CHAPTER 1

INTRODUCTION TO DATA SCIENCE

Data science is the process of analyzing data and extracting the relevant information from the observed data. Data science has roots in statistics and computer science. Data science is also referred to as "big data" or "analytics", however big data actually refers to the data set rather than the analysis of the data. The use of data science has become more common due to the ease of data collection, gains in computing power, and the need to make decisions based on data.

The data science process is an iterative process that begins with framing the problem. The first step is to frame the problem and determine the important questions of interest. Key questions include

- What is the goal of the process?
- What information is desired from the process?
- What is the data available for analysis?
- Is the data complete?
- IS there sufficient data to draw inferences from?
- Is the data representative? Current?
- Are there other sources of data that can be used?
- How accurate are the inferences?

Typical goals in a data science application are

- Prediction of some attribute from other attributes of the data. Regression and time series analysis are often used in a prediction analysis.

- Classifying a data point to one of several groups. Logistic regression, discriminant analysis, classification trees, and nearest neighbor methods are used in a classification analysis.
- Identifying clusters of data points that may be similar in some regard. Cluster analysis and nearest neighbor methods are often used in a cluster analysis.
- Profiling an observation from current and past information. Many different methods are used if a profile analysis.

Questions that are typically asked of a data set are

- Is variable (feature) Y related to variable X or variables $X_1, \ldots, X_n$. Correlation and regression are the many tools used to address these questions.
- Are subpopulations (subsets) A and B different. Exploratory statistics are often used to investigate differences in subpopulations.
- Does variable X provide any useful information. Many different approaches to answering this question.
- Are there similar subpopulations in the data set? Cluster analysis is used here.
- Which category does an observation fall in. This is a classification problem.

■ **EXAMPLE 1.1**

Prediction might be used to predict future sales, demand, weather, enrollment, voting patterns, etc.

■ **EXAMPLE 1.2**

Classification might be used to identify political affiliation, whether a student applicant will enroll or not, whether an email is spam or non-spam, etc.

■ **EXAMPLE 1.3**

Clustering might be used to find potential sub-populations in a particular population.

■ **EXAMPLE 1.4**

Profiling might be used to determine the interests of voters, Amazon shoppers, students who prefer a university, etc.

Once the problem has been framed the source of the data must be determined. It is important for the data to be complete and representative of the environment framing the problem. A complete data set will include all of the relevant variables for the problem at hand. A data set is representative when it contains data that is current, relevant, and contains representative sample of the target population. It is critical that good data is used in data science since "garbage in, garbage out." Data that is not representative or incomplete will lead to bad inferences no matter how extensive the data analysis process is.

The data used in data science is often messy and must be cleaned up before it can be used. Data cleaning is the process of detecting and correcting or removing inaccurate or incomplete observations in a data set. Most data scientists claim that 70-80% of their time is spent cleaning the data before being able to analyze the data. Typical problems encountered in a data set include

- missing data
- incorrect data
- old data that is no longer relevant

- incomplete records
- anomalies in the data set
- duplicative records
- formatting issues

A variety of different approaches are used in cleaning a data set. Keep in mind, the data set is often extremely large and cannot be easily investigated record by record.

Once the data set has been cleaned and is now usable, data summarization and exploratory data analysis (EDA) are often the first step in an analysis. Data summarization and EDA includes computation of basic summary statistics such as the mean, median, quantiles, standard deviation, inter-quartile range, and correlations. Graphical statistics such as histograms, boxplots, scatterplots, bar charts, and quantile plots (think normal plot) are important exploratory methods that often reveal a great deal about the data.

Initial analysis of the data often leads to important insights for the model building and inference stages of a data science analysis.

Modeling is often used for the processes of prediction, classification, clustering, and profiling. Regression, logistic regression, generalized regression are often used for prediction and classification. Special models are available for clustering and profiling, also. The models must be assessed for accuracy. Accurate models will be updated as new data is collected.

After analysis, the problem sometimes is reframed and the process begins anew addressing the new problem.

Finally, the results are presented in clear and concise format. Graphics are often a key component in presenting the results of the analysis. Graphics must be clearly labeled, informative, and must present the relevant information in an easy to see fashion.

Terminology

- **Big Data** - Big data refers to data that are too big or complex for standard data processing. Big data is also used to refer to the process of extracting information from large data sets (data science).
- **Machine Learning** - (From Wikipedia) Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed. Machine learning evolved through computer science and AI and many of the machine learning techniques are reinvented statistical methods, but not all. Machine learning uses a computer to solve problems intractable by hand, takes data as input, uses data that represents a sample from a population, and the data is tabular.
- **Statistical Learning** - Statistics based on statistical procedures.

  Machine learning tends to be data driven whereas statistical learning tends to be model driven.
- **Supervised Learning** - Supervised learning is machine/statistical learning where there are input and output variables (explanatory and response variables).
- **Unsupervised Learning** - Unsupervised learning is machine/statistical learning where the variables are unlabeled as input or output variables.
- **Quantitative Data** - Data taking on numerical values where the difference or ratio of two values has meaning.
- **Qualitative Data** - Data where the value represents a quality of the object and has no magnitude.
- **Ordinal Data** - Qualitative data where the qualities have a natural ordering.

- **Count data** - Quantitative data where the value represents the number of times an event has occurred.
- **Time Series** - A time series is a series of observations indexed by time. Time series analysis is used when the data is time oriented and the goal is to predict future events.

The type of data being collected will dictates the statistical analysis and models that can be used.

