Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 2: Overview of Probability

As stated in Lecture 1, the ultimate motivation for data analysis is often the desire to deduce certain properties of the underlying population from which the data were sampled. Before we introduce a sampling of the statistical methods that can be used to draw inference about a variety of population parameters and/or structures using the information contained in the data, we must first examine the rules of probability that govern the underlying population. Familiarity with these rules is an important prerequisite for understanding the derivation of and assumptions required by the methods of statistical inference that will be presented in Lectures 3-8.

(I.) Review of Elementary Set Theory

Prior to examining the basic rules of probability, a quick review of elementary set theory may be instructive. To begin, a "set" is defined as a collection of items, and the items that are members of a set are referred to as its "elements." An example of a set is F = {apple, pear, orange}. A set may contain an infinite number of elements, or it may contain no elements, in which case it is referred to as the "null set" or the "empty set" and designated by the symbol $\varnothing$. A "subset" of a set (say set A) is any set whose elements are all members of set A; the expression $B \subset A$ means that set B is a subset of set A. For instance, C = {apple, orange} is a subset of set F. Technically, $\varnothing$ is a subset of every set, and every set is a subset of itself.

(II.) Basic Probability

In general, we can view any phenomenon that we might want to investigate by collecting and analysing data as an experiment of some sort, whether that experiment occurs by design or naturally. A "designed experiment" is planned ahead of time by a researcher and carried out under carefully controlled and monitored conditions; this type of experiment is particularly common in the natural sciences, in animal psychology, and in medicine (i.e., clinical trials). However, in fields such as economics, anthropology, and human psychology, it is commonly impossible, for a variety of reasons (e.g., ethical considerations), to answer certain questions using a designed experiment; in these cases, researchers are often forced to collect data from naturally occurring experiments in order to address those questions. In general, naturally occurring experiments are vulnerable to more potential biases than are designed experiments (e.g., the self-selection problem), and these biases may invalidate the conclusions reached by statistically analysing the data from such naturally occurring experiments.

However, the problems associated with naturally occurring experiments aside, both kinds of experiments can be treated identically for the purposes of this lecture. For both types of experiments, before the experiment has been carried out, its outcome is unknown and can be predicted from existing theories or from the results of past experiments. We can view each possible "outcome" of an experiment as occurring with a certain probability (before the experiment is carried out), and a "probability model" can be used to describe how likely the occurrence of each potential outcome of the experiment is. More precisely, the set of possible outcomes of an experiment and the corresponding set of probabilities is termed a "probability distribution." Before proceeding, we should note that we are referring to the population probability distribution because we are talking about all the experimental outcomes that could potentially occur (before the experiment is performed), and not the outcome(s) that actually does (do) occur in our sample (after the experiment has been carried out).

(II.a.) Sample Space

The set of possible outcomes of an experiment is known as the "sample space" and denoted by Ω, where $\Omega = \{\omega_1, \omega_2, ...\}$ and the ω_i s are the potential outcomes. For instance, if the experiment consists of tossing a coin once, then $\Omega = \{H, T\}$. As another example, if we are interested in the number of individuals in a village with the flu on one particular day, the sample space is $\Omega = \{0,1,2,...,N\}$, where N is the total number of people in the village.

(II.b) Events

An "event" is any subset of the sample space. For instance, in the previous village example, possible events include 'at least five cases,' 'at most two cases,' and 'exactly five cases.' Individual outcomes can be referred to as events since a set containing one outcome is a subset of the sample space. However, there are clearly more possible events than possible outcomes for an experiment. Returning to the basics of set theory, we introduce the idea of the "complement of event A," which is denoted by A^C and means that 'A does not occur.' More specifically, the event A^C contains all outcomes that are members of Ω but not of A. In addition, if we are considering two events at once (say A and B), we can speak of their "union," which is denoted $A \cup B$ and means that 'either A or B or both happen(s).' The result of $A \cup B$ is an event consisting of the outcomes contained in set A or in set B or in both. Alternatively, for events A and B, we can speak of their "intersection," which is denoted $A \cap B$ and means that 'both A and B happen.' The result of $A \cap B$ is an event consisting of the outcomes contained both in A and in B. If the intersection of two events is the null set (i.e., they contain no common outcomes), these events can be referred to as "mutually exclusive" or "disjoint." Of course, the concepts of union and intersection can be generalised to apply to more than two events at once.

(II.c.) Probability

Each possible event for a given experiment has an associated "probability value." The probability value of an event *A* can be interpreted in either of two manners. In the first, which is called the "frequentist" interpretation, we assume that, in theory, the experiment in question can be repeated an infinite number of times under exactly the same conditions. Under the frequentist interpretation, the probability of event *A* is the proportion of times that the event would occur if the experiment were repeated an infinite number of times. Alternatively, since we will only be considering infinite populations in this course, we can think of the probability as the proportion of the population that satisfies the condition corresponding to event *A*. In the "Bayesian" interpretation of probability, we view the probability of event *A* as the strength of our belief that event *A* will occur; this interpretation does not assume that the experiment can be repeated an infinite number of times under the same conditions, which is obviously a theoretical impossibility in many cases. As an illustration of these two different interpretations of probability, suppose that it is known that the proportion of male births in a population is $^{105}\!/_{200}$. We can consider this number to represent the fraction of males in the population of new-born children, or we can think of it as our degree of belief that a particular birth will result in a boy (slightly more than 0.5).

We will use the symbol *P(A)* to denote the value assigned to the probability that event *A* will occur. Three basic axioms govern probability values:

(1) $0 \le P(A) \le 1$

(2) $P(\Omega) = 1$

(3) If *A* and *B* are mutually exclusive (i.e., $A \cap B = \varnothing$), then $P(A \cup B) = P(A) + P(B)$. This last axiom means that if events *A* and *B* share no common outcomes, then the probability of either or both of them happening equals the sum of their individual probabilities.

These rules imply the following properties of probability values:

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

(2) $P(\varnothing) = 0$

(3) If $B \subset A$, then $P(B) \le P(A)$.

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

As an example of these properties, consider the experiment that consists of randomly selecting a person from a particular village and has the sample space Ω ={Female with flu, Female with no flu, Male with flu, Male with no flu}. This sample space is represented in

Figure 1. Suppose that we know that the elements of Ω have the probabilities {0.15, 0.45, 0.10, 0.30}.

