An Introduction to Prescriptive and Predictive Modeling

Contents

Chapter 1

Prescriptive Modeling

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. [6, 3].

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

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:

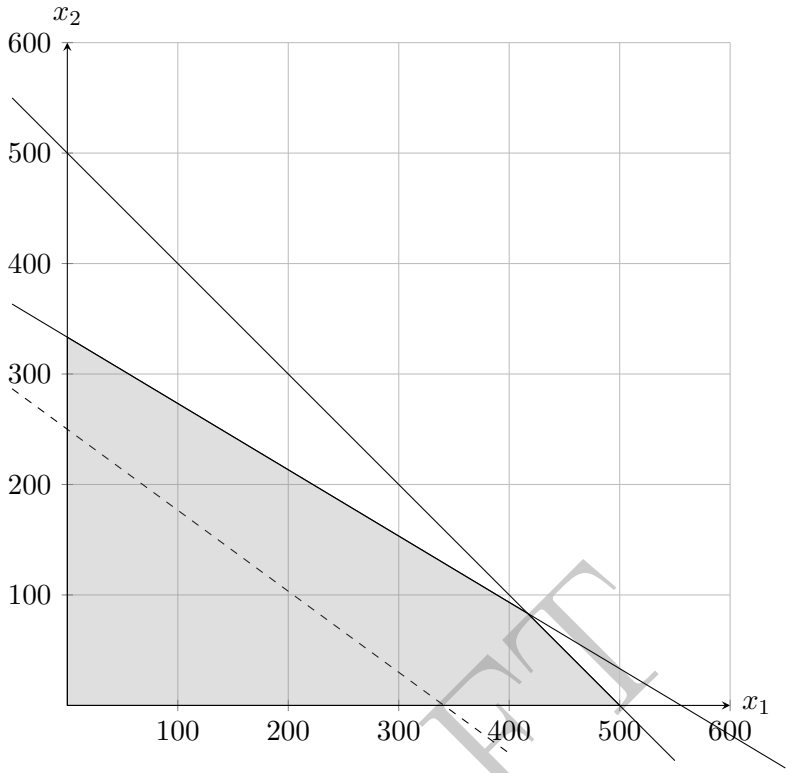


Figure 1.1: Feasible region for the farm problem.


```
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 3a. 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 3b 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 [7]. 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 `*` [2].

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 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`.

```
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`)

```
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`)

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. 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 [5].

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).

```
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 [4].

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

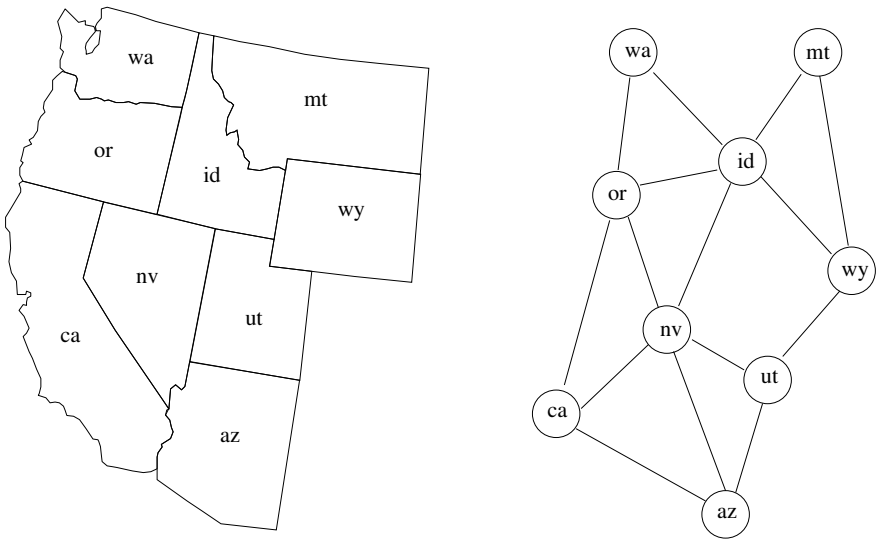


Figure 1.8: Western states and adjacency graph

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 [4] for an implementation of a heuristic to obtain an upper bound on the required number of colors.

