and otherwise $a_n = 0$. Therefore in this case, an explicit solution to the equation is known;

$$u(x,t) = \sum_{n=1,3,5,\ldots} \frac{4}{\pi n}e^{-n^2t}sin(nx).$$

Suppose for the moment we did not know this solution. We could attempt a solution as a linear combination of finitely many of the basis vectors $v_n(x) = \frac{sin(nx)}{n}$, $n = 0,1,2,\ldots 5$ so the attempted solution might take the form $\bar{u}(x,t) = a_0(t) + \sum_{n=1}^{5} a_n(t)v_n(x)$. Proceeding as in the last example, if we require that the residuals are orthogonal to the basis vectors, we obtain the equations

$$a_0(t) = 1$$

$$\int_0^\pi [\frac{\partial}{\partial t}\bar{u} - \frac{\partial^2}{\partial x^2}\bar{u}]v_j(x)dx = 0, \quad j = 0,1,\ldots 5 \tag{2.14}$$

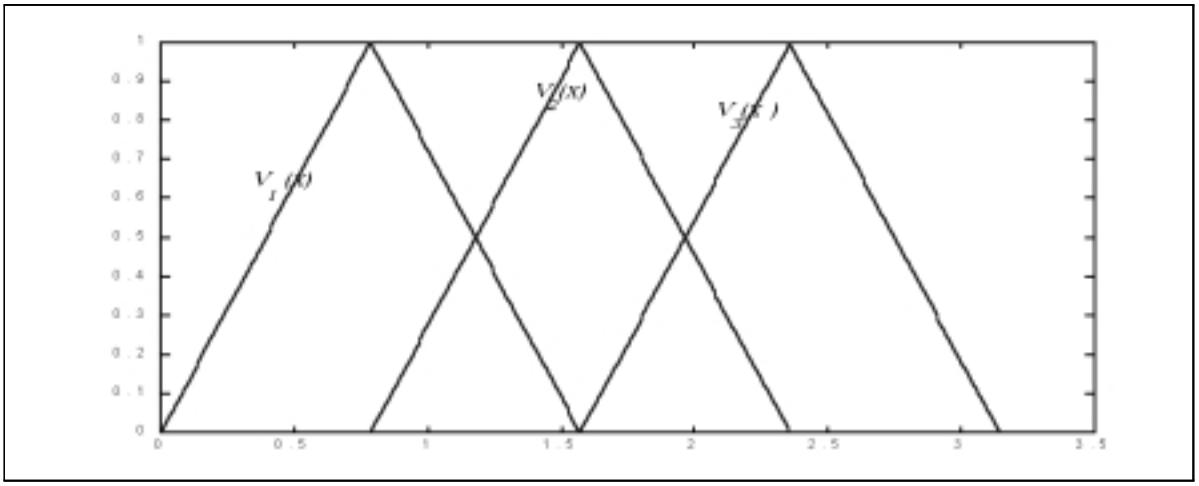

Figure 10.2:

which reduces a coupled system of first order differential equations that may be solved for the coefficients $a_j(t)$. Often the system of equations (2.14) is simplified further by discretizing time, i.e. replacing $\frac{\partial}{\partial t}\bar{u}$ by a first forward difference for small (time) step size h . Then the equations are simply linear equations, with no derivatives involved, and they are solved sequentially in $t = jh,\ j = 1, 2,$.

There is an alternative for the base functions v_n commonly used and often referred to as a finite element method. Suppose, for example, we use the three simple spline functions represented by the triangles below.

Clearly these functions are continous,

and indeed have piecewise continuous first derivatives. For example, $v_j(x)$ can be recovered by integrating its first derivative. Unfortunately, the second derivatives do not have this property and if we express $\bar{u}$ as a linear combination of these base functions, we still not be able to apply the operator A to it, since this requires second derivatives with respect to x. This seemingly harmless failure has led to a mountain of mathematical literature providing a weak interpretation of differential equations. Roughly speaking, functions are defined by the result upon integration after multiplication by well-behaved test functions. For example, the dirac delta function could be defined either as a measure, or as the "function"δ providing $\int \delta(x)\phi(x)dx = \phi(0)$ for all smooth functions ϕ. Under a weak interpretation, in order to solve for the coefficients above, we require integrals of the form $\int[-\frac{\partial^2}{\partial x^2}v_n(x)]v_j(x)dx$ and this can be interpreted using integration by parts. For example if we are integrating over a strip $a < x < b$ and if the basis vectors $v_j(x)$ are zero at $x = a, x = b$ (because