- **Dataframe** - A dataframe is a rectangular array of data where each row represents an observation. The columns of a dataframe are the variables recorded on each record.
- **R** - Statistical software package often used by data scientists. R is open source software and very powerful and used by most statisticians.

CHAPTER 2

PROBABILITY MODELS

There are three types of probability models that will be discussed in this class. Namely, *discrete models*, *continuous models*, and *mixed models*. The definitions of discrete and continuous random variables are given below.

Definition 2.1 *A random variable is said to be a* **discrete random variable** *if it can take on at most a countable number of values.*

Note: A set is said to be *countable* if it can be put into a 1-1 correspondence with the set of counting numbers. That is, the elements of the set can be counted and have $1st, 2nd, 3rd, \cdots$ members. That is, there are obvious gaps between each of the successive members. The following two sets are countable sets:

$$\{1, 2, 3, 4, \cdots\} \quad \text{and} \quad \{\cdots, -3, -2, -1, 0, 1, 2, 3 \cdots\}$$

Definition 2.2 *A random variable is said to be a* **continuous random variable** *if it can take on any value in one or more intervals.*

Definition 2.3 *A random variable is said to be a* **mixed random variable** *when it has a discrete component and a continuous component.*

■ **EXAMPLE 2.1**

Suppose the lifetime of an electronic component is 0 for 1% of the production (failures) and for successful components the lifetime is between 0 and 100,000 hours. This is a mixed random variable.

■ **EXAMPLE 2.2**

Determine which of the following random variables are discrete and which are continuous.

- A=Age in years of a student
- E=Number of errors in a baseball game
- W=Weight of a rainbow trout
- P=Points scored on midterm I
- T=Time between two consecutive phone calls
- C=Number of children in a family
- B=Amount of beer a person can drink

The two most important characteristics of a random variable are its *probability distribution* (pdf) and its *cumulative distribution function* (CDF) both of which completely describe a population.

Definition 2.4 *The* **probability distribution** *of a discrete random variable is a graph, table, or formula that explicitly specifies the probability associated with each possible value that the random variable can take on. The set of values of a random variable is called the* **support** *of the random variable and is denoted by* $\mathcal{S}$.

Definition 2.5 *The* **probability density function** *(pdf) is the probability function used to assign probabilities to a value of a random variable. The pdf is usually denoted by* $f(x)$.

Definition 2.6 *For a random variable* X*, the* **cumulative distribution function** *(CDF) associated with* X *is defined to be* $F(x) = P(X \leq x)$ *for* $x \in \mathbb{R}$.

2.1 DISCRETE PROBABILITY MODELS

Definition 2.7 *The* **probability density function** *(pdf), is often used to describe the probability distribution of a discrete r.v.* X*. The pdf is denoted by* $p(x)$ *and is a function satisfying*

1. $p(a) = P(X = a)$ *for* $x \in \mathcal{S}$
2. $p(x) \geq 0$ *for every value of* $x \in \mathcal{S}$
3. $\sum_{x \in \mathcal{S}} p(x) = 1$

Discrete random variables that are commonly used include the binomial, Poisson, geometric, negative binomial, and multinomial.

Recall, the pdf and CDF completely describe the distribution of a random variable X. Furthermore,

$$\begin{aligned} F(x) &= \sum_{X \leq x} f(x) \text{ when } X \text{ is discrete} \\ F(x) &= \int_{-\infty}^{x} f(u)\, du \text{ when } X \text{ is continuous} \end{aligned}$$

Also, $F(x)$ is a non-decreasing function and is right-continuous; $F(x)$ is a continuous function when X is a continuous random variable and a step function when X is discrete.

Figure 2.1 CDF for the discrete random variable in Example 2.3.

■ **EXAMPLE 2.3**

Let X be a discrete random variable with pdf $f(x) = \frac{x}{15}$ for $x = 1, 2, 3, 4, 5$. Then,

$$P(X = 1) = f(1) = \frac{1}{15}$$

and the CDF is given by

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{15} & 1 \leq x < 2 \\ \frac{3}{15} & 2 \leq x < 3 \\ \frac{6}{15} & 3 \leq x < 4 \\ \frac{10}{15} & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

A plot of the CDF is given in Figure 2.1.

2.2 CONTINUOUS PROBABILITY MODELS

A continuous random variable is a random variable that can take on any value in one or more intervals. Because a continuous random variable is defined on intervals the probabilities associated with a continuous random variable will be areas under a continuous curve defined by a pdf. With a continuous random variable probability is only assigned to intervals of values and integration can be used to find probabilities concerning a continuous random variable.

Definition 2.8 *The* **probability density function** *(pdf) associated with a continuous random variable is a function that specifies the probability associated with each possible sub-interval of values that the random variable can belong to. The probability density function of a continuous r.v. X, denoted by $f(x)$, must satisfy*

1. $f(x) \geq 0$, *for every* $x \in \mathcal{S}$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X < b) = \int_a^b f(x)dx$

Note that for a continuous random variable

$$P(X = a) = \int_a^a f(x)dx = 0$$

for every value of $-\infty < a < \infty$. Thus,

- $P(X \leq x) = P(X < x)$
- $P(X \geq x) = P(X > x)$
- $P(a \leq X \leq b) = P(a \leq X < x) = P(a < X \leq b) = P(a < X < b)$

Figure 2.2 CDF for the continuous random variable in Example 2.4.