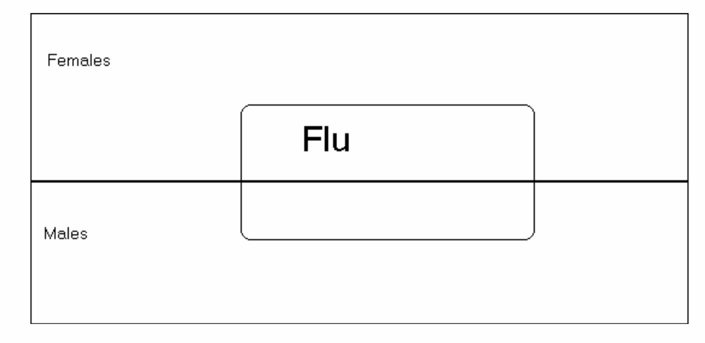

***Figure 1**: Events relating to gender and flu in a village*

We will denote the events 'Female,' 'Male' and 'flu' with the letters *F*, *M*, and *f*, respectively. Note that F^C=*M* and that M^C=*F*. Also, note that $F = (F \cap f) \cup (F \cap f^C)$, $M = (M \cap f) \cup (M \cap f^c)$, and $f = (F \cap f) \cup (M \cap f)$. Therefore, $P(F) = 0.6$, $P(M) = 0.4$, $P(f) = 0.25$, and $P(f^C) = 0.75$.

Conditional Probability

Often, the probability of an event depends on the occurrence (or non-occurrence) of other events. For instance, females may be more prone to a particular disease; if this is so, then our prediction of whether a person has that disease should be modified accordingly if we know his/her gender. We denote the probability of the occurrence of event *A*, conditional on the occurrence of event *B*, as $P(A|B)$, and define it as

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

The symbol $A|B$ is usually read as '*A* given *B*.' Continuing with the example of Figure 1, we see that the probability $P(F|f)$ refers to the event that 'the selected person is female, given that the person has flu.' Note the difference between the events $P(F \mid f)$ and $P(F \cap f)$. The former refers to the probability that a person is female given that we already know that he/she has the flu, and the latter refers to the probability that a person is female and has the flu. Returning to our example, we have that $P(F|f) = \frac{P(F \cap f)}{P(f)} = 0.60$, which reflects the fact that females comprise 60% of the people with the flu. Here, we should note than, in general, $P(A \mid B)$ does not equal $P(B \mid A)$. For instance, in our example, $P(F|f) = 0.6$ and $P(f|F) = 0.25$.

If knowing that *B* has occurred makes no difference to our knowledge of whether *A* occurs, and viceversa, we say that *A* and *B* are "independent" events. Mathematically, independence can be stated as $P(A|B) = P(A)$ or as $P(B|A) = P(B)$. These equations imply that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \quad \text{and} \quad P(A \cap B) = P(A) \times P(B).$$

All independent events satisfy the second equation above; conversely, if this equation is true for two events, then they are independent. Before proceeding, we should note that being independent is <u>not</u> the same as being mutually exclusive; in fact, if events *A* and *B* are mutually exclusive, then they are definitely not independent since the occurrence of event *A* clearly affects our knowledge of the occurrence of event *B* (because the former event precludes the latter), and vice versa.

Bayes' Theorem

Before we leave our discussion of basic probability, we will examine a probability theorem that is employed extremely often in statistics. This theorem, which is entitled "Bayes' Theorem," provides us with the ability to go from probabilities that are conditional in one direction to probabilities that are conditional in the other direction (e.g., from $P(A \mid B)$ to $P(B \mid A)$). This ability is an extremely useful one in cases where we are given $P(A \mid B)$ but are actually more interested in $P(B \mid A)$. These cases occur quite frequently in practice, an example being a situation where we know the probability with which a given disease results in a certain symptom, but instead desire to know the probability that the disease is present for individuals exhibiting the symptom. Bayes' Theorem can be stated in the following form:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^C)P(B^C)}.$$

As an illustration of the use of this theorem, consider (yet again) our flu example. If we are interested in finding out the probability that a person is female given that he/she has the flu, this probability is

$$P(f \mid F) = \frac{P(F \mid f)P(f)}{P(F \mid f)P(f) + P(F \mid f^C)P(f^C)} = \frac{0.60 \cdot 0.25}{0.60 \cdot 0.25 + 0.60 \cdot 0.75} = 0.25.$$

(III.) Random Variables and Univariate Probability Distributions

A "random variable" (r.v.) is a function that assigns a (possibly non-unique) real number value to each of the outcomes in the sample space for a given experiment. For instance, consider an experiment that consists of making two sequential and independent tosses of a coin, so that $\Omega = \{HH, HT, TH, TT\}$ is the sample space. A variable *X* that counts the number of heads that occur in the two tosses, so that $X(HH) = 2$, $X(HT) = X(TH) = 1$, and $X(TT) = 0$, is a random variable. Note that the random variable assigns the same (and thus a non-unique) real number value (i.e., 1) to the outcomes *HT* and *TH*; this is the case because the random variable *X* is not concerned with the *order* in which the heads/tails sequence occurs. For a given experiment, there is more than one associated random variable; often, one of these potential variables will be more commonly used as the variable of interest in practice. For instance, in the coin tossing experiment above, a variable *Y* that counts the number of tails, so that $Y(HH)=0$, $Y(HT)=Y(TH)=1$, and $Y(TT)=2$, is also a random variable, as is a variable *Z* that counts ¾ the number of heads plus ½ the number of tails, so that $Z(HH)=1.5$, $Z(HT)=Z(TH)=1.25$, and $Z(TT)=1$. In practice, probably either the first or the second, but not the third, random variable introduced above would be the variable of interest.

The "range" of a random variable *X*, which is denoted R_X, is the set of values that X can possibly assume. For instance, if *X* measures the number of heads in 20 coin tosses, then the range is {0,1,2,…,20}; if *X* measures the weight in kgs of healthy new-born children, then the range is the interval [2.3, 7.5], say. A random variable *X* is "discrete" if its range consists of a set that can be enumerated; if *X*'s range instead consists of a real-valued continuum, then *X* is "continuous" and can take on an uncountably infinite number of values. Examples of the former type of random variables are the number of siblings a person has, the number of heads resulting from tossing a coin 10 times, or the number of cars passing in front of a house between 9 a.m. and 10 a.m. Measures of length, weight, area, volume, time, etc. are examples of continuous random variables.

Here, an important point is that, when we collect a sample of values for an underlying random variable in an experiment, the data we collect are always discrete in the strict sense of the word since we can only perform measurements to a finite degree of precision. However, if the measurements of the random variable can still take on a relatively large number of possible values, we will think of it as continuous even if the set of possible measurements is not actually any value in a continuum. For instance, consider the distribution of incomes in a population. This variable is almost always modelled as continuous even though its range is not strictly continuous since we measure incomes only up to pence amounts. The fact that there are so many theoretically possible values lends plausibility to the use of a continuous distribution to model incomes.

