An Introduction to Prescriptive and Predictive Modeling

Darin England
Department of Industrial and Systems Engineering
University of Minnesota, Minneapolis

Contents

Preface

Goals

The book provides a practical introduction to the use of mathematical modeling and statistical computing techniques to solve decision problems that arise in various industrial settings. Examples are drawn from manufacturing, finance, sports, healthcare, retail, transportation, and other areas. In realistic settings, proficiency with mathematical and statistical computing software is required to solve these problems. We use the algebraic modeling language GAMS for solving mathematical programming problems and we use the R programming language for obtaining answers to statistical computing problems; however, the methodologies that are developed stand on their own and the book can be used with other programming langauges. GAMS code and listings are provided via github in the gams/ directory. R code, output, and discussion are provided in the form of Jupyter notebooks in the src/ directory. The github repository for the book is `https://github.com/bx549/IPPM`. We hope that students as well as practicing engineers and analysts will find the book to be useful.

Acknowledgments

I want to acknowledge the following undergraduate students in the Department of Industrial and Systems Engineering who created content and checked solutions to end-of-chapter exercises. Hannah Kleist created scenarios for deterministic optimization problems in Chapter 1 and implemented solutions in GAMS. She also created content for exercises on stochastic processes in chapter 2. Emily Sayles created content for end-of-chapter exercises on deterministic inventory models in chapter 1, probabilistic inventory models in chapter 2, and decision problems in chapter 3. Braeden Greseth created content and implemented solutions in R for exercises in chapters 2, 4, and 5.

Chapter 1

Deterministic Models

1.1 Linear Programming

The best way to learn the art of mathematical modeling is through practice. We begin with a simple two-variable Linear Programming model whose solution will inform the owner of a business how to allocate scarce resources to achieve maximum profit. Additional information returned with the solution will be used to gain insight on obtaining additional resources and an idea of the robustness of the solution to small changes in the data. Along the way we will illustrate various features of the GAMS algebraic modeling language. [1]

Why use GAMS? Well, it's not GAMS in particular, but rather algebraic modeling languages in general that are extremely useful when *formulating* mathematical programming problems [2] An algebraic modeling language allows one to easily specify and understand objective functions, constraints, and logical relationships among variables. The syntax of the GAMS language is simple, yet expressive enough to enable many different types of problems to be directly modeled. We should note that GAMS is not a general-purpose programming language, although it does contain constructs for conditional execution of statements and for iterative looping. Rather, GAMS is an example of a *domain-specific* language. Its purpose is to facilitate translation from a problem description into a form that is suitable for a solver to process.

A resource allocation problem. A local farmer maintains 500 acres on which she can grow feed corn for animals and/or organic corn for local millers. As the planting season approaches, she must decide how many acres to allocate to each type of corn. Her operating budget is $50,000. The costs and returns for each type of corn are as follows.

type	cost per acre	revenue per acre
feed corn	$90	$200
organic corn	$150	$300

A natural objective is to maximize total profit, and organic appears to be more attractive; however, the farmer must respect both her operating budget and the available acres. If we let $x1$ and x_2 represent the number of acres to allocate to feed corn and organic corn,

[1] We do not provide a complete introduction to GAMS. Excellent tutorials already exist. [12, 5].

[2] AMPL is another popular algebraic modeling language with a clear and expressive syntax [4].

respectively, then the farmer's decision problem is to

$$\begin{array}{lrcrcrl} \text{maximize} & 110x_1 & + & 150x_2 & = & & z(\text{profit}) \\ \text{subject to} & 90x_1 & + & 150x_2 & \leq & 50{,}000 & (\text{budget}) \\ & x_1 & + & x_2 & \leq & 500 & (\text{acres}) \\ & & & x_1, x_2 & \geq & 0 & \end{array} .$$

The operating budget and the available acres are constrained resources in the sense that if the farmer could obtain a little more of either resource, then she could earn more profit. Later in the section, will make this notion more precise. Since this problem has two decision variables, the inequalities that describe the constraints can be shown graphically as in Figure 1.1. Any combination of x_1 (feed corn) and x_2 (organic corn) within the shaded area represents a feasible solution for the farmer. The optimal solution is the combination that maximizes total profit.

The constraints in Figure 1.1 were drawn by 1) representing the inequality as a line, and then 2) determining on which side of line the inequality is valid. The dashed line represents the objective function, but it is drawn at an arbitrary place in the figure. The important thing to note is the slope. Writing the objective as

$$x_2 = \frac{z}{150} - \frac{110}{150}x_1$$

lets us easily see that the slope is $-11/15$. Now, to find the values of the decision variables (x_1 and x_2) that maximize the expression

$$z = 110x_1 + 150x_2$$

while still satisfying the constraints, imagine sliding the dashed line up and to right until the point just before the dashed line leaves the feasible region. This will occur at the point where the two lines representing the constraints intersect, i.e. $x_1 = 416.67$, $x_2 = 83.33$. The optimal solution to any linear programming problem will always occur at a "corner point". We can determine which corner point by comparing the slope of the objective function to the slopes of the lines representing the constraints that define the feasible region. In this case the slope of the objective function $-11/15$ is "steeper" that the slope of the budget constraint $= 9/11$, but not as steep as the slope of the acres constraint -1. Typically, a linear programming problem will have many variables and constraints, but the same ideas apply. Indeed, the kernel of the idea behind the Simplex algorithm, the workhorse algorithm that is commonly used to solve linear programming problems, is to move from corner point to corner point, each time increasing the value of the objective function until no further improvements can be made.

We now show how to model the farm problem in GAMS. The complete problem formulation is shown in figure 1.2.

Since this problem is small, we include all data parameters directly in the model specification. However, it is usually good practice to separate the data from the model, and AMPL encourages this separation by providing `param` declarations and the ability to have a separate `data` section. We will illustrate these features of AMPL in the other problem formulations in this article. AMPL doesn't actually solve the mathematical programing problem, but rather passes a description of the problem to a solver, which returns information about the solution (if any solution was found) to AMPL. An AMPL session for the forestry problem follows:

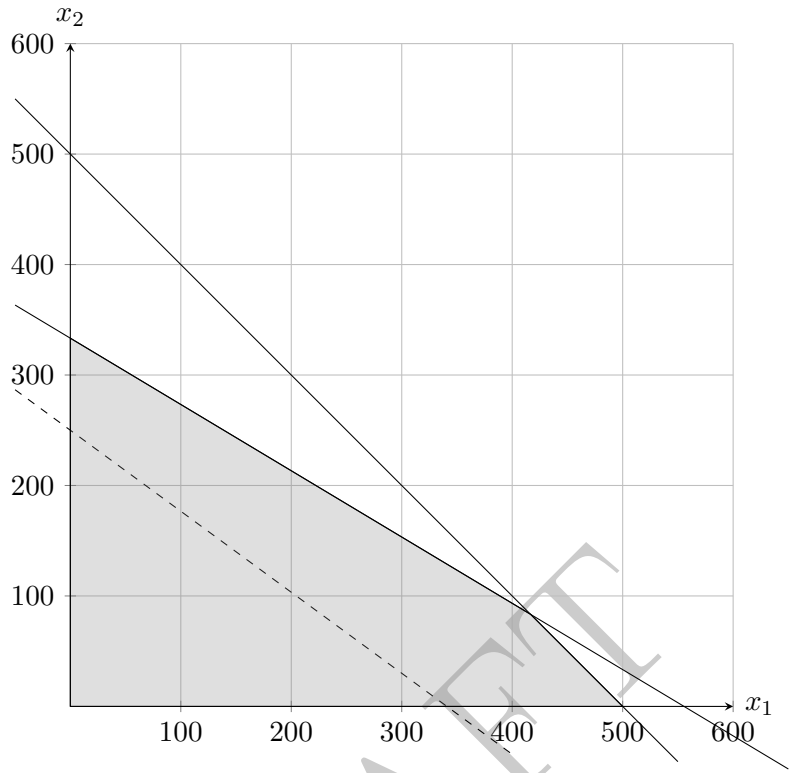


Figure 1.1: Feasible region for the farm problem.


```
variables
x1 'acres of feed corn'
x2 'acres of organic corn'
z;
positive variables x1, x2;

equations
 profit  'objective function'
 budget  'available cash'
 acres   'available acres';

profit.. z =e= 110*x1 + 150*x2;
budget.. 90*x1 + 150*x2 =l= 50000;
acres.. x1 + x2 =l= 500;

model farmer / all /;

solve farmer using LP maximizing z;
```

Figure 1.2: Farm problem (`farmer.gms`)

```
ampl: model forester.mod;
ampl: solve;
MINOS 5.5: optimal solution found.
2 iterations, objective 6250
ampl: display x1,x2;
x1 = 25
x2 = 75
```

The optimal solution indicates that the forester should harvest 100 acres, but let 25 acres regenerate naturally and plant 75 acres with pine, for a profit of \$6250. Notice that the entire budget of \$4000 is exhausted because $10 \times 25 + 50 \times 75 = 4000$. When a resource capacity constraint such as `acres` or `budget` holds with equality in the optimal solution (as in this example), then the constraint is *binding*; all of the resource associated with the constraint is being consumed. If more/less of the resource were available, then profit would be increased/decreased. The value of the associated dual variable indicates the amount by which profit would be affected for small changes in the supply of the resource. This is called the marginal value of a resource, or the shadow price.

```
ampl: display acres, budget;
acres = 32.5
budget = 0.75
```

Displaying the shadow prices in AMPL indicates that an additional acre of hardwood timber would increase profit by \$32.50, given the same budget of \$4000. Modifying the right-hand side of `acres` to 101 and re-solving shows that this is indeed the case.

```
ampl: reset;
ampl: model forester.mod;
ampl: expand acres;
subject to acres:
        x1 + x2 <= 101;

ampl: solve;
MINOS 5.5: optimal solution found.
2 iterations, objective 6282.5
ampl: display x1,x2;
x1 = 26.25
x2 = 74.75
```

The `expand` command displays the full form of a set of constraints. (This will be useful for indexed expressions.) Notice that the values of the decision variables have necessarily changed and that the solution is no longer integer-valued. This is typical of resource allocation problems. We are in fact assuming that the forester is able to execute the decisions on partial acres. For the `budget` constraint, an additional one dollar increase/decrease in the right-hand side would increase/decrease profit by \$0.75, given the same 100 acres of hardwood. It stands to reason, then, that the forester could take out a loan and apply the funds to her operation. As long as the interest rate is less than .75, profit will increase.

Shadow prices are valid for *limited* increases/decreases in the right-hand sides of the constraints. The ranges over which the these values are valid is a topic of sensitivity analysis. Here, we will show how to extract this information from AMPL. First we need to tell AMPL to use a solver that is able to return sensitivity information along with the optimal solution. We will use the solver `cplex`. Then we set a solver-specific option to make the sensitivity range information available.

```
ampl: option solver cplex;
ampl: option cplex_options 'sensitivity';
ampl: solve;
CPLEX 11.2.0: sensitivity
CPLEX 11.2.0: optimal solution; objective 6250
2 dual simplex iterations (1 in phase I)

suffix up OUT;
suffix down OUT;
suffix current OUT;
```

We are now able to display the ranges over which the current shadow prices of \$32.5 per acre and \$0.75 per dollar of budget are valid. To do this we simply add a suffix to the name of the constraint, separated by a ".". The suffix `.current` displays the current value of a constraint's right-hand side, while `.down` and `.up` display the lower and upper limits, respectively. In the AMPL session follows, the upper limit of 5000 on the right-hand side of `budget` indicates the forester should only consider loans of \$1000 or less to be valued at an incremental marginal value of 0.75.

```
ampl: display acres.down, acres.current, acres.up;
acres.down = 80
acres.current = 100
acres.up = 400

ampl: display budget.down, budget.current, budget.up;
budget.down = 1000
budget.current = 4000
budget.up = 5000
```

Sensitivity ranges may also be obtained for the decision variables. In this case the lower and upper limits represent the valid ranges on the objective function coefficients over which the current solution remains optimal. The suffix `.current` refers to the current value of the coefficient. Displaying this information for the forestry problem reveals that the current solution value of `x1 = 25` remains optimal as long as its coefficient in the objective function is in the range 14 to 70, holding all other parameters at their current values.

```
ampl: display x1.down, x1.current, x1.up;
x1.down = 14
x1.current = 40
x1.up = 70

ampl: display x2.down, x2.current, x2.up;
x2.down = 40
x2.current = 70
x2.up = 200
```

A Diet Problem. Dwight is an elementary school teacher who also raises pigs for supplemental income. He is trying to decide what to feed his pigs. Considering a combination of pig feeds available from local suppliers, he would like to feed the pigs at minimum cost while also making sure each pig receives an adequate supply of calories and vitamins. The cost, calorie content, and vitamin content of each feed are given in the table below.

Contents	**Stark County Coop Pig Feed**	**Pioneer Pig Feed**
Calories (per pound)	800	1000
Vitamins (per pound)	140 units	70 units
Cost (per pound)	\$0.40	\$0.80

Each pig requires at least 8,000 calories per day and at least 700 units of vitamins. A further constraint is that no more than one-third of the diet (by weight) can consist of Stark County Coop Pig Feed, since it contains an ingredient which is toxic if consumed in too large a quantity. First, let's write down the mathematical programming formulation without regard to AMPL. Let x_1 be the pounds per day of Stark County Coop feed to purchase. Similarly, let x_2 be the pounds per day of Pioneer feed to purchase. Then the problem is to

$$\begin{array}{lrcrcrl} \text{minimize} & .4x_1 & + & .8x_2 & & & \\ \text{subject to} & 800x_1 & + & 1000x_2 & \geq & 8000 & \text{(calories)} \\ & 140x_1 & + & 70x_2 & \geq & 700 & \text{(vitamins)} \\ & 2x_1 & - & x_2 & \leq & 0 & \text{(toxicity)} \\ & & & x_1, x_2 & \geq & 0 & \end{array} .$$

To get the toxicity restriction on the amount of Stark County Coop feed, begin with the relation

$$x_1 \leq \frac{1}{3}(x_1 + x_2).$$

Then collect terms and simplify. Since this problem has only two variables, we can easily specify the model in AMPL by creating each variable, the objective function, and the three constraints as shown in Figure 1.3. This model should be saved in a text file with a `.mod` extension.

The following AMPL session shows how to read the problem into AMPL, invoke the default solver, and print the solution. The minimum cost feed mixture is 2.86 pounds of Stark County feed and 5.71 pounds of Pioneer feed per pig per day, for a total cost of \$5.71 per pig per day.

```
var x1 >= 0;  # stark
var x2 >= 0;  # pioneer

minimize total_cost: .4*x1 + .8*x2;

subject to cal: 800*x1 + 1000*x2 >= 8000;
subject to vit: 140*x1 + 70*x2 >= 700;
subject to tox: x1 <= (1/3)*(x1 + x2);
```

Figure 1.3: AMPL model (`pigs.mod`)

```
ampl: model pigs.mod;
ampl: solve;
MINOS 5.51: optimal solution found.
2 iterations, objective 5.714285714
ampl: display x1, x2;
x1 = 2.85714
x2 = 5.71429
```


For problems of any reasonable size it is necessary (and desirable) to separate the model and data. This will allow us to solve different problem instances (i.e. different sets of data) without changing the model. Figure 1.4 shows the complete problem formulation and data with separate `model` and `data` sections in the same text file (`pigs2.mod`). It is also possible (and you will usually want to do this) to put the data section in its own text file. In that case we would name the data file with a `.dat` extension, e.g. `pigs2.dat`.

Most AMPL models are specified by declaring sets, parameters, and variables, and then by writing the objective function and the constraints that make use of the sets, parameters, and variables. Parameters are typically numeric values, although they can also be logical or symbolic values. They can be thought of as "data" values, e.g. the cost per pound of feed. Variables represent the decisions, e.g. the amount of feed to purchase. Sets specify the objects with which parameters and variables are associated. We declared the feed suppliers to be a set because the decisions for how much feed to purchase as well as cost and nutritional attributes of the feed are logically associated with each supplier. We say that the parameters and the variables are *indexed* over the set of suppliers.

Giving short, meaningful names to sets, parameters, and variables makes a model readable. I like to capitalize the first letter of set names so that they are easily distinguished from parameters. In the AMPL book, you will see that set names are in all upper-case letters. This is just a convention, and the language does not require it; however, good programming habits will make your code easier to maintain.

Here is the output from an AMPL session using the model and data in `pigs2.mod`. When working with AMPL, if you read in a model and subsequently make changes, or if you read in a new model after having worked with an initial model, you will need to use the `reset` command before reading in the new model. Note that when the decision variables are indexed over a set, we display all of them at once by typing the variable symbol.

```
model;
set Supplier;

param calories {Supplier};  # per pound
param vitamins {Supplier};  # units per pound
param cost {Supplier};      # dollars per pound

var x {Supplier} >= 0;  # pounds to purchase from each supplier (per pig)

minimize total_cost: sum {i in Supplier} cost[i]*x[i];

s.t. cal: sum {i in Supplier} calories[i]*x[i] >= 8000;
s.t. vit: sum {i in Supplier} vitamins[i]*x[i] >= 700;
s.t. tox: x['stark'] <= (1/3)* sum {i in Supplier} x[i];

data;
set Supplier := stark pioneer;

param : vitamins cost calories:=
stark   140      .4    800
pioneer  70      .8   1000;
```

Figure 1.4: Separation of model and data (`pigs2.mod`)

```
ampl: reset;
ampl: include pigs2.mod;
ampl: solve;
MINOS 5.51: optimal solution found.
2 iterations, objective 5.714285714
ampl: display x;
x [*] :=
pioneer  5.71429
  stark  2.85714
;
```

Factory Planning. This example is taken from [14]. A factory makes seven products that require various amounts of time on four different types of machines. The factory owns four grinders, two vertical drills, three horizontal drills, one boring machine, and one planing machine. For each product manufactured, the company can either sell the product (subject to market limitations) or hold the product in inventory at a cost of .5 per unit per month. We would like to develop a production and inventory plan for each of the next six months. We will not consider the sequence of machine operations; however, there is a fixed maintenance schedule that specifies when and how many of each machine type will be unavailable. The factory operates two eight-hour shifts each working day. There are 24 working days each month.

Refer to the data in figure 1.6 as well as the model in figure 1.5 while reading this description. Notice that the set `Month` is declared to be of type `ordered`, indicating a defined ordering among its (symbolic) members. This means that we can refer to the members of `Month` by their relative position. For example, in our data the expression `first(Month)` refers to the member `'jan'`. One common reason for declaring a set to be `ordered` is the need to refer to the previous and/or the next member in an indexed expression. A typical example of this use of an `ordered` set is illustrated by the `balance` constraints.

A defining feature of this model is use of three different variables to represent the quantities and timing for making, selling, and holding each product. The relationship among these variables is stated in the `balance` constraints. An upper bound of 100 units on the `hold` variable represents a storage limitation for each product in each month.

Our objective is to maximize the total profit of the factory over a period of six months. A unit profit is accrued for each item sold and a unit cost of .5 is incurred for each item held in inventory in a given month. Note that the parentheses surrounding the subtraction are necessary because the `sum` operator has higher precedence than `-`. As a side note, if we were to remove the holding cost from consideration in the objective then no parentheses would be required around the `profit[i]*sell[t,i]` term because the `sum` operator has lower precedence than `*` [4].

Each product requires a certain amount of processing time (possibly zero) on each type of machine. These requirements are specified in the data by the parameter `time_required` that is indexed over `Product` and `Machine` We need to specify constraints on the amount of machine time available each month. There are `work_hours` production hours available each month on each machine unless a machine is down for maintenance. Because the maintenance schedule is specific to each machine and month, the machine capacity constraints are indexed over `Month` and `Machine`. The total number of machine hours for all products must not

```
set Product;
set Machine;
set Month ordered;

param profit {Product} >= 0;
param time_required {Product,Machine} >= 0;
param num_available {Machine} integer, >= 0;
param downtime {Month,Machine} integer, >=0;
param market_limit {Month,Product} integer, >= 0;
param work_hours := 2*8*24;
# number of working hours in a month: 2 shifts of 8 hours each, 24 days/month

var make {Month,Product} >=0;  # how much of each product to make in each month
var sell {Month,Product} >=0;  # how much to sell
var hold {Month,Product} >=0, <=100;  # how much to hold

maximize total_profit:
  sum {t in Month, i in Product} (profit[i]*sell[t,i] - 0.5*hold[t,i]);

s.t. capacity {t in Month, m in Machine}:
sum {i in Product} time_required[i,m]*make[t,i]
  <= work_hours*(num_available[m]-downtime[t,m]);

s.t. marketing {t in Month, i in Product}: sell[t,i] <= market_limit[t,i];

s.t. balance {t in Month, i in Product : ord(t) > 1}:
  hold[prev(t),i] + make[t,i] = sell[t,i] + hold[t,i];
# on-hand inventory plus number produced must equal number
# sold plus number held in inventory for the next period

s.t. balance0 {i in Product}:
  make[first(Month),i] = sell[first(Month),i] + hold[first(Month),i];
# there is no inventory held over from december

s.t. end_inventory {i in Product}: hold[last(Month),i] = 50;
# stipulate that 50 of each product are to be held over from june
```

Figure 1.5: Model for factory planning problem (`factory_planning1.mod`)

exceed the time available in any particular month (respecting the `downtime` schedule.) There is an upper bound on the amount of each product that the market will absorb (i.e. that can be sold) each month.

We have three different variables to represent the decisions to make, sell, and hold product. The logical relationship among these variables is that in any particular month the amount sold plus the amount held in inventory must equal the amount produced plus any inventory from the previous month. When specifying these product balance constraints we must pay attention to the boundary conditions when indexing over `Month`. In our problem, we cannot refer to the previous month when the dummy index evaluates to `'jan'`. We can handle this in the main `balance` equations by placing a condition on the indexed set such that the order of the member is greater than 1 (and so `'jan'` is omitted.) This is possible because we declared the set `Month` to be of type `ordered`. We then need to specify a separate set of constraints for the first month (`balance0`). There is no beginning inventory and so the amount produced in `'jan'` equals the amount sold plus the amount held.

Instead of using the expression `make[first(Month),i]`, it would have been legitimate to refer to `make['jan',i]`; however, it's better practice to separate the model from the data. The model may then be applied to other problem instances with the same structure, but perhaps with a different starting month. Finally, we would like to end June with 50 of each product type in inventory. This is specified by the equality constraints named `end_inventory`.

1.2 Integer Programming

Perfect Matching. There are 10 students to be assigned to 5 dorm rooms. Each room holds exactly two students. For each pair of students, a value that indicates the desirability of placing that pair in the same room has been determined. Higher values correspond to better pairings. We would like to pair the students in such a way that the sum total of the values of each assigned pair is maximized. More generally, the perfect matching problem is to assign n pairs for $2n$ objects (with or without an objective function). The model and data are presented in figure 1.7.

`Pair` is a compound set with dimension two. Each member of this set is an ordered pair of students. We use the term "ordered" because the member (3,7) is distinct from the member (7,3). The `within` phrase tells us that the only allowed members in `Pair` are ordered pairs from the set `Student`. Such restrictions in the declaration of sets are encouraged in order to help detect errors in the data. For example, the following alternative declaration is perfectly acceptable, but using it would not allow the AMPL translator to recognize invalid pairs of students in the data section.

```
set Pair dimen 2;
```

The declaration of `Pair` is not strictly necessary. We could formulate the problem using only the set `Student`. However, using the compound set `Pair` makes the model easier to read and to maintain, particularly the indexing expressions. We specify a parameter `value` that is indexed over `Pair` to represent the desirability of matching up a particular pair of students. Our decision variables `x` are binary variables that are also indexed over `Pair`. `x[i,j]` will equal one if student `i` is matched with student `j`, and will equal zero otherwise.

```
data;
set Product := 1 2 3 4 5 6 7;
set Machine := grinder vdrill hdrill borer planer;
set Month := jan feb mar apr may jun;

param profit := 1 10 2 6 3 8 4 4 5 11 6 9 7 3;

param num_available :=
grinder 4 vdrill 2 hdrill 3 borer 1 planer 1;

param time_required (tr) :
         1   2   3   4   5   6   7 :=
grinder .5  .7   0   0  .3  .2  .5
vdrill  .1  .2   0  .3   0  .6   0
hdrill  .2   0  .8   0   0   0  .6
borer  .05 .03   0 .07  .1   0 .08
planer   0   0 .01   0 .05   0 .05;

param downtime :
     grinder vdrill hdrill borer planer :=
jan    1       0     0       0     0
feb    0       0     2       0     0
mar    0       0     0       1     0
apr    0       1     0       0     0
may    1       1     0       0     0
jun    0       0     1       0     1;

param market_limit :
       1     2     3     4     5     6     7 :=
jan  500  1000   300   300   800   200   100
feb  600   500   200     0   400   300   150
mar  300   600     0     0   500   400   100
apr  200   300   400   500   200     0   100
may    0   100   500   100  1000   300     0
jun  500   500   100   300  1100   500    60;
```

Figure 1.6: Data for factory planning problem (`factory_planning1.mod`)

```
set Student;
set Pair within Student cross Student;

param value {Pair};

var x {Pair} binary;

maximize total_value: sum {(i,j) in Pair} value[i,j] * x[i,j];

s.t. perfect_match {i in Student}:
        sum {(i,j) in Pair} x[i,j] + sum {(j,i) in Pair} x[j,i] = 1;

data;
set Student := 1 2 3 4 5 6 7 8 9 10;

param: Pair: value:
   1  2  3  4  5  6  7  8  9 10 :=
1  .  .  .  .  .  .  .  .  .  .
2  3  .  .  .  .  .  .  .  .  .
3  5  8  .  .  .  .  .  .  .  .
4  1 -4  7  .  .  .  .  .  .  .
5  2 -1  9  2  .  .  .  .  .  .
6  2  5  3  2  9  .  .  .  .  .
7  8  2  1  1  3 -2  .  .  .  .
8  2  3  3  4  5  1  1  .  .  .
9 13 -1  3  4  4 -5  2  2  .  .
10 1  2  6  6  7 -4  5  6  1  .;
```

Figure 1.7: Model and data for perfect matching problem (`roommates.mod`)

Notice the `binary` modifier in the declaration of `x`. This means that our decision variables may *only* take on the values zero or one and turns our formulation into one of pure integer programming [8].

The problem formulation itself consists of the objective function and a single set of constraints, one for each student. Our objective is to maximize the overall value of the assignments. It provides a good example of how to iterate over a compound set in AMPL. Notice that we must provide a pair of dummy variables `(i,j)` to index into `Pair`.

The constraints `perfect_match` state that each student must be matched with exactly one other student. To understand how this is accomplished note that the scope of the index `i` extends from its introduction in `i in Student` until the end of the statement marked by the semi-colon. In the expression `sum {(i,j) in Pair} x[i,j]`, `i` is held constant for a particular student (row of `value`) and the summation is over all room mates such that the pair of students represented by `(i,j)` exist in the set `Pair`, i.e. the summation is over columns. In the second expression with `i` and `j` interchanged, the summation is over rows. This is necessary due to the way that the data for `Pair` are structured. Notice that only the lower left portion of `value` is filled in.

This example demonstrates how to simultaneously define a set (`Pair`) and a parameter (`value`) that is indexed over that set. The "." symbols indicate missing values, e.g. there exists no member `(1,1)` in the set `Pair`. An AMPL session for solving this problem follows. Although not strictly necessary for this particular problem, we will instruct AMPL to use a solver that can handle integer programming problems, such as the solver `gurobi`.


```
ampl: option solver gurobi;
ampl: model roommates.mod;
ampl: solve;
Gurobi 3.0.0: optimal solution; objective 39
14 simplex iterations
ampl: display x;
x [*,*]
:    1   2   3   4   5   6   7   8   9    :=
2    0   .   .   .   .   .   .   .   .
3    0   1   .   .   .   .   .   .   .
4    0   0   0   .   .   .   .   .   .
5    0   0   0   0   .   .   .   .   .
6    0   0   0   0   1   .   .   .   .
7    0   0   0   0   0   0   .   .   .
8    0   0   0   1   0   0   0   .   .
9    1   0   0   0   0   0   0   0   .
10   0   0   0   0   0   0   1   0   0
;
```

The large number of zeros makes the solution difficult to read. A useful option is to only display the decision variables with nonzero values. Then we can easily see that the pairs are (3,2),(6,5),(8,4),(9,1), and (10,7).

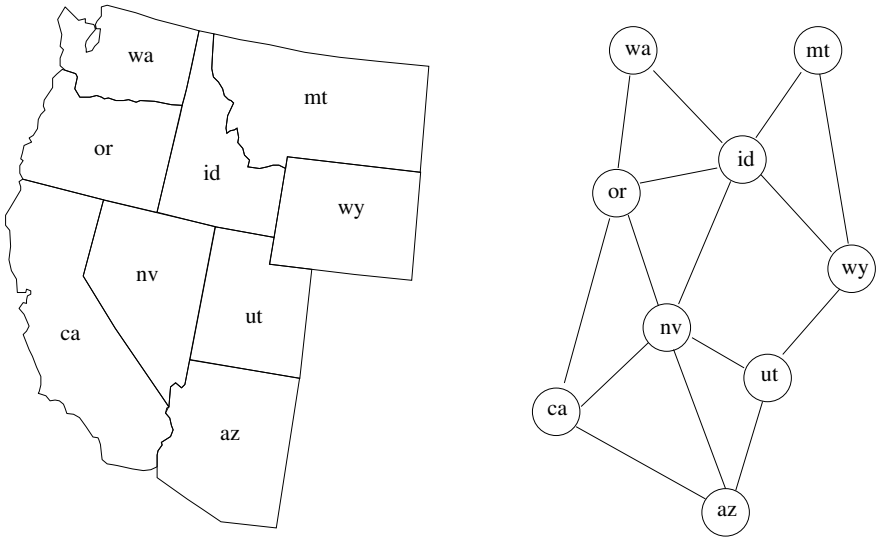


Figure 1.8: Western states and adjacency graph

```
ampl: option omit_zero_rows 1;
ampl: display x;
x :=
3  2   1
6  5   1
8  4   1
9  1   1
10 7   1
;
```


Map Coloring. The map coloring problem is to assign a color to each area on a map (e.g. state or county) in such a way that adjacent areas (i.e. those areas that share a border) are assigned different colors. Moreover, we would like to use as few colors as possible. The formulation for this problem was adapted from code written in GNU MathProg by Andrew Makhorin [6].

We may create a representation of the adjacency information by constructing a graph wherein each area on the map is represented by a node. An edge is drawn between two nodes if the respective areas on the map share a border. As an example, figure ? shows the map and the corresponding adjacency graph for the western United States. The graph representation will better facilitate the requirement that adjacent areas must be assigned different colors.