■ **EXAMPLE 2.4**

Let X be a continuous random variable with pdf $f(x) = 2e^{-2x}$ for $x \geq 0$. Then,

$$P(X > 1.5) = \int_{1.5}^{\infty} 2e^{-2x}\, dx = e^{-3}$$

and the CDF is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-2x} & x > 0 \end{cases}$$

A plot of the CDF is given in Figure 2.2.

Commonly used continuous probability models include the normal, exponential, log-normal, gamma, beta, cauchy, and Weibull.

2.3 PARAMETERS OF A PROBABILITY MODEL

Each probability model is governed by one or more parameters. The parameters of a distribution control the shape of the distribution and all of the characteristics of a probability model.

Definition 2.9 *A* **parameter** *is a numerical value that describes a characteristic of a population.*

The parameters of a distribution are the unknowns in a data science/statistics project and must be estimated from the data. Commonly estimated parameters of a population are the mean, median, mode, quantiles, standard deviation, and variance.

Definition 2.10 *The mean of a random variable is a weighted average of the values the random variable takes on. The weighting function used is the pdf of the random variable. In particular, the* **mean** *or* **expected value** *of a random variable, denoted by $\mu = E(X)$, for a discrete random variable is the mean is*

$$E(X) = \sum_{x in \mathcal{S}} x \cdot p(x)$$

and the mean of a continuous random variable X is

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Definition 2.11 *The* **qth quantile** *of a random variable is the value of the random variable where $F(x_q) = q$. The median is the 0.5th quantile or the 50th percentile.*

The quantiles summarize how the distribution is distributed over the range of the X values, however, the quantiles do not completely describe the distribution of a random variable. Commonly used quantiles include

- the pecentiles $x_{0.01}, x_{0.02}, \ldots, x_{0.99}$.
- the deciles $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$.
- the quintiles $x_{0.2}, x_{0.4}, x_{0.6}, x_{0.8}$.
- the quartiles $x_{0.25}, x_{0.5}, x_{.75}$. The 25th percentile ($x_{0.25}$ is called the first quartile and the 75th percentile ($x_{).75}$) is called the third quartile.

- the median $x_{0.5}$

The mean and the median are both measures of the center of a population. The mean is pulled out toward the the extremes in a population whereas the median ignores the extremes of a population. A median quite a bit smaller than the mean indicates that the population extends far out to the right (larger values); A median quite a bit larger than the mean indicates that the population extends far out to the left (smaller values).

Definition 2.12 *The* **variance** *of a random variable is* $\sigma^2 = E\left[(X_\mu)^2\right]$ *and the* **standard deviation** *is* $\sigma = \sqrt{\sigma^2}$.

The variance and standard deviation measure the variability in a population. The larger the standard deviation, the more spread out the population is. The mean, median, and standard deviation are often used to summarize a population.

Definition 2.13 *The* **inter-quartile range** *is the difference between the 75th and 25th percentiles.*

The inter-quartile range also measure the variability in the population but ignores the tail regions of the population. Fifty percent (50%) of the population fall between the 25th and 75th percentiles.

2.4 COMMON DISCRETE MODELS

2.4.1 The Binomial Model

The binomial probability model is often used to model chance experiments problems involving repeated trials of a dichotomous experiment or when sampling a population with replacement. For example, the binomial is often used to model the number of heads in n coin tosses, the number of hits in n at bats, the number of wins on a poker machine in n plays, and the number of guilty votes in the initial jury vote in a jury of n individuals. The binomial probability model can be used to model the number of successes in a chance experiment consisting of n trials under the following conditions.

Binomial Setting

- The chance experiment consists of n independent trials.
- Each trial results in one of two outcomes, say success (S) and failure (F).
- The probability of a success on each of the n trials is constant with $P(S) = p$ and $P(F) = 1 - p = q$.
- The random variable X is the number of successes in the n trials.

A random variable X having a binomial distribution is denoted by $X \sim Bin(n, p)$, where $n \in \mathbb{N}$ is the number of trials and p is the probability of success. The pdf of $X \sim Bin(n, p)$ is

$$f(x) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, \ldots, n;\ p \in [0, 1],\ p + q = 1$$

When $n = 1$ (i.e., $X \sim Bin(1, p)$), a binomial random variable X is called a *Bernoulli random variable* and the pdf of a Bernoulli random variable is

$$f(x) = p^x q^{1-x},\ x = 0, 1; p \in [0, 1], p + q = 1$$

The R command used for computing the pdf of a binomial random variable $X \sim Bin(n, p)$ at $x \in \mathcal{S}$ is `dbinom(x,n,p)`, and `pbinom(x,n,p)` is used to compute the value of the CDF at $x \in \mathcal{S}$. That is, for $x \in \mathcal{S}$, `dbinom(x,n,p)` $= f(x) = P(X = x)$ and `pbinom(x,n,p)` $= F(x) = \sum_{i=0}^{x} \binom{n}{i} p^i q^{n-i}$.

Theorem 1 *If* $X \sim Bin(n, p)$*, then*

(i) $E(X) = np$.

(ii) $Var(X) = npq$.

■ **EXAMPLE 2.5**

Suppose a basketball player makes 40% of their three-point shot attempts. If the player shoots 10 three-point shots in a game and X is the number of three-point shots made, then assuming the shots are independent and the probability of making each shot is 0.40, $X \sim Bin(10, 0.4)$. The probability the player makes five of his 10 shots is

$$f(5) = \binom{10}{5} 0.4^5 (1 - 0.4)^{10-5} = 0.2001$$

The mean and variance of the random variable X are $\mu = 10(0.4) = 4$ and $\sigma^2 = 10(0.4)(0.6) = 2.4$.

