CHAPTER 2

Statistics and the Law of Probability

OBJECTIVE

The purpose of chapter 2 is to review some of the most popular discrete and continuous probability models. The practical motivations of these models will be particularly emphasized.

CONTENTS

2.1 Overview

The exact law of probability associated with many natural phenomena that researchers study in practice is generally unknown due to complexity, or limited information, since observed data simply gives you a snapshot of the phenomenon at a given point in time. Moreover, the experiment that led to these observations may even have been contaminated by external factors beyond the experimenter's control, distorting the very nature of the subject under investigation. What statisticians have done overtime to get around this difficulty was to develop a large collection of probability models that were thoroughly investigated. You would then identify one specific model that appears to match your data reasonably well, and use that data to fully specify the model during a model-fitting process. The fully specified model will then be available for conducting a more sophisticated analysis of the variation underlying the variable of interest.

Model-fitting is a critical step for going from observed data towards exploring the larger universe that produced these observations. The purpose of this chapter is to review the most useful of these theoretical probability models that you will need in the later chapters of this book, devoted to statistical inference. The discussion on model-fitting strategies is delayed until chapter 3, devoted to point and interval estimation.

In the next few sections, I am going to consider the probability model to be a two-element set (X, f) where X is a random variable, and f a family of probability distributions of which the distribution of X is assumed to be a possible member. In section 2.2, the probability models are associated with discrete random variables. A random variable is discrete when its set of possible values is either finite or countable (i.e. all possible values can be explicitly listed). Section 2.3 will be devoted to continuous ran-

dom variables. These are random variables whose possible values lie in a continuum[1] such as an interval. Examples of continuous variables include the weight, the height.

2.2. Probability Models for Discrete Variables

The 5 discrete laws of probability that I am considering in this section are Bernoulli, Binomial, Hypergeometric, Poisson, and Multinomial. These are among the probability laws that I have found most useful in practice.

2.2.1 The Bernoulli Distribution - $\mathcal{B}er(p)$

Suppose that a researcher wants to study the incidence of a disease in a given region. The experiment will consist of selecting an individual and diagnosing the disease. Two outcomes are possible from such an experiment: the presence or the absence of the disease. One may consider a random variable X defined as follows:

$$X = \begin{cases} 1 & \textit{if a patient is diagnosed with the disease,} \\ 0 & \textit{otherwise.} \end{cases}$$

The probability law of X assigns a probability of occurrence to each of the 2 possible values 0 and 1 that X can take. The probability that the random variable X takes value 1 for example, is denoted by $P(X = 1)$ or by $p(1)$. Likewise, you will have $p(0)$ or $P(X = 0)$. The general form of the Bernoulli family of proba-

[1] One characteristic of a continuum is that one cannot list 2 consecutive numbers from a continuum without omitting more numbers in between

bility distributions is defined as follows :

$$p(x) = \begin{cases} \alpha & \text{if } x = 1, \\ 1-\alpha & \text{if } x = 0, \\ 0 & \text{otherwise}, \end{cases} \tag{2.1}$$

where α is a number greater than 0 and smaller than 1 ($0 < \alpha < 1$). It represents the parameter of the Bernoulli family of probability distributions, that must be estimated from observed data. The parameter α may represent the prevalence of a disease in a particular region for example. A value such as $\alpha = 0.05$ will define a specific member of the Bernoulli family of probability distributions. $p(x)$ is often called the *Probability mass function* (p.m.f.), or the *Probability distribution function* of X.

To conceptualize the notion of probability, it is convenient to see the law of probability as a mass system, where a total probability mass of 1 must be distributed across a number of mass points. In the case of a Bernoulli probability model, there are two mass points[2], which are 0 and 1, the portion of the unitary probability mass going to 1 is α, and the portion going to 0 is $1 - \alpha$.

The Swiss mathematician Jacob Bernoulli (1654-1705) was first to formally study the properties of this distribution. The expectation $E(X)$ and variance $V(X)$ of a Bernoulli variable X are given by:

$$E(X) = \alpha \text{ and } V(X) = \alpha(1-\alpha). \tag{2.2}$$

The Bernoulli probability model labeled as $\mathcal{B}er(p)$ has found an important application in the study of logistic regression (see chapter 11) used to predict the outcome of dichotomous variables.

[2]A mass point is simply a point in the sample space that is to receive a positive probability mass.

2.2.2 Binomial Distribution - $\mathcal{B}(n, p)$ -

A situation of practical interest often occurs when a Bernoulli trial is performed n times, and the experimenter observes the number of successes[3] obtained. Let X be a random variable representing the number of times the "Success" outcome is observed. A classical hypothetical example is where the same coin is tossed n times with heads representing success (S) and tails representing failure (F). X for this experiment will be the number of tails observed. The problem to be resolved is to determine the probability mass function associated with X.

The statistical model applicable to the experiment described above requires the use of $n+1$ mass points. Why ? Because there is a possibility to have 0, 1, or up to n successes. The sample space of X is given by:

$$\mathcal{A}(X) = \{0, 1, \cdots, n\}. \tag{2.3}$$

In the model to be constructed, we require the probability of success to be the same on each Bernoulli trial. A sufficient condition for this to be achieved is to require the $n = 20$ trials to be performed independently, and under the exact same conditions. The *independence* condition guarantees that the outcome of one trial does not affect that of the next one, while *homogeneity* of experimental conditions guarantees uniqueness of the success probability, denoted as p.

The p.m.f. of X denoted by $b(x; n, p)$ is the probability that

[3]Success in this context is achieved when the Bernoulli variable takes value 1, the value 0 representing failure.

X equals x, and is formulated as follows:

$$b(x;n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, 1, \cdots, n, \\ 0 & \text{otherwise.} \end{cases} \quad (2.4)$$

with n and p being the two parameters of the statistical model. To understand why the p.m.f of the Binomial model is formulated as shown in equation 2.4, note that you would observe x successes following n trials if you observe x successes and $n-x$ failures. This will occur with a probability of $p^x(1-p)^{n-x}$. However, the x successes may occur on the first x trials or on any other grouping of x trials. It can be established that the number of ways that x successes and $n-x$ failures can be observed is given by the number of combinations of x among n. That is,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

where $n!$ (read n factorial) is defined as $n! = 1 \times 2 \times 3 \times \cdots \times (n-1) \times n$. The symbol $\binom{n}{x}$ is read n choose x. To summarize,

A Binomial Model is characterized by a random variable X that represents the number of successes observed during an experiment, which conforms the following list of requirements :

- (a) The experimenter performs a predetermined number n of trials, resulting each to a success (S) or to failure (F).
- (b) All trials are identical (i.e. performed under the same conditions) so that the probability of success p stays the same from trial to trial.
- (c) The trials are independent in the sense that the outcome of one trial does not affect that of another trial.

The family of probability distributions corresponding to this model is the p.m.f. associated to X and given by equation 2.4

The Binomial probability is used later in this book in the statistical test of hypotheses regarding proportions based on small sample sizes.

Example 2.1

When a car comes to a dead-end intersection, it may either turn left or turn right. If the cars arrive at the intersection under similar traffic conditions, then one may consider that they will be turning left or right independently of one another. Suppose that it is known from past experience that about 60% of the cars coming at that intersection under current traffic conditions turn left. If the experimenter decides to observe the next 5 cars, he may apply the binomial model to the number of cars turning left. Figure 2.1 shows the 6 mass points that are part of the Binomial model, as well as an example of what their respective probability masses could be.

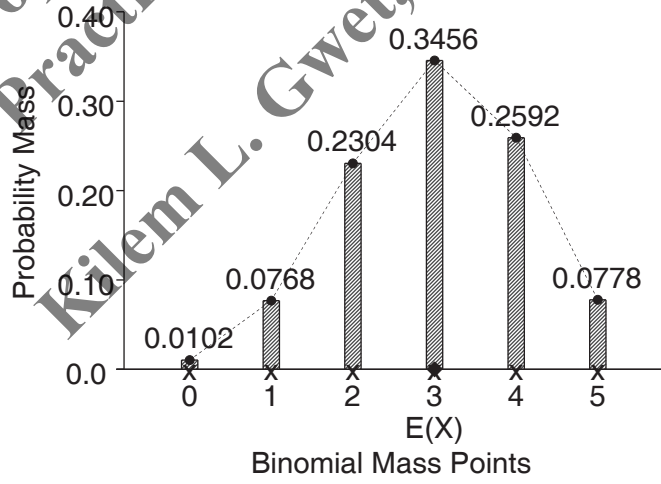

Figure 2.1. Graphical representation of Binomial statistical model

The unitary probability mass is distributed among the 6 mass points of the sample space, according to the Binomial law of probability $\mathcal{B}(5, 0.6)$ formulated in equation 2.4. The mass point 0 for example is assigned a probability mass of 0.0102 (i.e. $p(0) = 0.0102$), while mass point 4 is assigned a probability mass of 0.2592. All 6 probability masses sum to the total probability mass of 1.

Expectation and Variance

The mathematical expectation of X denoted by $E(X)$, which represents the position on the horizontal axis of the center of mass is calculated as the sum of the mass points weighted by their respective probability masses. This definition is not convenient for computational purposes. In fact, the center of mass is more conveniently calculated using the fact that X is the sum of n independent Bernoulli variables $X_1, X_2 \cdots X_n$. Because of independence, the center of mass of the Binomial variable equals the sum of the centers of mass of all n Bernoulli variables $E(X_i)$. That is,

$$E(X) = \sum_{i=1}^{n} E(X_i) = np, \tag{2.5}$$

since $E(X_i) = p$ for a Bernoulli X_i as seen before. Likewise, the variance of the Binomial distribution is the sum of the n Bernoulli variances $V(X_1), \cdots, V(X_n)$. That is,

$$V(X) = \sum_{i=1}^{n} V(X_i) = np(1-p). \tag{2.6}$$

The center of mass of the Binomial model of example 2.1 is given by $E(X) = 5 \times 0.6 = 3$, and its variance given by $V(X) = 5 \times 0.6 \times 0.4 = 1.2$.