boundary conditions imply that the solution is also zero on this boundary),

$$\int [-\frac{\partial^2}{\partial x^2} v_n(x)] v_j(x) dx = \int [\frac{\partial}{\partial x} v_n(x)][\frac{\partial}{\partial x} v_j(x)] dx \tag{10.2}$$

even when the second derivative fails to exist. Notice that this identity also shows that the linear operator $A = -\frac{\partial^2}{\partial x^2}$ satisfies $(Av.v) \geq 0$ and is therefore positive definite.

Summary

We have seen several alternative methods that can be used to numerically solve a system of the form

$$\frac{\partial}{\partial t} u + A(u) = f$$

for positive definite operator A. We may discretize both time and space, using a method such as Crank-Nicolson. We may discretize space and not time, resulting in a system of first order differential equations in t. Alternatively, we may discretize time. Putting $u_j(x) = u(j\Delta t, x)$ this requires solving a system of the form

$$\frac{u_j(x) - u_{j-1}(x)}{\Delta t} + A u_j(x) = 0$$

with the inherited boundary conditions. This may result in a system of second order differential equations in x which may be solved analytically or numerically (e.g. by finite element methods).

10.0.8 Solution of the Diffusion Equation.

In this section we consider the general solution to the diffusion equation of the form (1.15), rewritten as

$$\frac{\partial V}{\partial t} = r_t V - r_t S_t \frac{\partial V}{\partial S} - \frac{\sigma^2(S_t, t)}{2} \frac{\partial^2 V}{\partial S^2} \tag{2.15}$$

where S_t is an asset price driven by a diffusion equation

$$dS_t = a(S_t, t)dt + \sigma(S_t, t)dW_t, \tag{2.16}$$

$V(S_t, t)$ is the price of an option on that asset at time t, and $r_t = r(t)$ is the spot interest rate at time t. We assume that the price of the option at expiry T is a known function of the asset price

$$V(S_T, T) = V_0(S_T). \tag{2.17}$$

Somewhat strangely, the option is priced using a related but not identical process (or, equivalently, the same process under a different measure). Recall from the

backwards Kolmogorov equation (1.2) that if a related process X_t satisfies the stochastic differential equation

$$dX_t = r(X_t, t)X_t dt + \sigma(X_t, t)dW_t \tag{2.18}$$

then its transition kernel $p(t, s, T, z) = \frac{\partial}{\partial z}P[X_T \leq z | X_t = s]$ satisfies a partial differential equation similar to (2.15);

$$\frac{\partial p}{\partial t} = -r(s, t)s\frac{\partial p}{\partial s} - \frac{\sigma^2(s, t)}{2}\frac{\partial^2 p}{\partial s^2} \tag{2.19}$$

For a given process X_t this determines one solution. For simplicity, consider the case (natural in finance applications) when the spot interest rate is a function of time, not of the asset price; $r(s, t) = r(t)$. To obtain the solution so that terminal conditions is satisfied, consider a product

$$f(t, s, T, z) = p(t, s, T, z)q(t, T) \tag{2.20}$$

where

$$q(t, T) = exp\{-\int_t^T r(v)dv\}$$

is the discount function or the price of a zero-coupon bond at time t which pays 1\$ at maturity.

Let us try an application of one of the most common methods in solving PDE's, the "lucky guess" method. Consider a linear combination of terms of the form (2.20) with weight function $w(z)$. i.e. try a solution of the form

$$V(s, t) = \int p(t, s, T, z)q(t, T)w(z)dz \tag{2.21}$$

for suitable weight function $w(z)$. In view of the definition of p as a transition probability density, this integral can be rewritten as a conditional expectation:

$$V(t, s) = E[w(X_T)q(t, T)|X_t = s] \tag{2.22}$$

the discounted conditional expectation of the random variable $w(X_T)$ given the current state of the process, where the process is assumed to follow (2.18). Note that in order to satisfy the terminal condition **??**, we choose $w(x) = V_0(x)$. Now

