# A TOUR OF PROBABILITY AND STATISTICS FOR JDEP 384H

**Thomas Shores**
**Department of Mathematics**
**University of Nebraska**
**Spring 2007**

Contents



**Note:** This really is a *brief* tour. You will find additional details in the excellent probability and statistics review in Appendix B of our textbook. Everyone should read through this appendix. I'm also including some other material that we'll need on stochastic processes.

## 1. Probability

We'll begin with a few simple examples, one with a discrete set of outcomes and the other a continuous set.

**Example 1.1.** Consider the experiment of randomly selecting an individual out of the entire population of a certain species of animal for the purpose of some measurement. The selection of a particular individual could be thought of as an *outcome* to this random experiment. Selection of a male would amount to an *event* $E$ , and the probability of selecting a male would be a number $P(E)$ between 0 and 1.

**Example 1.2.** Consider the experiment of throwing a dart at a dart board. We assume that the throw always hits the dart board somewhere. Here the outcome of this experiment is to locate the dart on some point on the dart board, so we can think of these points as outcomes. One event of interest is the event $E$ of hitting the bulls eye region with the dart. Again, the probability of doing so would be a number $P(E)$ between 0 and 1.

Here are some of the key concepts of probability theory. You should relate these to the two experiments just described.

- **Sample Space:** A set $S$ of possible outcomes from a random experiment or sequence thereof.
- **Event:** Any subset of the sample space $S$, i.e., of outcomes. (In some cases, there might be limitations on what subsets are admissible.)
- **Probability measure:** A way of measuring the likelihood that an outcome belongs to event $E$. This is a function $P(E)$ of events with natural properties: $0 \leq P(E) \leq 1$, $P(S) = 1$ and for disjoint events $E_i$,

$$\sum_{i=1}^{\infty} P(E_i) = P\left(\bigcup_{i=1}^{\infty} E_i\right).$$

  *Simple consequence:* If $E$ is an event, then the probability of the complementary event $\overline{E}$ occurring is

$$P\left(\overline{E}\right) = 1 - P(E)$$

- **Conditional probability:** This is the probability that an event $E$ occurs, given that event $F$ has occurred. It is denoted and defined by the formula

$$P(E \mid F) = \frac{P(EF)}{P(F)}.$$

  Note the notation $EF$, which means the event of the occurrence of *both* $E$ and $F$. Another way of expressing this event is the set-theoretic notation $E \cap F$.
- **Independent events:** Events $E$ and $F$ such that

$$P(EF) = P(E)\,P(F)$$

  in which case the conditional probability of $E$ given $F$ is

$$P(E \mid F) \equiv \frac{P(EF)}{P(F)} = P(E).$$

- **Law of Total Probability:** Given disjoint and exhaustive events $E_1, E_2, \ldots, E_n$, and another event $F$,

$$P(F) = \sum_{i=1}^{n} P(F \mid E_i)\,P(E_j)$$

- **Bayes' Theorem:**

$$P(E \mid F) \equiv \frac{P(F \mid E)\,P(E)}{P(F)}.$$

Some writers identify Bayes' Theorem as a combination of the Law of Total Probability and the above, namely, with notation as in the LTP and index $k$,

$$P(E_k \mid F) \equiv \frac{P(F \mid E_k) P(E_k)}{\sum_{i=1}^{n} P(F \mid E_i) P(E_j)}.$$

## 2. Univariate Statistics

2.1. Random Variables. Once we have randomly selected an individual outcome $\omega$ in an experiment, we can observe some relevant quantity and call it $X(\omega)$. This function $X$ is called a **random variable**, *and a particular value observed in an experiment is customarily denoted as* $x = X(\omega)$.

Let's review the standard notations of this statistical framework.

- **Random variable:** a function $X$ (abbreviate to r.v.) mapping outcomes to real numbers. A particular value is denoted by lower case $x$.
- **Probability density function:** a function $p$ mapping the range of a random variable to probabilities (abbreviate to p.d.f.):
  In the case the r.v. is *discrete*, say has values $x_1, x_2, ...$ then
  $$P(a \leq X \leq b) = \sum \{p(x_i) \mid a \leq x_i \leq b\}.$$
  In the discrete case, $p(x)$ is also referred to as a *probability mass function* (p.m.f.).
  If the r.v. is continuous, then the density function $f$ satisfies
  $$P(a \leq X \leq b) = \int_a^b f(x)dx.$$
  In this case $f$ really is a density function with units of of probability per length. Note: since an experiment always results in *some* value of the r.v. $X$, we must have $\int_{-\infty}^{\infty} f(x)\, dx = 1$ and a similar result for discrete r.v.'s
- The **(cumulative) distribution function** (abbreviate to c.d.f.) associated to the r.v. is
  $$p(x) = P(X \leq x) = \sum \{p(x_i) \mid x_i \leq x\}$$
  for discrete r.v.'s and
  $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds$$
  for continuous r.v.'s. Note: from properties of the p.d.f., we see that
  - $F(x)$ is a monotone increasing function, i.e., if $x \leq y$, then $F(x) \leq F(y)$.
  - $\lim_{x \to \infty} F(x) = 1$.

**Example 2.1.** Consider the experiment of Example_1.2. Once we have thrown the dart and landed on $\omega$, we might observe the score $X(\omega)$ we earned according to the portion of the dart board on which our dart landed. Here $X$ will take on a finite number of values. Let us further suppose that there are only two areas on the board: the center bulls-eye of area $A$ (winner, value 1) and an outer area $B$ (loser, value 0.) Suppose that the probability of hitting one area is proportional to its area. Then the probability of hitting the bulls-eye is

$$p = \frac{A}{A+B}$$

and the probability of losing is $q = 1 - p$. The p.d.f. is given by $f(0) = q$ and $f(1) = p$. The c.d.f. is given by $F(0) = q$ and $f(1) = 1$.