Both the model and the data are presented in figure 1.9. In this model we must make the cardinality of the set `Color` large enough to effect a feasible solution. For a large problem the minimum number of colors needed will not be obvious (otherwise there is no need to solve the problem.) However, if the solver finds that the problem is infeasible we can always add more members to the set `Color`. See Andrew Makhorin's code [6] for an implementation of a heuristic to obtain an upper bound on the required number of colors.

```
set Node;
set Edge within (Node cross Node);
set Color;
# the cardinality of the set Color needs to be large enough
# to find a feasible solution

var x {Node,Color} binary; # x[i,c] = 1 means that node i is assigned color c
var u {Color} binary;      # u[c] = 1 means that color c is used

minimize num_colors: sum {c in Color} u[c];

s.t. assignment {i in Node}: sum {c in Color} x[i,c] = 1;
# each node is assigned exactly one color

s.t. different {(i,j) in Edge, c in Color}: x[i,c] + x[j,c] <= u[c];
# adjacent nodes must be assigned different colors

data;
set Color := red blue green yellow orange;

set Node := ca or wa mt id nv wy ut az;

set Edge:
   ca or wa mt id nv wy ut az :=
ca -  +  -  -  -  +  -  -  +
or -  -  +  -  +  +  -  -  -
wa -  -  -  -  +  -  -  -  -
mt -  -  -  -  +  -  +  -  -
id -  -  -  -  -  +  +  +  -
nv -  -  -  -  -  -  -  +  +
wy -  -  -  -  -  -  -  +  -
ut -  -  -  -  -  -  -  -  +
az -  -  -  -  -  -  -  -  -;
```

Figure 1.9: Model and data for map coloring problem (`kcolor.mod`)

Our binary decision variables `x` will indicate the particular color assigned to each node (area on the map). We also create binary modeling variables `u` to indicate that a color has been used in the solution. Now the objective is a simple summation over `u`. There are three dummy variables in the indexing expression for the `different` constraint. Notice that even though `x` is indexed over {`Node,Color`} the indexing expression for `different` will only create one constraint for each existing `Edge` and `Color` combination. Since `Edge` is declared to be `within (Node cross Node)` we may safely use the dummy indices `i` and `j` when indexing into `x`.

In the `data` section we see one way to specify the members of a two-dimensional set: the "`+`" symbols indicate membership while the "`-`" symbols indicate non-membership. Alternatively, we could have specified `Edge` by providing only the ordered pairs.

```
set Edge :=
(ca,or)   (ca,az)   (or,id)   (wa,id)   (mt,wy)   (id,wy)   (nv,ut)   (wy,ut)
(ca,nv)   (or,wa)   (or,nv)   (mt,id)   (id,nv)   (id,ut)   (nv,az)   (ut,az);
```

1.3 Deterministic Dynamic Programming

1.4 Deterministic Inventory Models

1.5 Exercises

Linear Programming

1. *Feasible region for an LP.* Indicate graphically whether each of the following linear programs has a feasible solution. Graphically determine the optimal solution, if one exists, or show that no optimal solution exists.

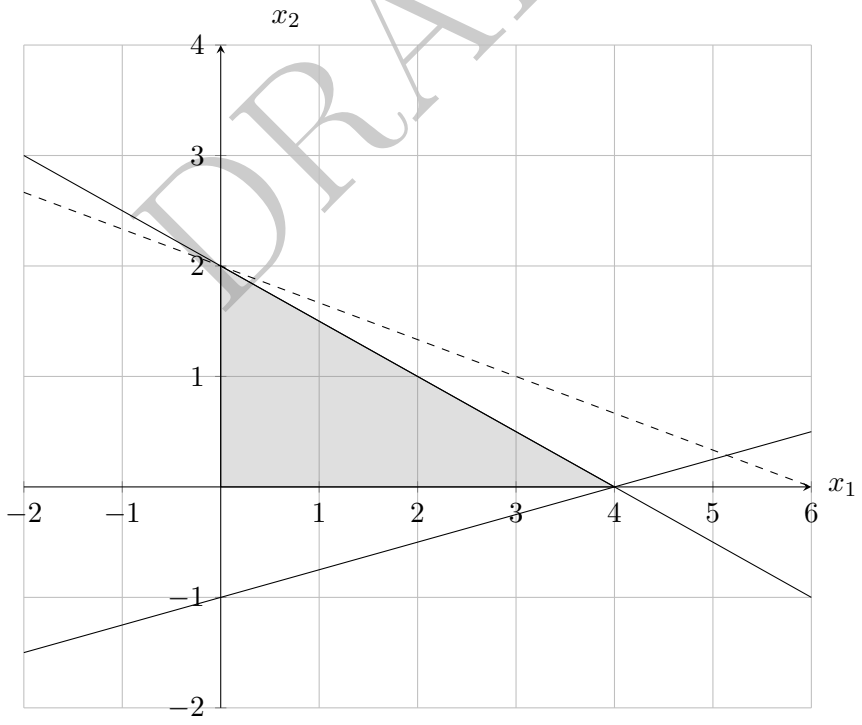
(a)

$$\begin{array}{lrcrcl} \text{maximize} & x_1 & + & 3x_2 & & \\ \text{subject to} & x_1 & - & 4x_2 & \leq & 4 \\ & x_1 & + & 2x_2 & \leq & 4 \\ & x_1, x_2 & & & \geq & 0 \end{array}$$

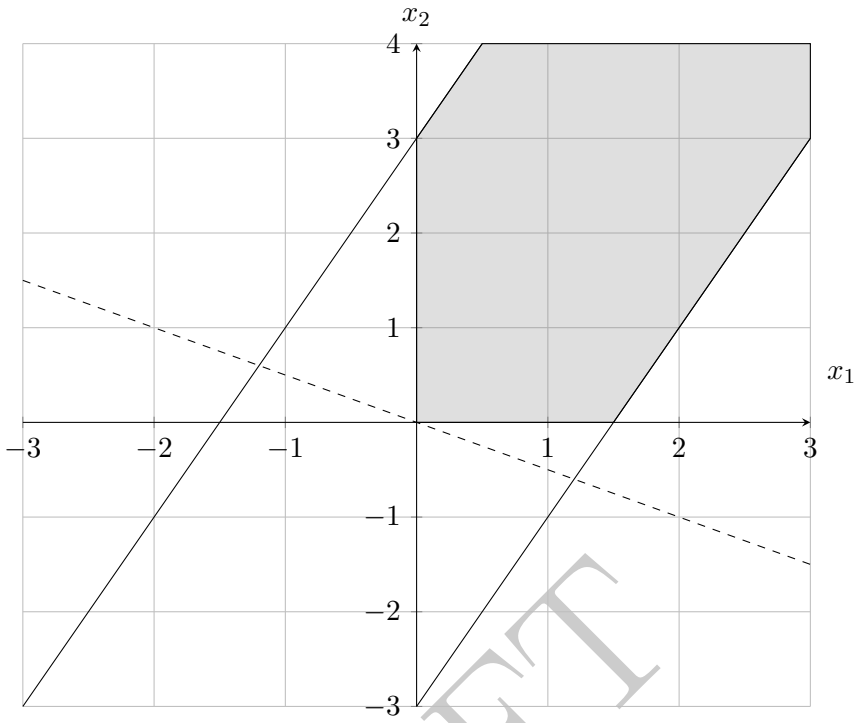
(b)

$$\begin{array}{lrcrcr} \text{minimize} & x_1 & + & 2x_2 & & \\ \text{subject to} & 2x_1 & - & x_2 & \leq & 3 \\ & 2x_1 & - & x_2 & \geq & -3 \\ & x_1, x_2 & & & \geq & 0 \end{array}$$

Solution. The feasible region for the linear program in part a) is shown below. The objective function is plotted as a dashed line. The optimal solution occurs at $x_1 = 0$, $x_2 = 2$, and value of the objective function at the optimal solution is $z = 6$.



The feasible region for the linear program in part b) is unbounded; however, since it is a minimization problem, it has an optimal solution at $x_1 = 0$, $x_2 = 0$. The value of the objective function at the optimal solution is $z = 0$.



2. *Karl's garden.* Karl is a local gardener who grows lettuce, broccoli, and carrots. His garden is 400 ft^2, and each individual plant of lettuce, broccoli, and carrots occupies 0.5 ft^2, 1.2 ft^2, and 0.25 ft^2, respectively. Karl recently learned that many of his neighbors are interested in purchasing his vegetables. After polling the neighborhood, he learned that he has demand for 350 lettuce plants, 350 broccoli plants, and 250 carrot plants. Karl estimates that he will earn a profit of \$9.50 per lettuce plant, \$12.80 per broccoli plant, and \$8.25 per carrot plant. How many plants of each type of vegetable should Karl grow to maximize profit? Solve this problem using optimization software.

   Solution. A GAMS model is provided in the file `karls-garden.gms`. The solution indicates that Karl should grow 350 lettuce plants, 135 broccoli plants, and 250 carrot plants for a total profit of about \$7120. Note that we are ignoring any fractional values of the decision variables.

3. *Workplace safety.* A manufacturing company has assembled a safety committee to reduce the number of injuries sustained by employees at work. The company has allotted the committee a budget of \$100 each week for purchasing items that will increase employee safety. After analyzing past incidents, the safety committee has concluded that the most common work hazard is exposure to loud noises. As a result, the committee would like to purchase (i) earplugs and (ii) other PPE (personal protection equipment). The relative value (utility) of the safety items was investigated, and it was determined that earplugs provide 1.2 units of value for each dollar spent on other types of PPE. The safety committee would like to determine how to spend the budget to maximize the safety of employees, but the company does not want to

spend more than \$70 on earplugs and \$50 on other PPE each week. The committee also has the option to save part of the money. Additionally, the company would like to know:

(a) How the total value of its expenditures for safety would change if there were only \$99 to spend.

(b) How the total value would change if \$75 could be spent on earplugs.

(c) Whether it would save any money if each dollar of savings would provide 1.1 units of value for each dollar spent on other PPE.

Formulate the safety committee's spending decision as a linear program. Sketch the feasible region and determine the optimal solution using the graphical method. Then solve the linear program using optimization software and obtain post-optimality output (i.e. a sensitivity report). Use this output to answer each of the questions 3a, 3b, 3c.

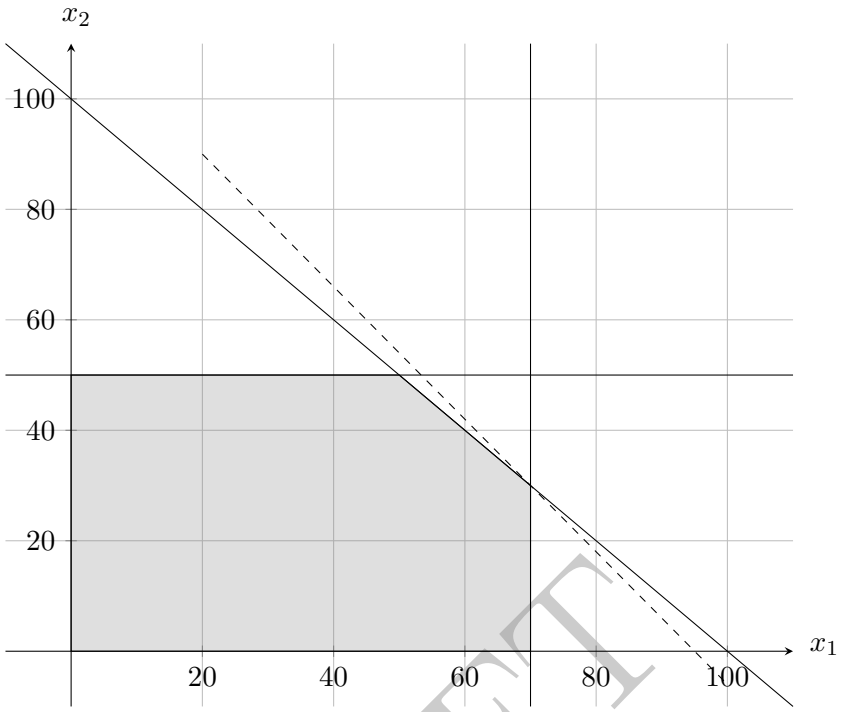
Solution. In the problem formulation to maximize total value, let

$$\begin{aligned} x_1 &= \text{\$ to spend on earplugs} \\ x_2 &= \text{\$ to spend on other PPE} \end{aligned}$$

Then the committee wants to solve the following problem.

$$\begin{array}{lrcrcr} \text{maximize} & 1.2x_1 & + & x_2 & & \\ \text{subject to} & x_1 & + & x_2 & \leq & 100 \\ & x_1 & & & \leq & 70 \\ & & & x_2 & \leq & 50 \\ & & & x_1, x_2 & \geq & 0 \end{array}$$

The feasible region is given below. The optimal solution is $x_1 = 70$ and $x_2 = 30$ for a total value of 114.



A GAMS model is provided in `workplace-saftey.gms`. Refer to the listing file `workplace-saftey.lst`, to answer questions 3a, 3b, and 3c. Regarding part 3a, the shadow price on the `budget` constraint tells us that if we decrease the RHS from 100 to 99, the value of the objective function will decrease to 113. Since earplugs provide more value than other PPE, the new solution will be $x_1 = 70$ and $x_2 = 29$. For part 3b, notice that \$75 is within the range for the shadow price on the `earplugs` constraint. So, increasing the RHS to 75 will increase the total value by

$$\$5 \times 0.2 \text{ units of value per dollar} = 1$$

for a total value of 115. Finally, for part 3c, saving one dollar requires \$1; however, the company gets 1.1 units of value for each dollar saved. Yes, the company would save *some* amount of money and spend less on other PPE. To use the shadow price to answer this question, notice that the shadow price on the `budget` constraint is 1. If we "price-out" the new activity of saving money, we see that it cost \$1 per unit, but it returns \$1.1 per unit in total value. So the company would save some amount of money. The question did not ask us to determine the new solution. It only asked whether *any* money would be allocated to savings.

4. *A blending problem.* A pet food manufacturer is developing a new dog food recipe with natural ingredients. It has been decided that the recipe will contain a combination of 4 ingredients: chicken, brown rice, vegetables, and corn meal. The cost and important nutritional information for each ingredient is summarized in the table below.

Ingredient	Cost ($/lb)	Protein (g/lb)	Fat (g/lb)	Fiber (g/lb)
(1) Chicken	3.00	125	60	0
(2) Brown rice	0.75	12	4	9
(3) Vegetables	1.80	14	1.5	12
(4) Corn meal	0.60	32	8	17

Industrial engineers have been asked to determine the most cost-effective mixture of ingredients given the following guidelines:

- Each pound of dog food must contain no more than 40 grams of fat.
- Each pound of dog food must contain at least 80 grams of protein.
- Each pound of dog food must contain at least 4 grams and no more than 10 grams of fiber.
- Each ingredient must comprise at least 10% of the mixture.

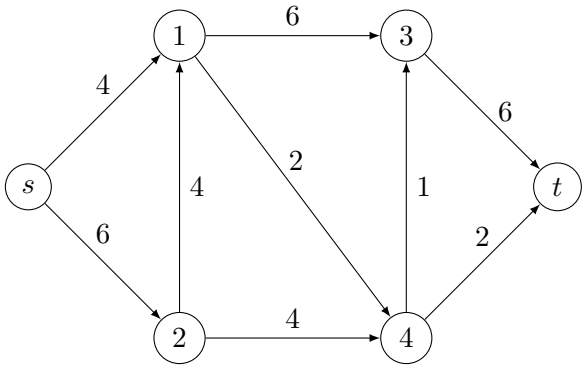
(a) Formulate an (algebraic) mathematical programming model using the information above and generate a solution using AMPL. Your problem formulation can be a model for the minimum-cost way to produce one pound of dog food. We can consider that to be the "recipe" that would then be scaled up for production purposes.

(b) Formulate an alternative (algebraic) mathematical programming model given the additional constraint that 80% of the mixture must contain a combination of chicken and vegetables. Generate a solution using AMPL and determine the added cost due to the new constraint.

Solution. For part 4a, an AMPL model is provided in the file `dog-fooda.mod`. The mixture should contain 55.7% chicken, 10% brown rice, 10% vegetables, and 24.3% corn meal for a cost of $2.07 per pound.

For part 4b, An AMPL model is provided in `dog-foodb.mod`. The added cost due to the additional constraint is $0.20. The mixture should contain 58% chicken, 10% brown rice, 22% vegetables, and 10% corn meal for a cost of $2.27 per pound.

5. *Maximum flow through a network.* An industrial engineer is analyzing the drainage system at a food processing plant. The arrows in the following network indicate the direction of flow in the system, and the numbers indicate the capacity of each pipe. Formulate a linear programming problem to determine the maximum flow from source node s to sink node t.



Solution. Add an artificial edge from sink node t to source node s that has unlimited capacity. Let x_{ij} be the amount of flow from node i to node j for each directed edge i, j that exists in the network. The problem formulation is

$$
\begin{array}{lrcrcrcl}
\text{maximize} & x_{ts} & & & & & & \\
\text{subject to} & & & x_{s1} & + & x_{s2} & - & x_{ts} & = & 0 \\
& x_{13} & + & x_{14} & - & x_{s1} & - & x_{21} & = & 0 \\
& & & x_{24} & + & x_{21} & - & x_{s2} & = & 0 \\
& & & x_{3t} & - & x_{13} & - & x_{43} & = & 0 \\
& x_{43} & + & x_{4t} & - & x_{14} & - & x_{24} & = & 0 \\
& & & x_{ts} & - & x_{3t} & - & x_{4t} & = & 0 \\
& & & & & & & x_{s1} & \leq & 4 \\
& & & & & & & x_{s2} & \leq & 6 \\
& & & & & & & x_{13} & \leq & 6 \\
& & & & & & & x_{21} & \leq & 4 \\
& & & & & & & x_{14} & \leq & 2 \\
& & & & & & & x_{24} & \leq & 4 \\
& & & & & & & x_{43} & \leq & 1 \\
& & & & & & & x_{4t} & \leq & 2 \\
& & & & & & & x_{3t} & \leq & 6 \\
& & & & & & x_{ij} & & \geq & 0 \quad \text{for all edges } i, j
\end{array}
$$

6. *Golf lessons.* Amy and Brian are professional golfers who have each agreed to donate 10 hours of private golf lessons to a charity auction. Three people have bid on the lessons, and their bids are shown in the table below. For example, Emma has bid \$28 per hour to receive lessons from Amy.

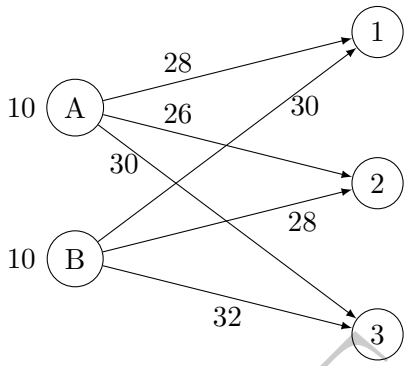
Bidder	Amy	Brian
Emma	\$28/hr	\$30/hr
Dan	\$26/hr	\$28/hr
Sam	\$30/hr	\$32/hr

In the spirit of fairness, the auction committee has decided that no bidder can win more than 8 hours of total instruction. Given the bid amounts, the committee must now decide how to allocate the 20 hours of available instruction time.

(a) Draw a diagram of the problem as a network flow model.

(b) Formulate a linear programming model to maximize the charity's revenue.
(c) Solve the problem using optimization software.

Solution. A diagram of the network flow model is shown below. Bidders 1, 2, and 3 correspond to Emma, Dan, and Sam, respectively.



Let x_{ij} be the number of hours of instruction bidder j receives from professional golfer i.

$$i \in \{\mathrm{A}, \mathrm{B}\} \qquad j \in \{1, 2, 3\}$$

The problem formulation is

$$\begin{aligned}
&\text{maximize} \\
&28x_{\mathrm{A1}} + 26x_{\mathrm{A2}} + 30x_{\mathrm{A3}} + 30x_{\mathrm{B1}} + 28x_{\mathrm{B2}} + 32x_{\mathrm{B3}} \\
&\\
&\text{subject to} \\
&\qquad x_{\mathrm{A1}} + x_{\mathrm{A2}} + x_{\mathrm{A3}} = 10 \\
&\qquad x_{\mathrm{B1}} + x_{\mathrm{B2}} + x_{\mathrm{B3}} = 10 \\
&\qquad x_{\mathrm{A1}} + x_{\mathrm{B1}} \leq 8 \\
&\qquad x_{\mathrm{A2}} + x_{\mathrm{B2}} \leq 8 \\
&\qquad x_{\mathrm{A3}} + x_{\mathrm{B3}} \leq 8 \\
&\qquad x_{ij} \geq 0 \quad \text{for } i \in \{\mathrm{A}, \mathrm{B}\},\ j \in \{1, 2, 3\}
\end{aligned}$$

The optimal solution is: bidder 1 (Emma) receives 6 hours of instruction with Amy and 2 hours of instruction with Brian, bidder 2 (Dan) receives 4 hours of instruction with Amy, and bidder 3 (Sam) receives 8 hours of instruction with Brian. The charity receives \$588 in revenue. A GAMS model and solution is provided in `golf-lessons.gms`. *Note:* There are multiple optimal solutions to this problem, so the values of the decision variables in your solution could differ, but the value of the objective function will be the same.

7. *A supply chain problem.* A company will produce the same new product at two different factories, and then the product will be shipped to two warehouses. Factory 1 can send an unlimited amount by rail to warehouse 1 only, whereas factory 2 can send an unlimited amount by rail to warehouse 2 only. However, independent truckers can can be used to ship up to 50 units from each factory to a distribution center (DC),

from which up to 50 units can be shipped to each warehouse. The shipping cost for each alternative is shown in the following table, along with the amounts to be produced and the amounts needed at the warehouses.

From \ To	Unit shipping cost			
		Warehouse		
	DC	WH1	WH2	supply
Factory 1	3	7	–	80
Factory 2	4	–	9	70
DC		2	4	
demand		60	90	

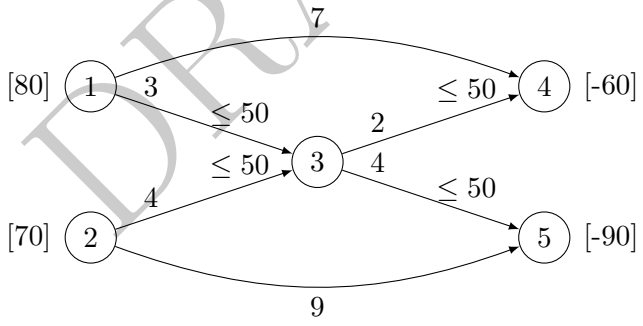
(a) Draw the network representation of this problem as a transshipment problem.

(b) Formulate the linear programming problem that minimizes total shipping cost.

(c) Solve this problem using an algebraic modeling language such as GAMS.

Solution.

To make the notation easier, assign numbers to the nodes in the network as follows: node 1 is factory 1, node 2 is factory 2, node 3 is the DC, node 4 is warehouse 1, and node 5 is warehouse 2. Let x_{ij} be the number of products shipped from node i to node j.



$$
\begin{array}{lrcrcrcrcrcl}
\text{minimize} & 3x_{13} & + & 4x_{23} & + & 7x_{14} & + & 9x_{25} & + & 2x_{34} & + & 4x_{35} \\
\text{subject to} & & & & & & & x_{13} & + & x_{14} & = & 80 \\
& & & & & & & x_{23} & + & x_{25} & = & 70 \\
& & & & & & & x_{34} & + & x_{14} & = & 60 \\
& & & & & & & x_{25} & + & x_{35} & = & 90 \\
& & & & & & & & & x_{13} & \leq & 50 \\
& & & & & & & & & x_{23} & \leq & 50 \\
& & & & & & & & & x_{34} & \leq & 50 \\
& & & & & & & & & x_{35} & \leq & 50 \\
& & & x_{13} & + & x_{23} & - & x_{34} & - & x_{35} & = & 0 \\
& & & & & & & & & x_{ij} & \geq & 0
\end{array}
$$

We note that the flow balance constraint at the DC is unnecessary because total supply equals total demand.

8. *A work-holding material for manufacturing.* A material compound is used in a manufacturing process to hold parts while they are being machined. The material flows when heated and becomes solid when cooled. An industrial engineer must determine a method for maintaining the quality of the material over a four-week period, while minimizing cost. The engineer knows that $r_j \geq 0$ pounds of the material will be required in each of the four weeks, $j = 1, 2, 3, 4$. This demand can be met by either purchasing new material at cost P dollars per pound or by filtering the old material and reusing it. Two options are available for filtering: normal service, which requires 1 full week at a cost of N dollars per pound, and expedited service, which allows material used during one week to be filtered by the start of the next week, at a cost of Q dollars per pound. Currently, the manufacturing company has no work-holding material available.

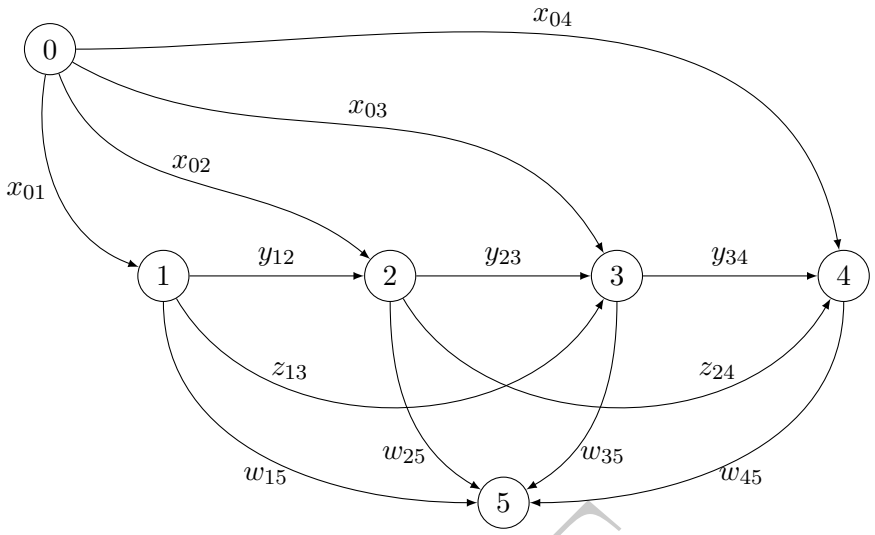
   (a) Formulate a network flow problem that will allow the company to satisfy the demand for the work-holding material at minimal cost. You should draw a diagram of the network to help you formulate the problem. Note that because the company can purchase new work-holding material, used material can be discarded at the end of a week, if it is economical to do so.

   (b) Now suppose that used material may be filtered and re-used only once. How does the model change?

   Solution. Define the following variables.

x_{0j}	pounds (lbs) of new material purchased for the beginning of week j
y_{ij}	lbs material for expedited filtering from end of week i to beginning of week j
z_{ij}	lbs material for normal filtering from end of week i to beginning of week j
w_{i5}	lbs material discarded at the end of week i

   The network flow can be depicted as follows. Node 0 represents the beginning of the four-week period. Nodes 1 through 4 represent weeks 1 through 4. Node 5 represents the sink node for discarded material.



The mathematical programming problem is to

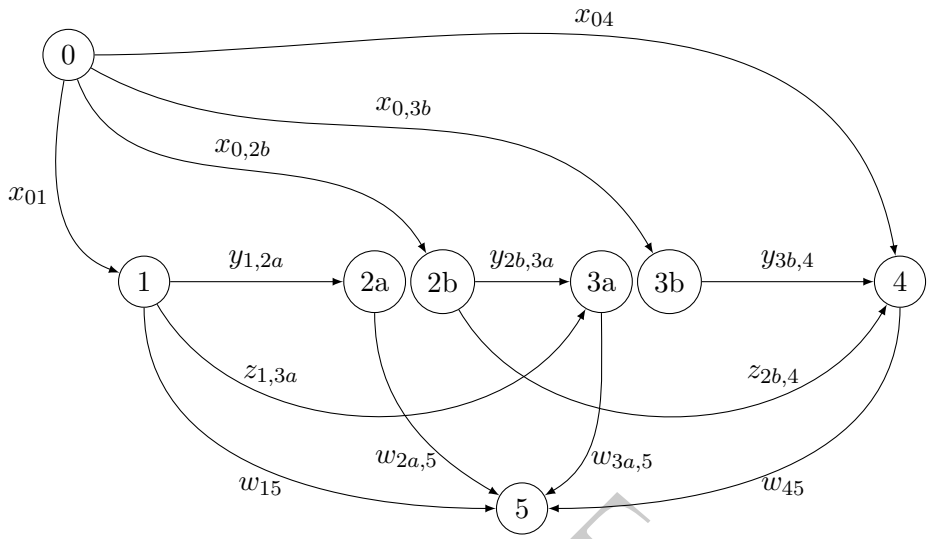
$$
\begin{array}{lll}
\text{minimize} & P(x_{01}+x_{02}+x_{03}+x_{04})+Q(y_{12}+y_{23}+y_{34})+N(z_{13}+z_{24}) & \\
\text{subject to} & x_{01} \geq r_1 & \text{(week 1)} \\
& x_{02}+y_{12} \geq r_2 & \text{(week 2)} \\
& x_{03}+y_{23}+z_{13} \geq r_3 & \text{(week 3)} \\
& x_{04}+y_{34}+z_{24} \geq r_4 & \text{(week 4)} \\
& x_{01} = w_{15}+y_{12}+z_{13} & \text{node 1} \\
& x_{02}+y_{12} = w_{25}+y_{23}+z_{24} & \text{node 2} \\
& x_{03}+y_{23}+z_{13} = w_{35}+y_{34} & \text{node 3} \\
& x_{04}+y_{34}+z_{24} = w_{45} & \text{node 4} \\
& x_{ij},\ y_{ij},\ z_{ij},\ w_{ij} \geq 0 \text{ for each edge } i,j \text{ in the network.} &
\end{array}
$$

The first four constraints pertain to the material requirements for weeks 1 through 4. The next four constraints are the flow balance constraints. They require that the material that enters the beginning of a week must exit at the end of the week.

If the work-holding material can be filtered only once, then we require that all incoming flow on edges that represent filtering (i.e. y_{ij} and z_{ij}) must be routed to node 5 (the discard node). To accomplish this, we can add the following constraints to the existing model formulation.

$$
\begin{aligned}
w_{25} &\geq y_{12} \\
w_{35} &\geq y_{23}+z_{13}
\end{aligned}
$$

Another idea is to split nodes 2 and 3 as depicted in the diagram below. All flow into nodes 2a and 3a must be discarded, but is allowed to help satisfy the demand requirements.