$$\begin{aligned}
\frac{\partial V}{\partial t} &= \frac{\partial}{\partial t}\int p(t, s, T, z)q(t, T)w(z)dz \\
&= \int [-r(S_t, t)S_t\frac{\partial p}{\partial s} - \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 p}{\partial s^2}]q(t, T)w(z)dz \\
&\quad + r(S_t, t)\int p(t, S_t, T, z)q(t, T)w(z)dz \text{ by (2.19)} \\
&= -r(S_t, t)S_t\frac{\partial V}{\partial S} - \frac{\sigma^2(S_t, t)}{2}\frac{\partial^2 V}{\partial S^2} + r(S_t, t)V(S_t, t)
\end{aligned}$$

where we have assumed that we can pass the derivatives under the integral sign. Thus the process

$$V(t,s) = E[V_0(X_T)q(t,T)|X_t = s] \tag{2.23}$$

satisfies both the partial differential equation (2.15) and the terminal conditions **??** and is hence the solution. Indeed it is the unique solution satisfying certain regularity conditions. The result asserts that the value of any European option is simply the conditional expected value of the *discounted payoff* (discounted to the present) assuming that the distribution is that of the process (2.18). This result is a special case when the spot interest rates are functions only of time of the following more general theorem.

Theorem(Feynman-Kac)

Suppose the conditions for a unique solution to (2.15,2.17) (see for example Duffie, appendix E) are satisfied. Then the general solution to (2.15) under the terminal condition **??** is given by

$$V(S,t) = E[V_0(X_T)exp\{-\int_t^T r(X_v,v)dv\} \mid X_t = S] \tag{2.24}$$

This represents the discounted return from the option under the distribution of the process X_t . The distribution induced by the process X_t is referred to as the *equivalent martingale measure* or *risk neutral measure*. Notice that when the original process is a diffusion, the equivalent martingale measure shares the same diffusion coefficient but has the drift replaced by $r(X_t,t)X_t$. The option is priced as if the drift were the same as that of a risk-free bond i.e. as if the instantaneous rate of return from the security if identical to that of bond. Of course, in practice, it is not. A risk premium must be paid to the stock-holder to compensate for the greater risk associated with the stock.

There are some cases in which the conditional expectation (**??**) can be determined explicitly. In general, these require that the process or a simple function of the process is Gaussian.

For example, suppose that both $r(t)$ and $\sigma(t)$ are deterministic functions of time only. Then we can solve the stochastic differential equation (2.22) to obtain

$$X_T = \frac{X_t}{q(t,T)} + \int_t^T \frac{\sigma(u)}{q(u,T)} dW_u \tag{2.25}$$

The first term above is the conditional expected value of X_T given X_t. The second is the random component, and since it is a weighted sum of the normally

distributed increments of a Brownian motion with weights that are non-random, it is also a normal random variable. The mean is 0 and the (conditional) variance is $\int_t^T \frac{\sigma^2(u)}{q^2(u,T)}du$. Thus the conditional distribution of X_T given X_t is normal with conditional expectation $\frac{X_t}{q(t,T)}$ and conditional variance $\int_t^T \frac{\sigma^2(u)}{q^2(u,T)}du$.

Problem.

Consider approximating an integral of the form $\int_0^T g(t)dW_t \approx \sum g(t)\{W(t+h) - W(t)\}$ where $g(t)$ is a non-random function and the sum is over values of $t = nh, n = 0, 1, 2, ...T/h - 1$. Show by considering the distribution of the sum and taking limits that the random variable $\int_0^T g(t)dW_t$ has a normal distribution and find its mean and variance.

Problem.

Give an example of a function $g(t, W_t)$ such that the random variable $\int_0^1 g(t, W_t)dW_t$ does not have a normal distribution but has larger tails than the normal distribution has.

The special case of (**??**) of most common usage is the Black-Scholes model: suppose that $\sigma(S,t) = S\sigma(t)$ for $\sigma(t)$ some deterministic function of t. Then the distribution of X_t is not Gaussian, but fortunately, its logarithm is. In this case we say that the distribution of X_t is lognormal.