The R command for computing $f(5)$ is `dbinom(5,10,0.4)` where $x = 5, n = 10$, and $p = 0.4$. The R command `dbinom(c(0:10),10,0.4)` computes the value of $f(x)$ for each $x \in \{0, 1, 2, \ldots, 10\}$. The R commands used to create plot the pdf of $X \sim Bin(10, 0.4)$ are

```
> s=c(0:10) # s is the support of X
> plot(s,dbinom(s,10,0.4),type="h",ylab="f(x)")
> abline(0,0)# puts in a reference line at f(x)=0
```

2.4.2 The Poisson Model

The Poisson probability model can be used to model the number of occurrences of a particular event. For example, the Poisson probability model is often used to model the number of tornadoes occurring in a tornado season, the number of grasshoppers in a field, and the number of hits on a website in a fixed time period. A random variable is a Poisson random variable if it has the following pdf

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x \in \mathbb{W};\ \lambda \in [0, \infty)$$

A random variable X having a Poisson distribution will be denoted by $X \sim Pois(\lambda)$.

The R command for computing the value of the pdf of $X \sim Pois(\lambda)$ at $x \in \mathcal{S}$ is `dpois(x,lambda)`, and `ppois(x,lambda)` computes the value of the CDF for $x \in \mathcal{S}$.

Theorem 2 *If $X \sim Pois(\lambda)$, then*

(i) $E(X) = \lambda$.

(ii) $Var(x) = \lambda$.

■ **EXAMPLE 2.6**

Suppose the number of grasshoppers in a one-square yard plot in a field is known to follow a Poisson distribution with $\lambda = 5$. Then, the probability that there are there are four grasshoppers in one-square yard plot in this field is

$$P(X = 4) = \frac{e^{-5}5^4}{4!} = \texttt{dpois(4,5)} = 0.1755$$

and the probability that there are more than 10 grasshoppers in one-square yard is

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \texttt{ppois(10,5)} = 0.0028$$

The expected number of grasshoppers in a one-square yard plot in this field is $E(X) = 5$ and the variance is $Var(X) = 5$.

The R commands used to create the plot of the CDF of $X \sim Pois(5)$ are

```
> x=c(0:25) effective support of X
> y=ppois(x,5)#CDF over effective support
> plot(x,y,type="s",ylab="CDF")
> abline(0,0)#adds an x-axis at height 0
```

2.4.3 The Negative Binomial Model

The negative binomial probability model can be used to model chance experiments where trials of an experiment are repeated until the rth success occurs. A random variable X will follow a negative binomial distribution under the following conditions.

Negative Binomial Setting

- A chance experiment which results in either a success or a failure will be repeated independently until the rth success occurs.
- The probability of a success on each of the trials is constant with $P(S) = p$.
- The random variable X is the number of failures preceding the rth success.

A random variable X having a negative binomial distribution is denoted by $X \sim NegBin(r, p)$, where r is the number of successes required to terminate the chance experiment and p is the probability of a success on each trial. The pdf of a negative binomial random variable $X \sim NegBin(r, p)$ is

$$f(x) = \binom{r+x-1}{x} p^r q^x,\ x \in \mathbb{W};\ p \in [0,1], p+q=1$$

The R command for computing values of the pdf of $X \sim NegBin(r, p)$ at $x \in \mathcal{S}$ is `dnbinom(x,r,p)` and `pbinom(x,r,p)` computes the value of the CDF for $x \in \mathcal{S}$.

When $r = 1$ a negative binomial random variable is called a *geometric random variable* which is denoted by $X \sim Geo(p)$. The pdf of a $X \sim Geo(p)$ is

$$f(x) = pq^x,\ x \in \mathbb{W};\ p \in [0,1], p+q=1$$

and for $x \in \mathbb{W}$ the CDF is $F(x) = 1 - q^{x+1}$. The R command for computing the value of the pdf of $X \sim Geo(p)$ is `dgeom(x,p)`, and `pgeom(x,p)` computes the value of the CDF at $x \in \mathcal{S}$.

Theorem 3 *Let $X \sim NegBin(p, r)$, then*

(i) $E(X) = \dfrac{rq}{p}$.

(ii) $Var(X) = \dfrac{rq}{p^2}$.

Since a geometric random variable is a negative binomial random variable with $r = 1$, it follows that the mean of a geometric random variable X is $E(X) = \frac{q}{p}$ and the variance is $Var(X) = \frac{q}{p^2}$.

■ **EXAMPLE 2.7**

A telemarketer will make phone calls until 25 successful contacts are made. Assuming each phone call is independent and the probability that a contact is made is constant with $p = 0.4$, the pdf of $X =$ the number of phone calls failing to make contact before the 25th contact, is

$$f(x) = \binom{25+x-1}{x} 0.4^{25} \times 0.6^x,\ \ x \in \mathbb{W}$$

The probability that 15 failed calls are made before the 25th contact is made is

$$P(X = 15) = \binom{39}{15} 0.4^{25} \times 0.6^{15} = \texttt{dnbinom(15,25,0.4)} = 0.0013$$

and the probability that at least 15 failed calls will be made before the 25th contact is made is

$$P(X \geq 15) = 1 - P(X \leq 14) = 1 - \texttt{pnbinom(14,25,0.4)} = 0.9980$$

The mean number of failed phone calls prior to making the 25th contact is $E(X) = \frac{25(0.6)}{0.4} = 37.5$ and the variance is $Var(X) = \frac{25(0.6)}{0.4^2} = 93.75$.