An interesting variation on the previous example is to repeat the experiment, say $n$ times. Now the random variable $X$ is your score: the number of times you hit the bulls-eye. The p.d.f. for this experiment (called a Bernoulli trial) is the so-called **binomial distribution**

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

2.1.2. *Continuous.*

**Example 2.2.** Consider the experiment of Example 1.1. Once we have selected an animal $\omega$, we might take its weight and call it the statistic $X(\omega)$. Note that $X$ could take on a continuous range of values. The p.d.f. and c.d.f. of a continuous random variable are more subtle and one often makes a priori assumptions about them.

Let's simplify our dart example, so we can obtain distributions more easily.

**Example 2.3.** Suppose that our target is not a two dimensional board, but a one dimensional line segment, say the interval of points $x$ such that $a \leq x \leq b$ or symbolically, $[a, b]$ Suppose further that there is no bias toward any one point. Then it is reasonable to assume that the p.d.f. is constant. Since it is defined on $[a, b]$ and the area under this function should be 1, we see that the p.d.f. is the function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

while the c.d.f. should be

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

This is the so-called **uniform distribution**.

Before we discuss further specific distributions, there are some more concepts we should develop.

2.2. **Expectation and Variance.** Key concepts:

- **Expectation** of a function $g$ of a r.v.:

$$E[g(X)] = \begin{cases} \sum_i g(x_i)p(x_i), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

- **Expectation** of $X$ (or mean, first moment): $\mu = \mu_X = E[X]$. One can show

$$\begin{aligned} E[\alpha X + \beta] &= \alpha E[X] + \beta \\ E[\alpha X + \beta Y] &= \alpha E[X] + \beta E[Y] \end{aligned}$$

- **Variance** of $X$: This is just

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right].$$

One can show

$$\begin{aligned} \mathrm{Var}(X) &= E\left[X^2\right] - E[X]^2 \\ \mathrm{Var}(\alpha X + \beta) &= \alpha^2 \mathrm{Var}(X) \end{aligned}$$

- **Standard deviation** of $X$: $\sigma = \sigma_X = \mathrm{Var}(X)^{1/2}$

Basically, the idea is this: the expected value is a kind of weighted average of values, so that one could say roughly that "on the average one expects the value of repeated experiments to be the mean." The variance and standard deviation are measures of the spread of the random variable. Note that the units of $\sigma$ are the same as the units of $\mu$, so that the standard deviation is a more practical measure of the spread of the random variable, but the variance $\sigma^2$ is more useful for some calculations and theoretical purposes.

**Standard Notation:** To save ourselves the inconvenience of always having to assign a name to the p.d.f. and c.d.f. of a given r.v. $X$, we adopt the convention that

$$\begin{aligned} f_X(x) &= \text{p.d.f. of the r.v. } X \\ F_X(x) &= \text{c.d.f. of the r.v. } X. \end{aligned}$$

2.3. Normality and the Central Limit Theorem**.** One of the most important single distributions in statistics is the This is a r.v. whose density function is the famous bell shaped curve

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

It can be shown that this really is a density function with mean $\mu$ and variance $\sigma^2$. Its corresponding distribution is

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(s-\mu)^2/2\sigma^2} ds, \ -\infty < x < \infty.$$

The **standard normal distribution** is the one with $\mu = 0$ and $\sigma = 1$, that is,

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

is the p.d.f. of the distribution. The c.d.f. for the standard normal distribution has the following designation, which we use throughout our discussion of statistics:

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds, \ -\infty < x < \infty.$$

The notation $N(\mu, \sigma^2)$ is used for a normal distribution of mean $\mu$ and variance $\sigma^2$. One sees phrases like "$X$ is $N(\mu, \sigma^2)$" or "$X$ $N(\mu, \sigma^2)$." We can pass back and forth between standard normal distributions because of this important fact: if $X$ has a distribution $N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ has the distribution $N(0, 1)$, the standard normal distribution.

Here is a key property of this important kind of distribution:

**Theorem 2.4.** *If $X$ and $Y$ are independent normal random variables with parameters $(\mu_1, \sigma_1^2)$, $(\mu_2, \sigma_2^2)$, then $X + Y$ is normal with parameters $(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

It follows that this Theorem is true for any finite number of independent r.v.'s.

In a limiting sense, sums of r.v.'s with finite mean and variance tend to a normal distribution. This is the Central Limit Theorem.

**Theorem 2.5.** ***(Central Limit Theorem)*** *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with a finite expected value $\mu$ and variance $\sigma^2$. Then the random variable*

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - \mu}{\sigma/\sqrt{n}}$$

*has distribution that approaches the standard normal distribution as $n \to \infty$.*

2.3.1. *Some Common Distributions.* Here are a few common distributions.

**Binomial:**

- $f(x) = \dfrac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$
- Mean: $\mu = np$
- Variance: $\sigma^2 = np(1-p)$
- Application: Bernoulli trials as in variation on Example 1.2.

**Poisson:**

- $f(x) = \dfrac{\mu^x e^{-\mu}}{x!}$, $x = 0, 1, \ldots$
- Mean: $\mu = \mu$
- Variance: $\sigma^2 = \mu$
- Application: A limiting case of binomial distribution. Used, e.g., to approximate binomial distributions with large $n$ and constant $\mu = np$ of moderate size (typically $< 5$.) There is a whole family of "Poisson processes" that are used in problems like manufacturing errors, etc.

**Gamma:**

- $f(x) = \dfrac{1}{\Gamma(\alpha)\,\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$, $0 < x < \infty$, $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s}\, ds$. Here $2\alpha = \nu$ is called the number of *degrees of freedom.*
- Mean: $\mu = \alpha\beta$
- Variance: $\sigma^2 = \alpha\beta^2$
- Application: An umbrella for other extremely important p.d.f.'s. For example, $\alpha = 1$, $\beta = 1/\lambda$ gives the family of exponential distributions and $\alpha = \nu/2$, $\beta = 2$ gives a chi-square distribution with $\nu$ degrees of freedom, which is denoted as $\chi^2(\nu)$. Also used in queueing theory.

**Normal:**

- $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$, $-\infty < x < \infty$.
- Mean: $\mu = \mu$
- Variance: $\sigma^2 = \sigma^2$
- Application: Many, e.g., random error. Also, a distinguished distribution by way of the Central Limit Theorem.