Lognormal Distribution

Suppose Z is a normal random variable with mean μ and variance σ^2. Then we say that the distribution of $X = e^Z$ is lognormal with mean $\eta = \exp\{\mu + \sigma^2/2\}$ and volatility parameter σ. The lognormal probability density function with mean $\eta > 0$ and volatility parameter $\sigma > 0$ is given by the probability density function

$$g(x|\eta, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} exp\{-(log\ x - log\ \eta\ - \sigma^2/2)^2/2\sigma^2\}. \tag{2.26}$$

The solution to (2.18) with non-random functions $\sigma(t), r(t)$ is now

$$X_T = X_t exp\{\int_t^T (r(u) - \sigma^2(u)/2)du + \int_t^T \sigma(u)dW_u\}. \tag{2.27}$$

Since the exponent is normal, the distribution of X_T is lognormal with mean $log(X_t) + \int_t^T (r(u) - \sigma^2(u)/2)du$ and variance $\int_t^T \sigma^2(u)du$. It follows that the conditional distribution is lognormal with mean $\eta = X_t q(t,T)$ and volatility parameter $\sqrt{\int_t^T \sigma^2(u)du}$.

We now derive the well-known Black-Scholes formula as a special case of (**??**). For a call option with exercise price E, the payoff function is $V_0(S_T) =$

$max(S_T - E, 0)$. Now it is helpful to use the fact that for a standard normal random variable Z and arbitrary $\sigma > 0, -\infty < \mu < \infty$ we have the expected value of $max(e^{\sigma Z+\mu}, 0)$ is

$$e^{\mu+\sigma^2/2}\Phi(\frac{\mu}{\sigma} + \sigma) - \Phi(\frac{\mu}{\sigma}) \tag{2.28}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. As a result, in the special case that r and σ are constants, (**??**) results in the famous Black-Scholes formula which can be written in the form

$$V(S,t) = S\Phi(d_1) - Ee^{-r(T-t)}\Phi(d_2) \tag{2.29}$$

where $d_1 < d_2$ are the values $\pm\sigma^2/2$ standardized by adding $log(S/E)+r(T-t)$ and dividing by $\sigma\sqrt{T-t}$. This may be derived by the following device; Assume (i.e. pretend) that, given current information, the distribution of $S(T)$ at expiry is lognormally distributed with the mean $\eta = S(t)e^{r(T-t)}$.

The mean of the log-normal in the risk neutral world $S(t)e^{r(T-t)}$ is exactly the future value of our current stocks $S(t)$ if we were to sell the stock and invest the cash in a bank deposit. Then the future value of an option with payoff function given by $V_0(S_T)$ is the expected value of this function against this lognormal probability density function, then discounted to present value

$$e^{-r(T-t)} \int_0^\infty V_0(x)g(x|S(t)e^{r(T-t)},\ \sigma\sqrt{T-t})dx. \tag{2.30}$$

Notice that the Black-Scholes derivation covers any diffusion process governing the underlying asset which is driven by a stochastic differential equation of the form

$$dS = a(S)dt + \sigma S dW_t \tag{2.31}$$

regardless of the nature of the drift term $a(S)$. For example a non-linear function $a(S)$ can lead to distributions that are not lognormal and yet the option price is determined as if it were.

Example.

Consider pricing an index option on the S&P 500 index an January 11, 2000 (the index SPX closed at 1432.25 on this day). The option SXZ AE-A is a January call option with strike price 1425. The option matures (as do equity options in general) on the third friday of the month or January 21, a total of 7 trading days later. Suppose we wish to price such an option using the lack-Scholes model. In this case, $T - t$ measured in years is $7/252 = 0.027778$. The annual volatility of the Standard and Poor 500 index is around 19.5 percent or 0.195 and the very short term interest rates approximately 3%. In *Matlab* we can value this option using

[CALL,PUT] = BLSPRICE(1432.25,1425,0.03,7/252,0.195,0)

CALL = 23.0381
PUT = 14.6011

Arguments of the function BLSPRICE are, in order, the current equity price, the strike price, the annual interest rate r, the time to maturity $T-t$ in years, the annual volatility σ and the last argument is the dividend yield in percent which we assumed 0. Thus the Black-Scholes price for a call option on SPX is around 23.03. Indeed this call option did sell on Jan 11 for \$23.00. and the put option for \$14 5/8. From the put call parity relation (see for example Wilmott, Howison, Dewynne, page 41) $S + P - C = Ee^{-r(T-t)}$ or in this case $1432.25 + 14.625 - 23 = 1425e^{-r(7/252)}$. We might solve this relation to obtain the spot interest rate r. In order to confirm that a different interest rate might apply over a longer term, we consider the September call and put options (SXZ) on the same day with exercise price 1400 which sold for \$152 and 71\$ respectively. In this case there are171 trading days to expiry and so we need to solve $1432.25 + 71 - 152 = 1400e^{-r(171/252)}$, whose solution is $r = 0.0522$. This is close to the six month interest rates at the time, but 3% is low for the very short term rates. The discrepancy with the actual interest rates is one of several modest failures of the Black-Scholes model to be discussed further later.