The Cumulative Distribution Function

In addition to the variance and the expectation, other quantities related to the Binomial distribution are often of interest. One of them is the cumulative distribution function (c.d.f.) $B(x; n, p)$, which associates each mass point x with the cumulative probability masses of x and all mass points below x. That is, $B(x; n, p) = P(X \leq x)$ and is defined as,

$$B(x; n, p) = \sum_{t=0}^{x} b(t; n, p) \quad \text{for } x = 0, 1, \cdots, n. \tag{2.7}$$

In other words, the cumulative distribution function represents the probability mass of the entire collection of mass points defined by the event $\{X \leq x\}$.

The Software Solution

Excel 2010 has two useful functions for dealing with the Binomial probability model. The first function is `BINOM.DIST` for computing the probability mass (equation 2.4) as well as the cumulative probability (equation 2.7) of a given mass point x. The second function is `BINOM.INV` for computing the percentiles of the Binomial distribution.

- BINOM.DIST$(x, n, p,$ FALSE) will calculate the p.m.f $b(x; n, p)$ of equation 2.4.
- BINOM.DIST$(x, n, p,$ TRUE) will calculate the c.d.f $B(x; n, p)$ of equation 2.7.
- BINOM.INV(n, p, α) will produce the α^{th} percentile of the Binomial distribution $\mathcal{B}(n, p)$.

Excel 2007 offers a single function BINOMDIST that is similar to BINOM.DIST. Calc too has the same function BINOMDIST with the same functionality.

As for the R package, see section F.2.1 of Appendix F for a detailed description of what is offered. The R package offers more options than Excel and Calc. In addition to the p.m.f and the c.d.f, you can simulate the Binomial distribution with R by generating a data series where each mass point has a frequency of occurrence that is close to the probability mass predicted by equation 2.4.

In real life, trials are rarely performed independently under the exact same conditions. Therefore the conditions of the Binomial model will often be met only approximately as shown in example 2.2.

Example 2.2

A small city has 10,000 workers, of whom 4,500 have a high school (HS) diploma. A sample of 20 ($n = 20$) workers is selected *without replacement* from the list of all workers. Suppose that the i^{th} trial is a success (S) if the selected worker has a high school diploma, and is a failure (F) otherwise. Define the random variable X = *number of workers in the sample of 20 who have a high school diploma.*

The experiment is the random selection of 20 workers without replacement and the question is whether the Binomial model is applicable. This experiment has been performed 20 times. Thus, condition *(a)* is satisfied. However, the 20 trials are neither identical nor independent. If the probability for the very first worker selected to have a HS diploma is known to be $p = 4,500/10,000 = 0.45$, the probability for the second worker selected to have a high school diploma depends on the outcome of the first trial. If the first worker happens to have a high school diploma, then there will be 4,499 workers left with HS diploma out of 9,999 yet to be selected. Therefore the second worker will have HS diploma with probability $4,499/9,999$. This probabi-

lity becomes $4,500/9,999$ if the first selected worker does not have HS diploma. Because the probability of success changes from trial to trial, conditions (b) and (c) of the Binomial model are not satisfied. However, if the total number of workers is reasonably large (e.g. $10,000$ or more), then the success probability will always remain close to 0.45 and it will still be safe to assume the Binomial model.

Example 2.2 shows an experiment that is approximately described by the Binomial statistical model. An interesting question is whether there is a statistical model that can describe with precision the experiment of Example 2.2. The answer is yes, and that model is referred to as the *hypergeometric distribution*.

2.2.3 The Hypergeometric Distribution - $\mathcal{H}(n, N_s, N)$ -

The hypergeometric model is suitable for any problem that amounts to selecting successively n individuals without replacement from a population made up of N individuals that are of 2 types. The characteristic of interest X is the number of selected individuals of the type labeled as "success". This experiment is called the *hypergeometric process*. Each selection of an individual is a trial, which yields either a success or a failure and the problem is to evaluate the probability to obtain x successes after n trials.

Although the hypergeometric process is very similar to the binomial, there are two main differences:

- The maximum number of trials is limited to N. It is not the case in a binomial model where there is no maximum set for the number of trials that can be performed.
- In the hypergeometric model, the trials are performed under different conditions as the population from which individuals are selected shrinks by 1 after each individual is

selected.

What is the number of mass points in a hypergeometric model? The answer depends on how the number of trials compares to the number of "failures" in the population. For example, in a population of 10 individuals of whom 3 are labeled as F (for "failures") and 7 as S (for "success"), 5 trials will necessarily lead to a minimum of 2 S's and a maximum of 5 (although there are 7 S's in the population), excluding thereby 0, 1, 6, and 7 as possible mass points in the hypergeometric system. Let N_s and $N_f = N - N_s$ be the number of successes and failures respectively in the population of N individuals. All mass points in the system are between $\max(0, n - N_f)$ and $\min(n, N_s)$. An integer value k is a mass point (i.e. belongs to the sample space[4] $\mathcal{A}(n, N_s, N)$) of a Hypergeometric distribution $\mathcal{H}(n, N_s, N)$ if,

$$\max(0, n - N_f) \leq k \leq \min(n, N_s). \tag{2.8}$$

The probability mass function of the hypergeometric process is defined as follows:

$$h(k; n, N_s, N) = \begin{cases} \dfrac{\binom{N_s}{k}\binom{N - N_s}{n - k}}{\binom{N}{n}} & \text{if } k \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Excel, R, and Calc provide functions for computing this p.m.f. However, only Excel 2010, and R provide functions for computing the c.d.f. I will describe these functions later.

[4]For the sake of simplicity, I will occasionally use $\mathcal{A}$ instead of $\mathcal{A}(n, N_s, N)$

Example 2.3

Suppose that in a small company of 40 employees, only 5 have an annual salary over \$90,000.00. If 15 employees are selected without replacement, then the number X of employees in the sample with an annual salary over \$90,000.00 follows a hypergeometric model. The associated probability mass function is given by :

$$h(k;15,5,40) = \begin{cases} \frac{1}{658,008}\binom{5}{k}\binom{35}{15-k} & \text{if } 0 \leq k \leq 5, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Expectation and Variance

The expectation $E(X)$ and the variance $V(X)$ of a hypergeometric variable X having a probability mass function $h(k;n,N_s,N)$ are given by :

$$E(X) = np \text{ and } V(X) = \frac{N-n}{N-1} \cdot n \cdot p(1-p), \tag{2.11}$$

where the success probability p is defined as $p = N_s/N$.

The hypergeometric and the binomial distributions can both be seen as the sum of n Bernoulli variables. That is, $X = X_1 + \cdots + X_n$, where X_i is a Bernoulli variable with a success probability p. However, there are some differences and Table 2.1 shows a comparison between these two distributions with respect to some of their characteristics.

Table 2.1 : Comparison of hypergeometric and binomial distributions

Characteristic	Binomial	Hypergeometric
Success probability p	p *(unknown)*	$p = N_s/N$ *(Known)*
Number of Bernoulli trials	n	n
Maximum number of trials	*Unlimited*	N
Dependency of the trials	*Independent*	*Dependent*
Expectation $E(X)$	np	np
Variance $V(X)$	$np(1-p)$	$\left(\frac{N-n}{N-1}\right)np(1-p)$

It follows from Table 2.1 that the differences between the two distributions do not affect the center of mass $E(X)$ of the distributions. However, the variance of the hypergeometric distribution is different from that of the binomial. This is essentially due to the fact that the hypergeometric distribution is characterized by two related conditions, which are the dependency of the bernoulli trials and the existence of a maximum N on the number of trials that can be performed. The variance of the hypergeometric variable can be rewritten as follows :

$$V(X) = \left(\frac{N-n}{N-1}\right)(1-f)np(1-p), \text{ where } f = n/N.$$

If the population size N is reasonably large and the *sampling fraction* f reasonably small (i.e. smaller than 5%), then the variance of the hypergeometric variable gets very close to that of the binomial. Thus, when the parameters n, N_s, N of the hypergeometric distribution are very large, computing the probability mass for the hypergeometric can be tedious and one can use the binomial approximation to solve the problem.

The Software Solution

Excel 2007 and Calc 3.2.0 only allow you to compute the probability masses (see equation 2.9) at the mass points with their

HYPGEOMDIST function. You will call = HYPGEOMDIST(k, n, N_s, N) with Excel 2007, or = HYPGEOMDIST($k; n; Ns; N$) with Calc.

However, you will need Excel 2010 or R to compute cumulative probabilities. Excel 2010 offers the HYPGEOM.DIST function (see section E.9 of Appendix E for a detailed description), and R offers the qhyper function (see section F.2 of Appendix F for a detailed description). Only R will allow you to compute the percentiles of the Hypergeometric distribution.

I have discussed the binomial distribution in section 2.2.2. I also indicated in section 2.2.3 that when the number of trials has an upper limit that is not very large, it may be necessary to use the hypergeometric distribution rather than the binomial. Now, I want to know what would happen if in a binomial experiment, the number of trials increases indefinitely. The interest of this inquiry lies in the practical desire to know about the chance of occurrence of an event of a particular type overtime.

2.2.4 The Poisson Distribution - $\mathcal{P}(\lambda)$ -

In a book published in 1837 and entitled *Recherches sur la probabilité des jugements em matière criminelle et en matière civile* (i.e. Researches on the probability of criminal and civil verdicts), the French mathematician Siméon Denis Poisson (1781-1840) stated that if the number of trials n is very large and the success probability p very small, the binomial probability mass function can be approximated by :

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ for } \lambda > 0 \text{ and } k = 0, 1, \cdots \tag{2.12}$$

Although Poisson presented equation 2.12 (actually a variant of it) simply as an approximation to the binomial p.m.f., this ex-

pression turned out to fulfill itself the mathematical conditions of a probability mass function. The problem was that for a longtime no known real-life experiment could be associated with Poisson's equation, making it impossible to fully describe a Poisson statistical model. It took several decades before an English mathematician named William Sealy Gosset (1876-1937) found a clear example of a practical experiment related to yeast fermentation where the Poisson law was applicable. This happenned when he was working as a chemist in the Dublin brewery Arthur Guinness and Son.