Since some outcomes in the sample space for a given experiment are more likely to occur than others, it makes sense that some of the values in the range of a random variable for that experiment will be more likely to occur than others. For a random variable *X*, the frequency (or probability) with which the values in R_X occur is described by X's probability distribution, which is generally referred to as *X*'s "pdf." More correctly, this distribution is termed the "probability density function" or "pdf" if *X* is continuous and the "probability mass function" or "pmf" if *X* is discrete. The probability distribution function for *X* is typically denoted $f_X(x)$, where *x* is any one "realisation" of the random variable *X* (i.e., one particular value in R_X). [Here, we should note that if we were to collect a <u>sample</u> of *n* values of the random variable, the empirical distribution of those values (discussed in Lecture 1) would be the sample counterpart of the population pdf. Hopefully, the empirical distribution would be a good representation of the underlying population pdf.] In addition to having a probability distribution, a random variable also has a "cumulative distribution function" or cdf, denoted $F_x(x)$. A random variable's cdf is the function that measures the probability accumulated by the values of X up to and including x. In other words, $F_X(x) = P(X \le x)$.

In our discussion of EDA techniques in Lecture 1, in addition to introducing the empirical distribution for the observed values of a given random variable, we also presented several numerical descriptions of the observed values of the variable, including the sample mean, the sample median, the sample mode, and the sample variance. Each of these statistics corresponds to an analogous population parameter that can be calculated using the information in the pdf (or cdf) for the underlying random variable. The "population mean" of a random variable X is also known as its "expected value" and will be denoted by E(X) or μ; the population mean can be thought of as the balance point of the population pdf for X. The "population median" is not commonly used, but it refers to the smallest value of X for which $F_x(x)$ is greater than or equal to 0.5 (i.e., the value of X for which the aggregate or cumulative probability of all smaller values in R_X exceeds ½). The "population mode" for X can be described as the value of X that occurs most frequently. The "population variance" of X is denoted by var(X) or σ^2 and is defined as $\text{var}(X) = E[(X-\mu)^2]$, i.e. the expected value of the squared differences between the values of X and their expected value. [As a point of interest, $\text{var}(X) = E[(X-\mu)^2]$ can be rewritten, using probability theory, as $\text{var}(X) = E[X^2] - \mu^2$, where $E[X^2]$ is the expected value of the square of the r.v.; this latter formula is often easier to calculate.] As would be expected, the population standard deviation is merely the square root of the population variance. Lastly, population analogues also exist for the sample skew, sample kurtosis, sample minimum and maximum, and sample kurtosis statistics introduced in Lecture 1.

Before proceeding, we should note that, in general, we cannot see the underlying population, and, thus, we do not know what form the population pdf and cdf take or what the values of the various population parameters are. In most cases, then, we will have to infer these properties of the underlying random variable of interest from the corresponding sample properties calculated using our sample of observed values of the variable; this is where statistical inference, the subject of Lectures 3-8, comes into play.

(III.a.) Discrete Random Variables

Suppose X is a discrete random variable with range R_X, and let x denote any one possible value in R_X. The pdf for a discrete r.v. is more properly called its "probability mass function" or "frequency function"; this function measures the probability with which X takes the value x (i.e., $f_X(x) = P(X = x)$) for all values of x in R_X. The cdf for the random variable X is, as usual, defined as $F_X(x) = P(X \le x)$; further, because X is discrete, it can be obtained from the probability mass function by summing the probability values for all of the values of X that are smaller than or equal to x. For a discrete r.v., both its probability mass function and its cdf can be represented in tabular form; for the former distribution, the table would contain probabilities (or, identically, relative frequencies) and, for the latter distribution, the table would contain "cumulative probabilities."

For a discrete random variable X, its population mean, μ, is the weighted sum of all the values of x in R_X, where the weight for each x is $f_X(x)$. Similarly, the population variance of X is the weighted sum of $(x-\mu)^2$ over all the values of x in R_X, where the weight for each $(x-\mu)^2$ is $f_X(x)$. The population mode for X is simply the value in R_X that has the highest associated probability, and the population median for X is, as usual, the smallest value in R_X for which $F_x(x)$ is greater than or equal to 0.5.

As an example, consider the set of books in the long-loan collection at Sussex University Library. Suppose that the collection contains only books that were borrowed at least once during a given year, but no books that were borrowed more than 14 times in one year. Let X denote the number of times a book was borrowed in one year, so that R_X={1,2,…,14}; its distribution is:

X	1	2	3	4	5	6	7
freq.	9674	4351	2275	1250	663	355	154
f_X	0.5131	0.2308	0.1207	0.0663	0.0352	0.0188	0.0082
F_X	0.5131	0.7439	0.8645	0.9308	0.9660	0.9848	0.9930

X	8	9	10	11	12	13	14
freq.	72	37	14	6	2	0	1
f_X	0.0038	0.002	0.0007	0.0003	0.0001	0	0.00005
F_X	0.9968	0.9988	0.9995	0.9998	0.9999	0.9999	1.0000

First, we should note that this example is concerned with a certain population of books. Normally, we would only have a sample of books from this population, and, as a result, we would not know the population frequency and cumulative distribution functions as we do above. In this example, the probability mass function is the probability that a book chosen randomly from the collection had been borrowed x number of times in one year; the cumulative distribution function measures the probability of a book having being borrowed at most x times in one year. Graphs of these functions can be seen below.

