A.1: Review of Univariate Random Variables

A univariate random variable is a single random variable *X*. If the domain of *X* is finite (or *denumerable*), we can characterize the uncertainty about *X* by listing for each possible value *x* of *X*—e.g., $x ? \{x_1, ..., x_m\}$—the probability that *X* has value *x*. We write this *probability distribution* of *X* as $p(X = x)$, or, to refer to the probability distribution of single values in general, $p(x)$. When the domain is finite, as in this case, the set of probabilities $\{p(x_1), ..., p(x_m)\}$ is often called a *probability mass function*. Note that the expression $p(X)$ refers to the set of *m* numbers $\{p(x_1), ..., p(x_m)\}$, and $p(x)$ refers to some (arbitrary) member of this set. The *cumulative distribution function* $P(x)$ of a random variable is the probability that it will take a value less than or equal to *x* (when the values *x* can be ordered).

Cumulative distribution functions can also be defined for continuous random variables (those that can take any value on an interval or the real line). In this case we will usually denote the cumulative distribution by $F(x)$ or $P(x)$, and the derivative of $F(x)$, the *probability density function* of *x* (often just "density function" for short), by $f(x)$ or $p(x)$. This function gives the probability that the observed value will lie in an infinitesimal interval surrounding *x*. Here, for the sake of simplicity, we will often provide descriptions only in terms of density functions, but analogous arguments apply to probability mass functions. Introductory texts on mathematical statistics provide more formal descriptions of these concepts, but these informal definitions will suffice for our purposes.

In terms of notation both $p(x)$ and $f(x)$ are often used to denote a probability density function on a continuous variable *x*. It should be clear from the context, depending on whether *x* is discrete or continuous, whether $p(x)$ is referring to a probability mass function or a probability density function for *x*.

The randomness of a random variable may arise for various reasons—essentially the sources of uncertainty: perhaps we have observed a randomly selected member of the population, perhaps there is measurement error associated with the value, perhaps *X* is not directly observable, and so on. We often approximate this randomness by assuming that the actual observed values have arisen from some well known distribution of possible values. Certain distributions are especially useful in data mining, and some of these are described in appendix A.2. They include the Normal (or Gaussian) distribution and the Poisson distribution.

We use the notion of the mean value (or *expected value*, or *expectation*) in chapter 2. For a sample (or finite population) the mean is the average value, obtained by dividing the sum of the values in the sample (or finite population) by the total number of values. More generally, suppose that value *x* occurs in the population with probability $p(x)$. Then the mean value of the variable *X* for the population is given by $?_x x p(x)$. However, if *X* can take a continuum of values, it is not meaningful to speak of the probability that any particular exact value *x* will occur, since exact values have zero probability of occurring. Instead we consider the probability that *X* lies in a small interval of width d*x* and find the limiting value of the sum $?_x x f(x) \mathrm{d}x$ as this width decreases toward zero. This leads us to replace summation with integration. If the probability density function of the continuous variable *X* is $f(x)$, the expected value is $?x f(x) dx$.

The notation *E* is often used to denote expectation, so that the expected value of a random variable *X* is $E[X]$. The Greek letter µ is also often used to denote a mean, or, if we need to make it clear that the random variable *X* is being discussed, μ_x may be used. More precisely, the expected value of *X* *with respect to* the density function $f(x)$ is denoted $E_{f(x)}[X]$. Note that we can define the expected value of a function of *X*, $g(x)$, with respect to $f(x)$, as $E_{f(x)}[g(x)] = ?g(x)f(x)dx$. If we let $g(x) = (x - E[x])^2$ we get the usual definition of variance σ_x^2.

Expectation is a linear operator. This is quite a useful general property. For example, it means that the expected value of a weighted sum of random variables is equal to the weighted sum of their expected values, regardless of whether the variables are dependent in any way (chapter 4 discusses more precisely what we mean by dependence of random variables).

The axioms of probability referred to above assign a probability of 0 to an event that cannot occur, and a probability of 1 to a certain event. If two events cannot occur together, the probability that one *or* the other will occur is the sum of their separate probabilities. Thus, in tossing a fair coin (with which the probability of obtaining a head is

1/2) the probability of obtaining either a head or a tail is 1/2+1/2 = 1. The situation starts to get more complicated—and more interesting—when events can occur together but are not certain to do so. This lead us to the notion of *multivariate* random variables, discussed in detail in chapter 4.

A.2: Some Common Probability Distributions

Above we discussed the general idea of a probability distribution. Here we describe some specific and important probability distributions that arise in data mining.

The Bernoulli Distribution
The Bernoulli distribution has just two possible outcomes. Situations which might be described by such a distribution include the outcome of a coin toss (heads or tails) or whether a particular customer buys a particular product or not. Denoting the outcomes by 0 and 1, let p be the probability of observing a 1 and (1-p) the probability of observing a 0. Then the probability mass function can be written as $p^x(1 - p)^{1-x}$, with x taking the value 0 or 1. The mean of the distribution is p and its variance is $p(1 - p)$. Note that this distribution has just a single parameter, namely p.