Example 2.4

Yeasts are organisms that are cultivated in jars of fluid and used in the manufacturing of beer. The amount of yeast used in the mash affects the beer quality and must be known. In jars, yeast cells multiply and divide continuously (at a unknown speed that decreases with time) so that at a given point in time the concentration of yeast cells is difficult to determine. Since the speed with which yeast cells multiply is neither known nor constant, the concentration of cells in the jar at time t is a random variable. If its probability distribution is known, then yeast cells concentration can be predicted with better accuracy. Gosset examined the data available and concluded that the Poisson distribution was suitable.

William Sealy Gosset is better known in the statistical community under the pseudonym of Student that he used to publish his work secretly as publishing was a forbidden activity at Arthur Guinness and Son to protect manufacturing secrets.

There are a few fancy mathematical assumptions of little statistical interest that are necessary to justify the derivation of the Poisson p.m.f (see Hogg, Craig, and McKean (2004)). What should be retained about a typical Poisson process is that it takes place overtime and,

- The relative number of times it generates a single new event of interest (i.e. "success") during a unit time is constant.
- The relative number of times it generates 2 successes or more during a unit time is negligible.
- The number of times success is observed in a given time interval does not affect the number of times that success will be observed in another nonoverlapping time interval. This is a form of independence.

Other examples of Poisson processes include the number of defects on a manufactured article, the number of car accidents at a given intersection, or the number of insurance claims, within a certain time period.

Example 2.5

Let X denote the number of automobile insurance claims during a specified time interval Δt from t_0 to t_1 (i.e. $\Delta t = t_1 - t_0$). Suppose that X has a Poisson distribution with $\lambda = 5$. This indicates that during a time interval Δt, there will be 5 claims on average. The probability mass function associated with X is depicted in figure 2.2.

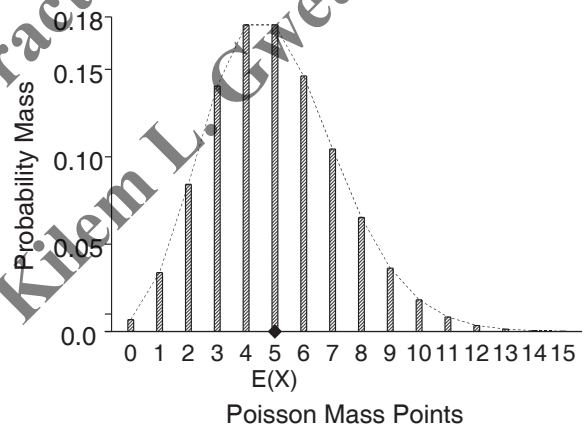

Figure 2.2. Graphical representation of a Poisson statistical model

The probability mass associated with having exactly two claims

during Δt is given by:

$$p(X = 2) = \frac{5^2}{2!}e^{-5} = 0.084.$$

The probability of having at least two insurance claims during time period Δt is

$$\begin{aligned} P(X \leq 2) &= \sum_{k=0}^{2} P(X = k), \\ &= \sum_{k=0}^{2} \frac{5^k}{k!}e^{-5} = \left(1 + 5 + \frac{25}{2}\right)e^{-5} = 0.125. \end{aligned}$$

The Poisson process in example 2.4 can be seen as an approximation of the binomial process where a trial is performed by verifying the filling of a insurance claim every second for instance. A success is observed if one finds a claim, otherwise it is seen as a failure. The number of trials in a time interval of 1hour is 3,600. Since the average number of claims is 5, one concludes that the probability of success (i.e. observing a claim) is about $p \approx 5/3600 = 0.0014$, which is very small.

The Poisson distribution can in general be used in practice to approximate the law of probability of a Binomial experiment with a large number of trials and a small success probability.

Expectation and Variance

If X follows a Poisson distribution $\mathcal{P}(\lambda)$, what is the center of mass $E(X)$ of the distribution, and what is the expected distance from the mass points to the center of mass measured by the variance $V(X)$? These two quantities are given by,

$$E(X) = \lambda, \ \textit{and}\ V(X) = \lambda. \tag{2.13}$$

This result is derived using basic results from introductory calculus.

You may also want to know that the Poisson distribution has found an important application in the development of log-linear regression also known as Poisson regression. Poisson regression is briefly introduced in chapter 11.

Software Solution

Excel (2007 & 2010), Calc 3.2.0, as well as R provide functions for computing the probability mass and the cumulative probability of the Poisson distribution. Only R however, will allow you calculate the percentiles of Poisson.

Excel 2007 and Calc 3.2.0 offer the POISSON function. In Excel 2007, POISSON(k,λ,FALSE), and POISSON(k,λ,TRUE) produce the probability mass and the cumulative probability respectively at mass point k. In Calc 3.2.0, POISSON(k;λ;0), and POISSON(k;λ;1) produce the probability mass and the cumulative probability respectively at mass point k.

For Excel 2010, see section E.9 of Appendix E, and section F.2 of Appendix F for R, to obtain more information on the functions available.

2.2.5 The Multinomial Distribution

In section 2.2.2, I presented the Binomial distribution characterized by the replication of n independent Bernoulli trials. This can be seen as the replication of n independent and identical trials, each of which has only 2 possible outcomes, 0 and 1. One can then define 2 random variables X_0 (*the number of failures*) and X_1 (*the number of successes*) representing the number of times the experiment produces 0 and 1 respectively. X_1 will be the Binomial variable as defined in section 2.2.2. You can also rephrase this by saying the pair (X_0, X_1) follows the Binomial distribution. This is not often done because $P(X_0 = x_0, X_1 = x_1)$

is identical to $P(X_1 = x_1)$ as long as $x_0 + x_1 = n$, which justifies using the simpler expression $P(X_1 = x_1)$.

The Multinomial probability model represents the replication of n independent and identical trials, each of which has r possible outcomes labeled as[5] $1, 2, \cdots, r$. You can then define r random variables $X_1, X_2, \cdots, X_r$ where X_2 for example represents the number of times the n trials produced outcome "2". In the Multinomial model, the r variables can only take values $x_1, x_2, \cdots, x_r$ that satisfy the following 2 conditions:

$$\begin{cases} x_i = 0, 2, \cdots, n, \text{ for } i = 1, 2 \cdots, r \\ x_1 + x_2 + \cdots + x_r = n. \end{cases} \tag{2.14}$$

Assume that on each of the n trials, any particular outcome i can be observed with probability p_i. The probability mass function of the set of variables $(X_1, X_2, \cdots, X_r)$ is defined as follows:

$$p(x_1, \cdots, x_n) = \begin{cases} \dfrac{n!}{x_1! x_2! \cdots x_r!} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}, & \text{if (2.14) met,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

The Multinomial probability model is used in hypothesis testing, when comparing equality of several population proportions when the number of observations per population is very small. I will address this issue in the chapter that discusses the chi-square test.

[5]The labeling of outcomes is subjective. Whether it ranges from 0 to $r-1$ or from 1 to r is irrelevant as long as you use one choice consistently.

2.3. Probability Models for Continuous Variables

This section deals with the quantification of chance for measurement variables. I indicated in the previous section that the calculus of probability for categorical variables is based on a system of mass where a probability mass is associated with each value that the categorical variable can take. Since the system has a total probability mass of 1, computing the probability of an event essentially comes down to summing the probability mass of all the mass points that can be associated with the particular event. For measurement variables, the calculus of probability is based on an entirely different computation model. The model is now based on a surface of a certain form[6] with a surface area of unity. Computing a probability comes down to assigning a portion of the baseline surface to an event and evaluating its surface area, which will necessarily be smaller than 1, a condition that all probability values must satisfy.

I want to illustrate this approach for quantifying probabilities by applying it to a practical problem. Consider the rectangle shown in figure 2.3. The experiment consists of selecting randomly, a point P from the rectangle's surface, and measuring the distance from P to the AC side. The random variable I am interested in is $X = Distance\ to\ Side\ AC$, which may take any value from 0 to 2. However, only those points P situated in the shaded area on the surface will lead to a distance that is below 0.5yd. In other words by choosing point P randomly, the probability $P(X \leq 0.5yd)$ that the distance from the selected location to the AC side is less that $0.5yd$, corresponds to the ratio of the shaded area to the total area of the rectangle. The area of the shaded surface of figure 2.3 equals $0.5 \times 1.5 = 0.75$ while the total rectangle

[6]The particular form of the surface will be specific to each probability model.

surface area is $2 \times 1.5 = 3$. Thus, $P(X \leq 0.5yd) = 0.75/3 = 0.25$. More generally, for any value d, the probability $P(X \leq d)$ can be expressed as follows :

$$P(X \leq d) = \begin{cases} 0, \ if\, d < 0, \\ d/2, for\ 0 \leq d \leq 2, \\ 1, if\, d \geq 2 \end{cases} \quad (2.16)$$

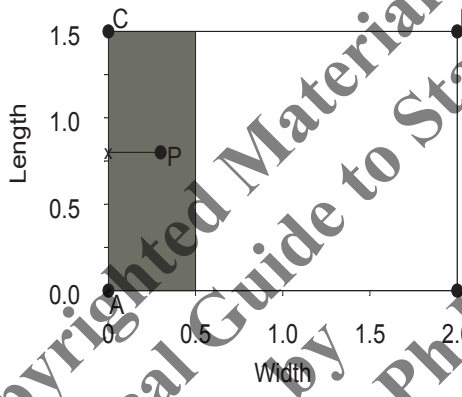

Figure 2.3. Rectangle's Surface

Although equation 2.16 describes the cumulative distribution function of X, deriving this function in practice will often be difficult for many complex real-life problems. It may even be impossible to derive the cumulative distribution function as a close expression. What is really needed, is a standard graphical display of the probability distribution of X that would be a more refined version of the histogram often used to depict the empirical probability of a measurement variable. Such a graph will help match your data with an existing and known theoretical probability model using histograms.