Problem

Verify that for any pair of constants $a \neq 0$ and $b > 0$

$$dX_t = (X_t^{-1} + ab)X_t dt + bX_t dW_t$$

does not have a solution in the form $X_t = f(t, Y_t)$, where $f(t, y)$ is, say, a real function and Y_t is a Gaussian process.

10.0.9 Black-Scholes with Transaction Costs.

We now modify the argument in section 2.8 to accommodate transaction costs. As in Leland (1985), Hoggard et al. (1993), we assume delta hedging and the transaction costs in each time interval is a constant proportion $k/2$ of the value of the trades in that interval. Suppose we have, at time t, u_t units of the security, and bank deposits or bonds, earning constant interest rate r, to a total value of B_t. Then the value of the portfolio $V_t = u_t S_t + B_t$. Therefore the change in value over a small time interval is of the form

$$dV_t = u_t dS_t + rB_t dt - (k/2)S_t |du_t|$$

plus terms of smaller order. The last term represents the transaction costs over this time interval. Using Ito's lemma on $V(S, t)$ we obtain

$$dV = \frac{\partial V}{\partial S} dS + \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} \right] dt$$

Equating terms, $u_t = \frac{\partial V(S_t,t)}{\partial S}$ resulting in delta-hedging and

$$rB_t dt - (k/2)S_t|du_t| = \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}\right]dt \tag{2.32}$$

Applying Ito's lemma to evaluating du_t will show that

$$du_t = \frac{\partial^2 V}{\partial S^2}dS + smaller \;\; order \; terms$$

Therefore

$$(k/2)S_t|du_t| \approx (k/2)S_t|\frac{\partial^2 V}{\partial S^2}||dS|. \tag{2.33}$$

Now we regard this differential equation as an approximation to a discrete process for which dS represents the change in the process over a discrete time interval of length dt. In this case, since approximately, $dS \sim N(a(S_t)dt, \sigma^2 S_t^2 dt)$ it follows that $E|dS| = \sqrt{2/\pi}\sigma S_t\sqrt{dt}$ and since, by the law of large numbers, the sum of many such increments converges to the expected value, the term $|dS|$ may be replaced in (2.33) by $\sqrt{2/\pi}\sigma S_t\sqrt{dt}$. Therefore, substituting in (2.32), we obtain the equation

$$rB_t dt - (k/2)S_t|\frac{\partial^2 V}{\partial S^2}|\sqrt{2/\pi}\sigma S_t\sqrt{dt} = \left[\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}\right]dt.$$

Collecting terms, and substituting $B_t = V_t - \frac{\partial V}{\partial S}S_t$, this reduces to

$$\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}(1 + A\; sgn(\Gamma))S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

where $\Gamma = \frac{\partial^2 V}{\partial S^2}$ and

$$A = \sqrt{\frac{2}{\pi}}\frac{k}{\sigma\sqrt{dt}}$$

is the so-called Leland number. This is exactly the same equation that was solved to give the Black-Scholes formula except that the volatility σ^2 is inflated (deflated) by the factor $(1 + A\; sgn(\Gamma)))$. In the case $0 < A < 1$, this equation has a solution for arbitrary payoff function V_0. In the case $A > 1$, if the payoff function is convex, then again the solution is given by Black-Scholes with inflated volatility. However, if the payoff function is non-convex, then the mathematical problem is ill-posed (see Avellanda and Paras (1994)).

10.0.10 Methods for American Options

The valuation of American options is what is known as a free boundary problem. Typically at each time t there is a value $S_f(t)$ which marks the boundary between two regions: to one side one should hold the option on the other side one should

exercise it. Since we don't know a priori $S_f(t)$, however, we cannot apply boundary conditions in the way we do for European options.

For an American put option, with value $P(S,t)$, the valuation problem can be written as a free boundary problem as follows. For each time t, we must divide the S axis into two distinct regions: the first, $0 \le D < S_f(t)$, is where early exercise is optimal and

$$P = E - S, \qquad \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP < 0,$$

the second, $S_f(t) < S < \infty$, where early exercise is not optimal and

$$P > E - S, \qquad \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP = 0.$$

The boundary conditions at $S = S_f(t)$ are that P and its slope are continuous:

$$P(S_f(t),t) = \max(E - S_f(t), 0), \quad \frac{\partial P}{\partial S}(S_f(t),t) = -1.$$

Chapter 11

Appendix C: Glossary.