■ **EXAMPLE 2.8**

Suppose a dart player will throw a dart until the bullseye is hit. Let X be the number of dart throws prior to the player finally hitting the bullseye for the first time. Assuming each dart toss is independent and the probability of hitting the bullseye on each throw is constant with $p = 0.1$, the random variable $X \sim Geom(0.1)$ and $f(x) = 0.1 \times 0.9^x$ for $x \in \mathbb{W}$. The probability it takes five throws (i.e., four misses) to hit the bullseye for the first time is

$$P(X = 4) = f(4) = 0.1 \times 0.9^4 = \texttt{dgeom(4,0.1)} = 0.0656$$

and the probability it takes less than 10 tosses to hit the bullseye is

$$P(\text{less than 10 tosses}) = P(X \leq 8) = F(8) = \texttt{pgeom(8,.1)} = 0.6126$$

since fewer than 10 tosses means fewer than 9 misses. The mean number of misses before the bullseye is hit for the first time is $E(X) = \frac{0.9}{0.1} = 9$ and the variance is $Var(X) = \frac{0.9}{0.1^2} = 90$.

■ **EXAMPLE 2.9**

In Example 2.8, X is the number of dart throws missing the bullseye before the first bullseye is hit. In this case, the number of dart throws required to hit the bullseye once is $Y = X + 1$ and

$$f_Y(y) = \binom{y-1}{1-1} 0.1^1 0.9^{y-1} = 0.1(0.9)^{y-1},\ y = 1, 2, \ldots$$

Thus, the probability that it takes 10 throws to hit the bullseye for the first time is $P(Y = 10) = 0.1(0.9)^9 = 0.0387$, and the mean number of throws taken to hit the bullseye once is $E(Y) = \frac{1}{0.1} = 10$.

2.4.4 The Multinomial Model

The multinomial probability model is a multivariate generalization of the binomial probability model. In particular, the multinomial is a multivariate probability model is associated with a k-dimensional random vector used to model chance experiments where n objects are allocated independently to k cells. A random vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ will have a multinomial distribution under the following conditions.

Multinomial Setting

1. A chance experiment with k possible outcomes, say $O_1, O_2, \ldots, O_k$, will be repeated independently n times.
2. The probability of outcome O_i on any trial is p_i.
3. $\sum_{i=1}^{k} p_i = 1$.
4. X_i is the number of trials resulting in outcome O_i, and the random vector is $\vec{X} = (X_1, X_2, \ldots, X_k)$.

A k-dimensional random vector $\vec{X}$ having a multinomial distribution is denoted by $\vec{X} \sim MNom_k(n, \vec{p})$, where k is the number of cells, n is the number of objects being allocated to the k cells, and $\vec{p} = (p_1, p_2, \ldots, p_k)$ is the vector of cell probabilities. The joint pdf of $\vec{X} \sim MulNom_k(n, \vec{p})$ is

$$f(x_1, x_2, \ldots, x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

for $x_i \in \mathbb{W}$, $\sum_{i=1}^{k} x_i = n$, $p_i \in [0, 1]$ for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} p_i = 1$.

The R command for computing values of the pdf of $\vec{X} \sim MulNom_k(r, p)$ at $\vec{x} \in \mathcal{S}$ is `dmultinom(x,n,p)` where $x = (x_1, \ldots, x_k)$ and $p = (p_1, \ldots, p_k)$. R does not have a command for computing the joint CDF for a multinomial random variable.

Theorem 4 *If $\vec{X} \sim MulNom_k(p_1, \ldots, p_k)$, then the marginal distribution of $X_i \sim Bin(n, p_i)$, for $i = 1, \ldots, k$.*

■ **EXAMPLE 2.10**

In a mayoral election suppose 40% of the electorate favors candidate A, 35% favors candidate B, and 25% favors candidate C. Suppose $n = 10$ voters are selected independently and surveyed, and let X_1 be the number of sampled voters voting for candidate A, X_2 the number voting for candidate B, and X_3 the number voting for candidate C. Then, the random vector $\vec{X} = (X_1, X_2, X_3)$ follows a multinomial distribution with $n = 10, \vec{p} = (0.4, 0.35, 0.25)$. The probability that $\vec{X} = (5, 3, 2)$ is

$$\begin{aligned} P\left(\vec{X} = (5,3,2)\right) &= \binom{10}{5,3,2} 0.4^5 \times 0.35^3 \times 0.25^2 \\ &= \texttt{dmultinom(c(5,3,2),10,c(0.4,0.35,0.25))} \\ &= 0.0691 \end{aligned}$$

The marginal distributions of the random variables X_1, X_2, and X_3 are $X_1 \sim Bin(10, 0.4)$, $X_2 \sim Bin(10, 0.35)$, and $X_3 \sim Bin(10, 0.25)$.

2.5 COMMON CONTINUOUS MODELS

Unlike the discrete probability models, the scenarios under which a continuous probability model will be appropriate are not obvious. Often, the appropriate continuous probability model is suggested by prior research or data. There are numerous continuous probability models, however, only the models frequently used in this text will be introduced and discussed in this section.