Figure 2.4 represents a rectangle whose surface area is 1 (i.e. $2 \times 0.5 = 1$). This surface area represents the overall probability mass of 1, which is uniformly distributed across all values

0 through 2 that the random variable X can take[7]. Now, the probability $P(X \leq 0.5)$ is the dashed-shaded surface area (the probability mass associated with the particular event $X \leq 0.5$), which is $0.5 \times 0.5 = 0.25$ (no need to divide by the overall probability mass, since it is 1).

The narrow solid-shaded strip on the other hand, represents the probability that X falls between $d = 0.6$ and $d + 0.1 = 0.7$, and is equal to $0.1 \times 0.5 = 0.05$. That is, if the random distance X increases by $\Delta = 0.1$ from an initial value d, the event probability mass will change by 0.05 for a change rate of $0.05/0.1 = 0.5$. You can then say that the density of probability or the *Probability Density* over the interval 0.6 to 0.7 is 0.5. Actually no matter how small the change Δ is, the probability density within the interval 0.6 to $0.6 + \Delta$ remains identical to 0.5. I can then conclude that the probability density at point 0.6 is 0.05. The function that links each point to the associated density is called the *Probability density function (pdf)*. In this case, the pdf is defined by,

$$f(x) = \begin{cases} 0.5, & \textit{if } 0 \leq x \leq 2, \\ 0, & \textit{otherwise.} \end{cases} \tag{2.17}$$

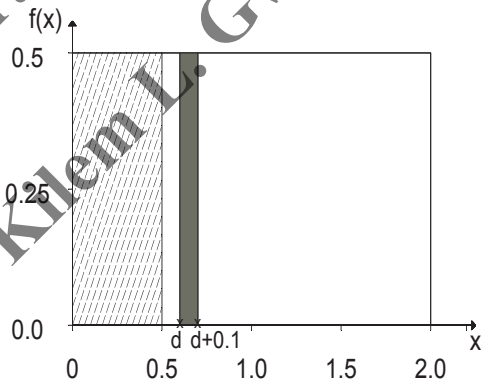

Figure 2.4. Probability Distribution of X

[7]The interval 0 to 2 is called the *Support* the probability distribution

The rectangle shown in Figure 2.4, is nothing else than the graphical representation of the probability density function. When this function is constant over the support of the probability distribution as is the case in equation 2.17, you may conclude that the probability mass has a uniform density across the support. The associated random variable X is said to belong to the family of the uniform distributions. The pdf completely specifies the probability distribution of a random variable, and most theoretical probability distributions discussed later in this chapter will be described by their pdf only.

Consider another experiment where a point P must be selected randomly from the triangle shown in Figure 2.5, and the distance from the selected point to the AC side of the triangle measured. X = *Distance from P to side AC* is the random variable I am interested in.

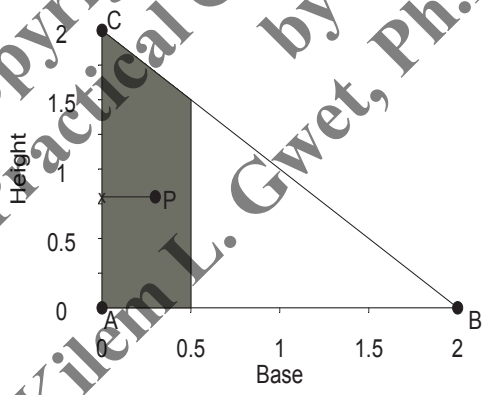

Figure 2.5. Triangle Surface

It follows from Figure 2.5 that the probability $P(X \leq 0.5)$ is calculated as the ration of the solid-shaded area to the overall triangle surface area. That is, $P(X \leq 0.5) = (2 \times 0.5 - 0.5^2/2)/2 = (1 - 0.125)/2 = 0.875$. More generally, for any value d, the pro-

bability $P(X \leq d)$ is calculated as follows:

$$P(X \leq d) = \begin{cases} d - d^2/4, & \text{if } 0 \leq d \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{2.18}$$

The standard representation of the law of probability of X in the form of the distribution of the overall unitary probability mass is shown in Figure 2.6. The surface area of this triangle is 1, and you may notice that this probability mass is distributed differently from that of Figure 2.4. The dashed-shaded area represents the probability for X to remain below 0.5. As for the solid-shaded strip, it represents the change in probability mass as a result of a small change in X from d to $d + 0.1$. This shaded area can be evaluated at $0.1(2 - d) - 0.1^2/2$. That is from d to $d + 0.1$, the probability mass changes at the rate of $0.1 \times (2 - d - 0.1/2)/0.1 = 2 - d - 0.1/2$, which represents the probability mass density in the interval d to $d + 0.1$.

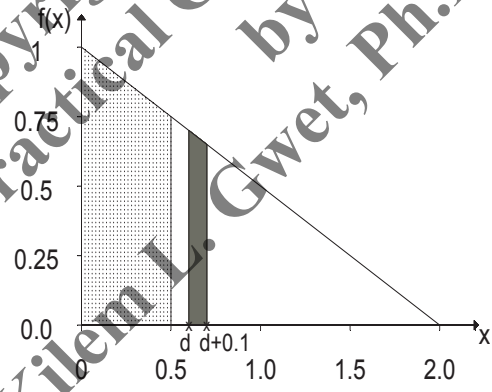

Figure 2.6. Graphical Representation of the Probability Distribution of Equation 2.20

Consequently, the area around 0 is one of high density of probability mass. Such areas are generally of great interest in practice as they are much more likely than any other area, to be the location where any given value of X will be found.

More generally, the probability mass density in an interval d to $d+\Delta$ for an arbitrary value Δ is given by $2-d-\Delta/2$. If you set $\Delta = 0$, you will get the probability mass density at point d, which is given by $2-d$. Consequently, to each value d you will associate the density function $f(d)$ defined as,

$$f(d) = \begin{cases} 2-d, & \textit{if } 0 \leq d \leq 2, \\ 0, & \textit{otherwise.} \end{cases} \tag{2.19}$$

In case you have been wondering how graphs such as that of Figure 2.6 are constructed, you may have already realized that they are obtained by graphing the probability density function. Histograms with very narrow intervals can give you a reasonably good approximation of the pdf associated with your variable of interest. Since the probability distributions in subsequent sections are characterized with their probability density function, it is interesting to know when a given function can be called a probability density function.

Definition. 2.1.

If X is a measurement variable then its probability mass distribution is fully specified by a probability density function f that satisfies the following conditions :

- $f(x) \geq 0$,
- The area under the curve representing f is 1.

Readers with some background in Calculus know that for any two numbers a and b, the probability that X takes on a value in the interval $[a, b]$ equals the integral of the p.d.f f over that interval. But the area under the curve and within the interval $[a, b]$ can generally be well calculated with a series of approximations using an Excel spreadsheet if necessary.

2.3.1 Center of Mass and Variance of a Measurement Variable

In section 2.2, I introduced the location of the center of mass as well as the variance (or the dispersion parameter) of a discrete random variable. The random variable was then seen as a system of indivisible and enumerable mass points, each of which having a predetermined probability mass. I found out that the location of the center of mass is determined by taking the sum of the locations of each mass point, weighted by their respective probability mass. The variance on the other hand was determined by calculating the differences from the mass points to their center of mass and by taking the sum of these squared differences weighted by the probability masses

When the random variable is continuous, its support is a continuum and can no longer be seen as an enumerable set of indivisible mass points. Instead, it is the overall unitary probability mass that is distributed across the support according to a certain law to be determined. An interesting question that arises is how to determine the location of the center of mass and its variance. I will describe the well-known method known as "the method of limits" to resolve this problem.

To fix ideas, consider a continuous random variable X whose probability density function f is given by :

$$f(x) = \begin{cases} (-3x^2 + 4x + 4)/8 & \text{if } 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

The graph of this density function is the smooth curve shown in figure 2.7(a). Since I already know from section 2.2 how to handle categorical variables, I am going to "categorize" the continuous variable X temporarily, so that I can use existing results derived in section 2.2. Let h be any number satisfying the condition $0 <$

$h < 1$ and let n be the smallest integer greater than $2/h$. The categorized version of X denoted by X' (say X prime) is defined as follows:

- X' can take any of the n values in the set $\{x_1, \cdots, x_n\}$ where $x_i = (i-1)h$.
- The probability mass point $p(x_i)$ of x_i is the surface area of the rectangle of length $f(x)$ and width h. The probability mass function associated with X' is defined as,

$$p(x_i) = \frac{D_{ih}}{D_h}, \text{ where } \begin{cases} D_{ih} = hf(x_i), \\ D_h = \sum_{j=1}^{n} D_{jh} \end{cases} \tag{2.21}$$

As an example, let us assume that $h = 0.25$. This creates the categorical variable X' that takes one of the $n = 8$ values in the set $\{0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$. The probability mass function calculated according to equation 2.21 and given in Table 2.2, is depicted in Figure 2.7(a) by the vertical bars. The mass points are identified on Figure 2.7(a) by the small ticks of the x-axis. The area of each vertical bar in Figure 2.7(a) represents the probability mass $p(x_i)$ of the associated points x_i as shown in Table 2.2.