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 Dynamic Programming

1.4 Inventory Models

```
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`)

1.5 Exercises

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.

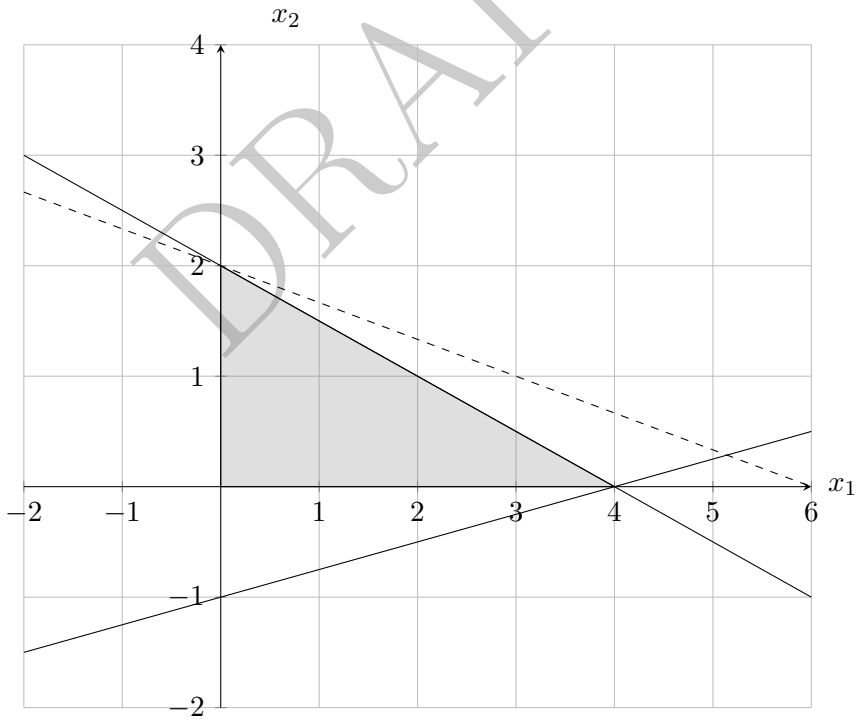
   (a)

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

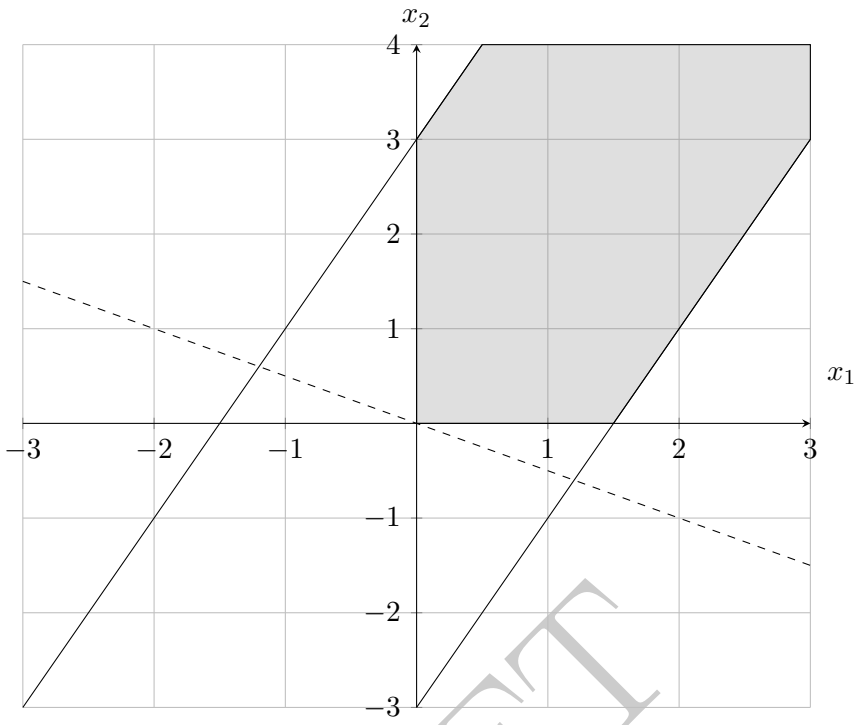
   (b)

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

   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.

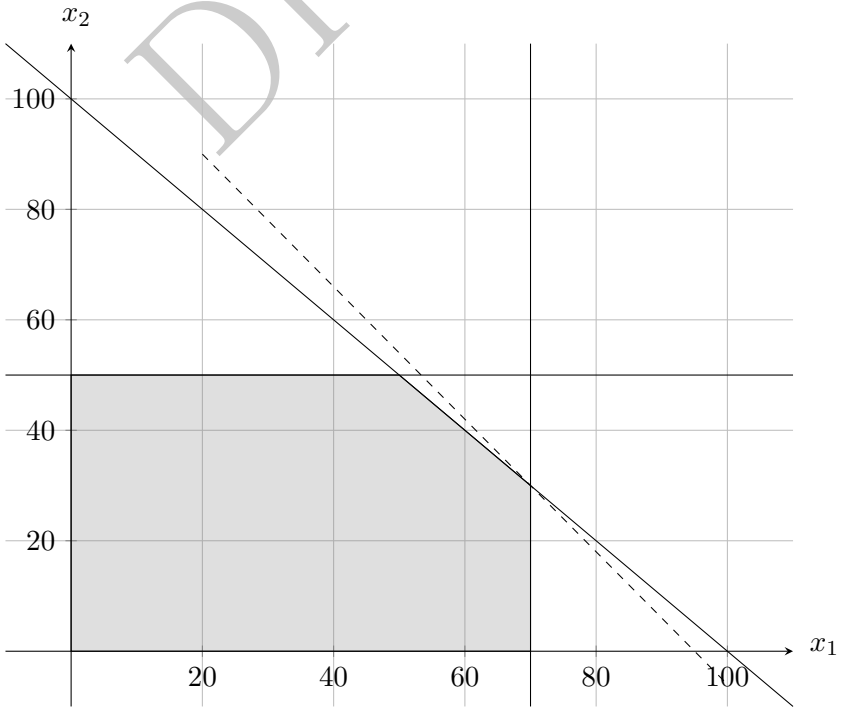
Solution. In the problem formulation to maximize total value, let

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

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 less 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 between 4 and 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 software.

(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 software and determine the added cost due to the new constraint.

Solution. For part 4a, a GAMS model is provided in the file `dog-fooda.gms`. 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, A GAMS model is provided in `dog-foodb.gms`. 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.

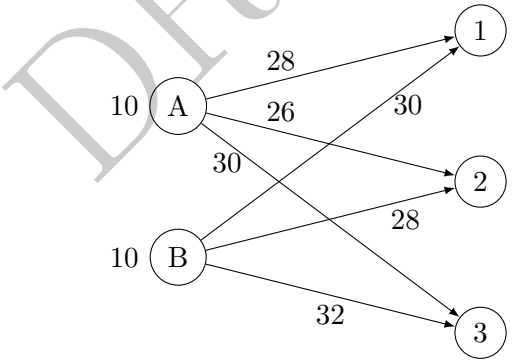
5. *Golf lessons.* Amy and Brian are professional golfers who have each agreed to donate 10 hours of private golf lessons to a charity auction. Three people have bid on the lessons, and their bids are shown in the table below. For example, Emma has bid $28 per hour to receive lessons from Amy.

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

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

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.

6. *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

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

7. *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.

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

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 $i = 1, 2, 3$.

Chapter 2

Probabilistic Modeling

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?

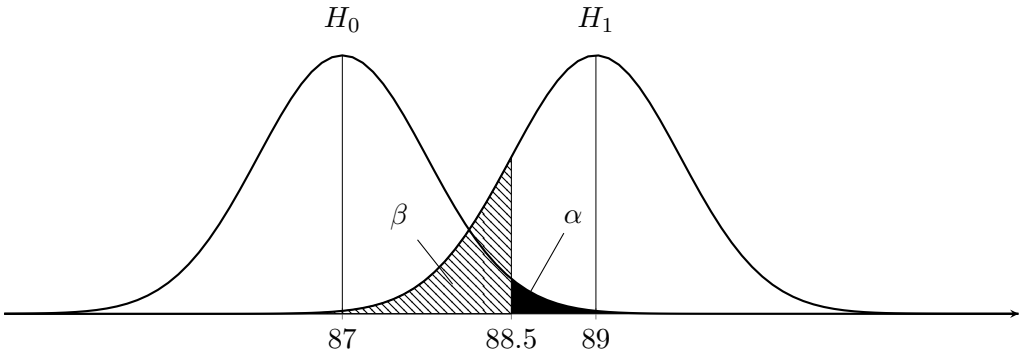
Suppose that a gas station owner has his own octane test kit and rule for accepting a tanker-truck of premium gasoline. The owner knows from past shipments that the distributions of octane ratings are

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

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,

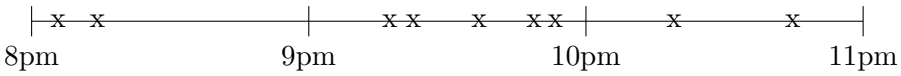
$$\begin{aligned}