Now the objective function and demand requirements are

$$\begin{array}{lrcl}
\text{minimize} & P(x_{01}+x_{0,2b}+x_{0,3b}+x_{04})+Q(y_{1,2a}+y_{2b,3a}+y_{3b,4})+N(z_{1,3a}+z_{2b,4}) & & \\
\text{subject to} & x_{01} \geq r_1 & & \text{(week 1)} \\
& x_{0,2b}+y_{1,2a} \geq r_2 & & \text{(week 2)} \\
& x_{0,3b}+y_{2b,3a}+z_{1,3a} \geq r_3 & & \text{(week 3)} \\
& x_{04}+y_{3b,4}+z_{2b,4} \geq r_4 & & \text{(week 4)} \\
& x_{01} = w_{15}+y_{1,2a}+z_{1,3a} & & \text{node 1} \\
& y_{1,2a} = w_{2a,5} & & \text{node 2a} \\
& x_{0,2b} = y_{2b,3a}+z_{2b,4} & & \text{node 2b} \\
& y_{2b,3a}+z_{1,3a} = w_{3a,5} & & \text{node 3a} \\
& x_{0,3b} = y_{3b,4} & & \text{node 3b} \\
& x_{04}+y_{3b,4}+z_{2b,4} = w_{45} & & \text{node 4} \\
& x_{ij},\ y_{ij},\ z_{ij},\ w_{ij} \geq 0 \text{ for each edge } i,j \text{ in the network.} & &
\end{array}$$

9. *Laundering clean-room garments.* A medical device company uses a cleanroom to assemble and inspect many of its products. Because the purpose of a cleanroom is to limit pollutants and contaminants, there are strict guidelines that must be followed. At this company, each employee and visitor who enters the cleanroom must wear a clean lab coat. As a result, there must be enough clean lab coats at the beginning of each day for all employees and visitors who enter the cleanroom. The company outsources cleaning to an laundry service. Normal laundry takes one full day at a cost of \$2.25 per lab coat. Overnight laundry service can be done at a cost of \$5 per lab coat. Under normal circumstances, the current supply of 40 lab coats is sufficient for complete dependence upon the normal laundry service. However, the normal laundry service will not be sufficient for the next three days due to a training program and audit that will take place. It is known that the requirements for the next three days will be 25, 35, and 30 lab coats. It is now mid-afternoon on the day before the

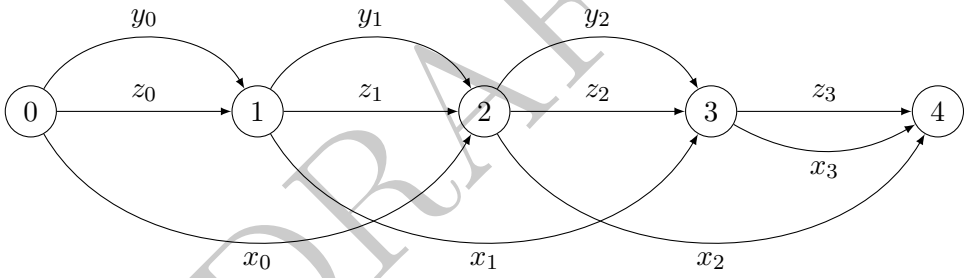
training program and audit. 15 lab coats are clean, and 25 are ready for laundry. It is against clean-room protocol to have used lab coats remain at the company overnight. The clean-room will be closed on the day after the training program and audit, so all used lab coats on the third day can be sent to normal laundry. An industrial engineer wants to determine a plan for laundering that will minimize total cost while satisfying demand and respecting the clean-room protocol.

Interpret this problem as one of network flow. Draw the network and formulate the LP (i.e. the decision variables, the objective function, and the constraints.)

Solution. Define the following variables.

x_i number of garments sent to regular laundry at the end of day i
y_i number of garments sent to overnight service at the end of day i
z_i number of clean (unused) garments at the end of day i

The network flow diagram can be depicted as shown below. Node 0 represents the current day (before the three-day event). Nodes 1, 2, and 3 represent the next three days of the higher-than-usual demand. Node 4 represents the day following the three-day event. For example, the flow labeled x_0 represents the garments that are sent to the normal laundry service today. Those garments will be available at the beginning of day 2.



The formulation of the mathematical programming problem is

$$
\begin{array}{lrcl}
\text{minimize} & 5\,(y_0 + y_1 + y_2) + 2.25\,(x_0 + x_1 + x_2 + x_3) & & \\
\text{subject to} & y_0 + z_0 + x_0 = 40 & & (1)\\
& y_0 + z_0 = y_1 + z_1 + x_1 & & (2)\\
& y_1 + z_1 + x_0 = y_2 + z_2 + x_2 & & (3)\\
& y_2 + z_2 + x_1 = x_3 + z_3 & & (4)\\
& y_0 + z_0 \geq 25 & & (5)\\
& y_1 + z_1 + x_0 \geq 35 & & (6)\\
& y_2 + z_2 + x_1 \geq 30 & & (7)\\
& y_0 + x_0 \geq 25 & & (8)\\
& y_1 + x_1 \geq 25 & & (9)\\
& y_2 + x_2 \geq 35 & & (10)\\
& x_3 \geq 30 & & (11)\\
& x_i, y_i, z_i \geq 0 & &
\end{array}
$$

Constraints 1 – 4 are flow balance constraints for nodes 0, 1, 2, and 3, respectively. Constraints 5, 6, and 7 are the requirements for the number of clean garments during

the three-day event. Constraints 8 – 11 model the clean-room protocol that used garments must be sent to a laundry service (either normal or overnight) at the end of the day. Note that at the end of day 3, it would not make sense to use the overnight laundry service.

Integer Programming

10. *Discrete purchases.* Sam is considering purchasing some combination of a new car, a washer, a dryer, and a high-definition television (HDTV). For practical reasons, Sam knows that he can only buy the dryer if he also buys the washer. It could still make sense to buy the washer without the dryer because Sam could dry his clothes on a clothesline. Due to political reasons with Sam's spouse, he can only purchase the HDTV if he does not purchase the car. Taking everything into consideration, Sam has estimated the values for each item. The item utilities (converted to dollars) and purchase prices in dollars are shown in the table below. Sam's total budget is \$16,000. Formulate a binay integer programming problem to maximize Sam's estimated value while respecting the practicality, spouse, and budget constraints mentioned above.

Item	Estimated value	Price
car	\$9,500	\$12,000
washer	\$3,000	\$2,000
dryer	\$3,000	\$2,000
HDTV	\$5,000	\$3,000

Solution.

$$\text{Let } x_j = \begin{cases} 1 & \text{if item } j \text{ is purchased} \\ 0 & \text{otherwise} \end{cases}$$

where the index $j = 1, 2, 3, 4$ corresponds to the car, washer, dryer, and HDTV. The problem is to

$$\begin{array}{lrcrcrcrcl} \text{maximize} & 9.5x_1 & + & 3x_2 & + & 3x_3 & + & 5x_4 & & \\ \text{subject to} & 12x_1 & + & 2x_2 & + & 2x_3 & + & 3x_4 & \leq & 16 \\ & & & & & & & x_3 & \leq & x_2 \\ & & & & & x_1 & + & x_4 & \leq & 1 \\ & & & & & & & x_j & \in & \{0,1\} \end{array}$$

11. *A set covering problem.* A general merchandise retailer is planning to expand into a new metropolitan area comprised of seven cities. The following table shows the distance (in miles) between each city.

	To						
From	City 1	City 2	City 3	City 4	City 5	City 6	City 7
City 1	0	15	20	38	20	28	22
City 2	15	0	5	23	16	12	10
City 3	20	5	0	18	12	13	14
City 4	38	23	18	0	31	19	27
City 5	20	16	12	31	0	7	19
City 6	28	12	13	19	7	0	22
City 7	22	10	14	27	19	22	0

The company assumes that customers will only visit a retail location if it is within 15 miles of the city in which the customer lives. Using this information, the company would like to construct the fewest number of stores while ensuring that they can serve every customer in the metropolitan area. Formulate an integer programming model to determine the cities where retail locations should be constructed.

Solution.

$$\text{Let } x_i = \begin{cases} 1 & \text{if a retail location is constructed in city } i, \\ 0 & \text{otherwise.} \end{cases}$$

To model the requirement that there is at least one retail location within 15 miles of city 1, we require that $x_1 + x_2 \geq 1$. The full problem formulation is

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^{7} x_i \\ \text{subject to} \quad & x_1 + x_2 \geq 1 \\ & x_1 + x_2 + x_3 + x_6 + x_7 \geq 1 \\ & x_2 + x_3 + x_5 + x_6 + x_7 \geq 1 \\ & x_4 \geq 1 \\ & x_3 + x_5 + x_6 \geq 1 \\ & x_2 + x_3 + x_5 + x_6 \geq 1 \\ & x_2 + x_3 + x_7 \geq 1 \\ & x_i \in \{0,1\} \quad i = 1, \ldots, 7 \end{aligned}$$

12. *A set partitioning problem.* A DSL (Digital Subscriber Line) Internet service provider is determining where to construct three central offices in a city. A customer's DSL connection works better when they are near a central office, so customers that live close to a new central office will experience faster Internet speeds. The city is divided into six geographical regions, and an industrial engineer must decide which regions should have central offices. The engineer must also determine which regions each central office will serve. The following table indicates the average internet speed in a region, depending on the region of the central office that provides service.

location of central office	avg internet speed in region 1	2	3	4	5	6
1	45	35	30	25	10	12
2	35	45	20	20	25	18
3	30	20	45	30	15	25
4	25	20	30	45	12	27
5	10	25	15	12	45	30
6	12	18	25	27	30	45

Formulate an integer programming model to determine the regions in which central offices are located and the regions to which each central is assigned so that the average internet speed in the city is maximized.

Solution. Let a_{ij} be the data describing the average internet speed in region j when served by a central office in region i. Also, let the decision variables be

$$x_{ij} = \begin{cases} 1 & \text{if central office in region } i \text{ provides service to region } j \\ 0 & \text{otherwise.} \end{cases}$$

and

$$y_i = \begin{cases} 1 & \text{if central office is constructed in region } i \\ 0 & \text{otherwise.} \end{cases}$$

The problem is to

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{6} y_i = 3 \\ & \sum_{i=1}^{6} x_{ij} = 1 \text{ for all } j \\ & x_{ij} \le y_i \text{ for all } i \text{ and } j \\ & x_{ij} \in \{0,1\} \\ & y_i \in \{0,1\} \\ & i = 1,\ldots,6 \\ & j = 1,\ldots,6 \end{aligned}$$

The objective function will maximize the average Internet speed of the six regions. We could include the constant 1/6, which would change the value of the objective function, but not the values of the decision variables.

13. *A problem with a fixed charge.* A bicycle manufacturing company can produce three types of bikes: mountain bikes, road bikes, and fat bikes. The company rents specific machinery for producing each type of bike. It costs \$150 per week to rent machinery for mountain bikes, \$250 per week to rent machinery for road bikes, and \$300 per week to rent machinery for fat bikes. A maximum of 120 hours of labor and 11,000

pounds of material (aluminum) are available each week for production. The material and labor requirements to produce one unit of each type of bike are shown in the table below. The unit variable cost and the unit selling price for each type of bike are also shown.

	Labor (hours)	Material (pounds)	Variable cost	Selling price
Mtn bike	2.5	28	\$350	\$550
Road bike	7	25	\$420	\$800
Fat bike	4.5	36	\$680	\$1200

The machinery needed to produce a specific type of bike is only rented if that type of bike is produced. Assume that the company can sell all of the bikes that it produces, regardless of type. Formulate a mixed integer linear programing problem to determine the weekly production quantities of mountain bikes, road bikes, and fat bikes that will maximize profit.

Solution. Let the decision variables x_1, x_2, x_3 be the number of mountain bikes, road bikes, and fat bikes, respectively, to produce each week. Machinery rental costs are only incurred if bicycles of that particular type are produced. We introduce binary decision variables y_i to apply the machinery rental costs when production x_i is positive.

$$\text{Let } y_i = \begin{cases} 1 & \text{if } x_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, 3.$$

The mathematical formulation of the Integer Programming problem is

$$\begin{array}{lrcrcrcl} \text{maximize} & 550x_1 & + & 800x_2 & + & 1200x_3 & & \\ & -\ 350x_1 & - & 420x_2 & - & 680x_3 & & \\ & -\ 150y_1 & - & 250y_2 & - & 300y_3 & & \\ & & & & & & & \\ \text{subject to} & 2.5x_1 & + & 7x_2 & + & 4.5x_3 & \leq & 120 \\ & 28x_1 & + & 25x_2 & + & 36x_3 & \leq & 11{,}000 \\ & & & & & x_1 & \leq & My_1 \\ & & & & & x_2 & \leq & My_2 \\ & & & & & x_3 & \leq & My_3 \end{array}$$

where M is a large number, $x_i \geq 0$, x_i are integer-valued, $y_i \in \{0, 1\}$, and $i = 1, 2, 3$.

14. *A maxi-min problem.* A college student has received a \$50 gift and is interested in four different investment options. One of three possible random events will occur in the next year that will impact the return on the investments. The return for each dollar invested in each of the four options for each event is shown in the table below. For example, if the student invests \$1 in option 3 and event B occurs, the student will receive a return of \$7. Investment amounts must be in increments of \$1. The student is risk-averse; his objective is to maximize the minimum return. Formulate an integer programming problem and obtain a solution. *Hint:* The decision variables are the amounts to invest in each option. Introduce another variable (say w) that represents the minimum return. Then, if w is the minimum return, the actual return for each possible outcome must be greater than or equal to w.)

	event		
option	A	B	C
1	-4	6	5
2	4	-2	-11
3	-5	7	12
4	12	6	-4

Solution. Let the decision variables x_i represent the amounts to invest in each option, where $i = 1, 2, 3, 4$. Also, let w represent the minimum return among the three events. The student wants to

$$\begin{array}{lrcrcrcrcrcr} \text{maximize} & w & & & & & & & & & & \\ \text{subject to} & -4x_1 & + & 4x_2 & - & 5x_3 & + & 12x4 & - & 50 & \geq & w \\ & 6x_1 & - & 2x_2 & + & 7x_3 & + & 6x4 & - & 50 & \geq & w \\ & 5x_1 & - & 11x_2 & + & 12x_3 & - & 4x4 & - & 50 & \geq & w \\ & & & x_1 & + & x_2 & + & x_3 & + & x_4 & = & 50 \end{array}$$

where x_i is integer-valued and non-negative, and w is unrestricted. An implementation and solution is provided in the files `maxi-min.gms` and `maxi-min.lst`. The solution indicates that the student should invest \$24 in option 3 and \$26 in option 4 for a guaranteed minimum return of \$134.

15. *Modeling with binary variables.* This problem (slightly altered) appears in [1], an excellent book that is freely available on-line. A business sells video games and movies. For each video game sold, the business earns a profit of \$1. For each movie sold, the business earns a profit of \$2. The store manager, who is a bit quirky, has imposed the following restrictions on customers' purchases.

    i) There is a limit of eight items (total) per customer.
    ii) The number of movies purchased minus the number of video games purchased cannot exceed 2.
    iii) The number of video games purchased minus the number of movies purchased cannot exceed four.
    iv) A customer can purchase zero, one, four, or six video games. No other quantities of video games are allowed. There is no such restriction on the quantity of movies.

    The manager would like to determine the combination of purchases by a customer that will maximize profit. Letting x_1 and x_2 represent the quantities of video games and movies, respectively, the problem can be formulated as follows.

$$\begin{array}{lrrcrcl} \text{maximize} & z & = & x_1 & + & 2x_2 & \\ \text{subject to} & & x_1 & + & x_2 & \leq & 8 \\ & & -x_1 & + & x_2 & \leq & 2 \\ & & x_1 & - & x_2 & \leq & 4 \\ & & & & x_2 & \geq & \text{0, and integer} \\ & & & & x_1 & = & \text{0,1,4, or 6} \end{array}$$

    (a) Reformulate the problem as an equivalent integer linear program.

(b) How would your answer to part 15a change if the objective function was instead

$$\text{maximize } z = x_1^2 + 2x_2?$$

Solution. Introduce the following binary variables. Let

$$y_i = \begin{cases} 1 & \text{if } x_1 = i \quad i \in \{0, 1, 4, 6\} \\ 0 & \text{otherwise} \end{cases}$$

Now the binary integer programming formulation is

$$\begin{array}{lrcrcrcrcl} \text{maximize} & & & y_1 & + & 4y_4 & + & 6y_6 & + & 2x_2 \\ \text{subject to} & y_0 & + & y_1 & + & y_4 & + & y_6 & = & 1 \\ & y_1 & + & 4y_4 & + & 6y_6 & + & x_2 & \leq & 8 \\ & -y_1 & - & 4y_4 & - & 6y_6 & + & x_2 & \leq & 2 \\ & y_1 & + & 4y_4 & + & 6y_6 & - & x_2 & \leq & 4 \\ & & & & & & & x_2 & \geq & 0 \text{ and integer} \\ & & & & & & & y_0, y_1, y_4, y_6 & \in & \{0, 1\} \end{array}$$

For part 15b, change the coefficients on the y_i variables in the objective function only to be the squares, i.e. $16y_4$ and $36y_6$. Note that by using this technique the objective function is transformed from a quadratic to a linear function of the variables.

Deterministic Dynamic Programming

16. *An optimal assignment of resources.* The number of crimes in each of a city's three precincts depends on the number of police cars assigned to each precinct. The number of crimes is shown in the table below. Three patrol cars are available. Use dynamic programming to determine how many patrol cars should be assigned to each precinct so that the total number of crimes is minimized.

	No. of patrol cars assigned to precinct			
	0	1	2	3
precinct 1	14	10	7	4
precinct 2	25	19	16	14
precinct 3	20	14	11	8

Solution. A natural representation for dynamic programming is that the stage is the precinct and the state is the number of patrol cars remaining. Let x_i represent the number of cars to assign to precinct i, $i = 1, 2, 3$. With one precinct to go it makes sense to assign all remaining cars. In this stage I am assigning cars to precinct 3.

state	x_3	cost	x_3^*
0	0	20	0
1	1	14	1
2	2	11	2
3	3	8	3

With two precincts to go, we are assigning cars to precinct 2.

state	x_2	cost	cost-to-go	total	x_2^*
0	0	25	20	45	0
1	0	25	14	39*	0
	1	19	20	39	
2	0	25	11	36	
	1	19	14	33*	1
	2	16	20	36	
3	0	25	8	33	
	1	19	11	30*	1
	2	16	14	30	
	3	14	20	34	

With three precincts to go we are assigning cars to precinct 1. At this stage no cars have been assigned.

state	x_1	cost	cost-to-go	total	x_1^*
3	0	14	30	44	
	1	10	33	43*	1
	2	7	39	46	
	3	4	45	49	

The optimal solution is to assign 1 car to each of the three precincts for a total of 43 crimes. You may have selected a different sequence of precincts for the stages, but the solution should be the same.

Deterministic Inventory Models

17. *Ordering lumber.* Lakeland Builders always has a full schedule of building projects. One of the most important resources required for their projects is lumber. To complete their projects, Lakeland has a constant demand for lumber of 5000 board feet per day. There is a $300 delivery fee every time the company orders more lumber. It costs $0.093 per week for the company to store each unused board foot (7 days in one week). Every order has a lead time of 2 days to account for order processing and delivery time.

    (a) What is the optimal order quantity (in board feet) for Lakeland Builders?

    (b) How frequently should Lakeland order to replenish the lumber supply? (i.e. What is the cycle time?)

    (c) Lakeland currently has space to store 20,000 board feet of lumber. Should they invest in more space to store lumber? Why or why not?

    (d) What is the reorder point?

    Solution. The cost per order is $C_0 = \$300$, the holding cost is $C_h = \$.093$ per board foot per week, the lead time is $m = 2$ days, the demand rate is $D = 5000$ board feet per day, and there are 7 days in a week. The demand per week is

$$5000 \text{ board feet per day} \times 7 \text{ days per week} = 35{,}000 \text{ board feet per week}$$

Using the formula for economic order quantity,

$$Q^* = \sqrt{\frac{2DC_o}{C_h}} = \sqrt{\frac{2(35{,}000)(300)}{0.093}} \approx 15{,}027 \text{ board feet.}$$

The cycle time associated with the economic order quantity is

$$T^* = \frac{Q^*}{D} = \frac{15{,}027}{5000} \approx 3 \text{ days.}$$

No, Lakeland Builders should not invest in more space to store lumber because the economic order quantity of 15,027 board feet is less than the current storage space capacity of 20,000 board feet.

The lead time is 2 days and the usage rate per day is 5000 board feet, so the reorder point is $2 \times 5000 = 10{,}000$ board feet.

Chapter 2

Probabilistic Models

2.1 Modeling with Probability Distributions

The normal rate of infection of a certain disease in cattle is 25%. Each animal becomes infected (or not) independently of other animals. A team of veterinarians would like to test a new vaccine. Which of the following two scenarios, A or B, provides more evidence that the vaccine is effective. *Hint:* Compute the probability that each scenario would occur under the hypothesis that the vaccine has no effect whatsoever.

A) 10 animals are vaccinated and none of them become infected.

B) 17 animals are vaccinated. At most one of the animals becomes infected.

Let X be the number of animals infected under the assumption that the vaccine is worthless. Under scenario A,

$$X \sim \text{Binomial}(n = 10, p = .25)$$

$$P(X = 0) = \binom{10}{0} .25^0 .75^{10} = .0563$$

Under scenario B,

$$X \sim \text{Binomial}(n = 17, p = .25)$$

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= \sum_{x=0}^{1} \binom{17}{x} .25^x .75^{17-x} \\ &= .0501 \end{aligned}$$

In the absence of a working vaccine, scenario B is less likely to occur, and so it provides a better test of the effectiveness of the vaccine.

At a facility that manufactures recreational sports vehicles (ATVs), each vehicle is subjected to a final inspection. The rate of defects during final inspection is $\lambda = 1.5$ defects per vehicle.

1. What is an appropriate probability distribution to model the number of defects?
2. What proportion of vehicles have more than 2 defects?
3. Management has set a new goal that the proportion of vehicles with no defects is .5. What rate λ would achieve this goal?

We are interested in the number of defects, which is discrete. The Poisson distribution makes sense because it is commonly used for count data. Moreover, we are given information for a single parameter, and the Poisson distribution has a single parameter. If we let the random variable X represent the number of defects per vehicle, then a reasonable distribution is

$$X \sim \text{Poisson}(\lambda = 1.5)$$

For part 2,

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) \\ &= 1 - \sum_{x=0}^{2} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= 0.191 \end{aligned}$$

For part 3, management's goal is that $P(X = 0) = 0.5$, or

$$P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = 0.5$$

Then

$$\lambda = -\ln 0.5 = 0.693$$

The number of bacteria colonies of a certain type in samples of polluted water has a Poisson distribution with a mean of 2 per cubic centimeter. If four 1–cubic–centimeter samples are independently selected from this water, find the probability that at least one sample will contain one or more bacteria colonies.

Let X be a random variable that represents the number of bacteria colonies in a 1 cm^3 sample of the polluted water. From the problem description,

$$X \sim \text{Poisson}(\lambda = 2)$$

First let's find the probability that any particular sample will contain at least one colony.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{e^{-\lambda}\lambda^0}{0!} \\ &= 1 - e^{-2} \\ &= .865 \end{aligned}$$

Now, the four samples are independent and the the probability that a sample contains one or more colonies is the same for each sample. Let Y be a random variable that represents the number of samples that contain one or more colonies. Then

$$Y \sim \text{Binomial}(n = 4, p = 0.865)$$

and we want to know $P(Y \geq 1)$.

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) \\ &= 1 - \binom{4}{0}(.865)^0(1 - .865)^4 \\ &= 0.9997 \end{aligned}$$

A refinery makes two grades of gasoline, regular and premium. The advertised octane ratings are 87 for regular gasoline and 89 for premium gasoline. The quality engineer at the refinery asks for 10 samples from one of the two types of gasoline. She does not know for sure whether the samples are from the regular batch or the premium batch. She devises the hypothesis test

$$\begin{aligned} H_0 &: \mu \leq 87 \\ H_1 &: \mu > 87 \end{aligned}$$

and sets the confidence level to be 0.995. Suppose that the mean of the 10 samples is 88.3 and the standard deviation is 1.0. What is her conclusion for the hypothesis test?

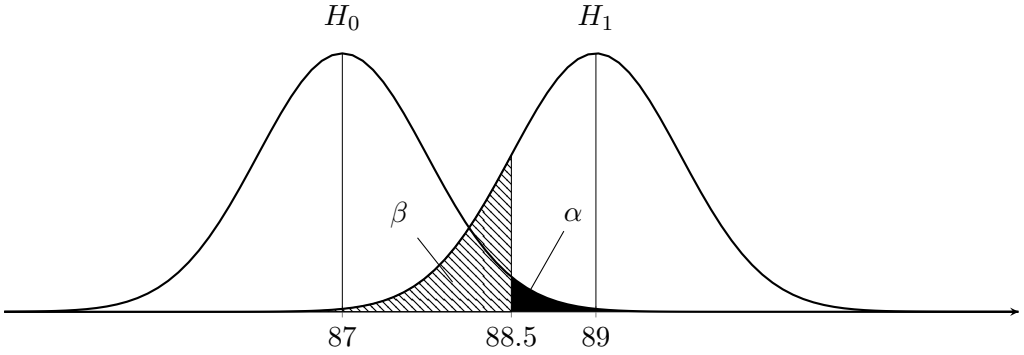
Suppose that a gas station owner has his own octane test kit and rule for accepting a tanker-truck of premium gasoline. The owner knows from past shipments that the distributions of octane ratings are

$$\begin{aligned} X_{\text{regular}} &\sim N(87, 1) \\ X_{\text{premium}} &\sim N(89, 1) \end{aligned}$$

Although the owner may not think about it explicitly, his hypothesis test is

$$\begin{aligned} H_0 &: \mu = 87 \\ H_1 &: \mu = 89 \end{aligned}$$

The owner takes one sample from the tanker-truck. If the octane measurement is greater than 88.5, then he will accept the shipment as premium gasoline. What is the probability that the owner accepts a shipment of regular gasoline as premium (i.e. what is α)? What is the probability that he declines a shipment of premium gasoline, claiming that he thinks it is regular (i.e. what is β)? Use the Normal distribution for this problem. The following diagram may help.



For the quality engineer, the test statistic is

$$t_0 = \left(\overline{X} - \mu_0\right) \frac{\sqrt{n}}{S} = (88.3 - 87)\,\frac{\sqrt{10}}{1} = 4.11$$

and since $t_0 > t_{\alpha,n-1} = 3.25$ ($\alpha = 0.005$) she will reject H_0 and conclude that the samples are from the premium batch of gasoline.

For the station owner,

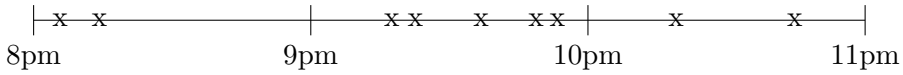
$$\begin{aligned}
\alpha &= P(X > 88.5 \mid H_0) \\
&= P\left(Z > \frac{88.5 - 87}{1}\right) \\
&= 1 - P(Z < 1.5) \\
&= 0.067
\end{aligned}$$

and

$$\begin{aligned}
\beta &= P(X < 88.5 \mid H_1) \\
&= P\left(Z < \frac{88.5 - 89}{1}\right) \\
&= P(Z < -0.5) \\
&= 0.309
\end{aligned}$$

2.2 Stochastic Processes

A Poisson Process. A statistician has observed the behavior of a Hollywood celebrity for about one year and has noted that between the hours of 8pm and 11pm this celebrity generates, on average, three tweets per hour and that the rate is approximately the same within each one-hour period. We can count the *number* of tweets that occur in a time interval t. We can also measure the *time* between tweets. Here is a depiction of the tweets from last night.



Let the random variable Y be the number of tweets from the celebrity in some time interval. When we say that the number of tweets in a time interval t follows a Poisson distribution with mean λt, we write

$$Y \sim \text{Poisson}(\lambda t)$$

If t is one hour, then we can write

$$Y \sim \text{Poisson}(\lambda = 3)$$

Stating that the number of tweets follows a Poisson distribution implies that the time between tweets follows an Exponential distribution (and vice versa). Let the random variable X be the time between tweets. Then

$$Y \sim \text{Poisson}(\lambda t) \iff X \sim \text{Exp}(\lambda)$$

Yes, it is the same λ in each distribution. The average number of tweets is $\lambda = 3$ per hour. The average time between tweets is $1/\lambda = 1/3$ hour (or 20 minutes). Recall that for the Exponential distribution

$$E(X) = \frac{1}{\lambda} = \frac{1 \text{ hour}}{3 \text{ tweets}} = 20 \text{ minutes per tweet on average}$$

Questions.

1. What is the probability that the celebrity sends out five or more tweets in one hour?
2. What is the probability that the celebrity sends out no tweets between 9pm and 11pm?

2.3 Queueing Models

2.4 Stochastic Dynamic Programming

2.5 Probabilistic Inventory Models

Monte Carlo simulation and Thanksgiving turkeys. Tom the grocer pays \$10 wholesale for his turkeys. Before Thanksgiving he sells them for \$18. After Thanksgiving he sells any remaining turkeys for \$5 each. In the past years Tom has noticed pre-Thanksgiving demand has varied uniformly between 50 and 100 birds. How many turkeys should he buy to maximize his expected profit?

Let's let x represent the number of turkeys that Tom should order. The random pre-Thanksgiving customer demand is D, and Tom's expect profit is P, where

$$P = \begin{cases} (18-10)x & \text{if } D \geq x \\ 18D - 10x + 5(x-D) & \text{if } D < x \end{cases}$$

Let's find an optimal value for x (our decision variable) via simulation. Here are the steps.

1. Set $x = 50,\ 51,\ 52,\ \ldots,\ 100$.
   (a) For each value of x generate $B \approx 200$ independent replications of D, where $D \sim U(50,\ 100)$. Call the realized values of demand D_i where $i = 1 \ldots B$.
   (b) Calculate the profit P_i for each D_i.
2. Estimate expected profit for each value of x as

$$\frac{1}{B} = \sum_{i=1}^{B} P_i$$

3. Choose the value of x that achieves the maximum profit.

```
Order <- 50:100
EP <- numeric(length(Order))  # will hold expected profit
B <- 200                      # number of replications

for (x in Order) {
    P <- numeric(B)   # will hold realized profits
    for (i in 1:B) {
        D <- sample(50:100, 1)
        if (D >= x) { # sell all turkeys
            P[i] <- 8*x
        } else {      # don't sell them all
            P[i] <- 18*D - 10*x + 5*(x-D)
        }
    }
    EP[which(x==Order)] <- mean(P)
}

plot(Order, EP)   # using base R graphics
```

Figure 2.1: Simulation to determine the number of turkeys to order so that expected profit is maximized.

An implementation in R is shown in Figure 2.1 and the output is shown in Figure 2.2. *Questions.*

1. Referring to Figure 2.2, why does it appear that the expected profit is becoming more varied as the order quantity increases? What can we do about it?
2. Can you solve this problem analytically?
3. Why use simulation?