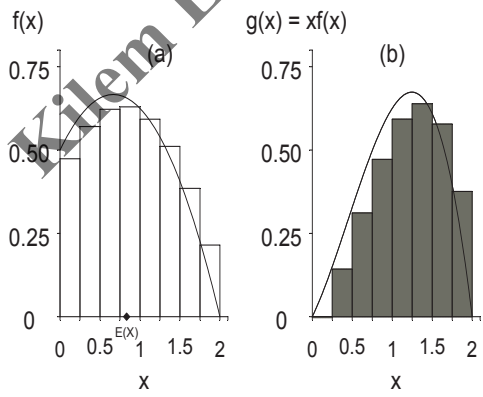

Figure 2.7. Graphical representation of the probability density function of equation 2.20

The bar area is obtained as the product $h \times l_i$ of its width h (the same for all bars) by its height l_i that varies for each mass points i and that is given by:

$$l_i = \frac{f(x_i)}{D_h}, \tag{2.22}$$

Table 2.2 : Probability distribution of the discretized random variable X'

x_i	l_i	$p(x_i)$
0	0.4741	0.1185
0.25	0.5704	0.1426
0.50	0.6222	0.1556
0.75	0.6296	0.1574
1.00	0.5926	0.1481
1.25	0.5111	0.1278
1.50	0.3852	0.0963
1.75	0.2148	0.0537
Total	4	1

Using the experience gained in section 2.2 with categorical variables, I can evaluate the center of mass of X' as follows:

$$\begin{aligned} E(X') &= \sum_{i=1}^{n} x_i p(x_i), \\ &= \sum_{i=1}^{n} h\Big[x_i f(x_i)/D_h\Big]. \end{aligned}$$

Note that $E(X')$ represents the shaded area in Figure 2.7(b), since $x_i f(x_i)/D_h$ is the height of i^{th} vertical bar.

If I gradually decrease the width h of the vertical bars towards 0, I will begin the inverse process of moving from the discrete variable X' back to the original continuous variable X. During that process, the number of vertical bars under the curve will grow as their width decreases (see Figure 2.8(a) and 2.8(b)). The areas under these curves will be covered with more accuracy. In Figure 2.8(a), the area under the curve will be the overall probability mass of 1, and in Figure 2.8(b), the area under the curve will be the center of mass, also called the expectation of X. It is denoted by $E(X)$, and reported in Figure 2.8(a). In the jargon of Calculus, one would say that the expected value of X is the integral of the function $xf(x)$ over the support of X (i.e. the interval 0 to 2), and denoted as follows:

$$\mu_{\text{X}} = E(X) = \int_0^2 xf(x)dx. \tag{2.23}$$

The variance of a measurement variable X is defined as the expectation of the variable $(X - \mu_X)^2$, and is a measure of how far you would expect a value taken by X to be with respect to its center of mass. It can be calculated using the method of limits used to evaluate the center of mass. Again, in the jargon of calculus, this variance is said to be the integral of the function $(x - \mu_{\text{X}})f(x)$ over the support of the measurement variable X, which is the interval 0 to 2. That is,

$$\sigma_{\text{X}}^2 = V(X) = \int_0^2 (x - \mu_{\text{X}})^2 f(x)dx. \tag{2.24}$$

The *Standard Deviation* is defined as the square root of the variance.

I like to say a word about the "method of limits". The first step is to determine not the exact magnitude of the quantity of

interest (such as the center of mass or the variance), but rather a rough approximation of it using simple arithmetic and algebra based upon enumerable objects. However, it is not just one approximation that should be made, but a whole series of them, the next one providing a more accurate result than the previous. The exact result is obtained by examining closely the series of approximation to detect what the limiting value looks like. That limited value will be a fixed constant that matches the exact value sought.

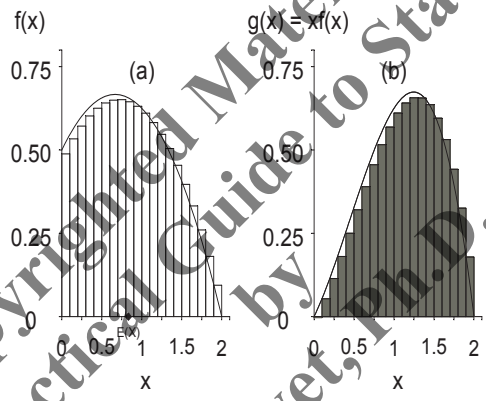

Figure 2.8: More accurate graphical representation of the probability density function of equation 2.20

There are several useful probability density functions that have been discovered during the course of the development of statistical science. Their usefulness is determined by the wide range of practical (random) phenomena they describe. In the next few sections, I will review the most popular of these density functions, and will discuss the conditions of their applicability.

During a statistical investigation, the researcher is often required to postulate a family of theoretical probability laws and to use collected data to determine the specific member of the family that can be used for statistical inference. Therefore, knowing the

probability laws that have been successfully used in practice is essential for a practitioner.

2.3.2 Uniform Distribution

One of the simplest continuous probability distribution is the *uniform distribution*. This distribution is characterized by its probability mass density being constant across the distribution support. Let X be a measurement variable that can take any value within an interval (a, b) with $a < b$. Strictly speaking, by saying that X can take any value within the interval (a, b), what I mean is that *there is a real chance that a value of X may be observed in any subinterval included in (a, b) following a random experiment*. The probability distribution of X is said to be *uniform*, if the associated density function f is defined as follows :

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

Equation 2.25 depends upon the 2 parameters a and b and therefore defines not one uniform probability distribution, instead it defines a family $\mathcal{U}(a, b)$ of uniform probability distributions. Each pair of values for (a, b) will specify one probability distribution.

Example 2.6

Suppose that my daily commute to work by train is expected to take one hour. However, because of a variety of circumstances the time that I actually spend in the train varies (randomly) from 45 minutes to 75 minutes. If X represents the commute time, then the set of possible values of X is the interval $(45, 75)$. If I believe that any subinterval included in $(45, 75)$ has the same chance to include the actual commute time of the next

train, then I may say that the probability distribution of X is uniform and its probability density function given by:

$$f(t) = \begin{cases} \frac{1}{30} & \text{if } 45 < t < 75, \\ 0 & \text{otherwise.} \end{cases}$$

In Example 2.5, I treated commute time as a continuous variable. Do I have to ? The answer is no, I do not. As discussed previously, continuity is a mathematical concept with no formal concrete representation. Commute time can be looked at as a discrete set of the 31 integers between 45 and 75, representing the number of minutes spent in the train. If this discretized version of commute time is uniformly distributed then the probability of observing any of the 31 values would be 1/31, which is close to the p.d.f shown in Example 2.5. Commute time may also be seen (like in example 2.5) as an entity not divided in parts (although divisible), that is a continuum allowing for fractional parts of one minute. The most important thing to retain here is that our model of continuity can sometimes provide a good approximation of a phenomenon that is discrete in nature. In Example 2.5, both discrete and continuous approaches can be used and will yield similar answers. However, there are situations where a discrete reality is difficult to describe and can usefully be approximated by a continuous model.

Going from a discrete structure to a continuous one has an important practical consequence. By taking commute T time as a continuous variable, the probability that it takes a specific value in the interval $(45, 75)$ is 0. This may a priori sound strange. The fact of the matter is that, if you believe there should be a "real chance" for the commute time to be exactly 52 minutes, then you have a discrete view of the problem that is incompatible with the continuity abstraction. A discrete model may then be more appropriate.

A continuum such as $(45, 75)$ is not made up of clearly distinguishable points such as 52. By allowing for unlimited accuracy in the determination of time, the concept of continuum takes away any meaning from the concept of exactness. If the observed commute time is $t_0 = 52$ minutes when time is seen as a continuous variable, then one could well assume that the actual time is $t_0 = 52.0000000001$ and not 52. One known thing is that the observed time lies between 51.5 and 52.5. Only intervals can be observed with a certain accuracy when the phenomenon under investigation is being explored with a continuous model in mind.

Cumulative Distribution, Expectation and Variance

- The cumulative distribution function (i.e. the probability $P(X \leq x)$ for an arbitrary x) of the uniform law of probability $U(a, b)$ is defined as follows:

$$F(x) = \begin{cases} 0, & if\ x \leq a, \\ (x-a)/(b-a), & if\ a \leq x \leq b, \\ 1, & if\ x \geq b \end{cases} \tag{2.26}$$

- The expected value $E(X)$ and the variance $V(X)$ of the uniform distribution are given by:

$$E(X) = (a+b)/2, \ \ and\ V(X) = (b-a)^2/12. \tag{2.27}$$

The Software Solution

The Uniform distribution is a very simple one to manipulate, even manually. This may be the reason why neither Excel nor Calc offer any command for manipulating this distribution. The R package however, offers some functions for the uniform distribution.

- `dunif(x, min=`a`, max=`b`)`. This function calculates the probability density of the uniform distribution $\mathcal{U}(a, b)$ at point x. The default values for the parameters a and b are 0 and 1 respectively.
- `punif(q, min=`a`, max=`b`, lower.tail = TRUE)`. This function calculates the cumulative probability distribution function of the uniform distribution $\mathcal{U}(a, b)$. If `lower.tail = TRUE` then the function calculates $P(U \leq q)$, and if `lower.tail = FALSE` then the function computes $P(U > q)$, which equals $1 - P(U \leq q)$.
- `qunif(p, min=`a`, max=`b`, lower.tail = TRUE)`. This function calculates the p^{th} quantile of the uniform distribution $\mathcal{U}(a, b)$. It represents the point x_p that satisfies the condition $P(U \leq x_p) = p$.
- `runif(n, min=`a`, max=`b`)`. This function generates a random number following the Uniform distribution $\mathcal{U}(a, b)$.