\alpha &= P(X > 88.5 \mid H_0) \\
&= P\left(Z > \frac{88.5 - 87}{1}\right) \\
&= 1 - P(Z < 1.5) \\
&= 0.067
\end{aligned}$$

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 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 power utility can supply electricity to a city from n different power plants. Each power plant fails with probability p, independent of the others.

   (a) Suppose that any one plant can produce enough electricity to supply the entire city. What is the probability that the city will experience a black-out?

   (b) Suppose that two power plants are necessary to keep the city from a black-out. Find the probability that the city will experience a black-out.

Solution. For part 3a, all n plants must fail for the city to have a black-out. Since failures are independent, the probability of a black-out is p^n.

For part 3b, let X be the number of failed plants. X has a Binomial distribution with parameters n and p. The probability of a black-out 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.* There are two clerks at the local post office. You enter the post office to find that both clerks are busy (i.e. each clerk is serving a customer), but that no one is in line. So, you are first in line and you are to be served by the first available clerk. Customers depart the post office as soon as they are finished being served by a clerk. If the service time distribution for each clerk is Exponential with rate λ, what is the probability that you are the last of the three customers to depart the post office?

   Solution. By the memoryless property of the Exponential distribution, the remaining time for each customer currently being served is identical. In particular, if we let Y be the remaining time of a customer, then the distribution of remaining time is

   $$P(Y \leq y) = 1 - e^{-\lambda y}$$

   Since the two customers have the same distribution for remaining time, the probability that customer 1 departs before customer 2 is 1/2 (likewise for customer 2 departing before customer 1). When you enter service, the memoryless property still applies. Regardless of how long the other customer has been in service, you have the same distribution for remaining time. So, the probability that you are the last to depart 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

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

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 are remaining.