Recall, a continuous random variable X is a random variable where

- the support, §, is the union of one or more intervals,
- the pdf $f(x)$ is a non-negative function for which $\int_{\S} f(x)\,dx = 1$,
- $E(X) = \int_{\S} x f(x)\,dx$,
- $Var(X) = E\left[(X-\mu)^2\right]$,
- and moment generating function $M_x(t) = \int_{\S} e^{tx} f(x)\,dx$.

2.5.1 The Uniform Model

The uniform probability model is used when each value in the support $\S = (\alpha, \beta)$ is equally likely and all of the sub-intervals of $\S = (\alpha, \beta)$ of width h have the same probability. A random variable X is said to have a uniform distribution over the interval (α, β) when the pdf of X is

$$f(x) = \frac{1}{\beta - \alpha},\ x \in (\alpha, \beta);\ \alpha < \beta,\ \alpha, \beta \in \mathbb{R}$$

The CDF of $X \sim U(\alpha, \beta)$ is

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

A uniform random variable on the interval (α, β) is denoted by $X \sim U(\alpha, \beta)$.

The R command for computing the pdf of $X \sim U(\alpha, \beta)$ at $x \in (\alpha, \beta)$ is `dunif(x,`α`,`β`)`, `punif(x,`α`,`β`)` computes the value of the CDF at x, and
`qunif(p,`α`,`β`)` computes the $100 \cdot p$th quantile of X for $p \in (0, 1)$.

Theorem 5 *If $X \sim U(\alpha, \beta)$, then*

(i) $E(X) = \frac{\alpha+\beta}{2}$.

(ii) $Var(X) = \frac{(\beta-\alpha)^2}{12}$.

■ **EXAMPLE 2.11**

Let $X_1 \sim U(0,2)$ and $X_2 \sim U(-1,3)$. The pdfs of X_1 and X_2 are given in Figure 2.3. The 25th percentiles of X_1 and X_2 are `qunif(0.25,0,2)`= and `qunif(0.25,−1,3)`=, respectively.

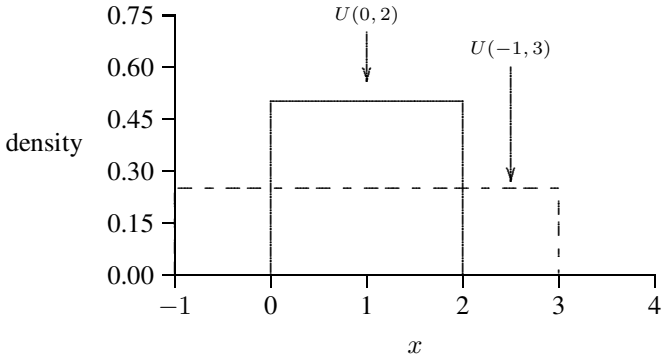

Figure 2.3 The pdfs of $X_1 \sim U(0,2)$ and $X_2 \sim U(-1,3)$.

■ **EXAMPLE 2.12**

Suppose $X \sim U(-1,1)$, then

$$P(0.5 \le X \le 0.8) \int_{0.5}^{0.8} \frac{1}{4}\, dx = \frac{0.8-0.5}{2} = 0.15$$

which can also be computed using the R commands

`punif(0.8,−1,1) − punif(0.5,−1,1)` $= 0.15$

The 78th percentile of X is `qunif(0.78,−1,1)`=0.56.

2.5.2 The Gamma Model

The family of Gamma distributions are flexible in shape, long-tail right distributions, and are often be used to model waiting times, life data, variables used in climatology and in finance. A random variable X has a gamma distribution when the pdf is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x \in (0,\infty); \alpha \in (0,\infty), \beta \in (0,\infty)$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ is called the *gamma function*.

A random variable X having a gamma distribution with parameters α and β is denoted by $X \sim Gamma(\alpha,\beta)$. When $X \sim Gamma(\alpha,\beta)$, α is called the *shape parameter* since it controls the shape of a gamma distribution, and β is called the *rate parameter*.

When $X \sim Gamma(\alpha,\beta)$, the R commands require the identification of the shape and scale parameters. In particular, `dgamma(x,shape=`α`,scale=`β`)` produces the value of the pdf at x, `pgamma(x,shape=`α`,rate=`β`)` produces the value of the CDF at x, and `qgamma(p,shape=`α`,rate=`β`)` produces the pth quantile of X.

Theorem 6 *If* $X \sim Gam(\alpha,\beta)$, *then*

(i) $E(X) = \frac{\alpha}{\beta}$.

(ii) $Var(X) = \frac{\alpha}{\beta^2}$.

■ **EXAMPLE 2.13**

Let $X_1 \sim Gamma(\alpha = 1, \beta = 0.25), X_2 \sim Gamma(\alpha = 2, \beta = 0.25)$, and $X_3 \sim Gamma(\alpha = 3, \beta = 0.25)$. The pdfs of X_1, X_2 and X_3 are given in Figure 2.4. Note that X_1, X_2 and X_3 differ only in their values of α and produce three different shapes of the gamma distribution. The means of X_1, X_2, and X_3 are $4, 8$, and 12, respectively, and the medians are `qgamma(0.5,1,4)`=2.77, `dgamma(0.5,2,4)`=6.71, and `qgamma(0.5,3,4)`=10.70.