**Student's t:**

- $f(x) = \dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \dfrac{1}{\sqrt{\nu\pi}} \left(1 + \dfrac{x^2}{\nu}\right)^{-(\nu+1)/2}$, $-\infty < x < \infty$. Here $\nu$ is the number of degrees of freedom.
- Mean: $\mu = 0$
- Variance: $\sigma^2 = \frac{\nu}{\nu-2}$
- Application: Approaches a standard normal distribution as $\nu \to \infty$. Also, given $n$ independent samples of normally distributed r.v.'s with a common unknown standard deviation $\sigma$, let the sample mean be given by $\overline{x} = (x_1 + x_2 + \cdots + x_n)/n$ and the sample variance by $s^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$, then the random variable

$$t = \frac{X - \overline{X}}{S/\sqrt{n}}$$

  has a Student's t distribution with $n-1$ degrees of freedom.

**Lognormal:**

- $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x - \nu)^2/\left(2\sigma^2\right)}$, $-\infty < x < \infty$.
- Mean: $\mu = e^{\nu + \sigma^2/2}$
- Variance: $\sigma^2 = e^{2\nu+\sigma}\left(e^{\sigma^2} - 1\right)$
- Application: This is really the density function for $e^X$, where $X$ is normally distributed. Equivalently, $X = \ln Y$. It is fundamentally important in modeling the dynamics of asset prices. Note: we could write $e^{-\left(\log x - \nu^2\right)/\left(2\sigma^2\right)}$ as well, since we use log for the natural log, like Matlab.

## 3. Joint Random Variables

For the most part, we'll restrict our discussion of joint random variables to continuous distributions, though all the ideas have a discrete counter part. Likewise, the ideas we are going to discuss extend to any number of r.v.'s, like $X_1, X_2, \ldots, X_n$, but we will mostly confine our attention to two r.v.'s, say $X$ and $Y$

3.1. **Joint Distributions.** In order to motivate the idea of joint distributions, let's consider Example 2.3 with a twist: we will throw a dart at our one dimensional dart board twice. With each throw, we will note the position of the outcome on the interval $[0, 1]$ and this number is our random variable. This gives us two random variables $X$ and $Y$ which share the same sample space of outcomes when viewed individually. Moreover, it makes the statistics of the new experiment more complicated than just numbers on the interval $0 \le x \le 1$. Now they are ordered pairs of numbers $(x, y)$ such that $0 \le x, y \le 1$; In other words, they belong to a unit square in the $xy$-plane. The event $X + Y \le 1$ can now be pictured as a subset of this square.

Now suppose we ask the question: what is the probability that $X + Y \le 1$? In order to answer this question, we need to understand how these variables behave jointly, so we will need a p.d.f. $f(x, y)$ that is a **joint distribution** of both random variables. Here "density" means probability per unit area, not length. Once we have such a function, we can describe the probability of an event $A$ occurring as a double sum in the case of discrete r.v.'s and as

a double integral in the case of a continuous r.v. Thus,

$$P(A) = \iint_A f(x,y)\, dA.$$

In most cases we can reduce these double integrals over plane regions to iterated integrals as in ordinary calculus. As an example of this, we can define a **joint cumulative distribution function** (c.d.f.) by the formula

$$F(x,y) = P(X \leq x, Y \leq y)$$

and obtain that

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,y)\, dy\, dx.$$

Now what about the p.d.f. of the example we have mentioned. This can get complicated. If both throws are random, and the p.d.f. for each r.v. separately is the uniform distribution, it is reasonable to expect that the joint p.d.f. should also be uniformly distributed, so we have $f(x,y) = 1$. But what if the throws are not independent? For example, if we play a game where the "score" of the throws, $x + y$, is close to a certain number, then where the first dart landed will affect where we throw the second one. So in this case we would expect $f(x,y)$ to express a more complicated relationship between $x$ and $y$.

**Standard Notation:** To save ourselves the inconvenience of always having to assign a name to the joint p.d.f. and c.d.f. of given r.v.'s $X$ and $Y$, we adopt the convention that

$$\begin{aligned} f_{X,Y}(x,y) &= \text{joint p.d.f. of the r.v.'s } X, Y. \\ F_{X,Y}(x,y) &= \text{joint c.d.f. of the r.v.'s } X, Y. \end{aligned}$$

3.2. **Expectation and Covariance.** Just as with p.d.f.'s of one variable, one can define some key concepts for r.v.'s $X$ and $Y$:

- **Expectation** of a function $g(x,y)$ of r.v.'s:

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dy\, dx.$$

- **Covariance** of $X$ and $Y$: This is just

$$\operatorname{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])].$$

One can show from the definitions that

$$\begin{aligned} \operatorname{Var}(X,X) &= \operatorname{Var}(X) \\ \operatorname{Var}(aX+bY) &= a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X,Y) \\ \operatorname{Cov}(X,Y) &= E[XY] - E[X]\,E[Y] \\ \operatorname{Cov}(X,Y) &= \operatorname{Cov}(Y,X) \\ \operatorname{Cov}(aX,bY) &= ab \operatorname{Cov}(X,Y) \\ \operatorname{Cov}(X,Y+Z) &= \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z) \end{aligned}$$

- **Correlation** of $X$ and $Y$:

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

One can show that $-1 \leq \rho(X,Y) \leq 1$ and that $\rho(X,X) = 1$. If $\rho(X,Y) = 0$ or, equivalently, $\operatorname{Cov}(X,Y) = 0$, we say that $X$ and $Y$ are **uncorrelated**.

- **Independent** r.v.'s $X$ and $Y$: means that for all $a$ and $b$,

$$P(X \leq a, Y \leq b) = P(X \leq a) P(Y \leq b).$$

One can show that if $X$ and $Y$ are independent and $g(x), h(y)$ are any functions then

$$\begin{aligned} F_{X,Y}(x,y) &= F_X(x) F_Y(y) \\ f_{X,Y}(x,y) &= f_X(x) f_Y(y) \\ E[g(X)h(Y)] &= E[g(X)] E[h(Y)]. \end{aligned}$$

Thus we see that if two random variables $X$ and $Y$ are independent, then

$$\operatorname{Cov}(X,Y) = E[X - E[X]] E[Y - E[Y]] = 0.$$

## 4. Vector Random Variables

A *random vector* is a vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ of random variables $X_1, X_2, \ldots, X_p$. Such variables are very common in practical statistics. For example, we might be interested in a portfolio of $p$ securities. The rate of return of each can be viewed as a random variable $X_i$, $i = 1, 2, \ldots, p$. At various times, we might sample these rates, so the resulting statistic is a vector of samples $(x_1, x_2, \ldots, x_p)$. These $p$ random variables may have correlations. We would be interested in studying the properties of the vector of random variables $\mathbf{X}$ taken as a whole, not simply the individual components. We want to define concepts analogous to the expectation and variance that we learned in the univariate case.