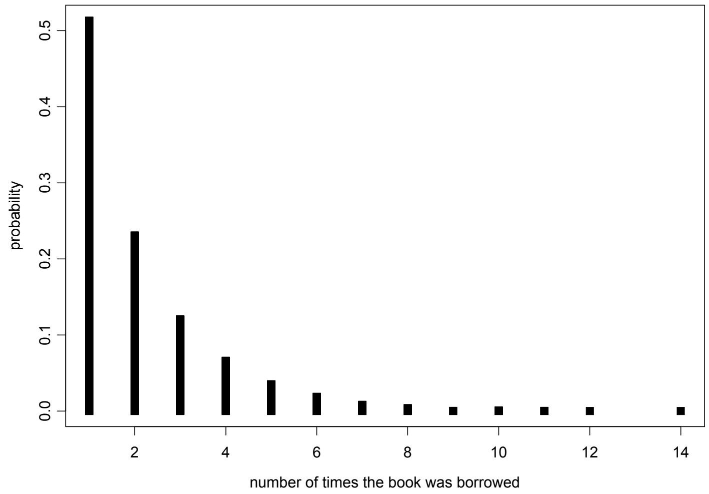

Figure 2a: *Population pdf for Sussex University books*

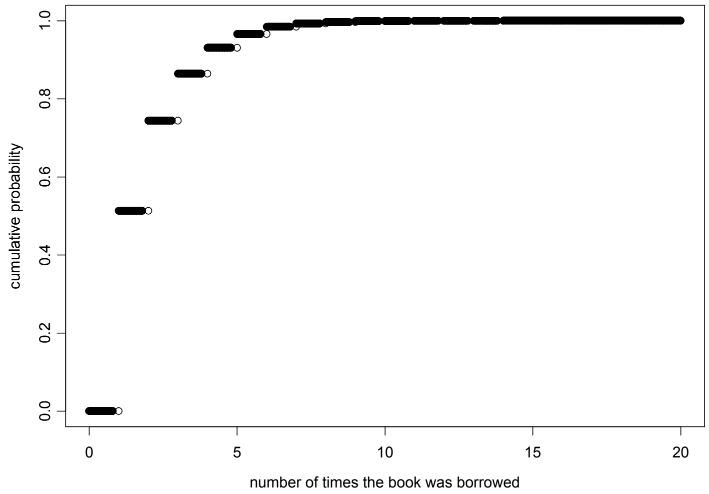

Figure 2b: *Population cdf for Sussex University books*

For this population of 18,850 books, we can calculate the (population) minimum, lower quartile, median, mean, upper quartile, maximum, standard deviation, skew, and kurtosis:

Min	Q1	Med	Mean	Q3	Max	SD	b1	b2
1	1	1	2.01	3	14	0.01	1.9	7.46

Note that the median number of times borrowed is 1 because 1 is the first value in the variable's range for which the cdf is greater than or equal to 0.50. In addition, the modal number of times borrowed is also 1 because 1 has the largest associated probability value (i.e., 0.5131) in the variable's pdf. However, the mean number of times borrowed is 2.01, which is calculated by weighting all the values in the variable's range by their associated probability values (i.e., $2.01 = 1 \cdot 0.5131 + 2 \cdot 0.2308 + ... + 13 \cdot 0 + 14 \cdot 0.0005$). The fact that the mean value is larger than the median value indicates a positively skewed distribution, as does the fact that $b_1 > 0$. Also, note that the standard deviation of times borrowed is 0.01; this value is obtained by taking the square root of the variance, which is calculated by summing $(1-2.01)^2 \cdot 0.5131 + (2-2.01)^2 \cdot 0.2308 + ... + (13-2.01)^2 \cdot 0 + (14-2.01)^2 \cdot 0.0005$. Lastly, note that the above quantities indicate a sharply peaked ($b_2 > 3$) distribution with at least 75% of the books being borrowed three or fewer times.

(III.b.) Continuous Random Variables

If *X* is a continuous random variable, then R_X consists of a real-valued continuum; by definition, then, *X* can take on any of an uncountably infinite number of values. Thus, the probability of observing exactly one particular value of *X* (say *x*) is 0, and the pdf for *X*, $f_X(x)$, does **not** measure probability. Rather, it indicates density, in terms of the frequencies of the measured values of *X* for units in the population, assuming that:

1) the population is infinite in size
2) the observations can be made with an infinite degree of precision

As an example of these assumptions, suppose that we are measuring fuel consumption (in miles per gallon) for a population of cars. The second assumption means that we could, in principle, distinguish between arbitrarily small differences in values of fuel consumption. The first assumption indicates that, even if we measured the values of *X* with infinite precision, we would not have gaps in the range of values due to a relatively small population size. Then, the pdf for a continuous variable can be thought of as the result of letting the number of classes and the number of observations in a true histogram tend to infinity. Figure 3 illustrates this interpretation of continuous variable pdfs by presenting histograms for samples of sizes 10, 100, 1,000 and 10,000 from an underlying population distribution that is normal (normal random variables are a commonly used example of continuous variables). The graphs show the true histograms for these samples of observations of the normal r.v., as well as curves indicating the theoretical population pdf for the normal random variable. Note that the 'optimal' number of classes used for the true histograms increases with the sample size.

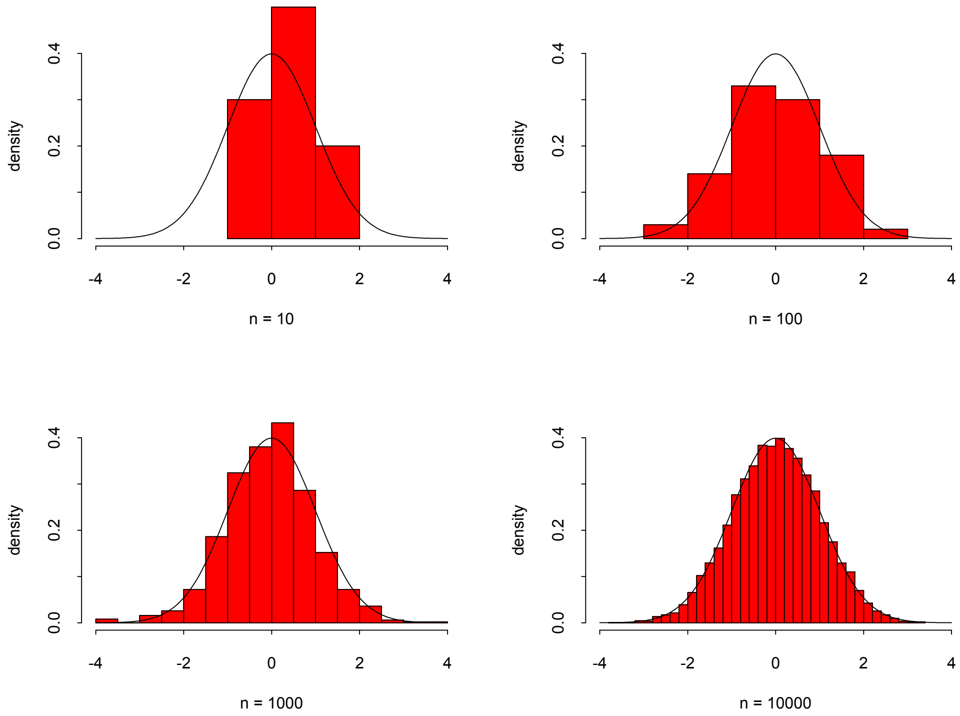

***Figure** 3: From true histograms to density functions*

We see that, as the number of observations in the sample and thus the number of bins in the histogram increases, the true histogram becomes closer and closer in shape to the underlying population pdf curve.