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 vending machine attendant 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 it is expected that the beverages will be purchased according to a Poisson distribution with a mean of 1 hour between purchases.

   (a) What is the probability that there are no beverages remaining in the vending machine when the attendant arrives?

   (b) On average, how many beverages will be remaining in the vending machine when the attendant arrives?

   (c) What is the probability that the attendant 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 distribution with a mean of one hour between purchases, the rate λ that beverages are purchased is 24 beverages per day. Therefore, 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 purchases from the vending machine each week is 168 beverages. Therefore, the expected number of beverages remaining at the end of the week is $180 - 168 = 12$.

For part 8c, 150 or more beverages will be replenished if 150 or more beverages are purchased before the end of the week. 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 car dealership keeps 200 used cars that are for sale in an uncovered parking lot. In the city where the dealership is located, hailstorms occur randomly throughout the month of June according to a Poisson distribution with a mean of 4 hailstorms. During a hailstorm, the number of cars that are damaged follows a Poisson distribution with a mean of eight damaged cars per 10 cars in the parking lot.

   (a) Determine the probability that there will be exactly 5 hailstorms at the dealership in June.

   (b) During a hailstorm, determine the probability that 170 or more cars will be damaged.

Solution. For part 9a, let X be a random variable that represents the number of hailstorms in June. The number of hailstorms in June follows a Poisson distribution with a mean of 4 hailstorms. Therefore, $X \sim \text{Poisson}(\lambda = 4)$. The probability that there are exactly 5 hailstorms is

$$P(X = 5) = \frac{e^{-\lambda}\lambda^5}{5!} = \frac{e^{-4}4^5}{120} \approx 0.16$$

For part 9b, the rate that cars are damaged is Poisson with a mean of 8 damaged cars per 10 cars in the parking lot. Let N be the number of damaged cars in the parking lot when 200 cars are present.

$$N \sim \text{Poisson}(\lambda = 160)$$

The probability that 170 or more cars will be damaged in a hailstorm is

$$\begin{aligned} P(N \geq 170) &= 1 - P(N \leq 169) \\ &= 1 - \sum_{n=0}^{169} \frac{\lambda^n e^{-\lambda}}{n!} \\ &\approx 0.22 \end{aligned}$$

Using R,