### 4.1. Expected Value and Covariance of a Random Vector.

Here is how they are defined.

**Definition.** The *expected value* (a.k.a. *mean*) of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is

$$E[\mathbf{X}] = (E[X_1], E[X_2], \ldots, E[X_p]).$$

**Definition.** The *covariance matrix* (a.k.a. *variance-covariance matrix*, *variance*) of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is the $p \times p$ matrix $\Sigma$ whose $(i,j)$th entry is $\sigma_{ij} = \operatorname{Cov}(X_i, X_j)$. This matrix is also written as $\operatorname{Var}(\mathbf{X})$ or $\operatorname{Cov}(\mathbf{X})$.

**Definition.** The correlation matrix of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is the $p \times p$ matrix $R$ whose $(i,j)$th entry is $\rho_{ij} = \operatorname{Corr}(X_i, X_j)$.

**Theorem 4.1.** *Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a random vector. Then*

1. $\operatorname{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T\right]$.
2. *If $D$ is the diagonal matrix whose diagonal entries are the standard deviations of $X_1, X_2, \ldots, X_p$ (equivalently, the square roots of the diagonal entries of $\operatorname{Cov}(\mathbf{X})$), then*

$$\operatorname{Corr}(\mathbf{X}) = D^{-1} \operatorname{Cov}(\mathbf{X}) D^{-1} \quad \text{and} \quad \operatorname{Cov}(\mathbf{X}) = D \operatorname{Corr}(\mathbf{X}) D.$$

.

Here are a few key facts from multivariate statistics which can be shown by applying matrix arithmetic to the basic definitions:

**Theorem 4.2.** *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ *be a random vector and* $C$ *any* $m \times p$ *matrix. Then*

(1) $E[C\mathbf{X}] = CE[\mathbf{X}]$.
(2) $\mathrm{Cov}(C\mathbf{X}) = C\,\mathrm{Cov}(\mathbf{X})\,C^T$.

A fundamental fact from matrix theory that is used in multivariate statistics is the principal axes theorem, which can be found in the LinearAlgebra-384h notes. From this fact the following key result can be deduced.

**Theorem 4.3.** ***(Principal Components Theorem)*** *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ *be a random vector.*

(1) *Then there exists an orthogonal matrix* $P$ *(this means* $p \times p$ *matrix* $P$ *is invertible and* $P^{-1} = P^T$*), such that* $P^T\,\mathrm{Cov}(\mathbf{X})\,P = D$*, a diagonal matrix.*
(2) *The diagonal entries of* $D$ *are the eigenvalues of* $\mathrm{Cov}(\mathbf{X})$ *and are non-negative.*
(3) *The columns of* $P$ *are orthogonal eigenvectors of unit length with the* $k$*th column corresponding to the* $k$*th diagonal entry of* $D$*.*
(4) *If* $\mathbf{P}_k$ *is the* $k$*th column of* $P$*, then the random variable* $Y_k = \mathbf{P}_k^T\mathbf{X}$ *is the* $k$*th principal component of* $\mathbf{X}$*.*
(5) *The principal components of* $\mathbf{X}$ *are random variables with variances the corresponding eigenvalues and covariances zero.*

4.2. **Multivariate Normal Distribution.** Here is a generic example of an extremely important multivariate distribution. In the case of two r.v.'s this type of distribution is called a *bivariate* distribution. These distributions are the "correct" analog in higher dimensions to the normal distributions in one dimension. In the following example we need the concept of a "**symmetric positive definite** matrix (SPD)". First, a square $n \times n$ matrix $A$ is **symmetric** if $A^T = A$. Secondly, $A$ is **positive definite** if $\mathbf{x}^T A\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x}$. Some useful facts:

- If $A$ is symmetric, then all the eigenvalues of $A$ are real.
- A symmetric matrix is positive definite if and only of all its eigenvalues are positive.
- If $A$ is symmetric, then there exists an orthogonal matrix $Q$ (i.e., $Q^T = Q^{-1}$) such that $Q^TAQ$ is diagonal and moreover the diagonal elements are exactly the eigenvalues of $A$.

**Example 4.4.** Suppose that we are given a vector $\mu$ and an $p \times p$ matrix $C = [c_{i,j}]$ that is SPD. Define the function

$$f(x_1, x_2, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(C)}} e^{-(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)/2},\ -\infty < x_i < \infty. \tag{1}$$

Then the following are true:

- The function $f(x_1, x_2, \ldots, x_p)$ is a joint p.d.f. for some r.v.'s $X_1, X_2, \ldots, X_p$.
- Each $X_i$ is normally distributed with mean $\mu_i$ and variance $c_{i,i}$.
- $c_{i,j} = \mathrm{Cov}(X_i, X_j)$.
- If $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, then $\mathrm{Cov}(\mathbf{X}) = C$.

**Notation:** If $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ has the joint p.d.f. of (1), then we say that $\mathbf{X}$ is distributed as $N_p(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_p)$ and $\Sigma = \mathrm{Cov}(\mathbf{X})$ and write $\mathbf{X} \sim N_p(\mu, \Sigma)$.

Finally, multivariate versions of the classical theorems for univariate statistics hold as well:

**Theorem 4.5.** *If* $\mathbf{a}$ *is a* $p$*-vector and the random vector variable* $\mathbf{X} \sim N_p(\mu, \Sigma)$*, then the random variable* $\mathbf{a}^T\mathbf{X} \sim N\left(\mathbf{a}^T\mu, \mathbf{a}^T\Sigma\mathbf{a}\right)$.