Auctioning a turkey. On the Wednesday before Thanksgiving, Tom decides to auction a turkey with the proceeds donated to charity. Tom will offer the turkey using a second-price sealed-bid auction. In a second-price auction (also valled a Vickrey auction), the highest bidder wins, but the pays the price of the second-highest bid. At the time of the auction it turns out that there are only two bidders. Tom doesn't want to give the turkey away so he sets a reserve price r. The reserve price modifies the rules of the auction as follows. If both bids are below r then neither bidder wins, and Tom collects no proceeds. If both bids are at or above r then the regular 2nd-price auction rules prevail. If only one bid is at or above r then that bidder wins the turkey and pays r. Both bidders agree to the rules.

Now, it is well-known that a 2nd-price is a truth-telling mechanism. That is to say, bidders will bid their true valuations for the item. Suppose each bidder has a value for the Turkey is independently and uniformly distributed between \$0 and \$20. What is the optimal reserve price r?

With some effort, this problem can be solved analytically. Alternatively, and with not as much effort, we can determine the optimal reserve price via simulation (see Figure 2.3.)

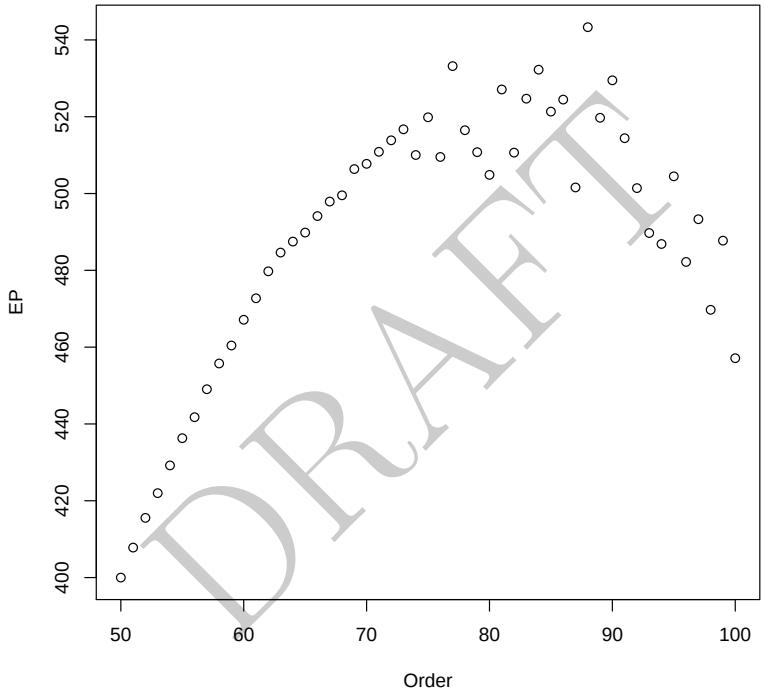


Figure 2.2: Plot of expected profit vs. turkeys ordered.

Now consider adding just a bit more complexity to the problem, e.g. a third-price auction, or multiple classes of bidders having different valuation distributions (Gamma, Lognormal, etc.) The problem can become intractable to solve analytically, but it is relatively easy to obtain an approximate optimal solution via simulation.

```
auction <- function(r) {
    b1 <- runif(1, 0, 20) # bidder 1's valuation
    b2 <- runif(1, 0, 20) # bidder 2's valuation, independent of bidder 1
    if (b1 < r && b2 < r) {
        rev <- 0
    } else if (b1 >= r && b2 >= r) {
        rev <- min(b1,b2) # regular 2nd price auction
    } else {
        rev <- r
    }
    rev
}

reserve <- seq(0, 20, by=.1)
expected.rev <- numeric(length(reserve))

for (i in 1:length(reserve)) {
    expected.rev[i] <- mean(replicate(10000, auction(reserve[i])))
}

plot(reserve, expected.rev, type="l")
```

Figure 2.3: Implementation of a 2nd-price auction with a reserve price r and independent valuations for two bidders. We would like to know the reserve price r that maximizes revenue for the seller.

2.6 Exercises

Modeling with Probability Distributions.

1. *Searching for an item.* Albert has 1176 Pokémon cards in total. Pokémon EX is a special type of card, and Albert has 39 EX-type cards. He is looking for an EX-type card, but all of the cards are completely mixed up and stored in a shoe box. His mother is calling him for dinner. What is the probability that Albert will have to look through no more than 25 cards before he finds an EX-type card?

   Solution. Consider finding an EX-type card to be a "success". Let X be a random variable that represents the number of cards that Albert has to handle up to and including the first success. Then

$$X \sim \text{Geometric}(p = \frac{39}{1176})$$

and

$$P(X \leq 25) = 1 - (1-p)^{25} \approx 0.57.$$

2. *Playing Pokèmon.* Albert is playing Pokémon cards with his friend. It's Albert's turn, and he decides to use Marowak. The card says the following.

   > *Flip a coin four times. The amount of damage done to your opponent's Pokémon is the number of heads times 40.*

   What are the odds that Marowak will do at least 120 damage to the opponent? One approach to answer this question is to use the Binomial distribution to compute the probability of doing at least 120 damage and then convert from probability to odds. You can take another approach if you prefer. In any case, assume that the coin is fair, i.e. the probability of getting heads on any particular toss is 1/2.

   Solution. In order to do at least 120 damage, we need either three or four heads out of the four coin tosses. Let X represent the number of heads obtained in four tosses of a fair coin. Then $X \sim \text{Binomial}(p = 1/2, n = 4)$.

$$\begin{aligned} P(X=3) + P(X=4) &= \binom{4}{3} p^3 (1-p)^1 + \binom{4}{4} p^4 (1-p)^0 \\ &= \frac{1}{4} + \frac{1}{16} \\ &= \frac{5}{16} \end{aligned}$$

   The odds are

$$\frac{p}{1-p} = \frac{\frac{5}{16}}{1 - \frac{5}{16}} = \frac{5}{11}$$

   or 5 to 11.

3. *System reliability.* A data warehouse has n servers. Each server will fail, independently, with probability p.

   (a) Suppose that a single functioning server can support the integrity of the data. What is the probability that data will be lost?

   (b) Suppose that two functioning servers are required for data integrity. What is the probability that data will be lost?

   Solution. For part 3a, all n servers must fail for data to be lost. Because failures are independent, the probability of data loss due to server failure is p^n.

   For part 3b, let X be the number of failed servers. X has a Binomial distribution

with parameters n and p. The probability of data loss due to server failure is

$$\begin{aligned}
P(X \geq n-1) &= \sum_{i=n-1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \binom{n}{n-1} p^{n-1}(1-p)^{n-(n-1)} + \binom{n}{n} p^n (1-p)^{n-n} \\
&= np^{n-1}(1-p) + p^n \\
&= np^{n-1} - np^{n-1}p + p^n \\
&= np^{n-1} - np^n + p^n \\
&= np^{n-1} + (1-n)p^n
\end{aligned}$$

4. *Evaluating a warranty.* A manufacturer of automotive batteries offers a one-year warranty. If the battery fails for any reason during the warranty period, it is replaced for free. The time to failure is distributed Exponential with rate $\lambda = .125$ failures per year.

   (a) What proportion of batteries fail within the warranty period?

   (b) The cost to manufacture a battery is \$50, and the profit per battery is \$25. What is the effect of the warranty replacement policy on profit?

   Solution. The question is asking for the theoretical proportion of batteries that fail within one year. Since all batteries have the same probability of failure, this proportion is equal to the probability that a single battery will fail within one year. Let X be a random variable that represents the time to failure.

   $$P(X < 1) = 1 - e^{-\lambda t} = 1 - e^{-.125} = 0.118$$

   Now, imagine that the manufacturer has, over time, sold many batteries and has kept data on how many batteries failed within one year. The empirical proportion is simply the number of batteries that failed divided by the number of batteries sold. The Law of Large Numbers tells us that when the number of batteries sold is large, the empirical proportion will be approximately equal to the theoretical proportion.

   Taking the warranty into account, the average profit per battery is

   $$\$25 - 0.118 \times \$50 = \$19.10$$

   So, the (average) effect of the warranty on profit is -\$5.90.

5. *Memoryless property of the Exponential distribution.* A hair salon has two hairdressers who provide haircuts on a walk-in basis. Sam arrives to the salon and finds that both hairdressers are busy with customers. However, no other customers are waiting, so Sam will begin his haircut (immediately) when the first space opens up. Suppose that the distribution of service times for both hairdressers is exponential with rate λ. What is the probability that Sam is the last of the three customers to complete a haircut?

Solution. By the memoryless property of the Exponential distribution, the distribution of remaining time for each customer currently receiving a haircut is exponential with rate λ. Since the two customers getting haircuts have the same distribution for remaining time, the probability that customer 1 finishes before customer 2 is 1/2 (likewise for customer 2 finishing before customer 1). When Sam enters service, the memoryless property still applies. Regardless of how long the other customers have been in service, Sam has the same distribution for remaining time. The probability Sam is the last to finish is 1/2.

6. *Donut giveaway.* A professional baseball team has just won a game that secured them a berth in the league's playoffs. To celebrate, a local donut shop will be giving away up to 200 free donuts during a two-hour period on the morning following the victory. All 200 donuts will be baked and decorated with a baseball theme before the giveaway starts. If there are any donuts remaining after the giveaway, they will be sold at a discounted price. Assume that customers will arrive at the giveaway according to a Poisson process at a mean rate of 100 customers per hour. Also, note that there is a limit of one donut per customer.

   (a) What is the probability that there will be donuts remaining after the giveaway?

   (b) What is the predicted number of donuts that will be remaining after the giveaway?

   For part 6a, present your answer as an expression for the probability that there will be donuts remaining. Then, use software, such as R, to compute a numerical answer.

   Solution. Given that customers arrive to the giveaway according to a Poisson process with a mean of 100 customers per hour, the number of customers that arrive during the two-hour period is Poisson distributed with a mean of 200. Let N be the number of customers arriving in a two-hour period. Then

$$N \sim \text{Poisson}(\lambda = 200)$$

   For part 6a, the probability that there will be donuts remaining after the giveaway is

$$\begin{aligned} P(N < 200) &= \sum_{n=0}^{199} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{n=0}^{199} \frac{200^n e^{-200}}{n!} \\ &\approx .49 \end{aligned}$$

   In R,

```
> sum(dpois(0:199,200))
[1] 0.4905966
```

   For part 6b, our calculations are all done in expectation (that is to say, on average). There are 200 customers in 2 hours, which means that 200 donuts are given away. So, on average, no donuts remain.

7. *Startup expenses.* Two friends are starting a small business selling ice cream. They applied for a grant and have received $1800 to help cover any startup expenses. The friends will incur expenses of $300 randomly throughout the first year, and the time between payments for these expenses is exponential with a mean of 2 months. Determine the probability that the friends will run out of grant money before the end of the year.

   Solution. Let X be a random variable that represents the time between payments. The mean time between payments, that is to say the expected value of X ($E(X)$), is two months. We know that for the Exponential distribution

$$E(X) = \frac{1}{\lambda}$$

   where λ is the rate (in units of payments per month). So,

$$X \sim \text{Exp}(\lambda = 1/2 \text{ payments per month})$$

   If the time between payments is distributed Exponential with rate λ, then the number of payments in t months is Poisson with mean λt. Let N be the number of payments in 12 months.

$$N \sim \text{Poisson}(\lambda t = \lambda \times 12 = 6)$$

   Now, the probability that the friends runs out of money is

$$\begin{aligned} P(N \geq 6) &= 1 - P(N \leq 5) \\ &= 1 - \sum_{n=0}^{5} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= 0.55 \end{aligned}$$

   You may have defined the event that the friends runs out of money as $P(N = 6)$. In other words, that there are exactly six payments during the first year. This is incorrect because we are modeling the spending activity as a Poisson process. In other words, the (unstated) assumption is that the number of payments is independent of the available funds.

8. *Stocking a vending machine.* A university cafeteria has a vending machine that is stocked with a variety of juices and sodas. A student employee replenishes inventory weekly so that there are 180 beverages in stock at the beginning of each week. The cafeteria is open 24 hours, 7 days a week, and beverages are purchased according to a Poisson process with a mean of 1 hour between purchases.

   (a) What is the probability that, at the end of any given week, the student employee will find the machine to be sold out?

   (b) On average, how many beverages remain in the vending machine when the employee arrives?

   (c) What is the probability that the employee will replenish 150 or more beverages?

Solution. For part 8a, there will be no beverages remaining in the vending machine if demand for beverages is at least 180. Because the beverages are purchased according to a Poisson process with a mean of one hour between purchases, the rate λ that beverages are purchased is 24 beverages per day. This means that the number of purchases in seven days is Poisson with mean λt. Let N be a random variable that represents the number of purchases in one week (seven days).

$$N \sim \text{Poisson}(\lambda = \lambda \times 7 = 168)$$

$$\begin{aligned} P(N \geq 180) &= 1 - P(N \leq 179) \\ &= 1 - \sum_{i=0}^{179} \frac{168^i e^{-168}}{i!} \\ &\approx 0.19 \end{aligned}$$

In R,

```
> 1 - sum(dpois(0:179,168))
[1] 0.1866995
```

For part 8b, the expected number of beverages purchased from the vending machine each week is 168. The expected number of beverages remaining at the end of the week is $180 - 168 = 12$.

For part 8c, The probability that 150 or more beverages are purchased during a week is

$$\begin{aligned} P(N \geq 150) &= 1 - P(N \leq 149) \\ &= 1 - \sum_{i=0}^{149} \frac{168^i e^{-168}}{i!} \\ &\approx 0.93 \end{aligned}$$

In R,

```
> 1 - sum(dpois(0:149,168))
[1] 0.9253016
```

9. *Car dealership.* A used car dealership parks their inventory on an uncovered lot. In the city where the dealership is located, the probability of a hailstorm during the month of June is 0.30. If a hailstorm occurs, the number of dents in a randomly chosen car follows a Poisson distribution with a mean of five dents.

   (a) If a car receives no more than one dent during a hailstorm, the dealership ignores it and hopes that the customer will not notice. Given that a hailstorm occurred, compute the probability that a car receives no more than one dent.

(b) The occurrence of a storm and the amount of damage after a storm are independent. (Of course damage is conditional on the occurrence of a storm.) If a car has more than five dents after a storm, then the dealership must file an insurance claim for the car. It is May 31 and there are 100 cars on the lot. Determine the expected number of cars for which the dealership will file insurance during June (assume that the inventory remains constant at 100 cars).

Solution. For part 9a, let N be a random variable that represents the number of dents in a car after a hailstorm. We know that

$$N \sim \text{Poisson}(\lambda = 5).$$

The probability that a car receives no more than one dent is

$$\begin{aligned} P(N \leq 1) &= P(N = 0) + P(N = 1) \\ &= \frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} \\ &\approx 0.04 \end{aligned}$$

Conditional on the occurrence of a hailstorm, the probability that any particular car receives more than 5 dents is

$$\begin{aligned} P(N \geq 6) &= 1 - P(N \leq 5) \\ &= \sum_{i=0}^{5} \frac{e^{-\lambda}\lambda^i}{i!} \\ &\approx 0.384, \end{aligned}$$

and the (unconditional) expected number of such cars is

$$0.30 \times P(N \geq 6) \times 100 \approx 11.5$$

Using R,

```
> 0.30 * (1 - sum(dpois(0:5,5))) * 100
[1] 11.52118
```

10. *A mining operation.* A dump truck at a mine takes ore to the railroad after 10 one-ton scoops have been loaded into the truck. The one-ton scoops are loaded from a large diesel-powered shovel at a mean rate of seven scoops per hour. The time between scoops from the shovel follows an Exponential distribution.

    (a) Find the probability that the time required to load the truck takes at least one hour.

    (b) It takes the dump truck 18 minutes to travel to the railroad, unload, and return. Suppose the truck returns and finds that no scoop is waiting to be loaded. What is the probability that the next scoop is ready within 5 minutes?

Solution. The time between scoop arrivals is distributed Exponential, so we know that the number of arrivals in a time interval is distributed Poisson. In particular, the number of arrivals in a one-hour period follows a Poisson distribution with mean $\lambda = 7$. In order for the time required to load the truck be at least one hour, the number of scoops in one hour be nine or less. Let N be the number of scoops in a one hour period.

$$P(N \leq 9) = \sum_{x=0}^{9} \frac{e^{-\lambda}\lambda^x}{x!} = .83.$$

For part 10b, we can invoke the memoryless property of the Exponential distribution. The remaining time until the next arrival is disitributed Exponential with rate 7 scoops per hour, regardless of how much time has elapsed since the last arrival. Let Y be the remaining time until the next arrival, and don't forget to convert from minutes to hours.

$$P(Y \leq 5) = 1 - e^{-7 \times \frac{5}{60}} = .44$$

11. *Blood pressure screening.* High blood pressure is an underlying health condition that makes people more susceptible to severe illness. A company conducted blood pressure screenings to determine the risk its employees have for severe illness. The systolic blood pressures (SBP) of 220 employees were measured. In the general population, SBP measurements follow a Normal distribution with mean $\mu = 135$ and with standard deviation $\sigma = 20$. The company doctor has created the following guidelines for determining which employees are at highest risk.

systolic blood pressure	risk
$\mu + 1.5\sigma < SBP$	very high
$\mu < SPB \leq \mu + 1.5\sigma$	high
$\mu - \sigma < SBP \leq \mu$	average
$SBP \leq \mu - \sigma$	low

How many employees of this company fall into each of the four categories?

Solution. Let $X \sim \mathcal{N}(\mu = 135,\ \sigma = 20)$ be a random variable that represents systolic blood pressure, and recall that $\Phi(z)$ indicates the CDF of the standard Normal distribution.

$$\begin{aligned}
\text{very high} &\quad 220 \times (1 - P(X \leq \mu + 1.5\sigma)) = 220 \times (1 - \Phi(1.5)) \approx 15 \\
\text{high} &\quad 220 \times (P(X \leq \mu + 1.5\sigma) - P(X \leq \mu)) = 220 \times (\Phi(1.5) - \Phi(0)) \approx 95 \\
\text{average} &\quad 220 \times (P(X \leq \mu) - P(X \leq \mu - \sigma)) = 220 \times (\Phi(0) - \Phi(-1)) \approx 75 \\
\text{low} &\quad 220 \times (P(X \leq \mu - \sigma)) = 220 \times \Phi(-1) \approx 35
\end{aligned}$$

12. *Time to failure.* The lifetimes of parts or components that are subjected to the environment (i.e. temperature, corrosion, stress, chemicals) are often modeled using a Lognormal distribution. Rather than being additive, the environmental factors that influence the time to failure are multiplicative. The Central Limit Theorem applies, but because the random effects are multiplicative on the time scale, they are additive on the log scale.

A certain component of a bridge is inspected annually to see if it needs to be replaced. The lifetime of the part follows a Lognormal distribution with parameters $\mu = 1.6$ and $\sigma = 0.25$.

(a) Determine the mean time to failure for this part.

(b) What is the probability that the part will last longer than seven years?

Solution. Let X be a random variable that represents the time to failure of a part. Then

$$X \sim \text{LogN}(\mu = 1.6, \sigma = 0.25).$$

The mean time to failure is

$$e^{\mu+\sigma^2/2} \approx 5.1 \text{ years.}$$

We wish to find $P(X > 7)$. Taking logarithms and standardizing,

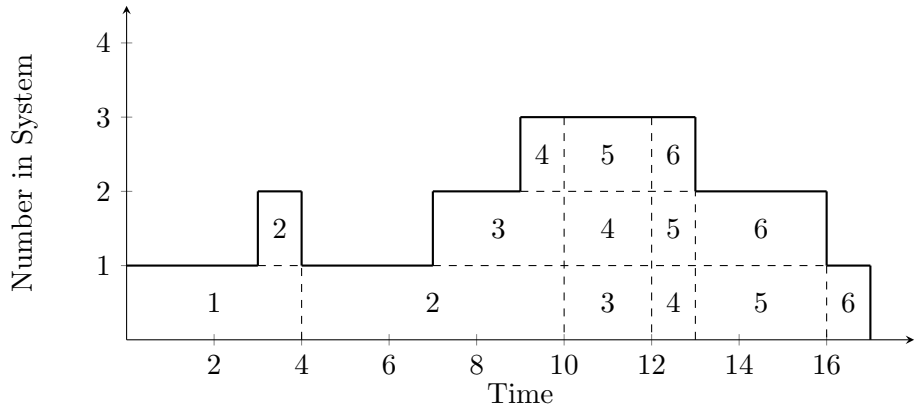
$$\begin{aligned} P(X > 7) &= P(\ln(X) > \ln(7)) \\ &= P\left(Z > \frac{\ln(7) - 1.6}{0.25}\right) \\ &= P(Z > 1.38) \\ &= 1 - P(Z \leq 1.38) \\ &= 0.084 \end{aligned}$$

Stochastic Processes

Queueing Models

13. *Performance metrics for a queueing system.* Consider a single server queueing system with FIFO queue discipline. For the particular day that this system was in operation, the arrival times and the service times of the first six customers were (0,3,7,9,10,12) and (4,6,2,1,3,1), respectively. Arrival times and service times are in minutes. Compute the average waiting time and the average number of customers in the queue for the first six customers. It will help to construct a diagram of number in system versus time.

Solution.



First note that the problem description does *not* tell us that the times between arrivals and/or the service times are exponentially distributed. So it is not an $M/M/1$ system. The total delay of all six customers is $0+1+3+3+3+4=14$. The average waiting time in the queue is the total delay divided by the number of customers.

$$W_q = \frac{14}{6} = 2.3333 \text{ min}$$

To compute the average number in the queue, weight the time in queue by the number of customers. In other words, compute the area under the curve but above one, and then divide by the total time.

$$L_q = \frac{14}{17}$$

14. *Justification for a capital expense.* An amusement park is known for having rides with short waiting times. A new ride has opened, and customers arrive to the ride according to a Poisson process with a mean of 20 arrivals per hour. Only one customer is allowed on the ride at a time, and customers can stay on the ride as long they wish. The length of time that each customer spends on the ride is exponentially distributed with a mean of 2.5 minutes. Management is willing to add an additional cart to the ride if, under the current system, 1) the average number of waiting customers is greater than four, and 2) the percentage of time that the ride is idle exceeds 10%. Can an additional cart be justified?

    Solution. The amusement ride can be modeled as an M/M/1 queuing system. Customers arrive at the rate

    $$\lambda = 20 \text{ customers per hour,}$$

    and the service rate is

    $$\mu = 24 \text{ customers per hour.}$$

    To see whether the first condition is satisfied, we compute

    $$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{20^2}{24(24-20)} = 25/6 > 3 \text{ customers}$$

    and so the first condition is satisfied. To see if the second condition is satisfied, we compute

    $$p_0 = 1 - \frac{\lambda}{\mu} = 0.167$$

    and so the second condition is also satisfied.

15. *Comparing system configurations.* Patients arrive to an emergency room according to a Poisson process at the rate of 10 patients per hour. Before a patient can see a doctor, they must have their medical history reviewed and their vitals measured. Currently, the hospital has one nurse who performs both of these tasks for each patient. The time that each patient spends with the nurse is exponentially distributed with a mean of five minutes. An Industrial Engineer at the hospital is considering two options for improving service.

(a) Hire a medical scribe to review each patient's medical history while the nurse takes vitals. The service rate would increase to 20 patients per hour with this improved single-server operation (service times are still exponentially distributed.)

(b) Hire a second nurse who also reviews medical history and takes vitals for each patient. The two nurses would each have a service rate of 12 patients per hour (each with exponential service times) with this two-server operation.

Consider the relative cost of each option and evaluate the improvements that would result. Then, determine which option the engineer should recommend.

Solution. We will look at the average number of patients in the queue L_Q and the average time in queue W_Q as metrics to judge improvement in system performance. Before any improvements are made,

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{10^2}{12(12-10)} = 25/6 = 4.167 \text{ patients}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{25/6}{10} = 5/12 \text{ hours} = 25 \text{ minutes}$$

If they hire a scribe to assist the nurse

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{10^2}{20(20-10)} = 0.5 \text{ patients}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{0.5}{10} = 0.05 \text{ hours} = 3 \text{ minutes}$$

and if they add an additional nurse we need to use the formulas for a two-server operation. First,

$$P_0 = \frac{1}{\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{k!}\left(\frac{k\mu}{k\mu-\lambda}\right)}$$

where $k = 2$ servers. Then

$$L_q = \frac{(\lambda/\mu)^k \lambda\mu}{(k-1)!(k\mu-\lambda)^2} P_0$$

and

$$W_q = \frac{L_q}{\lambda}$$

Plugging values, I got $P_0 = 0.4118$, $L_q = 0.1751$ patients, and $W_q = 0.0175$ hours or 1.05 minutes. Considering the relative improvement and the cost of adding a second nurse, I would recommend option 15a. That is, to hire a scribe to help the nurse. The average number in queue and the average time in queue appear to be acceptable for an emergency room setting.

16. *An $M/G/1$ queue.* A coffee shop has recently opened for takeout after being closed due to COVID-19. In order to follow public health guidelines, only one customer is allowed in the shop at a time. If a customer is being served, any customer that arrives must wait in a line outside of the coffee shop. Customers arrive at a rate of 12 customers per hour, and inter-arrival times are exponentially distributed. The shop has two baristas: one barista takes orders and the other prepares the beverages. The service time for each barista to complete their task is independently and exponentially distributed with a mean of two minutes. Determine the average number of customers waiting in line to be served, L_Q. A couple of helpful items:

    i) The variance of the sum of independent random variables is the sum of the variances,
    ii) the sum of IID exponential random variables is distributed Gamma,
    iii) the variance of an exponential distribution is the mean squared, and
    iv) for an $M/G/1$ queue,

$$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$$

    where ρ is the server utilization, μ is the service rate, and σ^2 is the variance of the service time distribution.

    Solution. The first thing to note is that the service time is the sum of two IID exponential random variables, and so we know that it has a Gamma distribution. Letting X represent the service time,

$$\text{Var}(X) = 2^2 + 2^2 = 8.$$

    The expected total service time for both baristas is 4 minutes, so $\mu = 1/4$ customers per minute. Now, in units of customers per minute, the arrival rate is $\lambda = 1/5$.

$$\begin{aligned} L_Q &= \frac{\frac{16}{25}\left(1+8\left(\frac{1}{16}\right)\right)}{2\left(\frac{1}{5}\right)} \\ &= 2.4 \text{ customers} \end{aligned}$$

Stochastic Dynamic Programming

Probabilistic Inventory Models

17. *Stocking a seasonal item.* Summer is approaching and a store is determining how many folding lawn chairs they should purchase from a supplier. Each chair costs \$15 for the store to purchase and can be sold for \$27.50. The store stocks lawn chairs only during the summer; any chairs left at the end of the summer are sold at a clearance price of \$10.50. Based on historical data, the seasonal summer demand for lawn chairs follows a normal distribution with $\mu = 205$ and $\sigma = 25$.

    (a) What is your recommended order quantity for the store?
    (b) What is the probability that the store will sell all the folding lawn chairs it orders before summer is over?

(c) How would the economic order quantity change if the clearance price for the chairs was only \$8? What happens to the store's order quantity as the clearance price is reduced?

Solution. Let D be the summer demand for folding lawn chairs. We know that $D \sim \mathcal{N}(\mu = 205,\ \sigma = 25)$. The penalty for ordering too few items is $c_u = 12.50$ and the penalty for ordering too many items is $c_o = 4.50$. We want to find the order quantity Q^* such that

$$P(D \leq Q^*) < \frac{c_u}{c_u + c_o} = \frac{12.50}{17.00} = 0.735$$

Standardizing,

$$P\left(z \leq \frac{Q^* - 205}{25}\right) = 0.735$$

implies that

$$\begin{aligned} \frac{Q^* - 205}{25} &= 0.63 \\ Q^* &= 205 + 0.63(25) \\ Q^* &= 220.75 \end{aligned}$$

The store should order 221 (or 220) folding lawn chairs. If we are using R, we can obtain the answer without standardizing by simply passing the mean and standard deviation to `qnorm`,

```
> qnorm(.735, 205, 25)
[1] 220.7002
```

For part 17b, the probability that all chairs will sell is

$$\begin{aligned} P(D \geq 221) &= 1 - P(D \leq 221) \\ &= 1 - P\left(z \leq \frac{221 - 205}{25}\right) \\ &= 1 - P(z \leq .64) \\ &= 1 - .74 \\ &= 0.26 \end{aligned}$$

Regarding part 17c, if the clearance price decreases to \$8 then the cost of being understocked will still be $c_u = 12.50$, but the cost of being over will increase to $c_o = 7.00$. The order quantity Q^* will change such that

$$\begin{aligned} P(D \leq Q^*) &= \frac{12.50}{19.50} = .641 \\ P\left(z \leq \frac{Q^* - 205}{25}\right) &= .641 \end{aligned}$$

which implies that

$$\frac{Q^* - 205}{25} = .36$$
$$Q^* = 205 + 0.36(25) = 214$$

With a clearance price of \$8, the new order quantity will be 214 chairs. The economic order quantity will continue to decrease as the clearance price decreases because the cost of being over will continue to increase.

18. *Programs for sale.* A college hockey team has game programs for sale at every home game they play. The demand D for these programs is a random variable that follows a Uniform distribution with a minimum of 1200 programs and a maximum of 2000 programs.

$$D \sim \text{Uniform}(1200,\, 2000)$$

Each program costs \$1 to produce and sells for \$3. The programs contain information specific to the game, so any unsold programs are thrown away after the game. How many game programs should the college print per game to maximize their revenue? You should use the CDF of the Uniform distribution to answer this question.

Solution. The CDF of Uniform distribution is

$$P(X \leq x) = F(x) = \int_{y=a}^{x} f(y)\, dy = \int_{y=a}^{x} \frac{1}{b-a}\, dy = \frac{x-a}{b-a}$$

for $a \leq x \leq b$. The cost per copy of being over/under is $c_o = \$1$ and $c_u = \$2$, respectively. Let Q^* be the optimal number of copies to produce. Using the critical ratio, we want

$$P\left(D \leq Q^*\right) = \frac{c_u}{c_u + c_o} = \frac{2}{3} \approx 0.667$$

Using the definition of the CDF for the Uniform distribution,

$$\frac{Q^* - 1200}{2000 - 1200} = 0.667$$
$$Q^* = 0.667 \times 800 + 1200$$
$$Q^* = 1734$$

The college should print 1734 copies.

19. *The risk of being out-of-stock.* The daily demand for milk cartons in a school cafeteria follows a Normal distribution with mean 600 and standard deviation 20. It costs \$.01 per day to hold one item in inventory. The fixed cost to make a replenishment order is \$50, and the lead time to receive an order is 2 days.