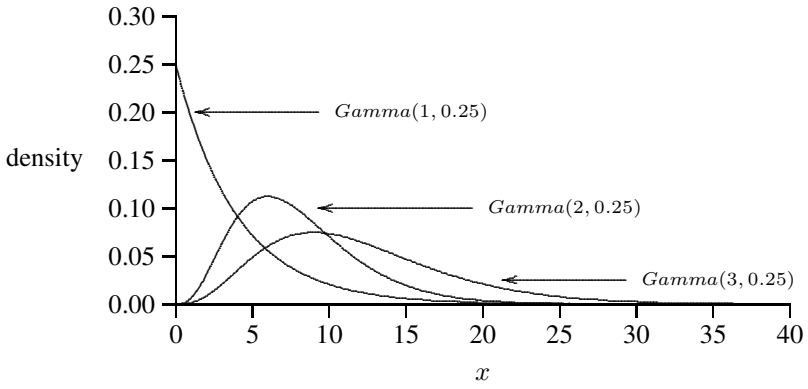

Figure 2.4 A plot of the pdfs of $X_1 \sim Gamma(1, 4), X_2 \sim Gamma(2, 4)$, and $X_3 \sim Gamma(3, 4)$.

‘

The *exponential distribution* is a gamma random variable with $\alpha = 1$ and is denoted by $X \sim Exp(\beta)$. Table 2.1 gives the pdf, mean, and variance of the exponential.

Table 2.1 Two special cases of the gamma distribution.

Distribution	pdf	Mean	Variance	MGF
Exponential	$\frac{1}{\beta}e^{-\frac{x}{\beta}}$	β	β^2	$(1-\beta t)^{-1}$

The R commands for the gamma distribution can be used for computing the pdf, CDF, and quantiles of an exponential, however, there are separate commands in R for the exponential. In particular, the R commands for the exponential given in terms of the rate are `dexp(x,`β`)`, `pexp(x,`β`)`, and `qexp(p,`β`)`.

■ **EXAMPLE 2.14**

Let $X_1 \sim Exp(0.5), X_2 \sim Exp(0.25)$ and $X_3 \sim Exp(0.15)$. The pdfs of X_1, X_2, and X_3 are shown in Figure 2.5.

■ **EXAMPLE 2.15**

Let X be the effective lifetime of a dose of a pain reliever used to treat migraine headaches, and suppose X follows an exponential distribution with mean $\beta = 0.25$ hours. The probability that the pain reliever has a lifetime of more than 6 hours is

$$P(X \geq 6) = \int_6^\infty 0.25e^{-0.25x}\,dx = e^{-1.5} = 0.2231 = \texttt{1 - pexp(6, 0.25)}$$

The median lifetime of the pain reliever is `qexp(0.5,0.25)`=2.77 hours.

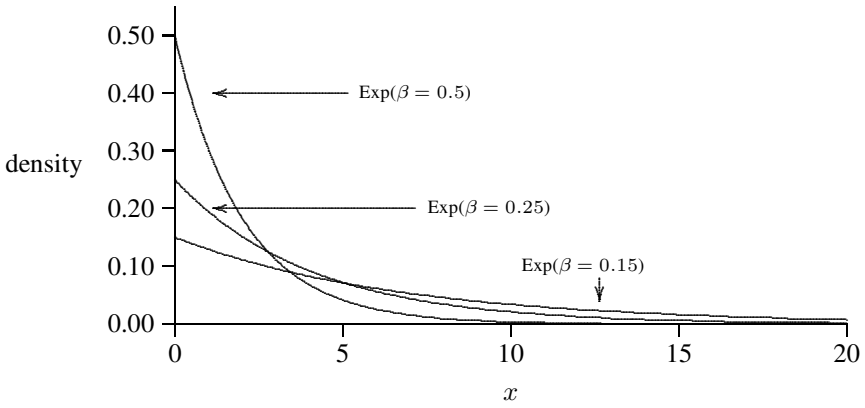

Figure 2.5 A plot of the pdfs of the exponential random variables X_1, X_2, and X_3.

2.5.3 The Normal Model

The normal probability model has a bell shaped distribution that is symmetric about its mean. The normal probability model is used to model many natural phenomena such as weights, heights, temperature, and modeling errors, however, the normal model is only appropriate when a random variable representing the phenomena is distributed symmetrically about its mean.

A random variable X has a normal distribution with parameters μ and σ when the pdf is

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

A normal distribution with $\mu = 0$ and $\sigma = 1$ is called a *standard normal distribution* and is denoted by Z.

The R commands for computing the pdf, CDF, and quantiles of $X \sim N(\mu, \sigma^2)$ are `dnorm(x,`μ,σ`)`, `pnorm(x,`μ,σ`)`, and `qnorm(p,`μ,σ`)`, respectively.

Theorem 7 *If $X \sim N(\mu, \sigma^2)$, then*

(i) $E(X) = \mu$.

(iii) $Var(X) = \sigma^2$.

■ **EXAMPLE 2.16**

Let W be the random variable associated with the weight in pounds of a one year old hatchery rainbow trout and suppose $W \sim N(0.8, 0.225)$. A plot of the pdf of W is given in Figure 2.6.

The probability that a one year old hatchery rainbow weighs between 1 and 1.5 pounds is

$$\begin{aligned} P(1 \leq W \leq 1.5) &= F(1.5) - F(1) \\ &= \texttt{pnorm(1.5,0.8,0.15)} - \texttt{pnorm(1,0.8,0.15)} \\ &= 0.9999985 - 0.9087888 = 0.09120969 \end{aligned}$$

The 95th percentile of the weights of one year old hatchery rainbow trout is `qnorm(0.95,0.8,0.15)` $= 1.05$.