Looking at the above graphs, we note that, in addition to the fact that the area contained in all the bars of a given histogram is 1, the area under the population pdf curves is always 1. In fact, a continuous variable pdf is required (for mathematical reasons) to have an area of 1 lying between it and the R_X-axis (and to be greater than or equal to 0 for all values in R_X). The former requirement leads to a second way of viewing the pdf for a continuous r.v. In this view, the areas under the pdf curve for continuous variable X correspond to probabilities. Specifically, the area under the curve between the vertical lines $x=a$ and $x=b$ is the probability that X will take on a value falling between a and b. So, for a continuous variable pdf, instead of the height of the curve at a specific X-value corresponding to the probability that X takes on that specific value, (in fact, as stated above, this probability is 0), the area under the curve for a certain range of X-values corresponds to the probability that X takes on a value in that range.

In the previous section, we noted that the probability mass function and the cdf for a discrete r.v. can be represented in tabular form, as was done for the library books example. However, although the pdf and cdf for a continuous random variable can be presented in tabular form if R_X is divided into classes or bins of values, continuous r.v. pdfs and cdfs are more commonly represented as mathematical functions, which are much easier to describe and analyse than tables of frequencies. Occasionally, it is not possible to describe the cdf or pdf for a continuous r.v. using a mathematical function because no suitable one exists, in which case, they can be represented in pictorial/graphical form.

If the pdf for a given continuous r.v. can be represented as a mathematical function, calculating the population mean, mode, etc. for that r.v. is relatively straightforward. For instance, the population mode for a continuous r.v., *X*, is the value of *X* corresponding to the highest point of the pdf (along the y or density-axis), and the population median is the value of *X* for which exactly 50% of the area under the pdf lies to its left. The population mean of *X* can be thought of a weighted average of all the *X*-values in R_X, where each value is weighted by its corresponding density (i.e., the height of the pdf curve at that value). These population quantities, as well as other population descriptions such as the variance, can be calculated from the mathematical function for $f_X(x)$ using calculus. The calculation of these quantities will not be covered in this course, but the formulas for doing so are presented in any textbook of mathematical statistics, such as Rice (1995).

(IV.) A Note on Joint or Multivariate Population Distributions

Above, we have discussed "univariate" pdfs and cdfs for random variables of the discrete and continuous varieties. However, if we are considering two or more random variables of either variety at the same time, then the ideas of univariate pdfs and cdfs generalise to "multivariate" or "joint" pdfs and cdfs for those variables. The joint pdf for two or more variables can be used to calculate various population quantities of interest, such as the population mean and variance for each of the variables and also the population analogue of the sample covariance (correlation) for any two of the variables.

(V.) Specific Probability Models

Returning to the consideration of only one random variable at a time, suppose that the pdf for the random variable of interest for our experiment has a particular form that can be expressed as a mathematical function. As alluded to above, knowing a pdf (in explicit mathematical form) for a random variable means that we also know (after some calculus) all population properties of interest for that variable, such as its sample mean, variance, etc.; further, knowledge of these quantities allows us to make very precise statements about the underlying population of values. For these reasons, we can say that the pdf for a random variable fully describes (in a statistical sense) or fully "characterises" the corresponding underlying population.

There are certain commonly used and extensively studied "families" of discrete and continuous pdfs, alternatively known as "pdf families," "families of distribution functions," or "distributional families." For each of these families, its pdf can be written in an explicit mathematical form and is completely known up to a small number (usually one or two) of undetermined parameters. Here, we should note that, in general, a parameter is any numerical characteristic of the population. For instance, for a population of walking times, the minimum time, the maximum time, the mean time, and the proportion of walking times smaller than 10 minutes are all parameters. The choice of the particular parameters for a pdf family affects the shape and/or location of the resulting pdf and also determines various population quantities for the random variable described by the pdf. Lastly, most pdf families are particularly appropriate for and often used with certain types of random variables. As an example of a continuous pdf family, most people are probably familiar with the "normal" family. For random variables in this family, their pdfs are known up to the parameters μ (the population mean) and σ^2 (the population variance) and take the following mathematical form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$$

The parameters μ and σ^2 can be chosen: the choice of μ affects the location of pdf for the random variable and, obviously, the population mean for that r.v., and the choice of σ^2 affects the shape of the pdf (a larger value makes the pdf appear more spread out) and, obviously, the population variance for that r.v.. The normal family is commonly used for continuous random variables that have symmetrical distributions and can take on any value along the real number line (or at least more or less any value along the positive part of that line), such as height.

The idea of univariate pdf families for one variable can be generalised for instances in which we are interested in the joint distribution of two or more variables: for these instances, there are a number of commonly used and extensively studied joint (i.e., multivariate) pdf families. In each of these multivariate distributional families, the mathematical form of the pdf is again known up to a small number of parameters; however, the number of parameters to be determined is usually somewhat higher than the number of free parameters in univariate pdf families (e.g., 5 for the bivariate normal family vs. 2 for the univariate normal family).

In practice, we will often assume that the pdf for the random variable(s) of interest in an experiment is a member of a particular family. This family would be chosen to be appropriate for our particular variable(s) of interest given our knowledge of the variable(s). Once a specific pdf family has been assumed, we need only calculate the few unknown parameters for that pdf family in order to completely determine the pdf for our random variable(s); these parameters can be calculated using the data contained in our sample of values for the random variable(s). Since complete determination of the pdf for the random variable of interest fully characterises the underlying population of interest (i.e., completely determines all population quantities of interest), assuming a pdf family and then estimating the necessary parameters enables us to make precise statements about the population of interest in our experiment.

Certain pdf or distributional families are utilised particularly often because they are appropriate for types of random variables that occur frequently in applications. In the following section, we will examine three such univariate discrete pdf families (the

Bernoulli, binomial, and Poisson distributions) and two such univariate continuous pdf families (the normal and exponential distributions).

(VI.) Examples of Commonly Used Univariate PDF Families

(VI.a.) Discrete distributions

The Bernoulli Distribution

Our first example concerns the simplest situation in which we can have uncertainty. Consider an experiment that has only two possible outcomes; without losing generality, we can refer to them as Success (S) and Failure (F). The random variable, *X*, that is most commonly applied to these outcomes takes a value of 1 for S and a value of 0 for F; note that the range of this random variable is R_X={0,1}. Continuing, suppose that the probability of observing a success in any single repetition of the experiment is *p*, where 0<*p*<1. For instance, consider an experiment that refers to a particular disease that occurs in a fraction *p* of a population. Suppose that we randomly select only one individual. We do not know, a priori, whether or not this person has the disease; instead, we will have to wait for the outcome of the experiment. If we take Success to mean that an individual has the disease and Failure to mean that he/she doesn't have the disease, then the pdf of the random variable *X* that measures presence of the disease can be stated mathematically as

$$P(X=1)=p; \quad P(X=0)=1-p\,,$$

or, identically,

$$f_X(x)=P(X=x)=p^x(1-p)^{1-x} \text{ for } x=\{0,1\}\,.$$

First, note that the range of a Bernoulli random variable, *X*, is $R_X=\{0,1\}$. Second, note that the pdf, which is known as the "Bernoulli distribution," depends on only one parameter (*p*), and that this completely determines the pdf's mathematical formula. If *X* is a random variable following this model, we write *X~Bern(p)*, which is read as '*X* is distributed Bernoulli with parameter *p*.'