    (a) Currently, the inventory policy for milk cartons is to order 3000 cartons whenever the inventory drops to 1225 units. Determine the probability of a stock-out under this policy.

(b) Running out of milk generates lots of complaints to the PTA. Recommend an inventory policy for the school when the probability of a stock-out cannot exceed .01. Your policy should include an economic order quantity and a re-order point.

(c) Compare the cost of the policy in part 19a with the cost of the policy in part 19b.

Solution. The distribution of demand during lead time X_m is

$$X_m \sim \text{Normal}(\mu_m = 1200, \sigma_m = \sqrt{800})$$

The probability of a stock-out is

$$\begin{aligned} P(X_m \geq 1225) &= P\left(z \geq \frac{1225 - \mu_m}{\sigma_m}\right) \\ &= 1 - P\left(z < \frac{1225 - 1200}{\sqrt{800}}\right) \\ &= 1 - P\left(z < 0.8839\right) \\ &= 1 - .8106 \\ &\approx .19 \end{aligned}$$

With the requirement that the probability of a stock-out not exceed 0.01, the optimal order quantity is

$$Q^* = \sqrt{\frac{2C_oD}{C_h}} = \sqrt{\frac{2 \times 50 \times 600}{.01}} \approx 2450$$

Without a buffer stock the optimal policy is to order 2400 milk cartons every $T_0 = Q^*/D \approx 4$ days. To compute the buffer stock, we want

$$P(X_m \geq B + 1200) \leq 0.01$$

or equivalently, we want $B \geq \sigma_m \times 2.33 \approx 66$. The best policy is to order 2400 milk cartons whenever the inventory drops to $1200 + 66 = 1266$ units.

On average, the total daily cost of the current inventory policy (not considering the frustration costs of a stock-out) consists of the holding cost and the ordering cost. The holding cost has two terms: one for the normal inventory and one for the buffer stock

$$\begin{aligned} &\left(\frac{Q}{2}\right)(C_h) + (25)\,(C_h)) \\ =&\left(\frac{3000}{2}\right)(0.01) + (25)(0.01) \\ =&\$15 + \$0.25 \\ =&\$15.25 \text{ per day.} \end{aligned}$$

The ordering cost of the current policy is

$$\left(\frac{600}{3000}\right)(50) = \$10 \text{ per day.}$$

Remember that demand is random and so these costs are "on average". The only difference in cost for the new inventory policy in part 19b is the increased holding cost for the larger buffer stock. This amounts to an increase of

$$(66 - 25)(0.01) = \$0.41 \text{ per day.}$$

Assuming the milk does not expire before it is used, the benefit of fewer stock-outs seems to be worth the increase in cost.

Chapter 3

Decision Problems

3.1 Games Against Nature

3.2 Games Against an Opponent

Nature as an adversary: a two-person zero-sum game. Merrill has a concession stand at Target Field for the sale of sunglasses and umbrellas. This entrepreneur likes to make sales regardless of the weather. When it rains can sell about 500 umbrellas. On a sunny day he can sell about 100 umbrellas and about 1000 sunglasses. Umbrellas cost him 50 cents and sell for $1. Sunglasses cost him 20 cents each and sell for 50 cents. Merrill is willing to invest $250 in the concession stand business. All unsold items represent a loss; there is no salvage value.

Formulate Merrill's problem as a two-person zero–sum game. Merrill is the row player and Nature is the column player. Merrill's strategy set is {buy inventory for rain, buy inventory for sun}. Nature's strategy set is {rain, sun}. The payoff entries represent the profit/loss. Find an equilibrium strategy for Merrill. That is to say, Merrill treats Nature as a strategic opponent and wants to find an optimal inventory strategy that will yield a maximum expected profit *regardless* of the weather.

Would Merrill necessarily need to invest all $250 into buying inventory exclusively for rain or sun? In other words, does it seem possible that Merrill could truly mix his two pure strategies and invest a portion of the $250 into each? The game is

		Nature: Rain	Nature: Sun
Merrill	Rain	250	-150
	Sun	-150	350

The best strategy for Merrill is to mix buying for rain and buying for sun in the ratio 5 to 4. These are the odds. To compute Merrill's expected profit (i.e. the value of the game) we use Merrill's equilibrium strategy against either of Nature's pure strategies. Here is the payoff for Merrill against Nature's strategy of Rain.

$$\frac{5 \times (250) + 4 \times (-150)}{9} = \$72.22$$

Merrill could play the odds and choose a pure strategy, but note that in this game it is possible for Merrill to physically mix the strategies. He could invest 5/9 of his \$250 in rainy–day inventory and invest 4/9 in sunny–day inventory. So he buys

$$\frac{5}{9}(500 \times .50) + \frac{4}{9}(100 \times .50) = \$161.11$$

worth of umbrellas and

$$\frac{4}{9}(1000 \times .20) = \$88.89$$

worth of sunglasses so that he enjoys a steady profit of \$72.22.

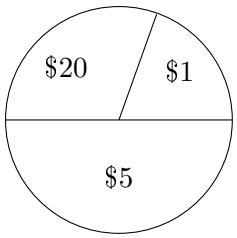
3.3 Utility Theory

The payoffs in the decision problems that we have discussed have been either monetary values or "utilities". So far, we haven't said much about them. In this section, we provide a bit more detail about the payoff values in a decision problem. The purpose of utility theory is to have a convenient way to represent personal preferences over outcomes, and by *convenient* we mean numerical [7], [10],[11]. Why is utility important? How can utility for outcomes be more useful than simply stating our preferences?

> Utilities are more useful than a list of preferences for the same reason that Arabic numerals are more useful than Roman numerals; they are better at facilitating the transfer of information.

It is important to note that we are not trying to explain why a decision-maker has certain preferences over outcomes. We are just trying to represent the preferences numerically. It turns out that utilities are measured on an interval scale, as opposed to a ratio scale. We cannot add, subtract, multiply, or divide utilities as we do with measurements such as length, weight and speed, and we cannot compare the utilities among different persons (unless they have the same utility function). Nevertheless, utilities capture attitudes toward risk and utility theory is fundamental to understanding decision problems.

Sometimes taking a simple expected value makes sense. Consider a gamble in which one of three outcomes will occur. The outcomes are worth \$1, \$5, and \$20 and the probabilities of the outcomes are .2, .5, and .3, respectively. For example, a wheel is spun and the probability of winning an amount corresponds to the area on the wheel. The expected monetary value (EMV) of the gamble is



$$\$1 \times .2 + \$5 \times .5 + \$20 \times .3 = \$8.70$$

However, there exist situations where EMV is not an appropriate indicator of "fair value". Consider another gamble in which a fair coin is tossed until the first head appears.

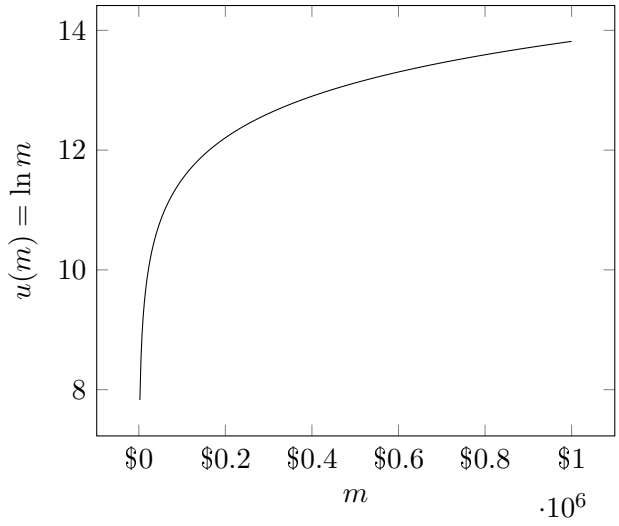


Figure 3.1: The diminishing rate of the value of money.

The gambler receives $\$2^n$ where n is the number of the tosses required until the first instance of heads appears. So there will be $n-1$ tails followed by heads. The probability of the first instance of heads occurring on toss n is $(1/2)^n$ and the EMV of the gamble is

$$2\left(\frac{1}{2}\right) + 4\left(\frac{1}{4}\right) + 8\left(\frac{1}{8}\right) + 16\left(\frac{1}{16}\right) + \ldots = \infty.$$

Few people (perhaps no one) would pay any amount to participate in the gamble. This is the St. Petersburg paradox stated by Bernoulli. To resolve the paradox, he said that the monetary value is not as important as the "intrinsic worth" of the money. For many people, an increase in the amount of money has increasing worth, but at a decreasing rate. Given an amount of money m, the logarithmic function captures the basic relationship between m and its "intrinsic worth" (see Figure 3.1).

However, there are many functions with this basic shape and they differ from person to person. How can the value (or utility) of money be specified for an individual? Also, expected value is commonly associated with value over the long run. What about a one-time gamble? Von Neumann and Morgenstern (VnM [9]) showed how to construst a utility that represents a decision-maker's preferences (numerically) among gambles. Note that if we were only concerned with choices among basic (certain) alternatives then the decision-maker would simply rank the alternatives in an ordinal manner. But we are concerned with decision-making under risk. When we allow gambles, and ask a person to express preferences over pairs of gambles (where the gambles are over basic alternatives) then VnM showed how to associate utilities to the basic alternatives in such a way that 1) the utility for an alternative is measured by the risk one is willing to take to receive it [13], and 2) if decisions are made based on expected utility, then the decision–maker is acting in agreement with her preferences. Of course, the decision-maker's preferences must have some amount of consistency. In other words, preferences must obey some rules that we will state later.

Here is the basic idea for the construction of a utility function. Suppose that among three different bands, Alice prefers Smashing Pumpkins to Yo La Tengo, and she prefers Yo

La Tengo to Wilco. Now any three numbers that decrease in magnitude will capture the ordinal preferences. But we are allowing gambles so that we can capture Alice's attitude toward risk. We allow Alice to choose between two alternatives 1) seeing Yo La Tengo for sure, or 2) a gamble where she sees Smashing Pumpkins with probability p or Wilco with probability $1-p$. If p is very close to one then Alice will choose the gamble. But if we decrease p, then at *some* point Alice will prefer the certain alternative of seeing Yo La Tengo. We ask Alice for the value of p at which she is indifferent between the certain alternative and the gamble. Suppose Alice indicates $p = 2/3$. Now, arbitrarily associate the value 1 with Smashing Pumpkins and associate the value 0 with Wilco. Then it seems natural to associate the value 2/3 with Yo La Tengo. Note that the value of Yo La Tengo equals the expected value of the gamble.

$$\frac{2}{3} = 1\left(\frac{2}{3}\right) + 0\left(\frac{1}{3}\right)$$

Instead of using (1, 2/3, 0) we could use *any* three numbers $a + c,\ (2/3)a + c,\ c$, where a and c are constants and $a > 0$, and not alter the preferences (including the preference over the certain outcome and the gamble).

Perhaps the most troublesome aspect of constructing utilities in this way is the natural inclination to think about the utility values in terms of ratios. You should avoid doing this. For example, it is *not* correct to say that Alice prefers seeing Yo La Tengo to seeing Wilco twice as much as she prefers seeing Smashing Pumpkins to seeing Yo La Tengo. No! The number 2/3 reflects Alice's attitude toward gambling, not her attitude toward the two intervals. A commonly used example [7] to illustrate this fact is the following. Suppose you like taking chances. You are indifferent between 1) receiving \$9 for sure, or 2) participating in a gamble that results in equal chances of receiving either \$10 or nothing. Your utilities for the three amounts \$10, \$9 and \$0 are 1, 1/2, and 0, respectively. It's not that you have equal preferences for going from \$0 to \$9 and going from \$9 to \$10. No. You simply like taking chances.

Another common mistake is to say that Alice prefers Smashing Pumpkins to Yo La Tengo because Smashing Pumpkins has higher utility. No! Alice's preferences (and your preferences) among basic alternatives and lotteries come first. A decision-maker knows her preferences. The point is that if Alice can state her preferences, then we can construct a numerical characterization of them. That is to say, we can associate a value (i.e. a number) to something that is inherently non-numeric. The expected utility theorem is a representation theorem.

Now for the rules of consistency. Let $L(p, x, y)$ represent a gamble (also called a lottery) in which you receive outcome x with probability p or you receive outcome y with probability $(1-p)$. Also, let $x \succ y$ indicate that outcome x is preferred to outcome y, and let $x \sim y$ indicate indifference between the outcomes x and y. A decision-maker's preferences must satisfy the following rules. (Note that the rules don't seem too objectionable).

1. $x \succ y$ or $y \succ x$ or $x \sim y$. (completeness)
2. if $x \succ y$ and $y \succ z$ then $x \succ z$. (transitivity)
3. $L\left(p, L(q, x, y), L(r, x, y)\right) \sim L(s, x, y)$ where $s = pq + (1-p)r$. (no fun in gambling)

4. given $x \succ y \succ z$, there is some lottery $L(p, x, z) \sim y$. (continuity)

5. if two lotteries are identical except for the first prize, then the lottery with the better first prize is preferred. (better prizes condition)

6. if two lotteries have the same alternatives as prizes, then the lottery that assigns higher probability of winning the best prize is preferred. (better chances condition)

If you can state your preferences according to the above rules, then an *interval* utility function $u()$ can be constructed in such a way that

i) $u(x) \succ u(y) \Longleftrightarrow x \succ y$

ii) $u(x) = u(y) \Longleftrightarrow x \sim y$

iii) $u\left(L(p, x, y)\right) = pu(x) + (1-p)u(y)$ (expected utility property)

iv) if $u'()$ is a positive linear transformation of $u()$ then $u'()$ also satisfies i, ii, and iii.

Items i through iv constitute the Expected Utility Theorem. When working through exercises, perhaps the most useful item is item iii, the expected utility property. It states that the utility of a lottery is equal to its expected utility. Do *not* confuse expected utility with the utility of the expected value, which is not of interest.

Retirement and risk. This example is inspired from the discussion on utility in Chernoff and Moses [3]. Alice currently has \$400,000 in her individual retirement account (IRA). Of course, it is always better to have more money for retirement, but Alice is not super-greedy; she wants to have enough money to pay bills and take a few backpacking trips each year. Her utility for an amount of money m in the range \$0 to \$1 million is given by the following function and is shown in Figure 3.2 on the next page.

$$u(m) = \frac{10}{1 + e^{-(\frac{m}{500{,}000} - 5)}}$$

Note that her current utility is $u(\$400{,}000) = 2.69$. Suppose that on a trip to Las Vegas, Alice is offered a gambling opportunity for gaining \$200,000 or losing \$100,000. The odds are 1-to-1 for gaining and losing. If Alice wins, she will have \$600,000, but if she loses, she will have \$300,000. Alice knows about the the expected utility property: that the utility of a lottery is equal to its expected utility. Converting from odds to probabilities, she computes the utility of the gamble as

$$\frac{1}{2}u(\$600{,}000) + \frac{1}{2}u(\$300{,}000) = 4.25,$$

which is higher than her current utility so she takes the gamble.

Suppose that Alice wins. She now had \$600,000 and her utility is $u(\$600{,}000) = 7.31$. Gaining confidence in her luck, Alice seeks and finds another opportunity. The new gamble has odds of 1-to-2 for winning \$200,000 or losing \$100,000. The probability p of winning is

$$p = \frac{\text{odds}}{\text{odds} + 1} = \frac{\frac{1}{2}}{\frac{1}{2} + 1} = \frac{1}{3}.$$

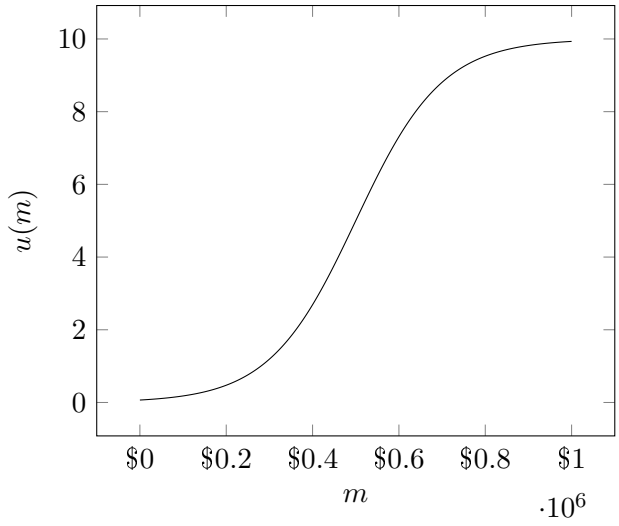


Figure 3.2: Utility function for the retirement funds.

Notice that the gamble is fair because

$$\frac{1}{3}u(\$200{,}000) - \frac{2}{3}u(\$100{,}000) = 0.$$

Alice's utility for the new opportunity is

$$\frac{1}{3}u(\$800{,}000) + \frac{2}{3}u(\$500{,}000) = 6.51,$$

which is less than her current utility and so, even though the gamble is fair, she declines.

We can represent the situation graphically by drawing a line segment from the utility when losing $u(\$500{,}000)$ to the utility when winning $u(\$800{,}000)$, and noting the point at one-third of the distance from $u(\$500{,}000)$ to $u(\$800{,}000)$ (see figure 3.3 on the following page). The corresponding monetary value is the expected value of the gamble, i.e. \$600,000, and the corresponding utility is the expected utility of the gamble, 6.51. With \$600,000 in retirement funds, Alice's behavior changes. We say that her utility function is risk-averse (or concave) in the range from \$500,000 to \$1,000,000 and that it is risk-seeking (or convex) in the range from \$0 to \$500,000.

Notice that if Alice had only \$100,000 in retirement savings, then her utility would be 0.180 and her utility for the gamble would be

$$\frac{1}{3}u(\$300{,}000) + \frac{2}{3}u(\$0) = 0.442.$$

With \$100,000 in retirement funds, Alice cannot fund the lifestyle that she desires and she would be willing to gamble even when it could be slightly unfavorable for her to do so[1]. At this point, one might object that most people do not gamble with retirement savings. However, we have not described the utility function for most people; we have

[1]Homeowner's insurance is another example where people engage in gambles (or lotteries) that are unfavorable in expectation.

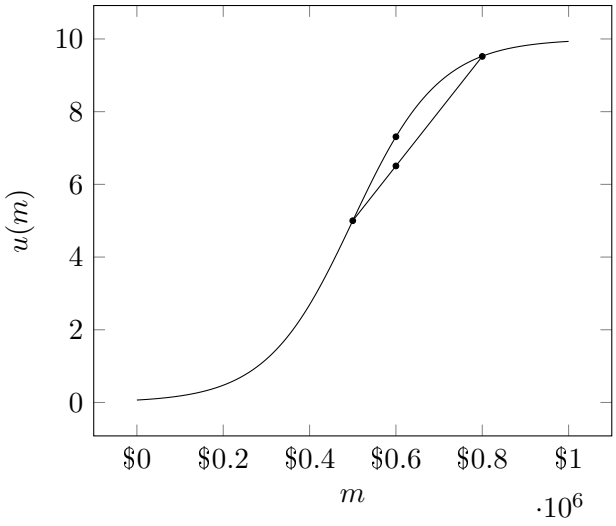


Figure 3.3: A graphic interpretation of the gamble.

described the utility function for Alice. Recall that a utility function is used to (numerically) represent a decision-maker's preferences over outcomes. If Alice was risk-averse with all of her retirement savings, then the shape of her utility function would be risk-averse over the entire domain.

3.4 Exercises

1. *Structuring a decision problem.* An online retail company is looking to purchase some robots to perform tasks in their warehouse. There are two types of helpful robots available. Robot A can collect items in the warehouse. Robot B seals boxes shut. Robot A costs \$6000 and Robot B costs \$3000. The business has \$15,000 available to spend and they value money. The money-equivalent value (payoff) they get from acquiring their first Robot A is \$14,000 and that of each additional Robot A is half the previous one (the second Robot A gives them a value of \$7000, the third \$3500 and so on). Similarly, the payoff they get from acquiring their first Robot B is \$5000 and that of each additional Robot B is half the previous one (\$2500, \$1250, ...). The robots do separate jobs, so the value from each Robot A acquired is not affected by how many of Robot B are acquired, and vice versa.

   (a) Draw the decision tree that is associated with this problem.

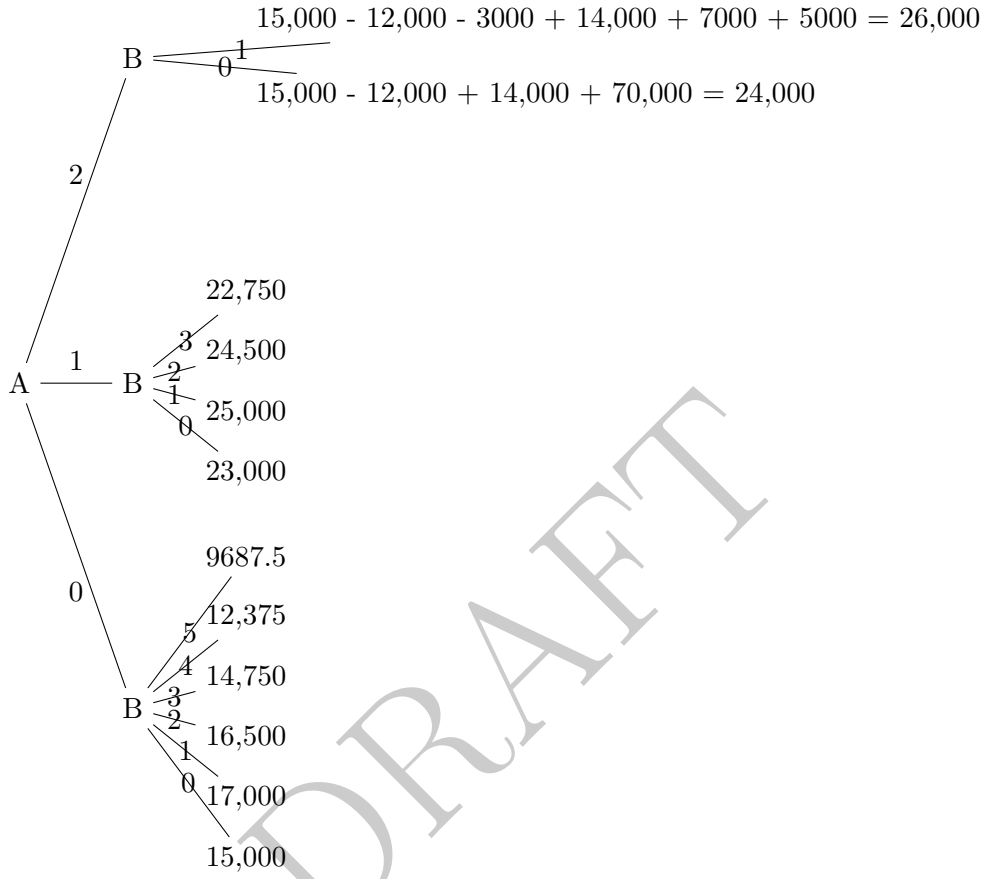
   (b) The business makes decisions based on maximizing their payoff. Should they spend all of their available money on these robots?

   Compute the payoff for each leaf (or end–node) of the decision tree. For example, if you purchase one Robot A and one Robot B, you will spend \$9000 and the payoff is

$$15{,}000 - 6000 - 3000 + 14{,}000 + 5000 = \$25{,}000$$

Note that I included the initial amount of available cash in the calculation because the problem states "... and they value money."

Solution. In the tree below, "A" indicates a choice for number of Robot A to purchase, and "B" indicates a choice for number of Robot B to purchase.



The company wants to make the most rational decision, so they choose the action that maximizes the payoff. The company should spend the entire $15,000 on robots for their warehouse. They should purchase two of Robot A and one of Robot B for a money-equivalent value of $26,000.

2. *Rules for decision–making under ignorance.* You have the opportunity to go on a blind date, but you are hesitant. You are lonely and would like to find the love of your life; however, you dislike awkward situations. Furthermore, you find it difficult to estimate the probability that this particular blind date will turn out to be the love of your life, but you know this probability is non-negligible. To be a little more precise, you have the following values: finding the love of your life is worth 1000, being in an awkward date situation (i.e. being on a date and knowing that you will not see the person again) is worth -10, and staying home watching Netflix is worth zero.

   (a) Formulate a decision problem for deciding whether to go on the blind date or to stay home.

   (b) Use the maximin rule to solve the problem.

(c) Use the minimax regret rule to solve the problem.

Solution. The decision problem can be represented with the following table.

	decision matrix	
	find love	lots of awkward moments
go on date	1000	-10
decline date	0	0

The maximin rule tells you to decline the date because it has the best of all the worst possible outcomes. To use minimax regret, we form the regret matrix.

	regret matrix	
	find love	lots of awkward moments
go on date	0	-10
decline date	-1000	0

Minimax regret tells you to go on the date because the possibility of not finding love has the most regret.

Games Against Nature

3. *Gardening against nature.* A family is considering growing their own garden to save money on fresh vegetables. They have space in their yard for the garden but would need to purchase seeds and gardening supplies. The family is excited to grow a garden, but they know there are a lot of hungry rabbits in their neighborhood that might eat their plants before the family can harvest any vegetables from them. Money saved by the garden is shown in the following table.

	State of Nature	
	s_1	s_2
	rabbits leave garden alone	rabbits eat garden
plant garden	\$400	-\$100
buy vegetables from store	0	0

(a) If the probability that the rabbits leave the garden alone is 0.3, what decision is recommended for the family? What are the expected savings?

(b) The family has the option to purchase fast-growing plant seeds (the fast-growing seeds are the same price as regular seeds but they must buy the fast-growing seeds now if they want them because they are in high demand). With these fast-growing seeds, the family can wait three more weeks to plant their garden. During that time, some scientists will finish their study on the appetites of the local rabbits, and the family will have a better idea about the probability that their garden is eaten by rabbits. They can return the seeds later for a partial refund if they do not use them. Let L represent the event the rabbits have large appetites and let S represent the event that rabbits have small appetites. Then

$$P(L) = 0.60, \quad P(s_1 \mid L) = 0.15, \quad P(s_2 \mid L) = 0.85,$$
$$P(S) = 0.40, \quad P(s_1 \mid S) = 0.79, \quad P(s_2 \mid S) = 0.21.$$

What is the optimal decision strategy if the family purchases the fast-growing seeds so they can wait and learn more about the rabbit appetites before making a decision?

(c) If \$40 of the fast-growing seed purchase is non-refundable, should the family purchase the fast-growing seeds? Why or why not? What is the maximum non-refundable amount the family should pay to get the fast-growing seeds?

Solution. For part a), the expected savings when planting the garden are

$$400 \times 0.3 - 100 \times 0.7 = \$50.$$

The savings from not planting the garden are \$0, so based on expected value, the best decision is to plant the garden.

For part b), if the rabbits have large appetites (L), then planting the garden would result in -\$25 of expected savings. If the rabbits have small appetites (S), then planting the garden will result in \$295 of expected savings.

If L,

$$\$100 \times 0.15 - \$100 \times 0.85 = -\$25$$

If S,

$$\$400 \times 0.79 - \$100 \times 0.21 = \$295$$

Not planting will always result in \$0 of savings. The optimal decision strategy is to plant the garden if S and buy vegetables from the store if L.

For part c), we use the optimal decision for each possible event L and S. The expected savings from purchasing the fast-growing seeds (but before actually purchasing the seeds) are

$$\$0 \times 0.60 + \$295 \times 0.40 = \$118$$

The maximum non-refundable amount that the family should be willing to pay for the fast-growing seeds is

$$\$118 - \$50 = \$68$$

4. *Using Baye's formula to update a prior belief.* Curling is a sport in which players slide a stone over ice toward a target. The association governing the sport has implemented drug testing. It is believed that 15% of all curlers use banned drugs to enhance performance. If a player uses banned drugs, the association may take away any prizes that the player has won; however, it undesirable to falsely accuse someone of using banned substances. The utilities for each decision and state of nature are

	drug use	no drug use
take away prizes	-100	-1000
do not	-600	0

Notice that there is a small dis-utility for taking prizes away from a drug user due to bad publicity for the sport. The test to detect drug use is less than 100% reliable. In particular, if D indicates that a player uses banned drugs, and $+/-$ indicate a

positive/negative test result, then the true positive rate and the true negative rate are

$$P(+ \mid D) = .97 \quad \text{and} \quad P(- \mid \overline{D}) = .97,$$

respectively, Given the utilities and the accuracy of the test, what is the best decision if a player has a positive test result? (The association wants to maximize expected utility.)

Solution. First we update the probability of drug use via Baye's formula.

$$\begin{aligned} P(D \mid +) &= \frac{P(D \cap +)}{P(+)} \\ &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid \overline{D})P(\overline{D})} \\ &= \frac{.97 \times .15}{.97 \times .15 + .03 \times .85} \\ &= .851 \end{aligned}$$

and then we can compute $P(\overline{D} \mid +) = 1 - P(D \mid +) = .149$. Using these posterior probabilities, the expected utilities are

$$\begin{aligned} E(\text{take away}) &= (-100)(.851) + (-1000)(.149) = -234 \\ E(\text{do not}) &= (-600)(.851) = -511 \end{aligned}$$

The best decision is to take away prizes when a player tests positive.

5. *Decisions under risk and sensitivity analysis.* The owners of a popular outdoor furniture company predict that their sales will double this coming year. The company is already producing the maximum amount of furniture possible in their current facility. They are considering expanding their manufacturing facility to accommodate the predicted increase in demand. If undertaken, the expansion will cost $500,000. If the demand doubles as predicted, revenue will increase by $800,000. If the predicted increase in demand proves to be too optimistic, revenue will increase by only $250,000. If the expansion is not undertaken, the company will lose $50,000 due to out-of-stock orders from agitated customers. The change in demand will be determined by next year's weather; more outdoor furniture is sold when the weather is nice. There is a 0.55 chance of good weather, which will result in a doubling of demand. There is a 0.45 chance of poor weather, which will result in only a slight increase in demand.

   (a) Should the manufacturing facility be expanded? The owners make decisions based on expected value.

   (b) How does the decision change with the probability of good/poor weather? To answer this question, you should perform a sensitivity analysis.

Solution. The decision table for this problem is

	0.55 good weather	0.45 poor weather
expansion	$300,000	-$250,000
no expansion	-$50,000	-$50,000

The expected payoff of each decision is

$$E(\text{expansion}) = 0.55 \times \$300{,}000 - 0.45 \times \$250{,}000 = \$52{,}500$$
$$E(\text{no expansion}) = -\$50{,}000$$

The company should expand the facility. As the probability of poor weather increases, the expected value of the expansion decreases. Let p represent the probability of poor weather.

$$\begin{aligned} E(\text{expansion}) &= 300{,}000(1-p) - 250{,}000p \\ &= 300{,}000 - 550{,}000p \end{aligned}$$

The company should expand the facility as long as

$$\begin{aligned} E(\text{expansion}) &\geq E(\text{no expansion}) \\ 300{,}000 - 550{,}000p &\geq -50{,}000 \\ -5{,}500{,}000p &\geq -350{,}000 \\ p &\leq \frac{7}{11} \approx .64 \end{aligned}$$

Expanding the facility is the best decision unless the probability of poor weather is greater than .64.