2.3.3 Normal Distribution/Bell-Shaped Curve

There is one probability law that proved useful in providing good approximations to a wide variety of probabilities involving averages of large collections of numbers. In fact, that probability law approximates very well the probability distribution of the mean of any large set of numbers, regardless of what probability law governs the initial raw data. The fact that the process that generates your data is unknown does not preclude you from having a reasonably good knowledge about the probability distribution of the average of these numbers. And as the collection of observed numbers grows the probability distribution of the mean tends to get closer and closer to a well-known limiting distribution that was understandably called the *Normal distribution*. The behavior of individual observations may be un-

predictable, but their average tends to have a "normal" behavior as the number of observations grows.

The Normal distribution has a long and glorious history. The French mathematician Abraham de Moivre (1667-1754) suggested a useful approximation of the Binomial distribution by the normal distribution. It is even thought that perhaps the Swiss mathematician Daniel Bernoulli (1700-1782) may have come accross the normal distribution earlier. Nevertheless, the normal distribution is often referred to as the Gaussian distribution as the german mathematician Carl Friedrich Gauss (1777-1855) was once believed to be the first man to formally discussed this distribution. The normal probability distribution is by far the most popular of all probability distributions. Its popularity is not simply justified by its ability to describe the large-sample probability distribution of the mean, but also stems from the fact that it simplifies the mathematics of probability calculus considerably. Once random variables are assumed to follow the normal distribution, the study of their statistical properties becomes much more tractable, although this simplification is not as important in this computer age as it once was.

The convergence of the mean's probability distribution to the normal probability distribution is a very powerful result that is usually referred to as the *Central Limit Theorem* (CLT). The CLT will be further examined later in this section.

Normal Distribution: Definition and Properties

Figure 2.9 depicts the probability mass function (represented by the vertical bars) of the Binomial distribution $\mathcal{B}(20, 0.25)$ with a total of n trials and a success probability $p = 0.25$. On the same graph is a bell-shaped curve that represents the continuous function that the French Mathematician Abraham de Moivre (1667-1754) used to approximate Binomial probabilities. This function turned out to be the probability density function of what was

later known as the normal distribution.

Today, the normal distribution is a well-defined and polished probability law that is usually introduced as a family of distributions, whose members are determined by a location parameter μ and a dispersion parameter σ^2.

Definition. 2.2.

Let X be a measurement variable that takes values in an interval (a, b) $(a < b)$. X is said to have a *normal probability distribution* with a location parameter μ and a dispersion parameter σ^2 $(\sigma > 0)$ if its probability density function is given by:

$$f(x; \mu, \sigma) = \frac{g(x; \mu, \sigma)}{I(a, b)} \text{ where } g(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

and $I(a, b)$ is a normalizing factor to ensure a mass probability of 1 over the interval (a, b).

In many statistical textbooks, measurement variables are typically assumed to take all possible values. Consequently, the interval (a, b) in the above definition is often replaced with $(-\infty, \infty)$, which should be read as "from negative infinity to infinity", and is a pure mathematical representation of all numbers. The value of the normalizing factor[8] $I(a, b)$ then becomes 1. Although such a presentation offers some mathematical elegance, it has the disadvantage of introducing the concept of infinity even when it is unnecessary, its value being essentially theoretical.

[8]The normalizing factor has only one purpose, which is to ensure that the surface area under the density curve, and limited to the interval (a, b) will be 1.

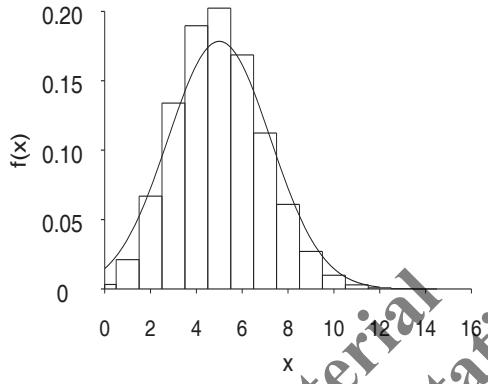

Figure 2.9. P.M.F. of a Binomial $\mathcal{B}(20, 0.25)$ with a normal approximation curve

Figure 2.10 shows the effect of the dispersion parameter σ on the form of the normal curve. This graph contains four normal curves $\mathcal{N}(\mu, \sigma^2)$ with the same location parameter $\mu = 12$ and a dispersion parameter σ that varies form 1 to 4. As it appears in Figure 2.10, the location parameter represents the center of symmetry of the normal distribution, while the dispersion parameter determines the narrowness of the bell under the curve.

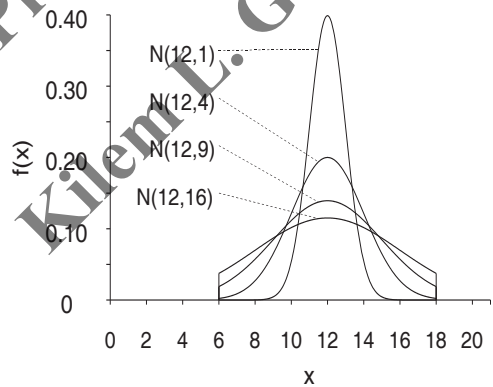

Figure 2.10. Graphs of normal density functions $\mathcal{N}(12, \sigma^2)$ for $\sigma = 1, 2, 3, 4$

The normal probability is used so often in statistical science that the practitioner is expected to be able to manipulate it with ease. This goal is achieved by mastering some of its basic and most important properties that will now be presented. The first thing to be noticed is that the density function of the normal distribution (see definition 2.2) is a complicated function. To evaluate cumulative probabilities or probability densities at various mass points, a natural solution could be to prepare numerical tables where various probabilities of the normal distribution are readily obtainable. But there is still a problem. The family of normal distributions is so broad (each value of μ or σ yields a different normal distribution) that the number of tables needed to cover them all will be excessive. Fortunately, the normal distribution has a nice property that helped resolve this problem.

Result 2.1.

Let us denote by Z a random variable that follows the normal distribution $\mathcal{N}(0,1)$ with a 0 location parameter and a dispersion parameter of 1. Let μ be an arbitrary number and σ a positive number. Then the random variable X defined by:

$$X = \sigma Z + \mu$$

follows the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with location and dispersion parameters given by μ and σ^2 respectively.

Why is this proposition important ? Suppose I like to compute the probability $P(X \leq x)$ that a random random variable X following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ take a value that is less than x. This probability can be expressed as,

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right), \tag{2.28}$$

where $Z = (X - \mu)/\sigma$ follows the normal distribution $\mathcal{N}(0,1)$.

Therefore, if we can tabulate the probabilities $P(Z \leq z)$ for various values of z, they can be used along with equation 2.28 to obtain the probabilities related to any normally-distributed random variable X. The normal distribution $\mathcal{N}(0, 1)$ is commonly referred to as the *Standard Normal Distribution* and plays a pivotal role in the family of normal distributions.

Cumulative Distribution, Expectation and Variance

The cumulative distribution function (often denoted by $F(x)$) of a random variable X is defined as the probability $P(X \leq x)$ that X takes a value smaller than or equal x. It is a widely-accepted convention to denote the c.d.f of the standard normal distribution as $\Phi(z)$ for any real number z. Equation 2.28 essentially stipulates that the c.d.f. F of any variable X that following the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is related to the c.d.f of the standard normal distribution by the following relation:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right). \tag{2.29}$$

Similarly, because we have that $P(\alpha \leq X \leq \beta) = F(\beta) - F(\alpha)$, the following relation is valid :

$$F(x) = \Phi\left(\frac{\beta - \mu}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu}{\sigma}\right). \tag{2.30}$$

The Software Solution

Although many statistics textbooks offer statistical tables that help compute various probabilities related to the normal distribution, I found it more convenient and more accurate to compute these probabilities using Excel, Calc, or R. The equations 2.28, 2.29, and 2.30 have gained importance overtime because they allow you to use the probabilities associated with the

standard Normal distribution in order to compute the probabilities of other non-standard normal distributions. Although it does not hurt to know about these equations, they are in my opinion only relevant to young mathematical statisticians who may need to study the math behind the normal distribution. For practitioners, these equations have become downright irrelevant in the computer age.

- The Excel 2010 function NORM.DIST($x, \mu, \sigma, cumulative$) applies to a random variable X that follows the Normal probability distribution $\mathcal{N}(\mu, \sigma^2)$. If cumulative = TRUE then NORM.DIST is the Normal cumulative distribution function, which calculates the cumulative probability $P(X \leq x)$. For cumulative=FALSE, it represents the Normal probability density function, which calculates the density of probability at x.

→ If X follows the Normal distribution $\mathcal{N}(12, 9)$ (i.e. $\sigma^2 = 9$) for example, then you may compute some probabilities using Excel 2010 as follows:

$$P(X \leq 13) = \text{NORM.DIST(13,12,3,TRUE)} = 0.63,$$
$$P(X > 13) = 1 - P(X \leq 13),$$
$$= \text{1-NORM.DIST(13,12,3,TRUE)} = 0.37,$$
$$P(10 < X \leq 13) = P(X \leq 13) - P(X \leq 10),$$
$$= \text{NORM.DIST(13,12,3,TRUE)}$$
$$\text{-NORM.DIST(10,12,3,TRUE)}$$

- Appendix C contains tables where you may find R, Excel 2007, or Calc functions that are equivalent to this Excel 2010 function.
- You may look at section E.9 of Appendix E for more Excel 2010 functions related to the Normal distribution.

Expected Value of the Normal Distribution

I have discussed in section 2.3.1 how the expected value and the variance of a measurement variable with probability density function f can be calculated. For the Normal probability distribution, these two parameters have already been calculated and are given in the following result:

Result 2.2.

If X follows the normal distribution $\mathcal{N}(\mu, \sigma^2)$ then its expectation $E(X)$ and variance $V(X)$ are given by :

$$E(X) = \mu \text{ and } V(X) = \sigma^2. \tag{2.31}$$

2.3.4 The Central Limit Theorem

If there was one single most important result in statistical science, it would probably be the *Central Limit Theorem* (CLT). It is because of the CLT that the normal distribution plays a central role in statistical inference today. I briefly mentioned in the beginning of the current section what the Central Limit Theorem was about. It states that the average of a large collection of observations is a random variable whose law of probability is well approximated by the Normal distribution, even if the probability distribution of original variable is non-normal.