```
> 1 - sum(dpois(0:169,160))
[1] 0.2245714
```

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 independently and at a mean rate of seven scoops per hour. The time between scoops from the shovel can be considered to follow an Exponential distribution.

    (a) Find the probability that the time between consecutive trips to the railroad will be 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 ready 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 between consecutive trips to the railroad to take at least one hour, we require that the number of arrivals in one hour is nine or less. Let X be the number of (scoop) arrivals in a one hour period.

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

    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 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 $\mu = 135$ 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
$\mu < SPB < \mu + 1.5\sigma$	high
$\mu - \sigma < SBP < \mu$	average
$SBP < \mu - \sigma$	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 sysolic blood pressure, and recall that $\Phi(z)$ indicates the CDF of the standard Normal distribution.

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

12. *Stock prices.* The prices of stocks are often modeled with a lognormal distribution. An investor is considering purchasing stock in one of two companies, A or B. The price of a share of stock today is \$1 for both companies. For company A, the value of the stock one year from now is modeled as lognormal with parameters $\mu = .05$ and $\sigma = .1$. For company B, the value of the stock one year from now is modeled as lognormal with parameters $\mu = .02$ and $\sigma = .2$.

    (a) Find the mean of the price of one share of company A one year from now.
    (b) Find the probability that the price of one share of company A one year from now will be greater than \$1.20.
    (c) Find the mean of the price of one share of company B one year from now.
    (d) Find the probability that the price of one share of company B one year from now will be greater than \$1.20.

Solution. Let X_A and X_B be the price of one share of company A and company B one year from now, respectively. We know from the problem description that $X_A \sim \text{LogN}(\mu = .05, \sigma = .1)$ and $X_B \sim \text{LogN}(\mu = .02, \sigma = .2)$.

For part 12a, the mean of X_A is

$$e^{\mu + \sigma^2/2} = \$1.06$$

For part 12b,

$$\begin{aligned}
P(X_A > \$1.20) &= P(\ln(X_A) > \ln(1.20)) \\
&= P\left(Z > \frac{\ln(1.20) - .05}{.1}\right) \\
&= P(Z > 1.323) \\
&= 1 - P(Z < 1.323) \\
&= .093
\end{aligned}$$

For part 12c, the mean of X_B is

$$e^{\mu+\sigma^2/2} = \$1.04$$

For part 12d,

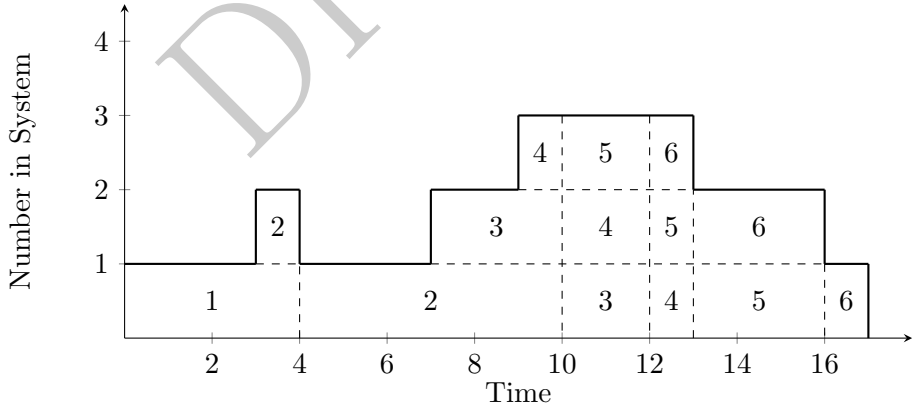
$$\begin{aligned} P(X_B > \$1.20) &= P(\ln(X_B) > \ln(1.20)) \\ &= P\left(Z > \frac{\ln(1.20) - .02}{.2}\right) \\ &= P(Z > .812) \\ &= 1 - P(Z < .812) \\ &= .208 \end{aligned}$$