6. *The value of perfect information.* You are taking a date to a movie. You like horror movies but you are unsure whether your date likes them. You have two choices for the movie, "The Burrowers" which, despite its title, you've heard is a good movie, and "A Star is Born", which is rather safe as far as date movies go. Your decision problem with associated payoffs is

	Your date: likes horror movies	Your date: hates horror movies
The Burrowers	10	1
A Star is Born	4	7

   (a) Let p represent the probability that your date likes horror movies. What is the optimal decision as a function of p?

   (b) At this point you have no information for your date's movie preference (even the trail of social media cannot help you) and so you assign $p = 1/2$ as your subjective prior probability. Suppose that the date's best friend will tell you for sure whether your date likes horror movies, but this person is not so nice and wants to be paid. If the payoffs represent dollar amounts then how much are you willing to pay this person for perfect information?

Solution. Choose "The Burrowers" if

$$\begin{aligned} 10p + (1-p) &> 4p + 7(1-p) \\ 9p + 1 &> -3p + 7 \\ 12p &> 6 \\ p &> 1/2 \end{aligned}$$

Regarding part 6b, if you knew that your date liked horror movies you would choose "The Burrowers" for a pyoff of \$10 and if you knew the opposite then you would choose "A Star is Born" for a payoff of \$7. The expected payoff with perfect information, that is, before you actually receive the information, is

$$\frac{1}{2}(10) + \frac{1}{2}(7) = 8\frac{1}{2}$$

Without the information your expected payoff is $\$5\frac{1}{2}$ (for either choice). You would be willing to pay at most

$$\$8.50 - \$5.50 = \$3$$

for the perfect information.

7. *Quality control and the value of perfect information.* A worker in a factory producing potato chips notices that a small bolt is missing off of a piece of equipment in the packaging area and determines that it fell into a bag before it was sealed. Two pallets worth of chips have been produced since the last routine check was performed determining that the bolt was in place. The worker does not want to throw away all bags of chips that the bolt could possibly be in, so he decides to search for the bolt. He has to decide if he should start searching bags on pallet one or pallet two. If the bolt fell into a bag on pallet one and he searches bags on pallet one, the probability of finding the bolt by the end of the day is 0.5. If the bolt fell into a bag on pallet two and he searches pallet two, the probability of finding the bolt by the end of the day is 0.8. The worker's prior probability that the bolt is in a bag on pallet one is 0.7 (he thinks it was loose from some work done earlier in the day and fell off right away).

   (a) The worker only has time to search one pallet per day. In which pallet, one or two, should he look for the bolt? The worker wants to maximize his probability of finding the bolt.

   (b) Suppose that at the end of the day when the worker goes home, he has not found the bolt yet. Which pallet should he search on day 2? *Hint*: First update the prior probabilities that the bolt is in pallet one/two.

   (c) Suppose that on the second day, just before the worker begins to search for the bolt, a security guard approaches and tells him that he has access to security camera footage that would reveal when the bolt fell into a bag (revealing which pallet the bolt is in). How much should the worker be willing to pay for this perfect information? Note that the "units" for the payment are in terms of probabilities.

Solution. Let A be the event that the bolt is in pallet one and let B be the event that the bolt is in pallet two. Also, let F_A represent the event that the worker finds the bolt in pallet one (and similarly define F_B). We are given the following.

$$P(F_A \mid A) = 0.5 \qquad P(\overline{F}_A \mid A) = 0.5$$
$$P(F_B \mid B) = 0.8 \qquad P(\overline{F}_B \mid B) = 0.2$$

The worker's decision problem is

		Nature	
		0.7	0.3
		A	B
Worker	Look in A	.5	0
	Look in B	0	.8

The best option is to look in pallet one. He will find the bolt on the first day with probability $0.7 \times 0.5 = 0.35$.

On the second day, the worker's belief that the bolt is in pallet one is updated to

$$\begin{aligned} P(A \mid \overline{F}_A) &= \frac{P(\overline{F}_A \mid A)P(A)}{P(\overline{F}_A \mid A)P(A) + P(\overline{F}_A \mid B)P(B)} \\ &= \frac{0.5 \times 0.7}{0.5 \times 0.7 + 1 \times 0.3} \\ &= 0.538 \end{aligned}$$

which implies that $P(B \mid \overline{F}_A) = 1 - 0.538 = 0.462$. On day two, the worker's decision problem is

		Nature	
		0.538	0.462
		A	B
Worker	Look in A	.5	0
	Look in B	0	.8

and his best option is to look in pallet two. The worker will find the bolt with probability

$$0.462 \times 0.8 = 0.37$$

With perfect information, but *before* receiving the information, the worker's expected value for the probability of finding the bolt is

$$0.5 \times 0.538 + 0.8 \times 0.462 = 0.639$$

and in terms of probability "units", he should be willing to pay

$$0.639 - 0.37 = 0.269$$

for the information.

8. *Planting crops and the value of imperfect information.* A farmer can plant one of four crops. The yield in bushels per acre depends on the particular crop planted and the weather during the growing season. Regarding crop yields, the weather can be categorized as dry, moderate, or wet. The estimated crop yields, commodity prices, and weather forecast are provided in the table below.

	dry	moderate	wet	$/bushel
crop 1	30	50	60	$3.00
crop 2	25	30	40	$4.50
crop 3	40	35	35	$3.00
crop 4	60	60	60	$1.50
forecast	.2	.5	.3	

   (a) If the farmer maximizes expected monetary value, which crop should she plant and what is her expected revenue per acre?

   (b) Suppose that the farmer could pay an oracle to tell her with certainty which category of weather would occur during the growing season. How much should the farmer be willing to pay for this perfect information? Your answer should be in terms of $/acre.

   (c) Now suppose a start-up company that specializes in predictive modeling announces a new machine learning algorithm for weather forecasting. The predictions from the model are good, but not perfect. In particular, the company advertises the following accuracy for the weather forecasting model.

		actual		
		dry	moderate	wet
model	dry	.9	.1	.05
	moderate	.05	.8	.05
	wet	.05	.1	.9

   For example, if the weather actually turns out to be dry, the model would have predicted dry weather with probability .9, moderate weather with probability .05, and wet weather with probability .05. How much should the farmer be willing to pay (again in $/acre) for this less-than-perfect prediction of the weather? Settle in with a cup of coffee: this part requires some work. Here is a strategy.

      i) Develop a notation, e.g., x indicates the weather during the growing season, and x' indicates the prediction, where x is dry, moderate, or wet.

ii) Use Baye's formula to compute the posterior probabilities for each weather category given the predictions, e.g. you want to know $P(d \mid d')$, $P(m \mid d')$, etc.

iii) Determine the expected payoff for each case that the prediction is dry, moderate, and wet.

iv) The expected value of the prediction (i.e. the expected value of the less-than-perfect information) is the difference between the expected payoff with the information and without it.

Solution. Considering the estimated commodity prices, the payoff table and expected values in \$/acre for each crop are

	dry	moderate	wet	E(payoff)
crop 1	90	150	180	\$147
crop 2	112.5	135	180	\$144
crop 3	120	105	105	\$108
crop 4	90	90	90	\$90

To maximize expected monetary value the farmer should plant crop 1 and expect \$147/acre of revenue. For part 8b, if she knew for certain that the weather would be dry, moderate, or wet, then she would plant crop 3 with revenue \$120 per acre, crop 1 with revenue \$150 per acre, and crop 1 (or 2) with revenue \$180 per acre, respectively. Prior to learning the perfect information, the farmer's expected payoff with the new information is

$$.2 \times \$120 + .5 \times \$150 + .3 \times \$180 = \$153$$

and so the value of the perfect information is worth no more than

$$\$153 - \$147 = \$6 \text{ per acre}$$

Now for part 8c, the likelihoods are provided as follows (using the notation given in the problem description).

$$\begin{array}{lll} P(d' \mid d) = .9 & P(m' \mid d) = .05 & P(w' \mid d) = .05 \\ P(d' \mid m) = .1 & P(m' \mid m) = .8 & P(w' \mid m) = .1 \\ P(d' \mid w) = .05 & P(m' \mid w) = .05 & P(w' \mid w) = .9 \end{array}$$

Using Baye's formula to compute the posterior probabilities, we get

$$\begin{aligned} P(d \mid d') &= \frac{P(d \cap d')}{P(d')} \\ &= \frac{P(d' \mid d)P(d)}{P(d' \mid d)P(d) + P(d' \mid m)P(m) + P(d' \mid w)P(w)} \\ &= \frac{(.9)(.2)}{(.9)(.2) + (.1)(.5) + (.05)(.3)} \\ &= .735 \end{aligned}$$

and in a similar manner

$$\begin{array}{ccc} & P(m \mid d') = .204 & P(w \mid d') = .061 \\ P(d \mid m') = .024 & P(m \mid m') = .941 & P(w \mid m') = .035 \\ P(d \mid w') = .03 & P(m \mid w') = .152 & P(w \mid w') = .818 \end{array}$$

The expected monetary values given the predictions are

	E(dry)	E(moderate)	E(wet)
crop 1	\$108	150	173
crop 2	121	136	171
crop 3	116	105	105
crop 4	90	90	90

If the prediction is dry, the farmer would plant crop 2 for an expected revenue of \$121 per acre. Similarly for predictions of moderate and wet, she would plant crop 1 with expected revenue \$150 per acre and crop 1 with expected revenue \$173 per acre, respectively. Prior to obtaining the weather prediction, the expected value of having the prediction is

$$\begin{aligned} E(\text{payoff}) &= \$121 \times P(d') + \$150 \times P(m') + \$173 \times P(w') \\ &= \$121 \times .245 + \$150 \times .425 + \$173 \times .33 \\ &= \$150.50 \end{aligned}$$

Note that the unconditional probabilities of the predictions are obtained using the formula for total probability and they were computed during the application of Baye's formula. The value of the less-than-perfect informtaion is

$$\$150.50 - \$147 = \$3.50 \text{ per acre}$$

9. *The break-up.* You are considering breaking up with your boy(girl)friend; however, you know that he(she) may break up with you. Your utility is higher if you end the realtionship, but it's Friday and you have joint plans for the weekend. Here is the situation: on Friday (and possibly Saturday) you need to decide to either stay in the relationship, or end it.

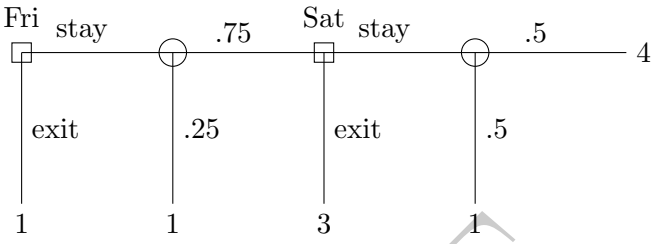
   i) *On Friday:* If you decide to end the relationship your payoff is one (and your done.) If you stay in the relationship then a random chance event happens where your boy(girl)friend ends the relationship with probability 0.25 and stays with probability 0.75.

   i) *On Saturday:* If you end the relationship your payoff is 3 (and your done.) If you stay then a chance event happens where your boy(girl)friend ends the relationship with probability 0.5 and stays with probability 0.5.

   If you stay in the relationship until Sunday, then you will definitely end it and your payoff is 4. If, on Friday or Saturday, your boy(girl)friend ends the relationship then your payoff is one.

Use backward induction to determine a policy for ending/staying that will maximize your expected payoff.

Solution. The decision tree is shown below. Working backwards, the expected value of "stay" on Saturday is 2.5, so the best decision is "exit", for a payoff of 3. On Friday, the expected value of "stay" is 2.5, which is better than "exit" with a payoff of 1. So, the policy that maximizes expected payoff is to stay in the relationship on Friday and exit the relationship on Saturday.

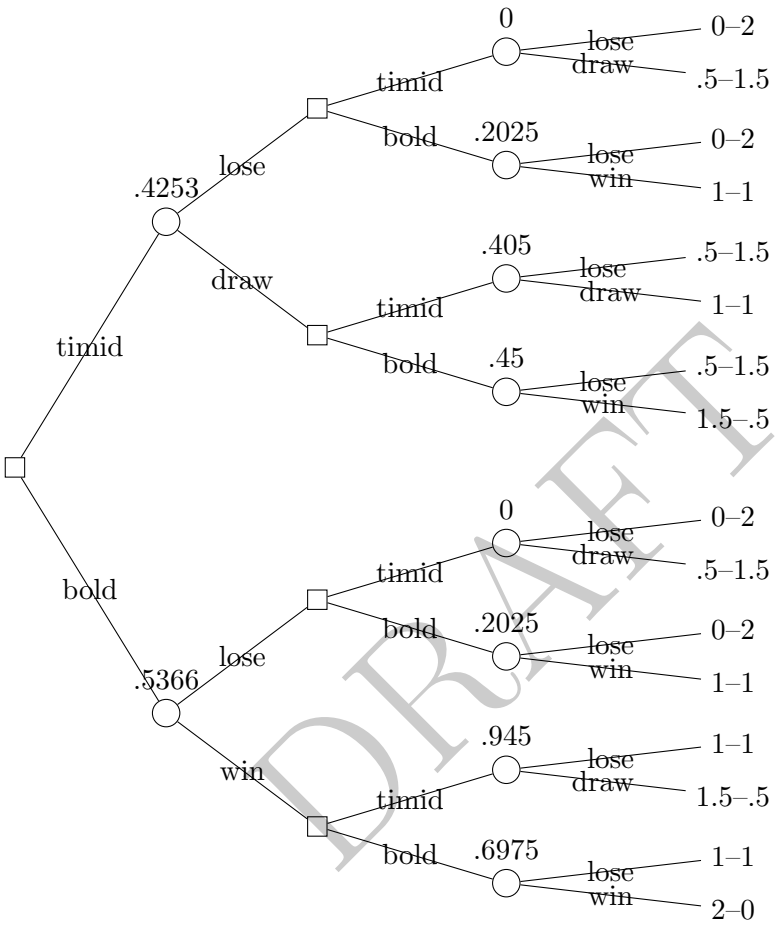


10. *Optimal play in a chess match.* Alice is preparing for a two–game chess match. In each game Alice can choose one of two strategies: timid play or bold play. The strategy selected in the second game can differ from the strategy selected in the first game. If Alice plays bold then she will win with probability p_w and lose with probability $1-p_w$. If Alice chooses timid play then the match will result in a draw with probability p_d and she will lose the match with probability $1- p_d$. Note that bold play results in a either a win or a loss, and that timid play results in either a draw or a loss. At the conclusion of each game, each player is awarded 1 point for a win, 1/2 point for a draw, or zero points for a loss. If the score is tied at the of two games, then the match goes into sudden death and the first person to win a game wins the match. Let $p_w = 0.45$ and let $p_d = 0.9$. Which strategy should Alice select for the first game of the match, timid or bold? What about the second game? In the second game, Alice's play is conditional on the outcome of the first game. Find the policy that will maximize Alice's probability of winning the match. You may want to draw the a decision tree and use backward induction.

The payoff for Alice is the probability of winning the match. This means that at the end of two games, a score of 2–0 is just as desirable as a score of 1.5–0.5 (where Alice's points are listed first). In both cases, Alice wins the match and the payoff is one. Instead, suppose that in the first game Alice plays bold and loses, and in the second game Alice plays timid and draws. Then the final score is 0.5–1.5, and Alice loses the match for a payoff of zero. Also, consider that in the event of a tied score at the end of two games Alice should play bold in sudden death. If she plays timid during sudden death then the best that she could hope for is to drag out the match indefinitely. The payoff for Alice for a tied score after two games is p_w.

Solution. The optimal policy for Alice is to play bold in the first game. If she wins the first game, then she should play timid in the second game. If Alice loses the first game, then she should play bold in the second game. In the diagram below, the the end of each decision branch has been labeled with the (expected) probability of

winning the match (from the chance node onward). They were computed by using backward induction and at each leaf node (i.e. after two games) treating a win as one, a loss as zero, and a tie as p_w. Notice that because Alice can adapt her playing style in the second game (and her opponent cannot), Alice's overall (expected) probability of winning the match is greater than 1/2 even though the probability of winning any single game is less than 1/2.



Games Against an Opponent

11. *Elimination of dominated strategies.* Two street vendors, A and B, are located near a major tourist attraction. The proportion of customers captured by each vendor depends on the merchandise sold by that vendor and by her competitor. A customer gained by one is lost to the other. Each vendor can stock one of the following: clothing, ice cream, or souvenirs. The possible strategies and proportion of customers captured are as follows.

If both shops sell souvenirs, A captures 25% of the customers.
If both shops sell clothing, A and B split the customers evenly.
If both shops sell ice cream, A and B split the customers evenly.
If B sells ice cream and A sells souvenirs, A captures 10%.
If B sells clothing and A sells ice cream, A captures 90%.
If B sells souvenirs and A sells clothing, A captures 10%.
If A sells clothing and B sells ice cream, A captures 100%.
If A sells souvenirs and B sells clothing, A captures 75%.
If A sells ice cream and B sells souvenirs, A captures 40%.

Model the decision of each vendor as two-person zero-sum game and find a solution by elimination of dominated strategies.

Solution. The game is

		B		
		clothing	ice cream	souvenirs
A	clothing	.50	1	.10
	ice cream	.90	.50	.40
	souvenirs	.75	.10	.25

For A, ice cream strictly dominates souvenirs and for B, souvenirs strictly dominates clothing, leaving a 2×2 game.

		B	
		ice cream	souvenirs
A	clothing	1	.10
	ice cream	.50	.40

Now for B, souvenirs dominates ice cream. Then, A will choose to sell ice cream over clothing. So, the best strategies for A and B are to sell ice cream and souvenirs, respectively. Using this pair of strategies, A will capture 40% of the customers.

12. *A two-person zero-sum game.* A professional football player believes that the team he plays for should be allocating more money to the salaries of the players, so he wants his contract to be changed to pay him more. His two options are to play in the upcoming season or not play in the upcoming season and hope the team will negotiate with him. The team knows that he is a valuable player but does not want to pay him more or go through the process of negotiations. The team has come up with three options to deal with the situation: negotiate, refuse to negotiate and play the season without him, or increase the player's salary by a set amount with no other negotiations. Keep in

mind that the player wants to maximize his salary, and the team wants to minimize their costs, which means keep salaries as low as possible. The utilities/payoffs to the player and to the team are described next.

If the player plays and the team does not negotiate, the player's salary will not change. If the player does not play and the team does not negotiate, the player will find a different job as a broadcaster for a payoff of 1 because he is such a well-known person. This is bad publicity for the team and hurts their jersey sales. If the player plays but the team still negotiates, the player will end up with a payoff of 3. If the player had chosen to not play and the team negotiates, the negotiations will go poorly and the player will end up with a payoff of -2 for having to deal with costs related to poor publicity. If the team decides to increase the player's salary with no negotiations, the player will end up with a payoff of 2 no matter what he chooses to do.

Formulate the 2-by-3 game and determine the best strategy for the player and for the team. Who is most likely to come out ahead in this situation?

Solution. The game is

		Team		
		negotiate	don't negotiate	increase salary
Player	play	3	0	2
	don't play	-2	1	2

The team's strategy of a set increase is dominated by the strategy to not negotiate, so there is no reason that they would chose to offer a pre-determined increase without negotiations. The reduced game is

		Team	
		negotiate	don't negotiate
Player	play	3	0
	don't play	-2	1

Since there is no saddle point, the best strategies for the player and for team are mixed. The player should mix the strategies "play" and "don't play" in the ratio 1:1. The team should mix the strategies "negotiate" and "don't negotiate" in the ratio 1:5. Using the player's mixing ratios against the team's strategy of "negotiate" the value of the game is computed as

$$\frac{1 \times (3) + 1 \times (-2)}{2} = 1/2.$$

The player is more likely to come out ahead.

13. *Marketing strategies.* Two peanut butter companies, Doodle's and Lola's, are deciding on their marketing strategy for the upcoming year. They know that they are each other's main competitor and that the demand for peanut butter is constant, so a gain in sales for Doodle's is a loss of sales for Lola's. Each company has two possible options for packaging their product: standard packaging and sponsored packaging (e.g. put an action hero on the package). If both companies choose to market only their sponsored packaging, Doodle's will gain an extra 2% of the market's sales. If both companies choose to market their standard packaging, Doodle's will loose 2% of the market. If Doodle's markets their sponsored product and Lola's markets their standard product, Doodle's will gain 10% of the market. If Lola's markets their sponsored product and Doodle's markets their standard product, Doodle's will gain 8% of the market.

    (a) Formulate this decision problem as a two-person zero-sum game and determine the optimal marketing strategy for each company. Note that they are able to change their packaging throughout the year, so a mixed strategy is possible.

    (b) Compute the value of the game.

    (c) Suppose that a member of Doodle's marketing team quits her job and goes to work for Lola's. She tells her new co-workers about the strategy that Doodle's is planning to use. She is even able to tell them the probabiliites with which Doodle's will market their standard product and their sponsored product. Lola's marketing team now knows that Doodles is more likely to market the sponsored product than the standard product. Armed with this knowledge, they choose to only market their standard product, thinking that this will improve their payoff. Is Lola's argument valid? In other words, does the value of the game change if Lola's knows Doodle's optimal strategy?

Solution. The game is

		Lola's: standard	Lola's: sponsored
Doodle's	standard	-2	8
	sponsored	10	2

Note that there is no saddle point, and so the best strategy is mixed. Doodle's should mix the strategies standard and sponsored in the ratio 4 to 5, while Lola's should mix their strategies of standard to sponsored in the ratio 1 to 2. The corresponding probabilities are (4/9, 5/9) for Doodle's and (1/3, 2/3) for Lola's.

To compute the value of the game, note that when Doodle's markets the standard product, they receive a payoff of -2 with probability 1/3 and payoff of 8 with proba-

bility 2/3. The value of the game is

$$\frac{1 \times -2 + 2 \times 8}{3} = \frac{14}{3} = 4.67$$

On average Doodle's comes out ahead.

Regarding part 13c), the team's thought process is not valid. As long as one player sticks to the optimal mixed strategy, the value of the game does not change.

14. *The birthday gift.* Liam's birthday is coming up and he can't wait to see what he will get as a gift. Liam's parents want the gift to be a surprise, but they always hide gifts in either the kitchen or the basement. Liam plans to search for the gift when his parents are busy, but he knows that even if he searches the room that contains the gift, he may not find it. If the gift is hidden in the kitchen and Liam searches the kitchen, he will find the gift with probability 0.75. If the gift in hidden in the basement and he searches the basement, then he will find it with probability 0.5. If he searches the wrong room, there is no way he will find the gift. Assume that the payoff to Liam for finding the gift early is the same as the payoff to the parents of keeping the gift a surprise. Formulate this game as a two-person, zero-sum game. Liam is the row player and his parents are the column player. Find the optimal strategies for both players.

Solution. Liam has two possible actions for this game: search the kitchen or search the basement. His parents also have two options: hide the gift in the kitchen or hide the gift in the basement. The game is

		Parents	
		hide in kitchen	hide in basement
Liam	search kitchen	3/4	0
	search basement	0	1/2

and the optimal strategy is the same for both players. Each should play a mixture of $(2/5,\ 3/5)$.

15. *Workforce staffing.* A salon is trying to decide how many stylists they should have available for walk-in customers. The hourly walk-in demand is specified by the following probability distribution, where p_n is the probability of having n customers in one hour.

n	1	2	3	4	5
p_n	.10	.20	.35	.25	.10

Each stylist costs the salon \$20 per hour. If a stylist has a customer that hour, the salon will charge the customer \$45 for service. Each customer requires about an hour

of time from a stylist. If a stylist does not have a customer that hour, he/she will complete a different task in the salon which adds $9 of value. If all stylists are busy with a customer and an additional customer arrives, that customer will be turned away and will go to a different salon. How many stylists should be available each hour in order to maximize profit for the salon?

Solution. Let Q represent the available number of stylists (quantity) and let z represent the demand. There are two situations. First, if $Q \geq z$

$$\begin{aligned} E(\text{profit}) &= 45z - 20Q + 9(Q - z) \\ &= 36z - 11Q \end{aligned}$$

Otherwise if $Q < z$, then

$$\begin{aligned} E(\text{profit}) &= 45Q - 20Q \\ &= 25Q \end{aligned}$$

We use the distribution of demand to compute the expected profit for each possible level of staffing.

		z					
		.1	.2	.35	.25	.1	
		1	2	3	4	5	E(payoff)
Q	1	25	25	25	25	25	25
	2	14	50	50	50	50	46.4
	3	3	39	75	75	75	60.6
	4	-8	28	64	100	100	62.2
	5	-19	17	53	89	125	54.8

The best decision is to have 4 stylists available.

Utility Theory

16. *Expected utility of a day off.* Utilities are the basis for making rational decisions. Sydney has the day off from work and she is trying to decide what to do. Her top three options and their utilities are:

option	utility
go to the beach	100
go hiking	75
stay home and watch movies	50

With only this information, Sydney would choose to go to the beach for the day. However, there is some uncertainty (or risk) in Sydney's decision. The weather and the people she runs into at the beach or while hiking impact her utility. Sydney does not enjoy being outside when it is raining. If it rains while she is at the beach, her utility is decreased by 80. If it rains while she is hiking, her utility is decreased by 30. On the other hand, Sydney enjoys running into friends. If she runs into a friend

at either the beach or on the hiking trail, it will increase Sydney's utility by a factor of 1.5 (after any disutility has been appplied).

At the beach, it is rainy 40% of the time and Sydney has a 30% chance of running into a friend. On the hiking trail, it is rainy 20% of the time and she has a 50% chance of running into a friend. Compute Sydney's expected utilities for each option. What is the "rational" decision for Sydney?

With this exercise, there are three points to be made:

(a) Rational choice means that agents (like Sydney) make decisions that maximize their utility (or payoff, or happiness). It doesn't mean that agents only care about themselves. Agents have preferences over states of the world. An agent might have a preference for making another agent happy, and that would be reflected in her utility for that state of the world.

(b) Each agent has a utility function that maps states of the world to real numbers that represent the agent's level of happiness.

(c) When there is uncertainty over which state will occur, then an agent's utility is her expected utility.

Solution. We must calculate the expected utility for each possible decision. The expected utility of going to the beach is:

$$\begin{aligned} E(\text{utility}) &= 100 - 80 \times 0.40 + (100 - 80 \times 0.40) \times 0.50 \times 0.30 \\ &= 78.20 \end{aligned}$$

The expected utility of going hiking is:

$$\begin{aligned} E(\text{utility}) &= 75 - 30 \times 0.20 + (75 - 30 \times 0.20) \times 0.50 \times 0.50 \\ &= 86.25 \end{aligned}$$

The expected utility of staying home watching movies is 50 because it is influenced by neither the weather nor running into friends. The "rational" decision for Sydney is to go hiking.

17. *Attitudes toward risk.* Consider the following scenarios for Alice and Bob. Then, draw reasonable utility functions for each person. Alice's utility function should be in the range zero to \$100K. Bob's should be from zero to \$40.

Alice has worked hard to save \$25K for a down payment on a house. She is ready to buy the house, but just before closing, she is offered the following hot tip. If she invest invests \$25K in a cryptocurrency, there is an 80% chance that she will make \$50K within one month, but a 20% chance that she will lose everything. The expected monetary value of the opportunity is

$$0.8 \times \$75\text{K} - 0.2 \times \$25\text{K} = \$55\text{K},$$

but Alice makes the \$25K down payment even though the EMV of the investment is higher. The certainty of getting the house is worth more to her than the uncertain gain from the investment.

Bob has \$20 in his possession. He *really* wants to see the show at First Avenue, but tickets cost \$40. Over on 7th Street, someone comes along and offers Bob the following gamble.

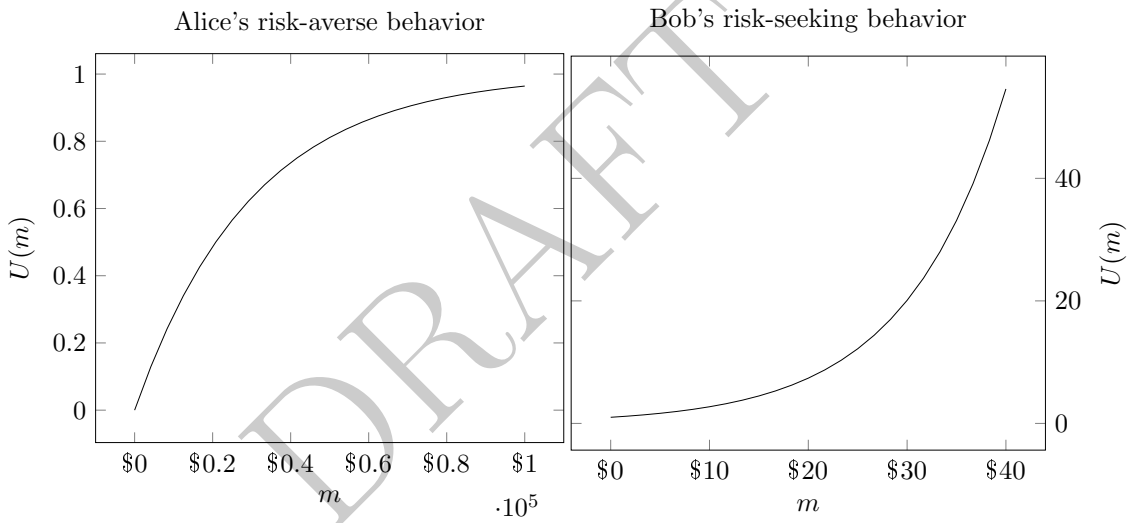
> Put in \$20. Roll a die. If the die comes up "6", then the player wins \$20.
> If the die comes up any other number the player loses the \$20.

If Bob is lucky enough to win the gamble, he would have \$40 for the ticket. The expected monetary value of the gamble is

$$\frac{1}{6} \times \$40 - \frac{5}{6} \times \$20 = -\$10,$$

but Bob takes the bet because the chance to see the show is worth more to him than having \$20 and *not* seeing the show.

Solution. The important point is the general shape of the utility functions, not the magnitudes of the utilities. Alice is risk-averse and Bob is risk-seeking.



18. *Registering for classes.* You are deciding to register for one of two classes, IE 5571 or IE 5591. If you register for IE 5571 you know that you will get a B for sure, but IE 5591 is a lottery because the department has not yet decided who will teach the class. If Professor England teaches the class then you will get an A, but if Professor Doroudi teaches the class then you will get a C. You will not know the identity of the instructor until class begins and you cannot withdraw from the course once registered. The points associated with the letter grades are

A	B	C	D
4.0	3.0	2.0	1.0

Now, you value having fun as much or more than improving your GPA. In fact, your utility for the points x associated with a letter grade is $u(x) = \sqrt{x}$. That is to say, your utility for getting a B is $\sqrt{3}$. The ISyE department head tells you that the probability that Professor Doroudi will teach IE 5591 is 2/5.