If *X~Bern(p)*, then it can be shown that $E(X)=p$ and that $\mathrm{var}(X)=p(1-p)$.

The Binomial Distribution

There is an immediate extension of the Bernoulli model to the case where the binary experiment is repeated, independently and identically, *n* times. If we now take the random variable *X* to be a count of the number of successes in the *n* repetitions, we can say that *X* has a "binomial distribution" with parameters *n* and *p*, which is denoted by *X~ B(n,p).* If *X~ B(n,p),* then the pdf for *X* can be stated mathematically as

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x = \{0,1,2,...,n\}.$$

First, note that the range of a binomial r.v., X, is $R_X = \{0,1,2,...,n\}$; in other words, between 0 and *n* successes may occur. Also, note that in the above formula, $\binom{n}{x} = \frac{n!}{(n-x)!\ x!}$ is the "binomial coefficient." [The number *x!* is the "factorial" of the integer *x* and is defined as $x! = x \cdot (x-1) \cdot (x-2) \cdots 2 \cdot 1$ for integers greater than zero and 0*!*=1 for zero.] The binomial coefficient is needed because *X* does not take account of the *order* in which the success/failure sequence occurs: as long as *x* successes occur, $p^x(1-p)^{n-x}$ has the same value regardless of what order the successes occurred in. Thus, $p^x(1-p)^{n-x}$ must be multiplied by the binomial coefficient, which gives the number of ways (orderings) in which *x* successes can occur in *n* experiments. For instance, suppose a coin is tossed four times and we are interested in the probability of seeing two heads (i.e, P(*x*=2)= $f_X(2)$). The two heads can occur in six possible ways (i.e., *HHTT*, *HTHT*, *HTTH*, *THHT*, *THTH*, *TTHH*), and, thus, the probability of seeing two heads in four tosses involves the binomial coefficient $\binom{4}{2}$, which has a value of 6.

Given that *n*, the number of repetitions of the experiment, is fixed, the only unknown parameter for this distribution is *p*; thus, the binomial distribution is a one parameter pdf family. Below, Figure 3 shows three different binomial distributions for *n*=10. A comparison of the spread of the three pdfs demonstrates the fact that the binomial distribution with *p*=0.5 has the greatest variance of all the possible binomial distributions for a given *n*.

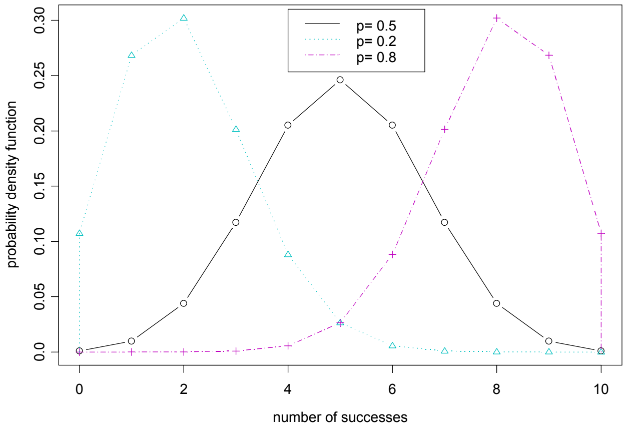

Figure 3: *Binomial pdfs with n=10 and p=0.5, 0.2, and 0.8*

It can be shown that that if *X~ B(n,p),* then the expected value or population mean of *X* is *np*, and the population variance is *np(1-p).* Using these properties, we can see that the expected values for the three distributions shown in Figure 3 are 5, 2, and 8, and that their variances are 2.5, 1.6, 1.6.

As an example for which the binomial distribution would be used, pretend that it is well known that the probability of a birth resulting in a girl is $95/200 = 0.475$ =*p*. Suppose that there are exactly *n*=50 births every day in a particular region; then, the number of girls, say *G*, is a binomial random variable, with range {0, 1, 2, …, 50}. The expected value of *G* is 23.8, its variance is 12.5, and its standard deviation is 3.5. [Note that this last number indicates the extent of *G*'s deviations from its expected value.] Thus, we would expect there to be 23.8 girl births (and thus 26.2 boy births) on a given day. However, there are, naturally, random fluctuations around these numbers every day. Given the value of the standard deviation for G, we would not be very surprised if one day the number of female births was 23 or 25; however, we would be surprised if this number were below 19 or above 28, for example.

The Poisson Distribution

The "Poisson distribution" is commonly used for random variables that count the number of occurrences of some type of rare event, where a large number of events could, theoretically, occur. Just as the binomial distribution represents a generalization of the Bernoulli distribution, the Poisson distribution can be derived from the binomial distribution. Specifically, a Poisson random variable can be thought of as a binomial random variable for which *n* is extremely large and *p* is extremely small, as long as we make the key assumption that the probability of success, *p*, remains the same for the large number of independent repetitions. More formally, let *X* be the random variable that counts the number of successes in those repetitions; further, denote the expected value of *X* as λ, where λ = *np* and is thus always greater than 0. If we let *n* go to infinity and *p* go to 0 in such a way that *np* remains equal to λ, then *X* will be a Poisson random variable, which can be denoted by $X \sim P(\lambda)$. Note that λ is often termed either the "rate" or the "mean" of the Poisson distribution.

The pdf of the Poisson distribution with parameter λ can be stated mathematically as

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad \text{x= \{0, 1, 2, ….\}.}$$

where the number *e*= 2.718282… is the base of the natural logarithms. As was true for the binomial distribution, one parameter (λ, in this case) is enough to completely describe the distribution. Also, note that the range for a Poisson random variable, *X*, is R_X ={0, 1, 2, ….}; in other words, any number of events between 0 and infinity could potentially occur.

The Poisson distribution has the interesting property that its population mean equals its population variance. Specifically, $E(X) = \text{var}(X) = \lambda$. This property means that for

Poisson distributions with higher expected values, the dispersion around the expected value will also be higher, as illustrated in Figure 4.