- Annuity A uniform series of payments or receipts over a specified period of time
- Asset A Physical or intangible item of value to a company or individual
- Bankruptcy A legal proceeding of the disposal of assets of a business or individual to satisfy creditors' claims in full or in part and protecting the debtor from further action.
- Beta A regression coefficient indicating the rate at which a given stock changes for a unit changge in the market as a whole.
- Bond A financial instrument representing a form of corporate long-term debt issued to investors.
- Bond rating A published ranking of a bond developed by fianancial organisations to express relative soundness on a defined scale.
- Book value the recorded value of an asset or liability as reflected in financial statements
- Call provision a provission permitting the issuing company to redeem part or all of a bond or preferred stock at a date determined by the company.
- Capital the funds committed to an ienterprise in the form of onership equity and long-term financing.
- Capitalization The sum of all long-term sources of capital of a company. The difference between current liabilities and total assets.
- Cash flow The postive inflow or negative outflow of cash caused by an activity over a specifies period of time.

- Common stock (common shares) Securities representing a derect ownership interest in a corporation and a residual claim on its assets.
- Compounding The process of calculating the growing value of a sum of money over time caused by the reinvestment of earned interest.
- Coupon rate The stated interest rate specifies on the interest coupons attached to bonds and calculated as a percentage of their face value.
- Credit the recognised ability of an individual or corporation to assume indebtetdness with the prospect of servicing such debt.
- Debt (libability) An obligation to pay amounts due under specified terms.
- Default failure to make a payment on a debt obligation when due.
- Discounting the process of calculating the reduced value of a future sum of money in proportion to the opportunity of earning interest over that period.
- Diversification the process of investing in a number of unrelated or partially independent assets or activities to establish a more satisfactory portfolio and reduce the volatility.
- Dividend payout the ratio of the amount of dividends distributed to the aftertax earnings o a corporation.
- Dividend yield The ratio of the total amount of dividends payable on one share over a specified period to the current market price per share.
- Earnings the difference between revenue and costs and expenses for a specifies period.
- Earnings per share; the total aftertax earnings divided by the number of common shares outstanding.
- Equity the recorded ownership claim of all common and preferred shareholders as reflected on the balance sheet.
- Fair market value the price for an asset on which two rational parties with sufficient information would agree in the absence of other factors.
- Financial Model the representation in a mathematical model or computer program relating the effect of various input factors to some measures of performance.
- Fixed-income security any security which provides a constant stream of interest or dividend over its life.
- Foreign exchange exposure The potential loss from changes in the exchange rate with one or more foreign currencies.

- Hedge a strategy to neutralize the risk of an investment by engaging in offsetting contracts whereby potential gains and losses will tend to cancel.
- Junk bond any bond with risk characteristics higher than normal investment grade.
- Leverage Any transaction which magnifies a given effect.
- Liability an obligation to pay a specified amount or perform a specified service at specifed times.
- Liquidity The degree to which a company is readily able to meet its current obligations, or the ease with which a security can be bought or sold.
- Market value The value of and asset as determined in an uncontrained market of many buyers and sellers.
- Net present value The difference between the present values of cash inflows and cash outflows
- Nominal amount Any quantity not adjusted for changes in maket conditions, purchasing powers.
- Option A contractual opportunity to purchase or sell an asset or security at a predetermined price, without obligation to do so.
- Par value the nominal value established by the issuer of a security. For a bond, the issuing company will pay the par value on maturity.
- Perpetuity a series of level periodic payments or receiptss expected to last forever.
- Portfolio a set of diverse investments held by and individual or company.
- Preferred Stock a special class of capital stock that receives a form of preference over common stock in its claim on earnings and assets.
- Present Value the value today of a future sum or a series of sums calculated by discounting the future amounts.
- Principal the original amount of a loan or bond (also called face value) on which interest is based.
- Risk analysis a process of integrating risk into an analysis.
- Risk aversion a subjective unwillingness to accept a given level of risk, unless there is a trade-off for higher average return.
- Risk-free interest rate The assumed yield obtainable on a guaranteed security.

- Risk premium The increased return required for an investment to compensate the holder for the level of risk involved.

-