(a) For which class will you register?

(b) In general, what is your risk profile toward letter grades? (A *brief* answer will suffice.)

(c) Your preferences satisfy the conditions of the Expected Utility Theorem. Furthermore, your utility function $u(x)$ was determined via questions pertaining to preferences between pairs of alternatives (including lotteries). Is it appropriate to say that receiving an A is twice as desirable as receiving a D?

Solution. The utility of registering for IE 5571 is

$$u(B) = \sqrt{3}$$

By the expected utility property, the utility of a lottery is its expected utility. So the utility of registering for IE 5591 is

$$\begin{aligned} u\left(L\left(\frac{3}{5},\ A,\ C\right)\right) &= \frac{3}{5}u(A) + \frac{2}{5}u(C) \\ &= \frac{3}{5}\sqrt{4} + \frac{2}{5}\sqrt{2} \\ &= \frac{6+2\sqrt{2}}{5} \end{aligned}$$

which is greater than $\sqrt{3}$, so you will register for IE 5591. In general you are risk-averse with respect to letter grades. It is *not* appropriate to say that receiving an A is twice as desirable as receiving a D. Utility functions are defined for interval scales, not for ratio scales.

19. *Preference orderings and utility functions.* This exercise is inspired from discussion in [11]. Suppose you have a goal to be a great long-distance runner. There are two choices involved with this goal in the form of $(x; y)$ where x represents the number of miles you are able to run in a day after you have trained for a month, anywhere from zero to 30 miles, and y represents the total number of miles ran to train over the course of the month, anywhere from zero to 500. Both x and y are continuous distances, and $x \leq y$. In general you prefer to be able to run farther with less training, but you always prefer to be able to run father no matter how much you have to train. For example,

$$(x = 25 \text{ miles}; y = 500 \text{ miles}) \succ (x = 24 \text{ miles}; y = 200 \text{ miles})$$

The symbol '$\succ$' means "is preferred to". Don't confuse that symbol with the mathematical inequality symbol '$>$' (although I suspect that the resemblence is intended). Note that, as defined, your preference ordering satisfies the completeness and transitivity properties (see the discussion on page 69). Is it possible to represent your preferences with a single (real-valued) number? That is to say, is there a function $u(x, y) : (x, y) \mapsto \mathbb{R}$ with the following property

$$(x_1, y_1) \succ (x_2, y_2) \implies u(x_1, y_1) > u(x_2, y_2)$$

To make things a little easier, you can restrict u to be a linear function of x and y. Support your answer with an explanation . . . a formal proof is great, but not required.

Solution. I think it is not possible to represent u as a linear function of x and y. Define $w = 500 - y$, so that we can represent an alternative as (x, w) with the interpretation that more is always better, but x still has priority. Now, a linear function will have the form

$$u(x, w) = Ax + Bw$$

where A and B are constants. It is enough to show a situation where $(x_1; w_1) \succ (x_2; w_2)$ but that $u(x_1, w_1) < u(x_2, w_2)$. We can write $x_2 = x_1 - \delta_x$ and $w_2 = w_1 + \delta_w$ where $\delta_x > 0$ (so that $x_1 > x_2$). Now,

$$u(x_1, w_1) = Ax_1 + Bw_1$$

and

$$\begin{aligned} u(x_2, w_2) &= u(x_1 - \delta_x, w_1 + \delta_w) \\ &= A(x_1 - \delta_x) + B(w_1 + \delta_w) \\ &= Ax_1 - A\delta_x + Bw_1 + B\delta_w \\ &= Ax_1 + Bw_1 + (B\delta_w - A\delta_x) \end{aligned}$$

We need only to show that $B\delta_w - A\delta_x > 0$.

$$\begin{aligned} B\delta_w - A\delta_x &> 0 \\ B\delta_w &> A\delta_x \\ \frac{\delta_w}{\delta_x} &> \frac{A}{B} \end{aligned}$$

Note that $\delta_w = w_2 - w_1 = y_1 - y_2 \leq 500$, but because x is continuous, we can choose δ_x and δ_w to satisfy the last inequality, which means that $u(x_1, w_1) < u(x_2, w_2)$.

Chapter 4

Data Analysis

4.1 Descriptive Statistics

Point estimate of a population proportion. Consider the pizza delivery data that is available on the class Moodle page. Construct a point estimate of the probability that the amount of tips received in a shift is greater than $60. What is the standard error of your point estimate? You can do the calculations by hand or use the software of your choice. If you use software, you can use it in any way you like. For example, I used R as a calculator to simply help with the required computations.

```
pizza <- read.table("pizza.txt", header=TRUE)
attach(pizza)
x <- sum(Tips > 60)
n <- length(Tips)

# follow the formula for the point estimate and the standard
# error of a sample proportion
phat <- x/n
se <- sqrt((phat*(1-phat))/n)
```

```
> phat
[1] 0.2413793
> se
[1] 0.0212373
```

Confidence Interval. An article in the *Journal of Heat Transfer* describes a method of measuring the thermal conductivity of high-purity iron. Using a temperature of 100°F and a power input of 550 W. The following 10 measurements of thermal conductivity (in Btu/hr – ft – °F) were determined.

$$41.60,\ 41.48,\ 42.34,\ 41.95,\ 41.86$$
$$42.18,\ 41.72,\ 42.26,\ 41.81,\ 42.04$$

A point estimate of the population mean thermal conductivity (at 100°F and 550 W) is the

sample mean

$$\overline{X} = 41.924$$

The standard error of the sample mean (i.e. the standard error of the point estimate) is

$$\text{se}(\overline{X}) = \frac{S}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

where S is the sample standard deviation. Notice that the standard error is about 0.2 percent of the sample mean, indicating a relatively precise point estimate of thermal conductivity. Occasionally you will hear people refer to the coefficient of variation (CV).

$$\text{CV} = \frac{S}{\overline{X}}$$

The CV is another measure of the spread of the data. *Question:* What are the units of the CV?

Now suppose we want to construct a 95% confidence interval for the population mean thermal conductivity μ. So, our confidence level will be $1 - \alpha = .95$, and $\alpha = .05$. Since there are only 10 sample data points, we will use the t distribution. From our discussion in class we know that a $(1 - \alpha)\%$ confidence interval for μ is

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

From the tabulated values of the t distribution, we see that $t_{\alpha/2,n-1} = t_{.025,9} = 2.262$. A 95% confidence interval for μ is

$$41.924 \pm 2.263 \times \frac{0.284}{\sqrt{10}}$$

or (41.721, 42.127).

Using R, we could do the following

```
> x <- c(41.60,41.48,42.34,41.95,41.86,42.18,41.72,42.26,41.81,42.04)
> alpha <- .05
> n <- length(x)
> xbar <- mean(x)
> se <- sd(x)/sqrt(n)
> cp <- qt(1-alpha/2,n-1)

> xbar + c(-1,1)*cp*se
[1] 41.72076 42.12724
```

To use the Normal distribution instead of the t distribution, obtain the critical point as

```
> cp <- qnorm(1-alpha/2)
```

Question: When I got the critical point in R, why did I use `qnorm(1-alpha/2)` and not `qnorm(alpha/2)`?

A model for the price of a stock. Consider the UNH stock price data that are posted on the class web site. Suppose that on November 19, 2018, your portfolio consists of 1 share of stock and one call option with a strike price of \$265 and a duration of six months (125 trading days). You are to obtain a 95% prediction interval for the value of the portfolio in six months. To estimate the price of the stock six months into the future, use the multiplicative model

$$S_n = S_0 e^{X_1+X_2+\ldots+X_n}, \quad n \geq 0$$

where S_i is the price of the stock on day i and X_i is a random disturbance on day i. The X_i are distributed normal with mean μ and variance σ^2. The policy for exercising the call option is to wait until the last possible day and then exercise the option if the stock price is greater than the strike price. Here are the steps you need to perform:

1. Using the historical data, estimate the mean μ and standard deviation σ of the random disturbance.
2. Use Monte Carlo simulation to estimate the price of the stock in six months (125 days). That is, six months from the last trading day in the data, which is Friday Nov 16, 2018.
3. Compute the cash flow from the call option.
4. Obtain a 95% prediction interval for the value of the portfolio (one share of stock plus one call option.)
5. Use the bootstrap technique to assess the accuracy of your estimate for the standard deviation σ that was computed in step 1. That is, use the bootstrap to compute the standard error of σ.

4.2 Descriptive Graphics

4.3 Exercises

Descriptive Statistics

1. *Taking averages.* Consider two datasets: 1, 5, 9 and 2, 4, 6, 8.
   (a) Denote the sample means of the two datasets by $\overline{x}$ and $\overline{y}$. Is it true that the average $(\overline{x}+\overline{y})/2$ of $\overline{x}$ and $\overline{y}$ is equal to the sample mean of the combined dataset with 7 elements?
   (b) Suppose we have two other datasets: one of size n with sample mean $\overline{x}_n$ and another dataset of size m with sample mean $\overline{y}_m$. Is it always true that the average $(\overline{x}_n+\overline{y}_m)/2$ of $\overline{x}_n$ and $\overline{y}_m$ is equal to the sample mean of the combined dataset with $n+m$ elements? If no, then provide a counterexample. If yes, then explain this.

(c) If $m = n$, is $(\overline{x}_n + \overline{y}_m)/2$ equal to the sample mean of the combined dataset with $n + m$ elements?

2. *Computing the sample variance.* The following rule is useful for the computation of the sample variance (and standard deviation). Show that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - (\overline{x}_n)^2$$

where $\overline{x}_n = \left(\sum_{i=1}^{n} x_i\right)/n$

3. *Summarizing a data set with statistics.* In the `datasets` package in R, there is a data set called `discoveries` which contains the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959.

   (a) From the data set of 100 observations, determine the following summary statistics: minimum, maximum, mean, median, mode, first quartile, third quartile, and standard deviation.

   (b) Create a table that shows the distinct values of the number of discoveries per year and their counts, i.e. the number of times that each value occurred.

4. *Point estimate of a population mean.* Suppose the following data points are a sample of a golfer's scores over his last 20 rounds. Construct a point estimate of his average score. What is the standard error of your point estimate?

```
73,69,65,70,67,67,78,72,74,71,70,69,70,67,68,73,70,77,72,69
```

5. *Standard error when estimating a proportion.* In a survey, a random sample of 1200 students are asked whether they prefer online or in-person classes. Out of the 1200 students, 424 said they prefer online classes. Compute a point estimate of the overall proportion of students that prefer online classes and calculate the standard error of your estimate.

6. *Lognormal distribution parameter estimation.* The file `component-lifetimes.txt` contains the time to failure for 1345 components (in hours). The times are known to come from a Lognormal distribution.

   (a) Estimate the parameters of the failure time distribution.

   (b) Use the parameters to estimate the mean time to failure.

   (c) Use the parameters to estimate the probability that a component lasts longer than 10,000 hours.

7. *Estimation of the size of a population.*

8. *Confidence interval for a mean.* Restaurants are making more use of their data on service times for planning purposes. The data in the file `restaurant-service-times.txt` contains 220 observations on the time in minutes from seating until departure in one particular restaurant.

(a) Construct a 95% confidence interval for the true mean service time. Remember that this data is just a sample from a larger population, the true distribution of which is unknown.

(b) Do you think that the Central Limit Theorem applies to this data? Why or why not?

9. *Validation of a simulation model.* A simulation model of a job shop was developed to investigate different scheduling rules. To validate the model, the scheduling rule currently used was incorporated into the model and the resulting output was compared against observed system behavior. By searching the previous year's records, it was estimated that the average number of jobs in the shop was 22.5 on a given day. The file `job-shop-wip.txt` contains the results for the average number of (simulated) jobs in the shop for 30 independent replications of the simulation model, each for 30 days of simulated time.

   One metric that can be used to validate the simulation model is to construct a 95% confidence interval for the true mean number of (simulated) jobs in the shop. If the confidence interval contains the value 22.5, then, the model captures the average work-in-process for the job shop[1]. Construct the confidence interval and comment on the result.

10. *Confidence interval for a proportion.*

11. *Comparison of system configurations.*

12. *Bootstrap confidence interval.* The data file `clinical-trial.txt` contains data on 20 patients. 10 patients were randomly assigned to receive Medicine A, and 10 were randomly assigned to receive Medicine B. The data represents the responses of the patients to their assigned medicines. Use the bootstrap technique to determine whether or not there is a difference in the median response between the two medicines. In order to do this, you should take B bootstrap samples of the data (where $B \geq 200$). For each bootstrap sample, compute the difference in median response between the two medicines. That is to say, for each bootstrap i sample you will compute

$$\text{median}(A_i) - \text{median}(B_i)$$

   where $\text{median}(A_i)$ is the median of the data associated with medicine A in the ith bootstrap sample (i goes from 1 to B). Determine a 95% confidence interval for the difference in median response by taking the .025 and .975 quantiles of the bootstrap replicates. If your confidence interval does not contain zero, then you should conclude that there is a difference. If that is the case, then state the direction of the difference. That is, state which medicine has a higher median response.

13. *Required sample size.* The following data are observations from a past study of hummingbird migration rates in miles flown per day. These observations are from 30 different birds. A researcher would like to construct a two-sided, 95% confidence interval for the average rate (in miles per day). The researcher would like for the width

[1] You would want to know that the baseline simulation model is accurate before simulating any proposed changes to the scheduling rules.

of the confidence interval to be no larger than 1 day. It is very expensive to attach identifiers to the birds, and so the researcher has asked you to determine the smallest sample size that will achieve the desired confidence interval. What sample size do you suggest?

```
17,17,22,18,19,21,21,23,21,25,
19,21,19,20,20,21,19,20,18,17,
18,20,19,23,18,22,18,22,18,24
```

Solution. The width W of a confidence interval is 2 times the half-width

$$W = 2 \times z_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

and the researcher would like $W \leq 1$. We can use the data to estimate a standard deviation $S = 2.133$. We do not know n. In fact, that is the question we are trying to answer. You can assume that you will have a sample size at least as big as, say, 30 birds. Using $n = 30$ and the tabulated values for the Normal distribution, we find that $z_{\alpha/2} = z_{.025} = 1.96$. Then, solving for n, the required sample size is

$$n \geq 4 \times (z_{\alpha/2} \times S)^2 = 4 \times (1.96 \times 2.133)^2 = 69.9$$

We would recommend a sample size no smaller than 70 observations. Note that you will want to use the 30 observations that you already have, so a confidence interval of width 1 day will require approximately 40 additional observations.

Descriptive Graphics

14. *College students and driving speed.* The file `speed_gender_height.csv` contains 1325 observations on gender, height, and the fastest speed ever driven (in mph) for a sample of college students.

    (a) Create a boxplot of speed by gender. That is to say, make one boxplot for males and one boxplot for females, but put them side-by-side on the same plot.

    (b) Make an x–y plot with height on the x–axis and speed on the y–axis. Color the plotted points according to gender. Place a legend that shows the color associations. Another option is to use different plotting symbols rather than color to distinguish males and females.

Chapter 5

Predictive Models

Predictive models help us summarize and understand data in order to 1) make predictions about the future, and/or 2) explain interesting relationships in the data. The ultimate purpose of a model can be for prediction or explanation, but we usually want some combination of both. In this chapter, the term "predictive model" includes the explanatory aspects. Predictive models are broadly categorized into two types: regression and classification. The difference is in the nature of the response variable, i.e. the "thing" that we are trying to predict or explain. Regression indicates a numeric response, e.g. a stock price, population growth, or sales of a product. Classification means that the response is categorical, e.g. gender, category of an email, or presence/absence of a disease.

Regardless of whether the model is one of regression or classification, modeling is a process that almost always involves iterating over the following steps [2].

1. obtaining data
2. choosing a candidate model
3. fitting the model, i.e. using software to estimate the model parameters
4. interpretation of the fitted model parameters
5. diagnostics to see in what ways the model *fails* to fit the data

The wide availability of high-quality, open-source software for statistical computing has made step 3 the easiest part. In this chapter, we concentrate on steps 2, 4, and 5.

5.1 Regression

Model Formulas

Regression (and classification) models are simplified descriptions of data that involve mathematical relationships. In addition to the data itself, a model consists of 1) a formula that specifies the mathematical relationships among the variables, and 2) a description of how well the data agree with the model. Let's start with an example of how model formulas are represented in the R programming language. Recall the data set from Chapter 4 on college students and driving speed. This data contains 1325 observations on gender, height, and the fastest speed ever driven (in mph) for a sample of college students. Do we really think that, on average, males drive faster than females (or vice versa)? Similarly, is there any relationship between height and speed? Consider the following model formula written in R.

```
speed ~ height + gender
```

You should read the formula above as "Speed is modeled as a linear function of height and gender". The term to the left of the "$\sim$" is the response (a.k.a dependent variable, output variable) and the terms to the right of the "$\sim$" that are separated by a "+" are the predictor variables (a.k.a explanatory variables, independent variables, features, input variables). Mathematically, the formula specifies the following model.

$$speed_i = \beta_0 + \beta_1 height_i + \beta_2 gender_i + \epsilon_i, \qquad i = 1, \ldots, n \tag{5.1}$$

The coefficients β_0, β_1, β_2 along with the variance σ_ϵ^2 of the error term ϵ are the *parameters* to be estimated from the data. The index i refers to the ith observation in the data set (in this case, the ith person). Notice that neither the coefficients nor the error term appear in the R formula. They are inferred and so we do not need to enter them. In particular the R formula does not include the intercept term β_0. It is there by default; this is usually what we want. The R formula above is equivalent to

```
speed ~ 1 + height + gender
```

If we do not want an intercept term in our model, perhaps because we want to force the fit to go through the origin, then we need to explicitly remove the intercept, like this:

```
speed ~ -1 + height + gender
```

"Fitting" a model means to obtain estimated values for the parameters (i.e. the coefficients) along with an indication of how well the data agrees with the fitted parameters. The data set on gender, height, and speed is from a sample of 1325 students. If we obtained the same data on a different sample of students (but drawn from the population of college students with similar attributes), we would expect the fitted parameter values from the second model to be close to, but not exactly the same as, those from the original sample. Imagine obtaining such a data set and fitting the model many times. For each coefficient, you would obtain a distribution of values. Usually, we are interested in whether this distribution of values could plausibly contain the value zero. If it does, then we say that there is no statistical relationship between the associated predictor variable and the response.

Understanding Regression Output

When fitting a model in the R programming language, the paradigm is to save the fitted model as an object. Then, we can extract explanatory information from the object and use that information to make inferences or predictions on future observations. For simplicity, let's ignore gender and fit a regression model with speed as the response and height as the only predictor.

```
1 Speed <- read.csv("../data/speed-gender-height.csv")
2
3 fm <- lm(speed ~ height, data = Speed)
4 summary(fm)
```

```
>
Call:
lm(formula = speed ~ height, data = Speed)

Residuals:
    Min      1Q  Median      3Q     Max
-98.859  -8.248   1.673  11.635  81.992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7145     9.8671  -0.072    0.942
height        1.3830     0.1489   9.287   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.75 on 1300 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.06222,        Adjusted R-squared:  0.0615
F-statistic: 86.25 on 1 and 1300 DF,  p-value: < 2.2e-16
```

Figure 5.1: Summary regression output from R for the model with height as the only predictor variable.

The call to `lm()` on line 3 fits the model and stores the information about the fir in an object named `fm`[1]. We then ask for summary information about the fit. The output from line 4 is shown in Figure 5.1.

The model formula is repeated in the summary output. Recall that, mathematically, this model is

$$speed_i = \beta_0 + \beta_1 height_i + \epsilon_i \tag{5.2}$$

where ϵ_i is the error term, i.e. the difference between the model's fitted value of speed for observation i and the actual value of speed for observation i. This is also known as the residual. One assumption about linear regression models is that the residuals are Normally distributed with mean zero and some variance σ_ϵ^2.

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \tag{5.3}$$

The output section entitled `Residuals:` provides a rough idea of this distribution. If the residuals are distributed according to 5.3, then we would expect the distribution to be symmetric, with median equal to zero and with first and third quartiles that are the same distance from zero. Of course, we are working with a sample of data, so we cannot expect perfect symmetry. The residuals for this particular model appear to be symmetric, or at least not too far from it.

Model 5.2 has three parameters: the intercept β_0, the coefficient on height β_1 (a.k.a the slope), and the variance of the residuals σ_ϵ^2. All three parameters are estimated from the

[1]You can name the object anything you like. I tend to use `fm` for "fitted model". It's nice and short.

data. The output section entitled `Coefficients:` provides a table of information about the fitted coefficients. The table has four columns:

`Estimate`	the estimated value of the coefficient
`Std. Error`	the uncertainty in the estimated value
`t value`	the test statistic in the test for whether $\beta_j = 0$
`Pr(>|t|)`	the associated p-value

Let's take the coefficient on height as an example. From the table, the estimate is $\hat{\beta}_1 = 1.383$. A regression coefficient is a statistic; it is a summary of data. Because a regression coefficient can be written as the sum of random variables, the Central Limit Theorem tells us that $\hat{\beta}_1$ is Normally distributed with mean equal to its true value and with variance equal to $\sigma^2_{\beta_1}$, or equivalently with standard deviation equal to σ_{β_1}. The standard deviation of a statistic is called the standard error, and from the table we see that $\hat{\sigma}_{\beta_1} = 0.1489$.

Recall that we are usually interested in whether $\beta_1 = 0$, i.e. whether there is a relationship between speed and height. The formal "hypothesis test" is

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned} \tag{5.4}$$

where H_0 is called the null hypothesis, which is sort of like the default position. H_A is called the alternative hypothesis. The question is whether the data provides enough evidence to reject the null hypothesis in favor of the alternative hypothesis. The test is set up so that if you reject the null hypothesis, then you have found something interesting to report (e.g. "On average, taller people drive faster!") The third column of the summary output reports the standardized value of β_1 when the null hypothesis is true. It is the test statistic for the hypothesis test 5.4.

$$t_0 = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}} = \frac{1.383 - 0}{0.1489} = 9.288$$

If the null hypothesis is true, the test statistic t_0 has a t distribution with degrees of freedom equal to the number of observations in the data set minus the number of estimated coefficients. A value of 9.288 falls very far into the right tail of the t distribution. The area to the right of 9.288 is called the p-value. It represents the probability that null hypothesis really is true, given this data set. This value is reported in the last column under `Pr(>|t|)`. In our case, the p-value for β_1 is extremely small, less than 2×10^{-16}. What we conclude from all of this is that the data provides strong evidence of a statistical relationship between speed and height.

The numeric interpretation of the value $\hat{\beta}_1 = 1.383 \neq 0$ is that, on average, a one-unit increase in height corresponds to a 1.383-unit increase in speed. Here, we need to pay attention to the units in the data. Speed is reported in miles per hour and height is reported in inches. So, the data indicate that for each one inch increase in height, the fastest speed ever driven increases by 1.383 miles per hour (again, this is "on average"). Although the coefficient on height is statistically significant, is it *practically* significant? Let's say that 3 inches represents a practical difference in height. Then our model tells us that (on average) the difference in speed is a $3 \times 1.383 \approx 4$ miles per hour. The average "fastest speed" is about 91 miles per hour. So, while the effect is statistically significant, the magnitude of the effect is mediocre at best.

The estimated value for σ_ϵ is provided in the last section of the output. We see that $\hat{\sigma}_\epsilon = 21.75$. It is reported as the residual standard error (i.e. standard deviation), rather than as the variance of the residuals. Notice that R reports that this estimate is based on 1300 "degrees of freedom". We can think of the degrees of freedom as the effective size of the data set. In this case, we have 1325 observations in the data set, but 23 observations contain either a missing value for speed or a missing value for height. Also, we lose a degree of freedom for each estimated coefficient. This leaves $1325 - 23 - 2 = 1300$ degrees of freedom.

For completeness, the last two lines of the output show the values for R^2 and the F-statistic for the overall hypothesis test of whether there is a relationship between any of the predictor variables and the response. Model 5.2 has a single predictor variable. In this case, the overall hypothesis test is equivalent to the hypothesis test 5.4[2]. We can think of R^2 as the proportion of variation in the data that is explained by the model. To be specific, if the actual response values are y_i and the fitted values from the model are $\hat{y}_i$, where $i = 1, \ldots, n$, then R^2 is computed as follows.

$$\begin{aligned} TSS &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ R^2 &= 1 - RSS/TSS \end{aligned} \tag{5.5}$$

where $\bar{y}$ is the sample mean of the response. TSS represents the total sum of squares around the sample mean. It is the total variation in the response (if we were to simply use the sample mean to make predictions). RSS represents the sum of squares around the regression line. The ratio RSS/TSS is the proportion of (squared) variation that is *unexplained* by the model.

The reported value of $R^2 = 0.06222$ seems low, but this typical of real-world data sets. Figure 5.2 shows the data with the regression line from model 5.2 over-layed. There is a detectable slope in the data, but there is also a lot of variation around the regression line. The code to produce Figure 5.2 is

```
ggplot(Speed, aes(x=height, y=speed)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE)
```

Keep in mind that R^2 is not an indication of model "correctness". Rather, it is a measure of the tightness of the data to the fitted model. Indeed, models with high values of R^2 are very often not all that interesting. Figure 5.3 shows the data and over-layed regression line for a model from the often-cited `mtcars` data set that is available in R. This data set contains 32 observations on various attributes of automobiles. The regression model with `mpg` (miles per gallon) as the response and `wt` (vehicle weight) as the only predictor variable provides an $R^2 \approx 0.75$, which is quite large for a regression model. The model may be useful when designing an automobile, but the result is unsurprising.

[2]In the case of a single predictor variable, the F-statistic is equal to the squared value of the t-statistic

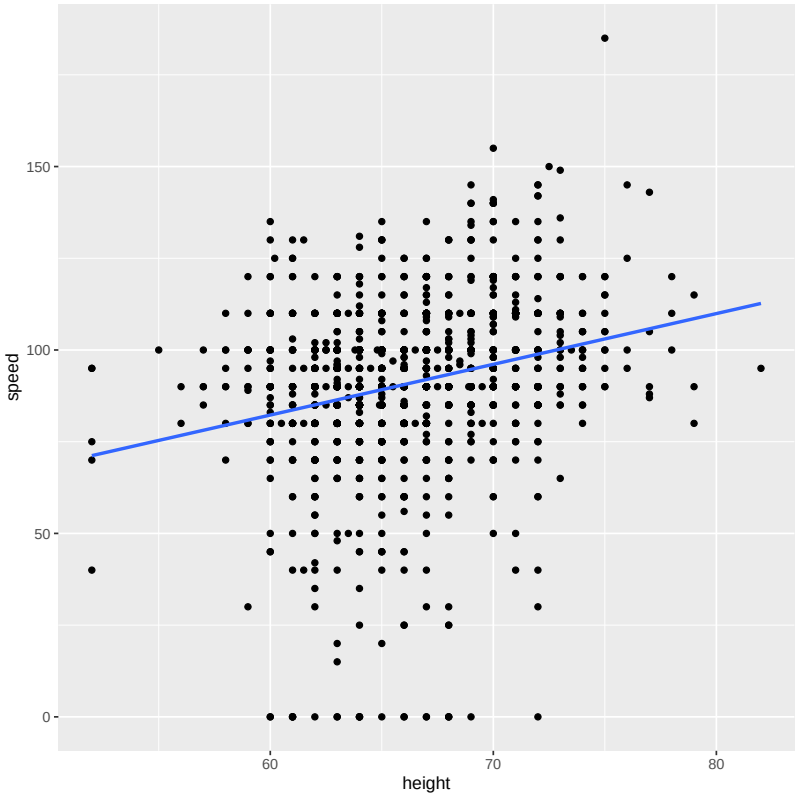


Figure 5.2: The fitted model `speed` $\sim$ `height`. Height is the single predictor variable. $R^2 \approx 0.062$.

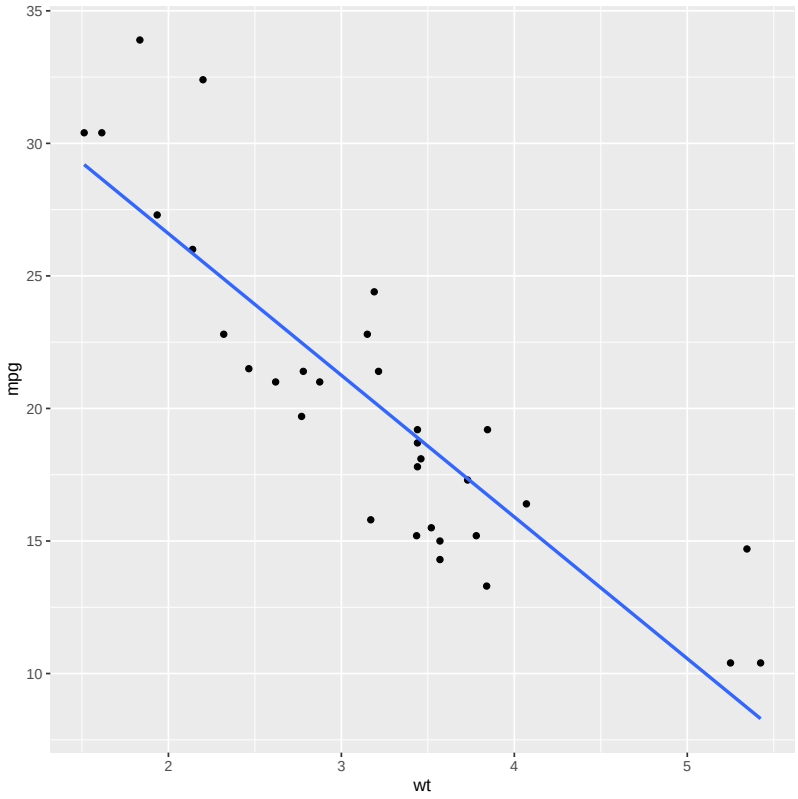


Figure 5.3: The model `mpg` $\sim$ `wt` from the `mtcars` data set. $R^2 \approx 0.75$.