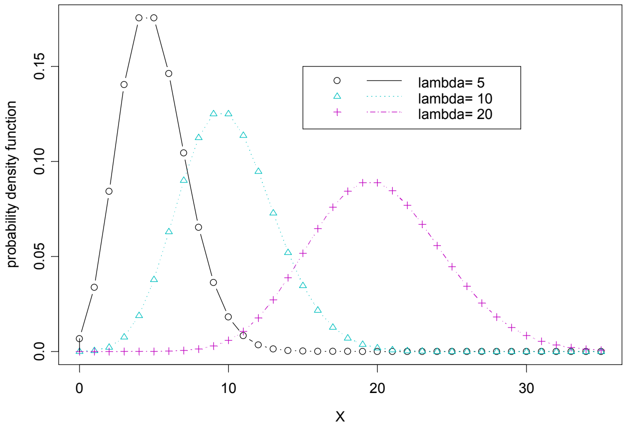

Figure 4: *Poisson pdfs with different rates*

As a classic example of the use of a Poisson distribution, consider the data recorded in 1898 by Von Bortkiewicz. This data consists of the number of fatalities that resulted from being kicked by a horse (a fairly rare event) for soldiers of 14 corps of Prussian cavalry over a period of 20 years (giving a total of 280 corps/year). The observed frequencies for the number of kicks can be seen in the second row of the following table:

Num. deaths/year	0	1	2	3	4
Observed frequency	144	91	32	11	2
Poisson prediction	139.0	97.3	34	8.0	1.4

The mean number of deaths/year is λ=0.7, which is also the variance of the distribution. In this case, we have a large number of independent repetitions of a binary event (each soldier being killed or not by a horsekick) with a small probability of 'success.' Therefore, the Poisson distribution should be a good model for this data, as is indeed shown by the closeness of the fitted Poisson model predictions in the third row of the above table to the observed frequencies in row 2. The fitted Poisson model predictions were calculated using $280 \cdot f_X(x)$, where λ=0.7 in $f_x(x)$, for x=0,1,2,3,4.

(VI.b.) Continuous Distributions

The Normal Distribution

As was mentioned before, the "normal" (or "Gaussian") distribution plays a key role in probability and statistics. Its pdf is a bell-shaped curve that is symmetric around the distribution's mean or expected value, μ. A second parameter, the distribution's variance (σ^2), determines how spread out this bell-shaped curve is. If X is a normally distributed random variable, and we denote its mean and variance by μ and σ², respectively, then we can write $X \sim N(\mu, \sigma^2)$. Recall that the square root of σ², denoted σ, is termed the standard deviation of the distribution. The following graph shows three normal pdfs with a common mean and different standard deviations.

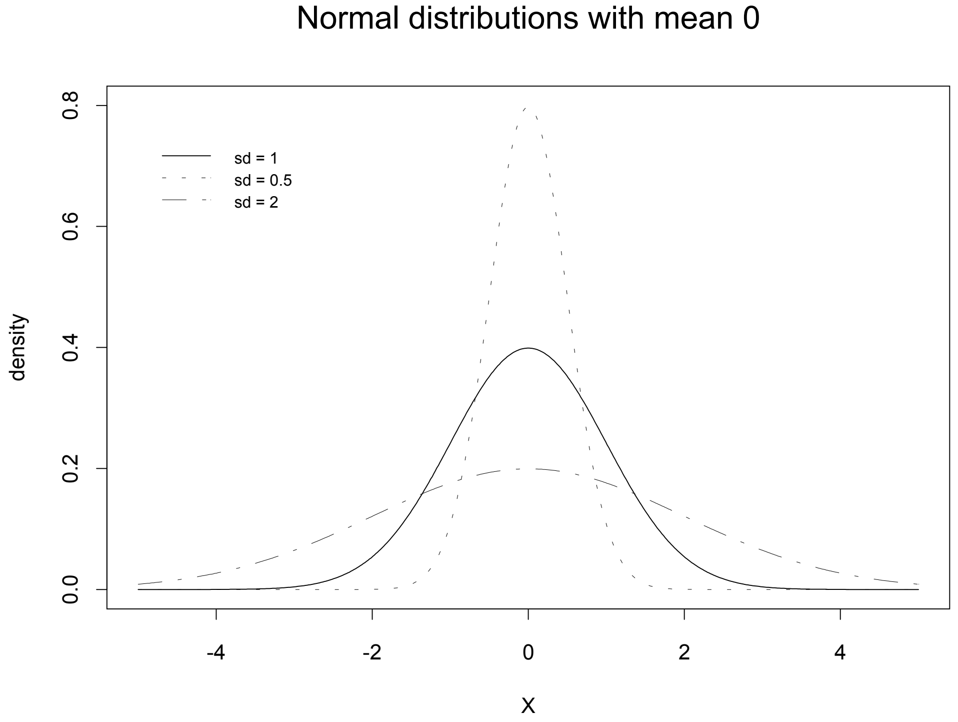

***Figure 5**: Normal pdfs with the same mean but different standard deviations*

We can see that, as σ increases (or, identically, as σ² increases), the pdf appears more spread out, reflecting the greater variance of the distribution.

The pdf curve for a random variable that is normally distributed with parameters μ and σ² can be stated mathematically as

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(x-\mu)^2} \text{ for } -\infty < x < \infty .$$

Note that the range of a normally distributed random variable, X, is $R_X = (-\infty, \infty)$; X can potentially take on any value on the real number line. In addition, we see that only two parameters (μ and σ²) are required to characterise any particular normal distribution; thus, the normal distribution is a two parameter family. Lastly, note that $E(X) = \mu$ and $\text{var}(X) = \sigma^2$, by definition.

An interesting fact about the normal distribution is that, for any normal distribution, 68.25%, 95.44% and 99.73% of its values fall within the intervals $[\mu \pm \sigma], [\mu \pm 2\sigma]$ and $[\mu \pm 3\sigma]$, respectively.

An important special case of the normal distribution is the "standard normal distribution," which has mean 0 and variance 1; random variables of this type are often denoted by the letter Z. For a variable that has a standard normal distribution, it is very easy to find out any of its cumulative probabilities: if one desires to know $\mathrm{P}[Z \le z]$, for any value z, where $Z \sim N(0,1)$, then the probability can be looked up in tables found in any introductory statistics text. However, in general, it is very difficult to calculate cumulative probabilities (i.e., $\mathrm{P}[X \le x]$) for a normal r.v. with general parameters μ and σ^2. Thus, in order to find specific cumulative probabilities for a normal r.v. with mean μ and variance σ^2, we will often transform it to a standard normal r.v. because the cumulative probabilities for Z are so readily available. More specifically, for $X \sim N(\mu, \sigma^2)$, if we use the transformation $Z = \dfrac{X - \mu}{\sigma}$, then Z will have a standard normal distribution. Note that the inverse of this transformation means that $X = Z\sigma + \mu$. As an illustration, if $X \sim N(10,4)$, and we want to calculate $P(9 \le X \le 12)$, we can use the following procedure:

$$\begin{aligned} P(9 \le X \le 12) &= P\left(\frac{9-10}{2} \le \frac{X-10}{2} \le \frac{12-10}{2}\right) \\ &= P(-1/2 \le Z \le 1) \\ &= P(Z \le 1) - P(Z \le -1/2) \\ &= 0.8413 - 0.3085 \\ &= 0.5328 \end{aligned}$$

where the values in the second to last line can be obtained from a cumulative probability table for the standard normal distribution. Figure 6 illustrates the calculation of the values on the second to last line:

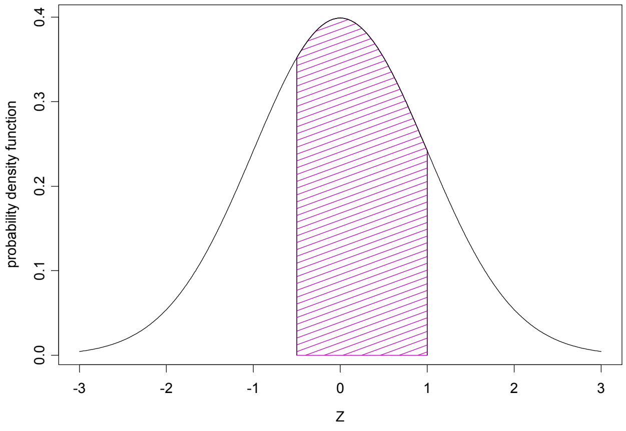

Figure 6: *P(-0.5 ≤Z≤1), with Z~N(0,1)*

The Exponential Distribution

Suppose now that instead of counting the number of occurrences of rare events over time, we measure the time between consecutive occurrences of events. The resulting random variable, *X*, that measures the time intervals between events is clearly continuous. Further, if a different random variable that counts the number of events that occur has a Poisson distribution with mean λ, then *X* will have an "exponential distribution" with parameter λ, which is denoted by $X \sim Exp(\lambda)$. This distribution is commonly used to model failure times in quality control and survival time in clinical studies.

The pdf of an exponential distribution with parameter λ can be stated mathematically as

$$f_X(x) = \lambda\, e^{-\lambda x} \text{ for } x > 0,$$

and its cdf is written as

$$F_X(x) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

Note that the range of an exponentially distributed random variable, *X*, is $R_X = (0, \infty)$; in other words, *X* can take on any positive value on the real number line. Note that the exponential distribution, like the Poisson and binomial distributions, has only one parameter, λ, which alone determines the shape of the pdf. The parameter λ is often termed the "rate" and should be defined in terms of number of events per unit of time.

Figure 7 shows the *pdf*s of three exponential random variables with different rates.

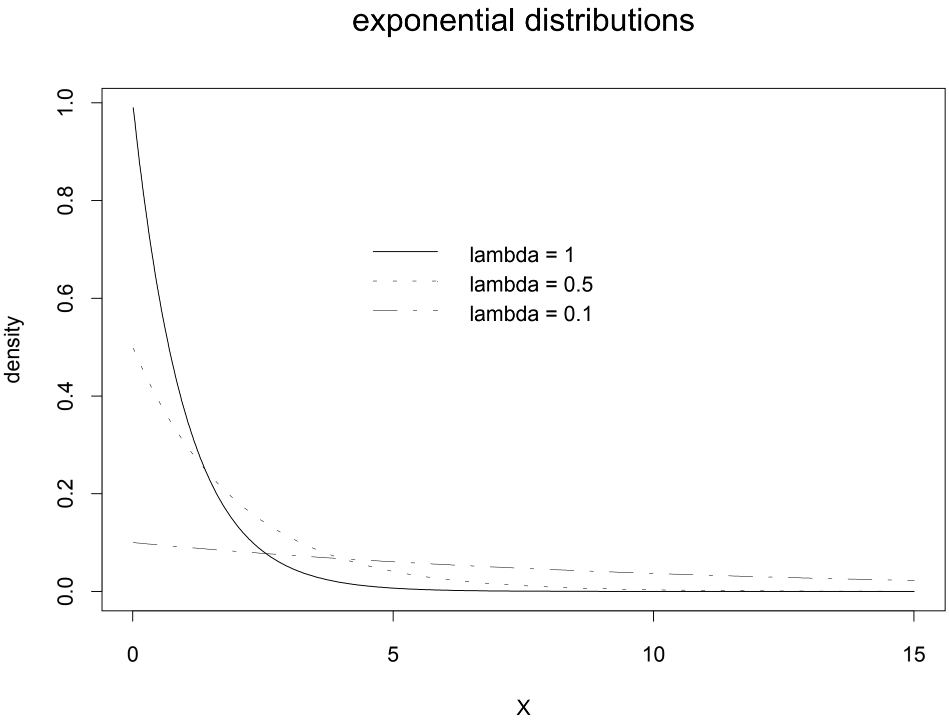

Figure 7: *Exponential distributions*

The expected value and variance of *X* are $1/\lambda$ and $1/\lambda^2$, respectively. The intuition behind these values is clear. If an event is rarer, then the time until its next occurrence should have a bigger expected value (and, perhaps slightly less intuitively, a bigger variance), which is the case for the formulas above since rarer events correspond to a smaller rates and both equations involve the reciprocal of the rate.

As an example of using the exponential distribution, suppose that it is known that a certain type of equipment works, without failure, for 1,200 days on average. In other words, $E[X]=1/\lambda=1{,}200$. We will denote the time until failure as *X* and assume that it is exponentially distributed. Thus, we can compute, for instance, the probability that the equipment would last for at least 1,000 days by using the exponential cdf formula with λ=1/1,200:

$$P(X>1000)=1-P(X\le 1000)=1-F_X(1000)=1-\left(1-e^{-1000/1200}\right)=0.4346$$

To calculate the probability of observing a failure between 1,500 and 1,600 days, we perform the following calculations:

$$P(1500\le X\le 1600)=F_X(1600)-F_X(1500)=(1-e^{-1600/1200})-(1-e^{-1500/1200})=0.0229$$

Note that if we had not assumed a particular pdf family for *X* (i.e., the exponential family), then we would not have been able to calculate probabilities like those just shown. In this example, we have also assumed that we know the value of the parameter, λ, that allows the distribution function to be specified completely. In practice, we would usually be able to identify, before seeing any data, which distribution function is adequate for modelling the random variable(s) of interest for the experiment; however, in practice, the distribution's parameters would have to be estimated from the sample data. Indeed, the main difference between probability and statistics is that, in the former, the parameter values are known, whereas, in the latter, estimation of parameters is one of our main concerns.

*MCB (*I-2000), *KNJ (*III-2001), *JIB (*III-2001)