**Theorem 4.6.** ***(Multivariate Central Limit Theorem)*** *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent and identically distributed random vector variables of length $p$ with a finite expected value $\mu$ and covariance matrix $\Sigma$. Then the random vector variable*

$$\sqrt{n}\left(\overline{\mathbf{X}} - \mu\right)$$

*has joint distribution that approaches $N_p(\mathbf{0}, \Sigma)$ as $n \to \infty$.*

## 5. Parameter Estimation

5.1. Confidence Intervals**.** The word "statistic" has different meanings to different folks. For us a **statistic** shall mean any definite function of one or more r.v.'s. Two important examples come from the notion of a random sample, which means a sequence of independent and identically distributed (abbreviated to i.i.d.) random variables, say with mean $\mu$ and variance $\sigma^2$:

- The **mean** of a random sample $X_1, X_2, \ldots, X_n$:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- The **variance** of the mean of a random sample $X_1, X_2, \ldots, X_n$:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

- The **variance** of a random sample $X_1, X_2, \ldots, X_n$:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2.$$

- The **sample variance** of a random sample $X_1, X_2, \ldots, X_n$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

It can be shown that the mean of a random sample is an unbiased estimator of $\mu$ and the special and sample variance are an unbiased estimators of the variance of the distribution; that is, the expected values of sample mean, special and sample variance are $\mu$, $\sigma^2$, respectively.

In the case of a vector random variable $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, the definitions are nearly the same. One has to be a bit careful about the notation here because the vector $(X_1, X_2, \ldots, X_n)$ is not the same thing as a univariate random sample as above. In this setting, $X_1, X_2, \ldots, X_n$ are random variables with a joint distribution and need not be independent. Rather, by a vector random sample we means a sequence of independent and identically jointly distributed (also abbreviated to i.i.d.) vector random variables with mean vector $\mu$ and covariance matrix $\Sigma$:

- The **mean vector** of a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$:

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i.$$

- The **covariance matrix** of the mean of a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$:
$$\mathrm{Cov}\left(\overline{\mathbf{X}}\right) = \frac{1}{n}\Sigma.$$
- The **covariance matrix** of a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$:
$$V = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mu\right)\left(\mathbf{X}_i - \mu\right)^T.$$
- The **sample covariance matrix** of a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$:
$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T.$$
- The **sample correlation matrix** of a random sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$:
$$R = D^{-1}SD,$$
where $D$ is the diagonal matrix whose diagonal entries are the square roots of the diagonal entries of $S$.

Analogous to the scalar case, it can be shown that the mean of a random sample is an unbiased estimator of $\mu$, variance and sample variance are unbiased estimators of the covariance matrix of the distribution, that is, the expected values of sample mean, special and sample variance are $\mu$, $\Sigma$, respectively.

**Key question:** How do we estimate the mean and and variance of the distribution of these random samples, or for that matter any other parameter associated with the distribution?

This question leads us to the notion of **confidence intervals**: given a probability $1-\alpha$ called a "**confidence coefficient**" use the data to construct the smallest possible interval $I$ of real numbers such that the probability that the true value of the parameter being in this interval is $1-\alpha$. The use of $1-\alpha$ is a matter of convenience in formulas. One ofter sees terms like the "95% confidence interval". This means the confidence interval found with $1-\alpha = 0.95$. We'll explore these ideas for the case of mean and variance.

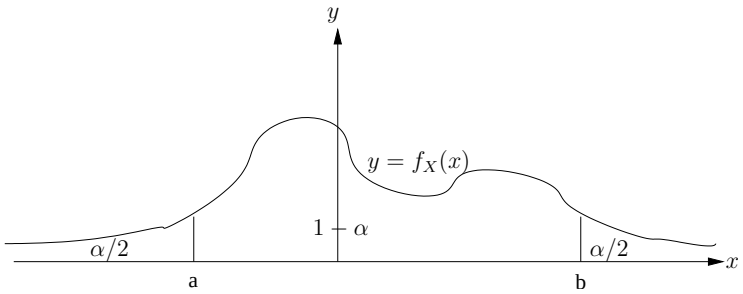
**Motivation:** What's so great about the mean?

**Example 5.1.** Suppose that we are attempting to measure a ideal and definite physical quantity, say the mass $m$ of the object. We do so by taking repeated measurements of the mass, say $m_1, m_2, \ldots, m_n$. What to do with these numbers? In the absence of any other information, we might average them out, in the hopes that errors will somehow cancel each other out. Is this realistic? Answer: sometimes.

Specifically, we'll describe the experiment more formally as a sequence $M_1, M_2, \ldots, M_n$ of r.v.'s. We may write
$$M_i = m + X_i$$
where $X_i$ is the error of the $i$th measurement. Certainly, it is reasonable to assume that these r.v.'s are independent. In many cases it is also reasonable to assume that the errors are normally distributed with mean 0 and standard deviation $\sigma$. It follows that the $M_i$ are normally distributed with mean
$$E\left[M_i\right] = E\left[m + X_i\right] = E\left[m\right] + E\left[X_i\right] = m + 0 = m$$
and variance
$$\mathrm{Var}\left(M_i\right) = \mathrm{Var}\left(E_i\right) = \sigma^2.$$

In particular, the number we're really interested in, $m$, is the expectation of a normal random variable. We'll see why the sample mean is useful by answering the key question.

The basic idea for computing confidence intervals is to find a statistic $X$ that is a "good" estimator of the desired parameter and has a known distribution, which can be used to compute a confidence interval. Here's how: we split the probability into half and construct an interval which has $\alpha/2$ area under the p.d.f. $f_X$ to the left of the point $a$ and area $\alpha/2$ under the p.d.f. to the right of the point $b$. Here's a picture:




We accomplish locating the points $a$ and $b$ as follows: solve the equations

$$\begin{aligned} F_X(a) &= \frac{\alpha}{2} \\ F_X(b) &= 1 - \frac{\alpha}{2}. \end{aligned}$$

If $F_X$ is continuous, we are guaranteed that solutions exist, and in fact, an inverse function to the c.d.f. $F_X$ exists.