The Binomial Distribution
This is a generalization of the Bernoulli distribution, and describes the number x of "type 1 outcomes" (e.g., successes) in n independent Bernoulli trials, each with parameter p. The probability mass function has the form $\binom{n}{x}p^x(1-p)^{n-x}$, where x can take integer values between 0 and n. The mean is np and the variance is $np(1 - p)$.

The Multinomial Distribution
The multinomial distribution is a generalization of the binomial distribution to the case where there are more than two potential outcomes; for example, there may be k possible outcomes, the ith having probability p_i of occurring, 1 = i = k. The probabilities p_i sum to 1, and the model has k - 1 parameters p_1, ..., p_{k-1} (since p_k = 1 - ? $_i p_i$).
Suppose that n observations have been independently drawn from a multinomial distribution. Then the mean number of observations yielding the ith outcome is np_i and its variance is $np_i(1 - p_i)$. Note that, since the occurrence of one outcome means the others cannot occur, the individual outcomes must be negatively correlated. In fact, the covariance between the ith and jth (i ? j) outcome is $-np_ip_j$.

The Poisson Distribution
If random events are observed independently, with underlying rate ?, then we would expect to observe ?t events on average in a time interval of length t. Sometimes, of course, we would observe none in time t, at other times we would observe 1, and so on. If the rate is low, we would rarely expect to observe a large number of events (unless t was large). A distribution which describes this state of affairs is the *Poisson distribution*. It has probability mass function (?t)^x $e^{-?t}$/x!. The mean and variance of the Poisson distribution are the same, both being ?.
Given a binomial distribution with large n and small p such that np is a constant, then this may be well approximated by a Poisson distribution.

The Normal (or Gaussian) Distribution

The probability density function takes the form

$$\frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$$

where µ is the mean of the distribution and s^2 is the variance. *The standard normal distribution* is the special case with zero mean and unit variance. The normal distribution is very important, partly as a consequence of the *central limit theorem*. Roughly speaking, this says that the distribution of the mean of a sample of n observations becomes more and more like the normal distribution as n increases, regardless of the form of the populations distribution from which the data are drawn. (Of course, mathematical niceties require this to be qualified for full rigor.) This is one reason why many statistical procedures are based on an assumption that various distributions are normal.

The normal distribution is symmetric about its mean, and 95% of its probability lies within ±1.96 standard deviations of the mean.

The Student's *t*-distribution

Consider a sample from a normal distribution with known standard deviation s. An appropriate test statistic to use to make inferences about the mean would be the ratio

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean. Using this, for example, one can see how far the sample mean deviates from a hypothesized value of the unknown mean. This ratio will be normally distributed by the central limit theorem (see *Normal distribution* above). Note that here the denominator is a constant. Of course, in real life, one is more likely to be in a situation of making inferences about a mean when the standard deviation is *unknown*. This means that one would usually want to replace the above ratio by

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

where *s* is the sample estimate of the standard deviation. As soon as one does this the ratio ceases to be normally distributed—extra random variation has been introduced by the fact that the denominator now varies from sample to sample. The distribution of this new ratio will have a larger spread than that of the corresponding normal distribution—it will have fatter tails. This distribution is called the *t*-distribution. Note that there are many—they differ according to how large is the sample size, since this affects the variability in *s*. They are indexed by (*n* - 1), known as the *degrees of freedom* of the distribution.

We can also describe this situation by saying that the ratio of two random variables, the numerator following a normal distribution and the square of the denominator following a chi-squared distribution (see below), follows a *t*-distribution.

The probability density function is quite complicated and it is unnecessary to reproduce it here (it is available in introductory texts on mathematical statistics). The mean is *n* - 1 and the variance is (*n* - 1)/(*n* - 3).

The Chi-Squared Distribution

The distribution of the sum of the squares of *n* values, each following the standard normal distribution, is called the *chi-squared distribution with n degrees of freedom*. Such a distribution has mean *n* and variance 2*n*. Again it seems unnecessary to reproduce the probability density function here—it can be readily found in introductory mathematical statistics texts it needed. The chi-squared distribution is particularly widely used in tests of goodness-of-fit.

The *F* Distribution

If *u* and *v* are independently distributed Chi-squared random variables with n_1 and n_2 degrees of freedom, respectively, then the ratio

$$F=\frac{u}{n_1}/\frac{v}{n_2}$$

is said to follow an *F distribution with* n_1 *and* n_2 *degrees of freedom*. This is widely used in tests to compare variances, such as arise in analysis of variance applications.

The Multivariate Normal Distribution

This is an extension of the univariate normal distribution to multiple random variables. Let **x** = (x_1, ..., x_p) denote a *p* component random vector. Then the probability density function of the multivariate normal distribution has the form

$$\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

where µ is the *p*-dimensional mean vector of the distribution and S is the *p* × *p* covariance matrix.