Regression with a Categorical Predictor Variable

Is it possible to improve the fit of the model to predict fastest speed ever driven (and perhaps increase R^2)? Typically, two options are available to address this question: 1) more (and/or better) predictor variables, and 2) a different model.. The latter option is appropriate when, for example, we observe a nonlinear trend in the data. We would then want to create a model that could capture the nonlinear relationship, e.g., a nonlinear regression model, a tree-based model, or a neural network. Regarding the first option, the data set contains one other variable: `gender`. Let's add it as the second predictor variable.

```
1 fm2 <- lm(speed ~ height + gender, data = Speed, na.action="na.exclude")
2 summary(fm2)
```

Two changes are made to the call to `lm()`: 1) we add gender to the model formula, and 2) we explicitly exclude observations with missing values. The output from the call to `summary()` is shown in Figure 5.4. The residuals still appear to be (roughly) symmetric about zero. Notice that the table of information on the coefficients now has a row for β_2, the coefficient on gender, and that the label for the coefficient is `gendermale`. Recall that gender is a categorical predictor variable. In R, the categories (i.e. male and female) are called *levels*.

```
> levels(Speed$gender)
[1] "female" "male"
```

The levels are simply labels; however, the underlying mathematical model knows nothing about the labels. In model 5.1, which is repeated here, the categorical variable $gender_i$ is an indicator variable that takes the value 0 or 1.

$$speed_i = \beta_0 + \beta_1 height_i + \beta_2 gender_i + \epsilon_i, \qquad i = 1, \ldots, n \qquad \text{(5.1 revisited)}$$

where $height_i$ is a numeric value in units of inches and

$$gender_i = \begin{cases} 0 & \text{if person } i \text{ is female,} \\ 1 & \text{if person } i \text{ is male.} \end{cases}$$

For categorical variables, one level is called the "reference level". The effect of all other levels on the response are with respect to the reference level. In this case, the reference level is `female`. You can think of the variable $gender_i$ as a switch that gets flipped on if person i is male. When $gender_i = 1$, the estimated value for β_2 is added to the response. In this way, we can interpret $\hat{\beta}_2 = 5.7276$ as the average difference between males and females for the fastest speed ever driven. Now, we are in a position to understand why R concatenates the non-reference level to the identifier for the coefficient for gender (i.e. `gendermale`). The mnemonic is "add 5.7276 to the predicted response if observation i has `male` as the level for gender". For categorical variables with more than two levels, the `Coefficients:` table will contain an entry for each level except the reference level.

$(86.246 = 9.287^2)$. Also notice that the p-value for the overall test matches the p-value for the coefficient on height, the only predictor variable.

```
>
Call:
lm(formula = speed ~ height + gender, data = Speed, na.action = "na.exclude")

Residuals:
     Min       1Q   Median       3Q      Max
-100.234   -7.717    2.283   11.313   81.857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.6755    12.1534   2.030 0.042525 *
height        0.9699     0.1885   5.145 3.09e-07 ***
gendermale    5.7276     1.6142   3.548 0.000402 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.66 on 1299 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.07122,        Adjusted R-squared:  0.06979
F-statistic:  49.8 on 2 and 1299 DF,  p-value: < 2.2e-16
```

Figure 5.4: Summary regression output from R for the model with both height and gender as predictor variables.

In effect, the model 5.1 contains two regression lines, one for males and one for females. The effect of height on speed (the slope) is assumed to be the same for both genders. A plot of the fitted regression model is shown in Figure 5.5. The separation between the lines for male and female is the value $\hat{\beta}_2 = 5.7276$. If we have reason to believe that the effect of height on speed varies by gender, then we would add an interaction term to the model. The addition of interaction term may be warranted, but doing so complicates the interpretation of the model. This modification to the model is explored in Exercise 5.

The code to produce Figure 5.5 is shown below. On line 1, we add the predicted (i.e. fitted) values from the model to the `Speed` data frame so that we can overlay the fitted regression lines for males and for females. If we did not use the predicted values from the call to `predict()`, then the call to `geom_smooth` would use height alone as the predictor (i.e. it would not include gender). This is a work-around so that we can show the fit from the full model.

```
1 Speed <- mutate(Speed, fit = predict(fm2, na.action=NULL))
2
3 ggplot(Speed, aes(x=height, y=speed, color=gender)) +
4     geom_point() +
5     geom_smooth(method="lm", mapping=aes(y=fit))
```

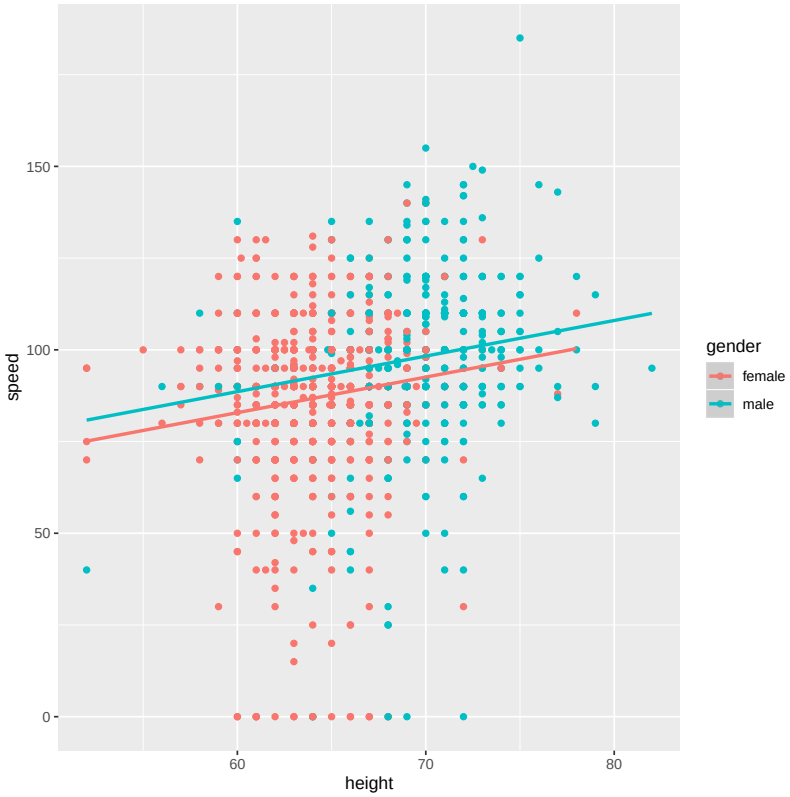


Figure 5.5: The fitted model `speed` $\sim$ `height + gender`. $R^2 \approx 0.071$.

Transformations on data.

Sometimes we can transform the data in order to model nonlinear relationships among variables, but in the framework of a linear model. With experience, you will be able to recognize situations when this technique may be appropriate. As an example, the `msleep` data set is available in R as part of the `ggplot2` package. This data set contains, among other things, the total hours of sleep per day for 83 species of mammals. To read about the data set, type

```
> ?msleep
```

at the R prompt. Suppose that an animal sciences researcher would like to explain the amount of sleep as function of brain weight. First, we plot total sleep in hours versus brain weight in kilograms.

```
ggplot(msleep) +
    geom_point(aes(x = brainwt, y = sleep_total))
```

The plot is shown in Figure 5.6. At first glance, the data does not appear promising for a linear model; however, just looking at the spread of the brain weight values along the x axis, we notice that the small values are "bunched up" and the larger values are spread out. This is an indication that a transformation may be appropriate. For mathematical (and biological) reasons, a logarithmic transformation is a common approach to make the data more amenable to the assumptions required by a linear regression model. By taking the logarithms of the brain weights, we penalize (or shrink) the large values toward the smaller values (on the log scale).

```
ggplot(msleep, aes(x = log(brainwt), y = sleep_total)) +
    geom_point() +
    geom_smooth(method="lm", se=FALSE)
```

Figure 5.7 shows total sleep versus the log-transformed brain weight along with a fitted regression line [3]. The pattern is now clear: on average, mammals with larger brains sleep less. Moreover, the residuals appear to be roughly symmetric about the regression line. For simplicity and reliability during the fitting process, we would like to use linear models whenever possible; however, the price we pay for transforming a predictor variable is increased complexity when we interpret the estimated parameters.

Fitting the model in R is straightforward.

```
fm <- lm(sleep_total ~ log(brainwt), data=msleep)
summary(fm)
```

Mathematically, this model is

$$sleep\_total_i = \beta_0 + \beta_1 \ln brainwt_i + \epsilon_i \tag{5.6}$$

[3] When taking a logarithmic transforming of a predictor variable, it's common to use the natural logarithm, but any logarithmic transformation will have a similar effect.

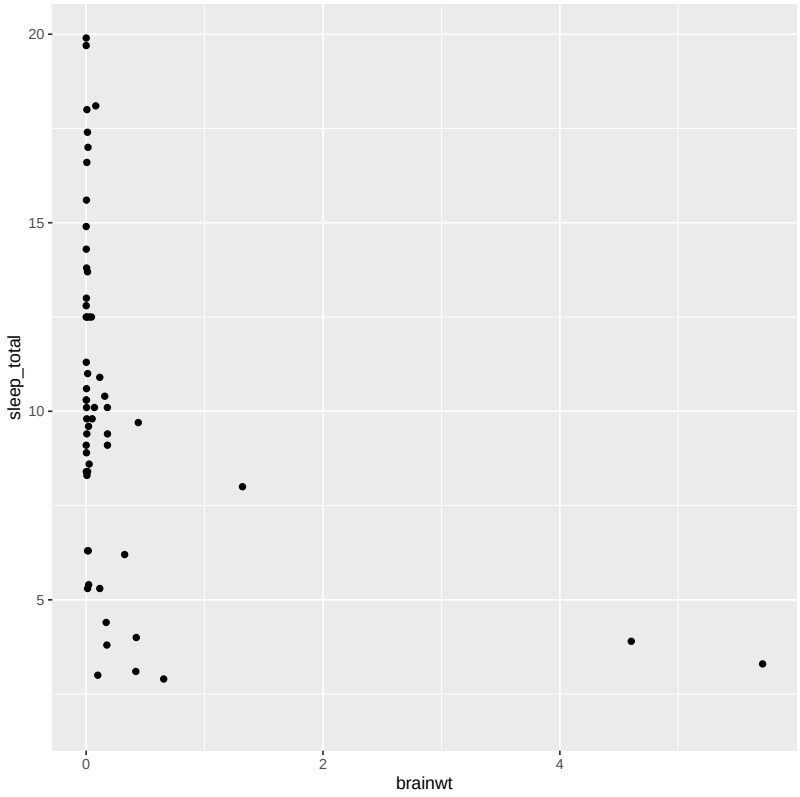


Figure 5.6: Total sleep in hours vs. brain weight in kg from the `msleep` data set.

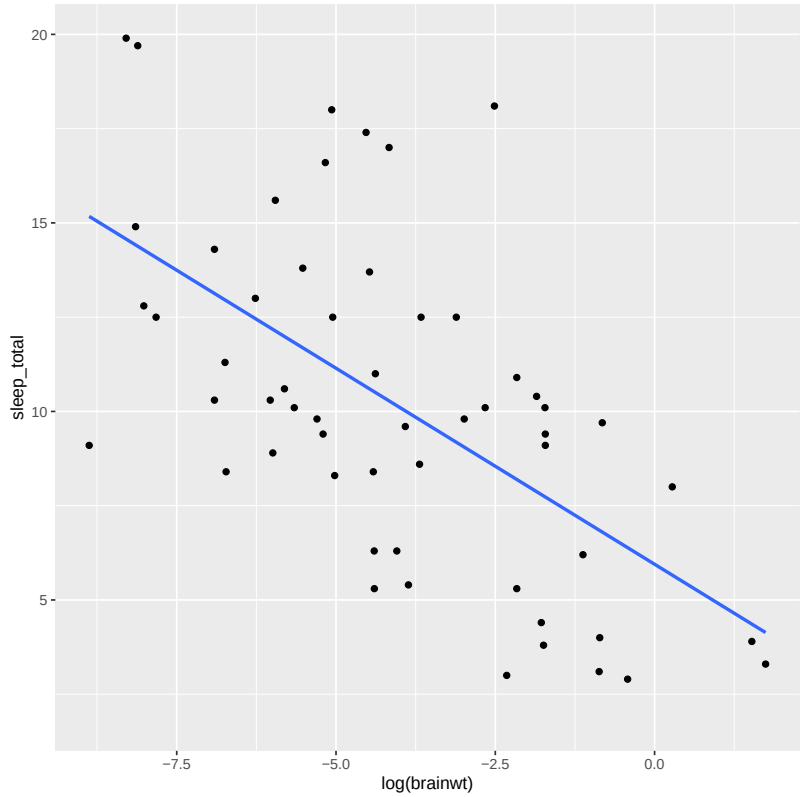


Figure 5.7: Total sleep in hours vs. logarithm of brain weight in kg from the `msleep` data set.

```
Call:
lm(formula = sleep_total ~ log(brainwt), data = msleep)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0734 -2.8483 -0.6479  2.4049  9.5395

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.9474     0.9135   6.510 2.57e-08 ***
log(brainwt)  -1.0397     0.1914  -5.432 1.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.588 on 54 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.3534,        Adjusted R-squared:  0.3414
F-statistic: 29.51 on 1 and 54 DF,  p-value: 1.361e-06
```

Figure 5.8: Summary regression output from R for mammalian sleep in hours with log-transformed brain weight as the predictor variable.

The summary output is shown in Figure 5.8. An understandable interpretation of $\hat{\beta}_1$, the coefficient on brain weight, is more difficult, but the basic linear relationship remains the same. A one-unit increase in the logarithm of brain weight corresponds to a decrease in total sleep time of about on hour (on average). Perhaps the easiest way to get an idea of the effect of brain weight on sleep time is to plug in some values and observe the difference. Suppose we are interested in comparing the difference in the amount of sleep for two mammals, one with a brain weight of 2 kg and one with a brain weight of 3 kg. The average difference in sleep time is about 0.421 hours, or about 25 minutes.

```
> (5.9474 -1.039*log(2)) - (5.9474 -1.039*log(3))
[1] 0.4212782
```

5.2 Classification

5.3 Time Series

5.4 Exercises

Regression

1. *Old Faithful.* In the `datasets` package in R there is a data set named faithful that contains data on the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The variables in the data set are

- `eruptions` the eruption time in minutes
- `waiting` the time in minutes until the next eruption

To view information about the data set, type

```
> ?faithful
```

at the R prompt. Use this data to perform the following exercises.

(a) Create histograms of `eruptions` and `waiting` (separately).

(b) Create a scatter plot (i.e. an x–y plot) with `eruptions` on the x axis and `waiting` on the y axis.

(c) From the histograms and the scatter plot that you created, what can you say about the behavior of the Old Faithful geyser?

(d) Fit a linear regression model using the `lm()` function with `waiting` as the response variable and `eruptions` as the only predictor variable. Print a summary of the results.

i. What is the interpretation of the intercept?

ii. What is your interpretation of the fitted coefficient on `eruptions`?

iii. You just observed an eruption of duration 4 minutes. Make a prediction on how long you will have to wait until the next eruption. Can you make any statement about the uncertainty in your prediction? In other words, can you give a range for the time until the next eruption? Don't worry about being exact with your range, just give something reasonable. .

2. *Cherry trees.* In the `datasets` package in R, the `trees` data set contains measurements on the girth (diameter) in inches, height in feet, and volume in cubic feet of 31 black cherry trees.

(a) Fit a regression model with volume as the response variable and girth as the predictor variable.

(b) Plot the data and overlay the fitted regression line.

(c) Provide an interpretation for the coefficient on girth.

(d) In your own words, state your interpretation of the p-value for the coefficient on girth.

3. *Fuel efficiency.* For this problem we will be using a dataset called `mtcars` from the `datasets` package in R. This dataset contains data about different types of cars. Fit a linear regression model using `lm()` with miles per gallon (`mpg`) as the response variable and the following predictor variables:

- number of cylinders (`cyl`)
- horsepower (`hp`)
- weight in thousands of lbs (`wt`)

So the model is

$$mpg_i = \beta_0 + \beta_1 cyl_i + \beta_2 hp_i + \beta_3 wt_i + \epsilon_i$$

Now do the following.

(a) Looking at the summary of the fitted model, the coefficient for weight $\beta_3 \approx -3.17$. What is the interpretation of β_3?

(b) How do the number of cylinders and horsepower affect fuel efficiency?

(c) Plot `mpg` as a function of `wt`. Overlay a fitted regression line from the full model onto the plot. When plotting the regression line you should show `mpg` at the average `cyl` and average `hp`. In other words, it's a two-dimensional plot, but for the other variables that are not shown, we compute `mpg` at their average values. So you want to overlay

$$mpg_i = \beta_0 + \beta_1 \overline{cyl} + \beta_2 \overline{hp} + \beta_3 wt_i$$

onto the data. You can use `coef()` to extract the coefficients from the fitted model object.

(d) Plot the actual `mpg` vs. the predicted (fitted) mpg. If your fitted model is stored in an object named `fm`, then you can get the predicted price as follows.

```
mtcars$pred <- fitted(fm)
```

or

```
mtcars$pred <- predict(fm)
```

(e) In the summary output of the fitted model, the estimated residual standard error is reported to be $\hat{\sigma}_\epsilon = 2.512$. Independently compute this quantity. In other words, use the actual values from the data and the fitted values from the model to compute the residual standard error yourself. The formula is

$$\hat{\sigma}_\epsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}}$$

where y_i and $\hat{y}_i$ are the actual and fitted values of observation i, respectively, n is the total number of observations, and k is the number of fitted parameters in the model. $n - k$ is the degrees of freedom.

(f) Do you think that a linear model is appropriate for this data?

4. *Linear regression with numeric and categorical predictors.* The following sales data were collected for one particular product from a company for the past 10 seasons. The data are the price of the product that the company itself charged, the price that its competitor charged (for the competitor's version of the same product), the corresponding sales of the company's product, and the season. This data is available in the file `sales.csv`.

company price	competitor price	sales (1000s)	season
\$10.2	\$9.9	71.1	winter
11.6	9.9	63.0	summer
9.8	11.7	71.7	winter
13.7	9.5	58.3	summer
12.0	8.9	61.8	summer
11.2	10.1	66.0	summer
10.2	11.1	71.2	winter
10.6	10.7	66.9	winter
9.5	12.6	72.5	winter
11.8	10.0	65.4	winter

(a) Create a visualization that shows all of the data on one plot. One idea is to plot sales vs. price, distinguish company and competitor price by symbol shape, and distinguish season by color.

(b) Fit a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where y represents the sales in 1000s of units, x_1 represents the company price in dollars, x_2 represents the competitor price in dollars, and x_3 is an indicator variable as follows

$$x_3 = \begin{cases} 0 & \text{if season is winter} \\ 1 & \text{if season is summer} \end{cases}$$

(c) Provide an interpretation each of the fitted parameters β_0, β_1, β_2, and β_3.

(d) Do you think that competitor price should be included in the model? Explain the reasoning for your answer.

(e) What is the expected company sales if the company and the competitor both set their prices to \$11 for the winter season? Use the full model, regardless of your answer to part 4d

5. *Adding an interaction term to a regression model.* Refer to model 5.1 where we express the fastest speed ever driven for a sample of college students as a function of height and gender.

$$speed_i = \beta_0 + \beta_1 height_i + \beta_2 gender_i + \epsilon_i, \qquad i = 1, \ldots, n \qquad \text{(5.1 revisited)}$$

The summary regression output is shown in Figure 5.4 and the fitted model is plotted against the data in Figure 5.5. In this model, the effect of height on speed is the same for both genders. Now, we want to investigate the possibility that the effect of height differs by gender. In other words, we want to allow the slope for height to be different for males and females (if the data indicates so). To do this, we add an interaction term to the model.

$$speed_i = \beta_0 + \beta_1 height_i + \beta_2 gender_i + \beta_3 height_i gender_i + \epsilon_i, \qquad i = 1, \ldots, n \quad (5.7)$$

Remember that in the mathematical model $height_i$ and $gender_i$ are both numeric. The term $\beta_3 height_i gender_i$ is literally the product of $beta_3$, $height_i$, and $gender_i$. The corresponding model formula in R is:

```
speed ~ height + gender + height*gender
```

(a) Fit this model in R. The summary output will contain an entry for the coefficent β_3 (the coefficient on the interaction term).

(b) Provide an interpretation for the effect of height on speed and also for the effect of gender on speed. *Hint:* When an interaction term is present, we can no longer interpret the coefficients in isolation. For example, the effect of height on speed will involve both β_1 and β_3.

(c) Plot the fitted model against the data. Your plot will be similar to Figure 5.5 except the slopes will differ by gender.

(d) Do you think that, at least statistically, the interaction term is appropriate for this model?

6. *Computing R^2.* Refer to model 5.2 and the output shown in Figure 5.1. Using the equations 5.5 as a guide, implement your own computation of R^2 for model 5.2. In other words, using R as a calculator, compute R^2 on your own.

7. *Logistic regression.* The file `harrell.csv` contains data on 40 people. The variables are `age` in years, `gender`, a categorical variable with two levels, `female` and `male`, and `response`, a 0/1 indicator variable for whether the person responded to a medical treatment (1 means that the person responded). Fit a logistic regression model with `response` as the dependent variable and `age` and `gender` as the independent variables.

   (a) How does the probability of response change for a 42-year-old male compared to a 52-year-old male?

   (b) Which gender has a higher probability of response to the medical treatment?

   (c) What is the effect on the odds of response for a one–year increase in age?

   (d) Make a plot of the probability of response as a function of age, with one curve for females and one curve for males.

8. *Transformations on data.* Suppose that a response variable u is related to an explanatory variable v as

$$u = \gamma_0 e^{\gamma_1 v}$$

where γ_0 and γ_1 are parameters to be estimated. Even though the relationship between u and v is nonlinear, explain how you could use simple linear regression to find estimates for γ_0 and γ_1.

Solution. We can transform the relationship into a linear one by taking logarithms.

$$\ln(u) = \ln(\gamma_0) + \gamma_1 v$$

Then we have the form of a linear regression model with $\beta_0 = \ln(\gamma_0)$ and $\beta_1 = \gamma_1$. After fitting the linear model,

$$\hat{\gamma}_0 = e^{\hat{\beta}_0}, \qquad \hat{\gamma}_1 = \hat{\beta}_1$$

Classification

Time Series

9. *Flattening the curve.* The file `us-daily-covid-cases.csv` contains data on the daily number of confirmed COVID-19 cases in the United States from March 1st to July 7th 2020.

   (a) Construct a single plot that shows the actual number of cases, a 7-day moving average, and exponentially smoothed average for the case where the smoothing parameter is $\alpha = 0.3$.

   (b) Briefly explain the main differences between the moving average values and the exponential smoothing values.

10. *Simulation of a simple trading strategy.* The file `BTC-USD.csv` contains one year's worth of price data on Bitcoin. The `Close` column is the price of Bitcoin in USD on `Date`.

    (a) Read the data into an R data frame.

    (b) Turn the `Date` field into a proper Date variable rather than a character string.

    (c) Make a time-series plot with `Date` on the x axis and `Close` on the y axis.

    (d) Make a scatter plot with `Volume` on the x axis and `Close` on the y axis. Do you see a pattern? Try making this plot using a logarithmic transformation on `Volume` and `Close`.

    (e) Create a one-step-ahead forecast for the closing price using simple exponential smoothing. Use a smoothing parameter value of $\alpha = 0.5$.

    (f) Using your forecast simulate a very simple trading strategy. The strategy is: if the forecast is for a price increase then buy, otherwise if the forecast is for a price decrease then sell. Don't worry about the bid/ask spread, trading fees, or any of the messy details. You are simply buying or selling one Bitcoin at the closing price. Keep track of your gain/loss. How did you do? Do you have any criticisms of this trading strategy?

    (g) Now plot your forecasted/predicted price and the actual price on the same plot. Do you see a general behavior in the forecast? Experiment with different values for α.

Appendix A

A Primer on Probability

The get-away. You plan to rob four banks and then escape to Mexico. In each robbery the probability of getting caught is 1/3, and the outcome of each robbery is independent of that of the others. What is the probability that you end up in jail?

Since the outcome of each robbery is independent, the probability of *not* ending up in jail is the probability that you never get caught.

$$P(\text{don't get caught}) = \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) = \frac{16}{81}$$

and so the probability of ending up in jail is $1 - \frac{16}{81} \approx 0.8$.

Suppose that the probability of exposure to the flu during an epidemic is 0.6. Experience has shown that a serum is 80% successful in preventing an inoculated person from acquiring the flu, if exposed. A person not inoculated faces a probability of 0.9 of acquiring the flu if exposed. Two persons, one inoculated and one not, are capable of performing a highly specialized task in a business. Assume that they are not at the same location, are not in contact with the same people, and cannot expose each other. What is the probability that at least one will get the flu?

Let A be the event that the inoculated person gets the flu, and let B be the event that the person who is not inoculated gets the flu. The probability that at least one of them gets the flu is

$$P(A \cup B)$$

From the problem description we can safely say that the events A and B are independent. Also note that a person's exposure to the flu is independent of inoculation, so that

$$P(A) = (.6)(.2) = .12 \qquad \text{and} \qquad P(B) = (.6)(.9) = .54$$

Then,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A)P(B) \\ &= .12 + .54 - (.12)(.54) \\ &= .5952 \end{aligned}$$

Conditional probability. In a population of 100,000 females, 89.835% can expect to live to age 60, while 57.062% can expect to live to age 80. Given that a woman is 60, what is the probability that she lives to age 80?

We can use the definition of conditional probability. Let E be the event that a woman lives to be 60, and let F be the event that a woman lives to be 80.

$$P(F \mid E) = \frac{P(F,E)}{P(E)} = \frac{P(E \mid F)P(F)}{P(E)} = \frac{1 \times .5706}{.8984} = .6352$$

Also, read this answer that was taken from Grinstead and Snell. It nicely describes the idea of conditional probability.

> The original sample space can be thought of as a set of 100,000 females. The events E and F are the subsets of the sample space consisting of all women who live at least 60 years, and at least 80 years, respectively. We consider E to be the new sample space, and note that F is a subset of E. Thus, the size of E is 89,835, and the size of F is 57,062. So, the probability in question equals 57,062/89,835 = .6352. Thus, a woman who is 60 has a 63.52% chance of living to age 80.

Baye's formula. An automobile manufacturer makes cars with three types of engines. Of all cars made by this manufacturer, 45% are hybrids (gasoline-electric), 35% are gasoline, and 20% are diesel. From past data, it is known that 5% of the cars with hybrid engines fail the emissions test, while 12% of cars with gasoline engines and 25% of cars with diesel engines fail the test. A record of a failed emissions test is selected at random, what is the probability that it is for a car with a diesel engine?

engine type	% of cars	% failed
hybrid	45	5
gasoline	35	12
diesel	20	25

Let H represent the event that a car has a hybrid engine. Similarly, G and D represent gasoline and diesel. Let F represent the event that a car failed the emissions test. We want to know $P(D \mid F)$.

$$\begin{aligned} P(D \mid F) &= \frac{P(D \cap F)}{P(F)} \\ &= \frac{P(F \mid D)P(D)}{P(F \mid H)P(H) + P(F \mid G)P(G) + P(F \mid D)P(D)} \\ &= \frac{(.25)(.20)}{(.05)(.45) + (.12)(.35) + (.25)(.20)} \\ &= .437 \end{aligned}$$

Updating a prior belief with new information or Bayesian updating. Your prior probability that a certain coin is biased to always land heads up is 0.1. Now you toss the coin three times and observe that it lands heads up every time. What is your posterior probability that the coin is biased to always land heads up? Use Baye's formula to compute the posterior probability. Use the Binomial distribution to compute the likelihood.

Let B indicate that the coin is biased, and let $3H$ indicate an outcome of three heads. We are given (or can determine)

$$P(B) = 0.1, \quad P(3H \mid \overline{B}) = \left(\frac{1}{2}\right)^3, \quad P(3H \mid B) = 1$$

We can use Baye's Theorem to compute the posterior probability.

$$\begin{aligned} P(B \mid 3H) &= \frac{P(B \text{ and } 3H)}{P(3H)} \\ &= \frac{P(3H \mid B)P(B)}{P(3H \mid B)P(B) + P(3H \mid \overline{B})P(\overline{B})} \\ &= \frac{(1)(0.1)}{(1)(0.1) + \left(\frac{1}{8}\right)(.9)} \\ &\approx 0.47 \end{aligned}$$

Appendix B

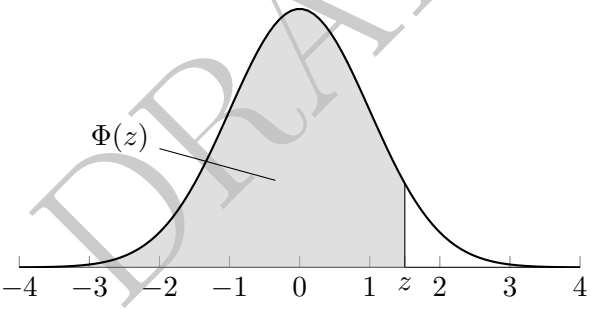
Standard Normal Distribution

Standard Normal random variables are denoted by an upper case Z.

$$Z \sim \mathcal{N}(0, 1)$$

The cumulative distribution function (CDF) is denoted $\Phi(z)$. It is represented by the area under the curve and to the left of z.

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$



The probabilities in the table on the opposite page were generated using R. For example, `pnorm(1.31)` returns 0.9049. Given an area, one can obtain the corresponding quantile by working backwards through the table. Using R, `qnorm(.9)` returns 1.28.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997

References

[1] Stephen P. Bradley, Arnoldo C. Hax, and Thomas L. Magnanti. *Applied Mathematical Programming*. `http://web.mit.edu/15.053/www/AMP.htm`. Addison-Wesley Publishing Company, 1977.

[2] John Chambers and Trevor Hastie. *Statistical Models in S*. Wadsworth and Brooks/-Cole, 1992.

[3] Herman Chernoff and Lincoln E. Moses. *Elementary Decision Theory*. Dover Publications, Inc., 1959.

[4] Robert Fourer, David M. Gay, and Brian W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Second. Brooks/Cole Publishing Company, 2003. ISBN: 0-534-38809-4.

[5] GAMS. *Quick Start Tutorial*. (accessed June 5, 2020). URL: `https://www.gams.com/latest/docs/UG_TutorialQuickstart.html`.

[6] *GLPK (GNU Linear Programming Kit)*. `www.gnu.org/software/glpk`.

[7] R. Duncan Luce and Howard Raiffa. *Games and Decisions*. Dover Publications, Inc., 1957.

[8] George L. Nemhauser and Laurence A. Wolsey. *Integer and Combinatorial Optimization*. John Wiley and Sons, 1988.

[9] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

[10] Martin Peterson. *An Introduction to Decision Theory*. Cambridge University Press, 2009.

[11] Michael D. Resnik. *Choices: An Introduction to Decision Theory*. University of Minnesota Press, 1987.

[12] Richard E. Rosenthal. *A GAMS Tutorial*. (accessed June 5, 2020). URL: `https://www.gams.com/latest/docs/UG_Tutorial.html`.

[13] Leonard Savage. *The Foundations of Statistics*. second revised. Dover Publications, 1972.

[14] Paul Williams. *Model Building in Mathematical Programming*. Fourth. John Wiley and Sons, 1999.