More formally, the CLT is stated as follows:

Result 2.3.

Let $X_1, \cdots, X_n$ be n (random) observations of a random variable X with expectation $\mu = E(X)$ and variance $\sigma^2 = V(X)$. When the number n of observations is *sufficiently* large, the probability distribution of the average $\overline{X}$ of the X_i's, is *approximately* normal $\mathcal{N}(\mu, \sigma^2/n)$. Similarly, the sum $S = X_1 + \cdots + X_n$ of the X_i's is *approximately* normal $\mathcal{N}(n\mu, n\sigma^2)$.

The sample mean that is based on a fixed number of observations will follow a certain unknown law of probability. But this law of probability will change as the number of observations at your disposal increases. However, this law of probability will not change indefinitely. After the body of observations grows to a certain level, the law of probability associated with the sample mean will take its final shape, its *limit* form, which is the Normal distribution. In this sense, Result 2.3 is a limit theorem. And it is because of its central role in statistical inference, it is referred to as the central limit theorem.

It follows from the CLT that if you want to calculate the probability that the average is smaller than a quantity x, you may accomplish that as follows:

$$P(\overline{X} \leq x) = \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right). \tag{2.32}$$

where Φ is the cumulative distribution function of the standard Normal distribution. Equation 2.32 can be used as long as the two parameters μ and σ are known. In chapter 3, I will discuss some practical methods for approximating these parameters based on the observations if necessary.

Example 2.7

The human body temperature varies from one individual to another, and its distribution among individuals is generally unknown. Suppose you know from past experience that such a temperature can be expected to be around $\mu = 98.20^oF$ and has a standard deviation $\sigma = 0.62$. Your problem is to determine how often the average body temperature falls between the two values 98.15^oF and 98.6^oF.

With no new data and using the CLT, the researcher can determine the probability $P(98.15 \leq \overline{X}_{100} \leq 98.6)$ that the average of 100 randomly collected body temperatures be in the specified interval. You know from the CLT that $\overline{X}_{100}$ follows approximately the normal distribution $\mathcal{N}(98.20, \sigma^2_{\overline{X}})$, where $\sigma_{\overline{X}} = 0.62/\sqrt{100} = 0.062$. Thus,

$$P\big(98.15 \leq \overline{X}_{100} \leq 98.60\big) = P\big(\overline{X}_{100} \leq 98.6\big) - P\big(\overline{X}_{100} \leq 98.15\big),$$
$$= 1 - 0.209991 = 0.79$$

You may compute $P\big(\overline{X}_{100} \leq 98.15\big)$ for example, using Excel 2010 as `=NORM.DIST(98.15,98.2,0.062,TRUE)`.

Example 2.6 shows how the CLT can be used to approximate certain probabilities without having to conduct an experiment. The CLT can be used to approximate probabilities whenever the sum or average of variables is being used. I discussed in section 2.2.2 how the Binomial variable $\mathcal{B}(n, p)$ can be seen as a sum of n Bernoulli variables. Thus, the CLT can be used to approximate probabilities related to the Binomial distribution provided n is large and p is neither too close to 0 nor too close to 1.

Although the CLT is a very powerful result, its routine use still poses two problems. The practitioner should normally be able to answer the following questions before deciding to apply the CLT with confidence.

(a) The CLT requires the number n of observations to be large. *How large should it be ?*

(b) The CLT provides an approximation of the law of probability of the mean. *How good is that approximation ?*

A rule of thumb often used in practice is to achieve a sample size of at least 30 before using the CLT. In reality, $n = 30$ will be unnecessarily high for some applications and dramatically low for others. The optimal n depends on the (usually unknown) probability distribution of the original data. Although the CLT is appealing for letting us ignore the distribution of our data, its performance still depends on it. Very skewed original data will require a bigger sample than non-skewed data. Although averaging normalizes the original probability distribution, the more distorted the original distribution, the harder and the longer the normalization.

It should clearly be understood that the normal distribution does not describe a concrete fact that can be observed from experience. It is an abstract concept that precisely represents what all averages $\overline{X}$ (all concrete variables and their associated actual probability distributions confounded) have in common. Although the CLT guarantees a wide applicability of the normal distribution, it does not really provide an accurate description of any particular situation. Therefore, the CLT should be expected to provide at best a partial answer to any real-life problem, a start from which further investigation can be conducted on the basis of our knowledge about the problem under investigation.

2.3.5 The Exponential distribution

In section 2.2.4, I indicated that the Poisson distribution was useful for modeling the number of occurrences of an event within a predetermined time period of length t. In such

an experiment, the time period length t is fixed and it is the number X of occurrences of the event within that period that is random. I indicated that in this case X follows a Poisson distribution $\mathcal{P}(\lambda t)$ where λ is the expected number of occurrences of the event in one time unit.

In this section, I am considering the reverse problem where the number of observed events of interest is fixed at 1. However, the time period length T necessary to observe that first event is random. Consider for example the number of patients calling a health insurance company call center to request a specific medical service. The time length T from the opening of the call center to the reception of the first call will vary from day to day. I am interested in the probability distribution of T. Since I do not want to impose a limitation on the precision with which the length of time T can be measured, I will take it to be a continuous variable (you as practitioner will want to verify that this assumption does not distort the nature of the problem under study).

The probability that T is smaller than a specified time t is given by:

$$F(t) = P(T \leq t) = 1 - P(T > t).$$

Remember that all I want is to observe the first event and to record the elapsed time T until that occurrence. For that elapsed time to be greater than t, there should be no event observed until time t. That is, the number of events X observed until time t, which follows the Poisson distribution $P(\lambda t)$, must be equal to 0. We have that,

$$F(t) = 1 - P(X = 0) = 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = 1 - e^{-\lambda t},$$

which corresponds to the *cumulative distribution function (C.D.F.)* of T. This leads to the following probability density function:

$$f_\lambda(t) = \lambda e^{-\lambda t}.$$

Definition. 2.3.

A random variable T has a probability distribution in the family of *exponential distributions* $\mathcal{E}(\lambda)$ if it has a probability density function, which is of the following form:

$$f_\lambda(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

where λ is a positive parameter to be determined from each particular application.

Figure 2.12 shows the graph of the exponential density function when the λ parameter takes the values 0.5, 1.5 and 2.5.

The exponential distribution was derived as the distribution of the time length until the occurrence of the first event. One may also be interested in the distribution of the time length T_k until the occurrence of the k^{th} event. This leads to a generalization of the exponential distribution discussed in the next sub-section.

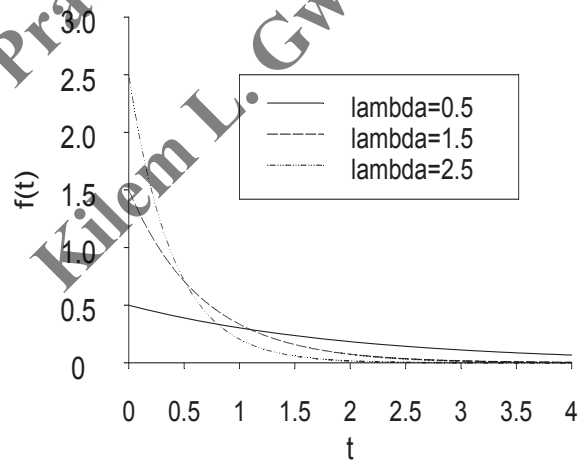

Figure 2.12. Exponential density functions for $\lambda = 0.5,\ 1.5,\ 2.5$

Cumulative Distribution, Expectation, and Variance

The cumulative distribution function of the Exponential probability distribution is defined as follows:

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

If a random variable X follows the exponential distribution $\mathcal{E}(\lambda)$, then its expected value $E(X)$, and variance $V(X)$ are given by:

$$E(X) = 1/\lambda \text{ and } V(X) = 1/\lambda^2. \tag{2.34}$$

The Software Solution

- Both versions of MS Excel 2010 & 2007 offer functions for calculating the probability density $f(t)$ as well as the cumulative probability $F(t)$. These calculations are done as follows:

  Excel 2007 $f(t)$ =EXPONDIST(t,λ,FALSE),
  $F(t)$ =EXPONDIST(t,λ,TRUE).

  Excel 2010 $f(t)$ =EXPON.DIST(t,λ,FALSE),
  $F(t)$ =EXPON.DIST(t,λ,TRUE).

- OpenOffice Calc 3.2.0 also offers the function EXPONDIST for calculating the probability density $f(t)$ and the cumulative probability $F(t)$ at a given point t. These quantities are calculated as follows:

  $f(t)$ =EXPONDIST(t;λ;0),
  $F(t)$ =EXPONDIST(t;λ;1),

- The R Package offers the dexp and pexp functions for calculating the probability density $f(t)$ and the cumulative probability $F(t)$ respectively at a given point t. These quantities are calculated as follows:

$$f(t) = \boxed{\mathsf{dexp}(t,\ \mathsf{rate} = \lambda)},$$
$$F(t) = \boxed{\mathsf{pexp}(t,\ \mathsf{rate} = \lambda}.$$

You may want to look at section F.2 of Appendix F for more R functions related to the exponential distribution

2.3.6 The Gamma Distribution

Suppose you want to obtain the probability distribution of the time length T_k until the occurrence of the k^{th} event. Starting with the smaller values of k and using the same approach I used to derive the p.d.f. of the exponential distribution, you can see that the probability density function of T_k is given by:

$$f_\lambda(t|k) = \begin{cases} \dfrac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.35}$$

Equation 2.35 is the probability density of a generalized exponential distribution where the number of occurrences of an event is an arbitrary integer k.