Finally, WHAT DOES ALL THIS MEAN, exactly??? It means that if you calculate a confidence interval based on data you have observed, and if all the hypotheses about i.i.d. normal r.v.'s is correct, then the true value of parameter you are estimating is in this interval with a probability of $1-\alpha$. Put another way: $100 \cdot (1-\alpha)$ times out of 100 this calculation will yield an interval containing the desired parameter.

.

5.2. Estimating Mean with Known Variance and Normal Distribution. Some simple facts about normal distributions play a key role here. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. normal r.v.'s with mean $\mu$ and variance $\sigma^2$. From various facts outlined in these notes we have:

- $X_1 + X_2 + \cdots + X_n$ has a normal distribution with mean $n\mu$ and variance $n\sigma^2$. (See p. 2.3.)
- So $\frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \overline{X}$ has a normal distribution with mean $n\mu/n = \mu$ and variance $n\sigma^2/n^2 = \sigma^2/n = (\sigma/\sqrt{n})^2$. (See p. 2.3.)
- Hence $Z = \left(\overline{X} - \mu\right) / (\sigma/\sqrt{n})$ has a standard normal distribution. (See p 2.3.)

Thus we have shown that

**Theorem.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean $\mu$ and variance $\sigma^2$. Then the statistic*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

*has a standard normal distribution.*

5.3. Estimating Mean with Unknown Variance and Normal Distribution. The Student's t distribution plays a key role here. The key theorem is as follows:

**Theorem.** *(Sampling Theorem) Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean $\mu$ and variance $\sigma^2$. Then the statistic*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

*has a Student's t distribution with $n-1$ degrees of freedom.*

*Estimating Variance with a Normal Distribution.* The chi-square distribution plays a key role here.

**Theorem.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. normal r.v.'s with mean $\mu$ and variance $\sigma^2$. Then the statistic*

$$Y = (n-1)\,\frac{S^2}{\sigma^2}$$

*has a chi-square distribution with $n-1$ degrees of freedom.*

We use these statistical facts as outlined above. If we are given sample data, we calculate the resulting test statistic and observe whether or not it falls in the confidence interval.

## 6. Stochastic Processes

6.1. **Basic Ideas.** Some random processes can be thought of as a sequence $X_n$, $n = 0, 1, \ldots$, of random variables. Examples are daily measurements of temperature at some fixed locale or values of a particular stock at the end of each day. If we try to pass from sequences of time instants $t_0, t_1, \ldots$ and discrete random variables $X(t_n)$, to a continuum of time values $t$ where $t$ ranges over some finite or infinite interval, we obtain a continuum of random variables $X(t)$, one for each time $t$. Let $T$ be any time set such as a finite or infinite interval, or a sequence of discrete times. Assume that there is an underlying sample space $S$ and probability measure $P$ on $S$. A **stochastic process** is a function

$$X : T \times S \to \mathbb{R},$$

that is, mapping ordered pairs $(t, \omega)$ of time $t$ and outcome $\omega$ to real numbers subject to the condition that for each fixed time $t$ , $X(t, \omega)$ is a random variable on the sample space $S$. Thus, if we select and fix an outcome $\omega \in S$, we obtain a function

$$X(\cdot, \omega) : T \to \mathbb{R}.$$

Such a function is called a **realization** (**sample path**, **trajectory**) of the stochastic process $X$. Plots of sequences of random values can be thought of as a graph of a sample path. For example, we could plot the value of a stock $S(t)$ against time and we would have a realization of the stochastic process $S(t, \omega)$. It is customary to suppress the outcome $\omega$. Indeed, the "outcomes" of these random experiments are rather nuanced: one could imagine selecting a single universe, where the stock took one sample path, whereas in another parallel universe, the stock behaved differently in time. Notice that it is possible that there is some correlation between the random variables $X(s)$ and $X(t)$, $s < t$, or that they are entirely independent of each other. We'll say more about this later.

**Examples.**

**(1)** The sure thing: experiment is that you pick a number by selecting a ball from an urn; the number written on the ball is the outcome...except that every ball has the same value. What are sample spaces, etc?

**(2)** A **binomial lattice**: We are going to track the price $S(t)$ of a stock through time, which we think of as a random variable, with some simplifying assumptions:

(1) Times measured are discrete in a fixed unit (period), e.g., days. Thus, the only values of $S(t)$ are $S = S_0 = S(0) > 0$, $S_1 = S(1)$, ..., $S_n = S(n)$
(2) At each stage, the stock will either move up with a return of $u$ or move down with a return of $d$ and does so with probabilities $p, q$.

**(3)** Random moment-to-moment fluctuations in a noisy quantity are modeled by a normal distribution of mean zero, variance $dt$ (roughly the time between measurements) and scale factor $\sigma$ (called the **volatility**.)

Now let's focus on (3). Let $X(t)$ be the function of random fluctuations that we are interested in. What on earth does this symbol mean? Well, for each particular time $t_0$, $X(t_0)$ is a random variable. The whole ensemble $X(t)$ of random variables is a **stochastic process** (or **random process**.)

The sample space for $X(t)$ is rather subtle: if we observe some statistic in time, e.g., a stock price, we get a realization $x(t)$ of $X(t)$. Let's draw a rough picture of what such a realization might look like: a very jagged graph that tends to move in a direction (up or down.)

Now $x(t)$ is just one function, and is super-highly discontinuous. But couldn't you imagine another universe, where the sampling of $X(t)$ led to a different realization? Sure! And there are infinitely many such possibilities.

Now this idea is so general that it is entirely worthless without further qualifications. So what we are going to do is to examine changes in random variables

$$\Delta X = X(t + dt) - X(t).$$

However, we are going to do this in a "limiting" sense, that is, we will let $dt$ diminish in size and pass to "differentials" $dX(t)$. This, too, is a stochastic process, and here is here we will lay down some conditions. BTW, the formal development of this topic would work with so-called stochastic integrals instead of differentials:

$$X(t + dt) - X(t) = \int_t^{t+dt} dX(t),$$

and though it is more rigorous, this is exactly what we are going to avoid. Just remember that every differential assertion we make has a formal integral justification.