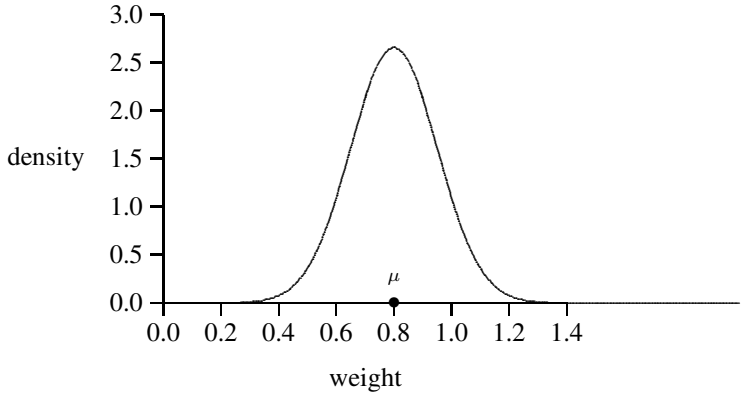

Figure 2.6 The distribution of weights of hatchery rainbow trout in Example 2.16.

2.5.4 The Beta Model

The beta distribution provides a family of distributions that are flexible in shape on the interval $(0, 1)$. In particular, a beta distribution can be u-shaped, j-shaped, long-tail right, or long-tail left, Moreover, beta distribution can be generalized to have support on any finite interval (a, b). A random variable has a beta distribution when its pdf is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in (0,1);\ \alpha \in (0,\infty), \beta \in (0,\infty)$$

A random variable X having a beta distribution with parameters α and β is denoted by $X \sim Beta(\alpha, \beta)$.

Both α and β control the shape of a beta distribution, and hence, α and β are shape parameters,. The integral

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is called the *beta function* and is denoted by $B(\alpha, \beta)$. The R commands associated with the pdf, CDF, and quantiles of the beta distribution are `dbeta(x,`α, β`)`, `pbeta(x,`α, β`)`, and `qbeta(p,`α, β`)`.

Theorem 8 *If* $X \sim Beta(\alpha, \beta)$*, then*

(i) $E(X) = \frac{\alpha}{\alpha+\beta}$.

(ii) $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

(iii) for $\alpha = \beta = 1$, $X \sim U(0, 1)$.

■ **EXAMPLE 2.17**

Let $X_1 \sim (2, 4)$, $X_2 \sim (2, 2.5)$, and $X_3 \sim (2, 2)$. The pdfs of X_1, X_2, and X_3 are given in Figure 2.7.

The means of X_1, X_2, and X_3 are $\mu_1 = 0.33, \mu_2 = 0.4$, and $\mu_3 = 0.5$, and their medians are $\widetilde{\mu}_1 = 0.314$, $\widetilde{\mu}_2 = 0.436$, and $\widetilde{\mu}_3 = 0.5$; the medians were computed using `qbeta(0.5,`α, β`)`.

■ **EXAMPLE 2.18**

Based on past experience teaching an introductory statistics course, midterm exam scores (S) were found to be approximately distributed as a beta distribution with $S \sim Beta(8, 2)$. A plot of pdf of S is given in Figure 2.8.

The 95th percentile of the midterm exams is `qbeta(0.95,8,2)`=0.96, the median is `qbeta(0.95,8,2)`=0.82, the mean is 0.80, and the probability of a passing score is $P(S \geq 0.70)$=`1-pbeta(0.70,8,20)` $= 0.804$.

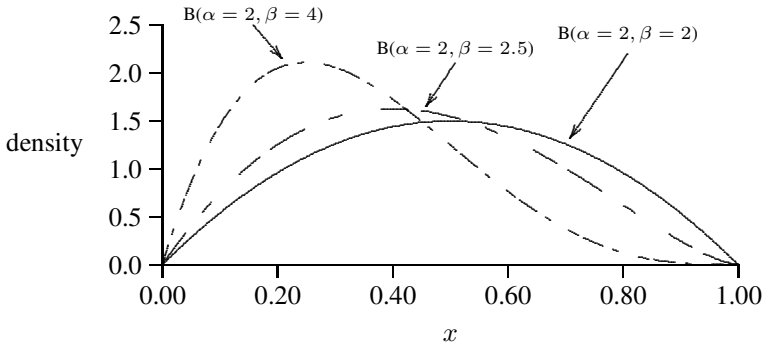

Figure 2.7 A plot of the pdfs of beta random variables X_1, X_2, and X_3.

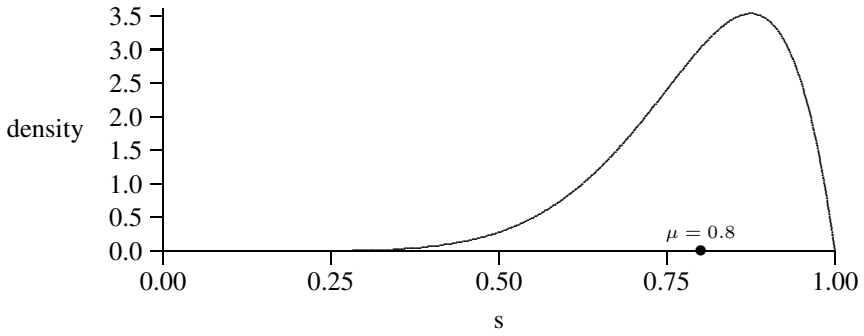

Figure 2.8 The distribution of exam scores in Example 2.18.