Let me consider an even more general situation where a homeowner spends a time duration T_α to mow an area of α yard square of his lawn. In this particular case, the phenomenon being observed during the time interval T_α is no longer discrete. It is continuous (the surface area of the lawn). Therefore, equation 2.35 cannot be used to describe the probability density of T_α. This problem is resolved by replacing k with α and $(k-1)!$ in the denominator with the surface area under the curve $g(t) = \lambda^\alpha t^{\alpha-1} e^{-\lambda t}$

for $t \geq 0$. This surface area is generally denoted by $\Gamma(\alpha)$ and referred to as the *Gamma Function*. This leads to the following definition:

Definition. 2.4.

A random variable T_α has a *gamma probability distribution* $\mathcal{G}(\alpha, \lambda)$ if its probability density function is given by:

$$f_\lambda(t|\alpha) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The Software Solution

- Both versions of MS Excel 2010 & 2007 offer functions for calculating the probability density $f(t)$ as well as the cumulative probability $F(t)$ for the Gamma distribution. These calculations are done as follows:

Excel 2007

$f(t)$ =GAMMADIST(x_0,α,$1/\lambda$,FALSE),

$F(t)$ =GAMMADIST(t,α,$1/\lambda$,TRUE).

Excel 2010

$f(t)$ =GAMMA.DIST(t,α,$1/\lambda$,FALSE),

$F(t)$ =GAMMA.DIST(t,α,$1/\lambda$,TRUE).

- OpenOffice.Calc 3.2.0 also offers the function GAMMADIST for calculating the probability density $f(t)$ and the cumulative probability $F(t)$ at a given point t. These quantities are calculated as follows:

$f(t)$ =GAMMADIST(t;α;$1/\lambda$;0),

$F(t)$ =GAMMADIST(t;α;$1/\lambda$;1).

▶ The R Package offers the `dgamma` and `pgamma` functions for calculating the probability density $f(t)$ and the cumulative probability $F(t)$ respectively at a given point t. These quantities are calculated as follows:

$$f(t) = \texttt{dgamma}(t,\ \alpha,\ \texttt{rate} = \lambda)\,,$$
$$F(t) = \texttt{pgamma}(t,\ \alpha,\ \texttt{rate} = \lambda)\,.$$

You may want to look at section F.2 of Appendix F for more R functions related to the gamma distribution

2.3.7 The Chi-Square Distribution

There is a special gamma distribution that is worth mentioning because of its importance in statistical inference. Many statistical procedures on hypothesis testing discussed in chapters 4 through 10 use this special gamma distribution called the *chi-square distribution*, which is defined as follows:

Definition. 2.5.

A random variable X follows the *Chi-square distribution* with parameter r (where r is a *positive integer*), if its probability distribution is a gamma $\mathcal{G}(r/2, 1/2)$. The corresponding probability density function is given by:

$$f(t|r) = \begin{cases} \dfrac{1}{2^{r/2}\Gamma(r/2)} t^{r/2-1} e^{-t/2} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This probability distribution is often denoted by χ^2_r.

The parameter r associated with the chi-square distribution is known in the statistical literature as the *Number of degrees of*

freedom. The name "number of degrees of freedom" reflects the fact that the chi-square distribution with r degrees of freedom often describes the law of probability of statistical aggregates in which r observations have the "freedom" to take any value, the remaining observations being derived from the information already available.

The software solution presented for the Gamma probability distribution can be used for the chi-square distribution as well. The Gamma parameters in this case will be $\alpha = r/2$, and $\lambda = 1/2$.

2.3.8 The Chi-Square, F, and Student t-Distributions

There are a few important results on probability distributions that have proved very useful in practice. I have selected a few of them to be presented in this section.

(a) The chi-square distribution has found numerous important applications in statistical inference. This partly stems from the fact that if $Z_1, Z_2, \cdots, Z_n$ are n independent standard Normal random variables, then the sum of their squares U is known to follow the chi-square distribution with n degrees of freedom.

$$If\ Z_i \sim N(0,1)\ then\ U = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2 \qquad (2.36)$$

The Chi-square distribution is used in chapter 7 for comparing two proportions based on data from two dependent samples. It is also used extensively in chapter 9 to test the equality of multiple population proportions.

(b) Let X and U be two independent random variables such that X follows the standard Normal distribution, and U the

chi-square distribution with n degrees of freedom. Then the random variable T derived as the ratio of X over $\sqrt{U/n}$ follows a probability distribution called the *Student's t Distribution* with n degrees of freedom and denoted by t_n. Mathematically, this is expressed as,

$$If\ X \sim N(0,1)\ \&\ U \sim \chi^2_n\ then\ T = \frac{X}{\sqrt{U/n}} \sim t_n \quad (2.37)$$

The t distribution has found numerous applications in statistical inference as will be seen in subsequent chapters. This probability distribution has been thoroughly investigated, and is well documented in the statistical literature.

Tables 2.3 and 2.4 provide Excel, Calc, and R functions for calculating probability densities and cumulative probabilities of the t distribution.

Table 2.3: Excel Solutions for Student's t_n Distribution

Software	*Probability Type*	*Commands*
Excel 2007	*Probability Density*	N/A
	Cumulative Probability	1-TDIST(x,n,1)
Excel 2010	*Probability Density*	T.DIST(x,n, FALSE)
	Cumulative Probability	T.DIST(x,n, TRUE)

Table 2.4: R and Calc Solutions for Student's t_n Distribution

Software	*Probability Type*	*Commands*
R	*Probability Density*	N/A
	Cumulative Probability	1-TDIST(x;n;1)
Calc	*Probability Density*	dt(x, n)
	Cumulative Probability	pt(x, n)

(*c*) If two random variables U_1 and U_2 are independent and follow the chi-square distribution with n_1 and n_2 degrees

of freedom respectively, the derived random variable $F = (U_1/n_1)/(U_2/n_2)$ follows a distribution function called F. This is expressed mathematically as,

$$\textit{If } U_1 \sim \chi^2_{n_1} \;\&\; U_2 \sim \chi^2_{n_2} \textit{ then } F = \frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2} \quad (2.38)$$

The F distribution has been thoroughly studied, and is often used in practice (see chapters 10 and 11 in this book). However, its probability density function is very complex and is not often used in practice. Nevertheless practitioners need to be able to compute the percentiles and the cumulative probabilities of the F distribution. Tables 2.5 and 2.6 show some software options that you can used to handle the F distribution. Section E.9 of Appendix E describes more Excel functions related to F distributions. Section F.2 of Appendix F will show you R functions for the F distribution.

Table 2.5 : Excel Solutions for F_{n_1,n_2} Distribution

Software	*Probability Type*	*Commands*
Excel 2007	*Prob. Density*	N/A
	Cum. Probability	1-FDIST(x,n_1,n_2)
Excel 2010	*Prob. Density*	F.DIST(x,n_1,n_2,FALSE)
	Cum. Probability	F.DIST(x,n_1,n_2,TRUE)

Table 2.6 : R and Calc Solutions for F_{n_1,n_2} Distribution

Software	*Probability Type*	*Commands*
R	*Probability Density*	N/A
	Cumulative Probability	1-FDIST(x ;n_1 ;n_2)
Calc	*Probability Density*	df(x,n_1,n_2)
	Cumulative Probability	pf(x,n_1,n_2)

(*d*) If $Z_1, Z_2, \cdots, Z_n$ are n independent variables that follow the normal distribution $N(\mu_1, 1), N(\mu_2, 1), \cdots, N(\mu_n, 1)$,

the sum of their squares follows a distribution that is called the *Non-central chi-square* distribution with n degrees of freedom and the *non-centrality parameter* $\lambda = (\mu_1^2 + \cdots + \mu_n^2)/2$.

$$\textit{If } Z_i \sim N(\mu_i, 1) \textit{ then } Q \sim \chi_n^2(n, \lambda), \tag{2.39}$$

where $Q = Z_1^2 + Z_2^2 + \cdots + Z_n^2$. Non-central chi-square distributions are often used to quantify the propensity for rejecting the assumption of equality among proportions when these are indeed not all equal.

The chi-square distribution described in (a) is often called the *Central Chi-Square* distribution to indicate that its centrality parameter is 0.

The mean and variance of the $\chi^2(n, \lambda)$ are $n+2\lambda$ and $2n+8\lambda$ respectively.

(e) Let X and U be two independent random variables such that X follows the Normal distribution $N(\mu, 1)$, and U the chi-square distribution with n degrees of freedom. Then the random variable T derived as the ratio of X over $\sqrt{U/n}$ follows the *Non-Central t Distribution* with n degrees of freedom, and non-centrality parameter μ.

$$\textit{If } X \sim N(\mu, 1) \textit{ and } U \sim \chi_n^2 \textit{ then } T = \frac{X}{\sqrt{U/n}} \sim t(n, \mu). \tag{2.40}$$

The t distribution has found numerous applications in statistical inference as will be seen in subsequent chapters. Among other applications, the non-central t-distribution is used to study the quality of the t-test as shown in chapter 6.

(f) If U_1 and U_2 are 2 independent random variables such that U_1 follows the non-central chi-square distribution $\chi^2(n_1, \lambda)$

and U_2 follows the central chi-square $\chi^2_{n_2}$ then the random variable F derived as the ratio of U_1/n_1 to U_2/n_2 follows a probability distribution called the *Non-central F-Distribution* and denoted by $F(n_1, n_2, \lambda)$.

$$If \left\{ \begin{array}{l} U_1 \sim \chi^2(n_1, \lambda), \\ U_2 \sim \chi^2_{n_2} \end{array} \right\} \; then \; F = \frac{U_1/n_1}{U_2/n_2} \sim F(n_1, n_2, \lambda), \tag{2.41}$$

Note that of all software options I am considering in this book, only the R package can handle non-central chi-square, non-central t, and non-central F distributions. See section F.2 for more details.