Consider a random process $X(t)$ which has fluctuation $dX$ which has these properties. Such a process is called **Brownian motion:**

(1) $dX$ has a normal distribution function (for any $t, dt$.)
(2) the mean of $dX$ is zero
(3) the variance of $dX$ is $dt$.

Let's think about the intuitive content of each of these:

(1) We'll, that's because we love normal distributions and in the absence of additional information, this is a reasonable approximation. In the absence of additional information, there's no reason to think the distribution function at any one time is different

from that of some other time. (In statistics parlance, the r.v.'s $X(t)$ are **identically distributed**.

(2) We are talking about random fluctuations; so why should they prefer up to down? Hence, expected value zero.

(3) If we sample at $t$ and $t+dt$, we get a difference $X(t+dt)-X(t)$ that has a certain variability, which is nicely measured by the variance. If we cut the time lag in half, isn't it reasonable to expect the variability of the difference to go down by a half? That's essentially what condition 3 is saying, though it's a little stronger than that.

Let the r.v. $\phi$ have the standard normal distribution and we can write this as

$$dX = \phi\sqrt{dt}.$$

Thus the random fluctuations that we consider here could be described as, when we add a **volatility** factor $\sigma$,

$$\sigma dX = \sigma\phi\sqrt{dt}$$

which is a normally distributed r.v. with mean zero and variance $\sigma^2 dt$.

**Reason:** Calculate

$$E\left[\sigma dX\right] = \sigma E\left[dX\right] = \sigma 0 = 0$$

and

$$\mathrm{Var}\left(\sigma dX\right) = \mathrm{Var}\left(\sigma\phi\sqrt{dt}\right) = \sigma^2 dt\,\mathrm{Var}\left(\phi\right) = \sigma^2 dt.$$

We seem to be a step ahead of ourselves in that so far, differentials of random variables has presented in a very informal way. We don't really have a precise idea of what differentials mean and how we would integrate them in the same way that we integrate ordinary differentials in calculus. Once we have the correct definitions in place, we'll see that Brownian motion is essentially the same thing as a Wiener process, which we'll describe below.

6.2. **Asset Price , Random Walks and Ito's Lemma.** We think of the price of an asset as random variable $S(t)$ that moves with time. It's often stated that asset prices must move randomly because of the efficient market hypothesis, which basically says two things:

- Past history is fully reflected in the present price, which does not hold any further information.
- Markets respond immediately to any new information about an asset.

Thus modeling of prices is about modeling the arrival of new information which affects the price. Unanticipated changes in the asset price can be modeled by a "Markov process" of which the Wiener process we describe above is a special case.

Now were ready for a model of the price of a stock as a function of time, $S(t)$. Understand: $S(t)$ is a stochastic process! If we sample this stochastic process at discrete points, say we obtain end-of-day values $S\left(t_1\right) = S_1$, $S\left(t_2\right) = S_2$, etc., we obtain what is called a **random walk**. Suppose that in a small increment of time $dt$ the stock price experiences a change $dS$.

**Note:** we're using the differential as the "limiting" form for $\Delta S$.

Rather than model $S(t)$ itself, it makes more sense to think about relative changes, which are exactly (well, nearly for small $dt$)

$$\frac{dS}{S}.$$

Let's discuss the deterministic and random parts of what this expression should be.

**Deterministic:** $\mu\,dt$, where $\mu$ is the relative rate of change of $S$ with respect to time. In the case of a bond you can think of $\mu$ as the (continuous) prevailing risk-free interest

rate. For our stock, it is called the **drift**. This factor is determined by the prospects of the company, quality of the management, etc.

**Stochastic (random):** $\sigma\, dW$, where the constant $\sigma$ plays the role of a **volatility** factor.

Thus, the equation describing variations in return over time is

$$\frac{dS}{S} = \sigma\, dW + \mu\, dt. \tag{2}$$

This equation is an example of a **stochastic differential equation**.

Our objective is to solve this differential equation for $S$ in some fashion. We could rewrite it in an alternate form

$$dS = \sigma S\, dW + \mu S\, dt \tag{3}$$

**What if there were no randomness?**

We mean that $\sigma = 0$. Solve this ordinary differential equation and obtain the familiar formula

$$S(t) = S(0)e^{\mu t}.$$

O.K., now back to the real world.

**Example.** The differential $dS$ is also a random variable. What is its mean and variance, given today's price?

We'll calculate the mean at the board. Variance is left as an exercise.

There is a key fact that we'll need that is, at first glance, not very intuitive:

**Key Fact:**

$$dW^2 \to dt \text{ as } dt \to 0.$$

We'll put an explanation of this off for a moment and get to a really fundamental fact.

**Ito's Lemma:** Let $f(S,t)$ be a smooth function and $S$ the random variable given by Equation (3). Then $df$ is a random variable given by

$$df = \sigma S\frac{\partial f}{\partial S}dW + \left(\mu S\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 f}{\partial S^2} + \frac{\partial f}{\partial t}\right)dt \tag{4}$$

We can see why this is reasonable in steps:

(1) Start with Taylor's theorem for $\Delta f$ and calculate

$$\Delta f = \frac{\partial f}{\partial S}\, dS + \frac{\partial f}{\partial t}\, dt + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}\, (dS)^2 + \frac{\partial^2 f}{\partial t^2}\, (dt)^2 + \frac{\partial^2 f}{\partial S \partial t}\, dS\, dt + \cdots$$

(2) Recall Tschebychev's inequality: For a random variable $Y$ with finite variance $\sigma^2$ and mean $\mu$, and for $k > 0$,

$$P\left(|Y - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

and note that $dX^2 = N^2 dt$ and $N^2$ has a chi-squared distribution of mean 1 and variance 2. Hence $dX^2$ has mean $dt$ and variance $\sigma = 4dt^2$. Thus if we take $k = 1/4dt$, we obtain

$$P\left(\left|dX^2 - dt\right| \geq dt\right) \leq 16dt^2$$

so that $dX^2$ approaches $dt$ in a probabilistic sense as $dt \to 0$.

(3) Finally, substitute $dS = \sigma S\, dW + \mu S\, dt$, and since we are letting $dt \to 0$, replace $\Delta f$ by $df$ and $dW^2$ by $dt$, do some algebra and we're done.

**Note:** (1) Ito's Lemma is actually a little more general than we stated it. It applies to differentials

$$dS = a(S,t)dW + b(S,t)dt. \tag{5}$$

yielding for $f(S,t)$ the formula

$$df = b\frac{\partial f}{\partial S}dW + \left(a\frac{\partial f}{\partial S} + \frac{1}{2}b^2\frac{\partial^2 f}{\partial S^2} + \frac{\partial f}{\partial t}\right)dt$$

(2) Ito's Lemma underscores the difference between deterministic and random variables. Were we dealing with deterministic variables and were given that we would simply have plugged this into the formula of Equation (2), done a little algebra and obtained

$$df = \sigma S\frac{\partial f}{\partial S}dW + \left(\mu S\frac{\partial f}{\partial S} + \frac{\partial f}{\partial t}\right)dt$$

The difference comes from the Key Fact:

$$dW^2 \to dt \text{ as } dt \to 0.$$

Take, e.g., $\epsilon = \sqrt{dt}$.

6.3. **Stochastic Integrals.** Equation (3) has something really nice going for it: we can actually integrate this stochastic differential equation. Before we do so, we need to interpret these differentials. Here is a differential-free version Brownian motion $W$. A **Wiener process** $W(t)$ is a stochastic process that has the following properties:

(1) $W(0) = 0$ and $W(t)$ has a normal distribution function for $t > 0$.
(2) The mean of $W(t)$ is zero.
(3) The variance of the increment $W(t) - W(s)$ is $t - s$ for $t > s$.
(4) Increments over non-overlapping intervals are independent.

Let's think about the intuitive content of each of these:

(1) We'll, that's because we love normal distributions and in the absence of additional information, this is a reasonable approximation. Also $W(0) = 0$ because we assume that whatever process we are interested in is determined at the one instant in time $t = 0$.
(2) We are talking about random fluctuations; so why should they prefer up to down? Hence, expected value zero.
(3) If we sample at $t$ and $t + dt$, we get a difference $W(t + dt) - W(t)$ that has a certain variability, which is nicely measured by the variance. If we cut the time lag in half, isn't it reasonable to expect the variability of the difference to go down by a half? That's essentially what condition 3 is saying, though it's a little stronger than that.
(4) This is a way of saying that past changes before time $t$ has no effect on the change $W(s) - W(t)$, $s > t$.

Now if we look back to the definition of the Brownian motion $dX$ we see how the properties of this motion can be deduced from the limiting idea

$$W(t + dt) - W(t) \to dW, \text{ as } dt \to 0.$$

For example, we required that $dW$ have variance $dt$, which is exactly the variance of $W(t + dt) - W(t)$. The fact that each $dW$ is normally distributed is a bit more subtle, as it requires the Central Limit Theorem. We'll leave it at that.

Before we try to integrate, let's have a look at the ACTUAL working definition of a stochastic integral:

**Definition:** Let $W(t)$ be a Wiener process and $b(W,t)$ a function of $t$ and $W$. Then we define

$$Y(t) - Y(0) = \int_0^t b(W(\tau), \tau)\, dW(\tau)$$

provided that $Y(t)$ is a stochastic process defined by

$$Y(t) - Y(0) = \lim_{m\to\infty} \sum_{j=0}^{m} b(W(t_j), t_j)\,(W(t_{j+1}) - W(t_j))$$

where $0 = t_0 < t_1 < \cdots < t_{m+1} = t$ and $\max_j (t_{j+1} - t_j) \to 0$ as $m \to \infty$. This integral with respect to a Wiener process is called an *Ito stochastic integral*.

Notice, I haven't said what these limits mean. Nonetheless, it's pretty clear from definition that the ordinary properties of the integrals from calculus, like linearity, hold for this object.

Now we can interpret the meaning of the differential expression

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

namely, we understand that this means that the stochastic process $Y(t)$ satisfies the integral equation

$$Y(t) - Y(0) = \int_0^t a(X,t)\ dt + \int_0^t b(X,t)\ dW$$

where $W(t)$ is a Wiener process and the second integral is the Ito stochastic integral.

This definition is good enough to calculate one integral:

$$W(t) - W(0) = \int_0^t 1 dW.$$

Let's do it....Warning: most stochastic integrals are MUCH tougher than this and the answers are not quite the ones we might expect.

Now handle $df$, where $f(S,t) = \ln(S)$: we'll get from Ito's Lemma that

$$\ln(S(t)) - \ln(S(0)) = \sigma(W(t) - W(0)) + \left(\mu - \frac{1}{2}\sigma^2\right) t.$$

Now $W$ is a Wiener process, so $W(0) = 0$ and $W(t)$ is a normally distributed r.v. with mean zero and variance $t$. This says that if $S(0) = S_0$ is given, then $\ln(S(t))$ is the sum of a constant and a normal distribution $W(t) - W(0) = W(t)$ of variance $t - 0 = t$. Thus, we have $\sigma W(t) = \sigma\sqrt{t}z$, where $z \sim N(0,1)$. It also implies that if we set $\nu = \mu - \frac{1}{2}\sigma^2$, then we can take exponentials of the above equation and obtain

$$S(t) = S(0)\, e^{\nu t + \sigma\sqrt{t}\cdot z}.$$

We'll calculate

$$E[\ln(S(t))] = \left(\mu - \frac{1}{2}\sigma^2\right) t + S(0)$$

and

$$\mathrm{Var}(\ln(S(t))) = \sigma^2 t.$$

Therefore, the probability density function for $S(t)$ is

$$f_S(s) = \frac{1}{\sigma\sqrt{2\pi t}} e^{-\left(\ln s - \left(\left(\mu - \frac{1}{2}\sigma^2\right)t + S(0)\right)\right)/(2\sigma^2 t)}.$$

When a r.v. $Y$ is such that $\ln Y$ is normally distributed, we say that $Y$ is *lognormally* distributed. Thus, stock price at a given time is modeled as a lognormal random variable. See Appendix B for a discussion.