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.* Cars arrive at the Lincoln Tunnel toll gate according to a Poisson process with an average rate of 90 cars per hour. The time for passing the gate is exponential with mean 38 seconds. Drivers complain of the long waiting time, and authorities are willing to reduce the average passing time to 30 seconds by installing automatic toll collecting devices, provided two conditions are satisfied: 1) the average number of waiting cars in the present system exceeds 5 and 2) the percentage of the gate idle time with the new device installed does not exceed 10%. Can the new device be justified?

    Solution. Cars arrive at the rate

    $$\lambda = 90 \text{ cars per hour} \times 1/3600 = 1 \text{ car}/40 \text{ seconds}$$

    We know that $\mu = 1/38$ for the current system and $\mu_{\text{new}} = 1/30$ for the new system. Condition 1 states that the average number of waiting cars in the current system must exceed 5.

    $$L_s = \frac{\lambda}{\mu - \lambda} = \frac{1/40}{1/38 - 1/40} = 19$$

    You may have also computed

    $$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{(1/40)^2}{(1/38)(1/38 - 1/40)} = 18.05$$

    Either way, the first condition is satisfied. The second condition states that the percentage of idle time in the new system cannot exceed 10%.

    $$p_0 = 1 - \frac{\lambda}{\mu_{\text{new}}} = 1 - \frac{1/40}{1/30} = .25$$

    The second condition is not satisfied. There is no justification for the new toll gate device.

15. *Comparing system configurations.* Pete's Market is a small local grocery store with one checkout counter. Shoppers arrive at the checkout lane according to a Poisson process, with an arrival rate of 15 customers per hour. The checkout service times are distributed Exponential with a service rate of 20 customers per hour. The manager is considering two options for improving service.

    (a) Hire a second person to bag groceries while the cashier is scanning and collecting money from the customer. With this improved single-server operation, the service rate would be increased to 30 customers per hour.
    (b) Hire a second person to operate a second checkout counter. The two-server operation would have a service rate of 20 customers per hour for each server.

Determine the improvements that would result, consider the relative cost of each option, and make a recommendation to the manager.

Solution. Let's look at the average number in the queue L_Q and the average time in queue W_Q as metrics to judge improvement in system performance. Because the interarrival times and the service times are exponentially distributed, we can use the results from section 11.2. Before any improvements are made,

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{15^2}{20(20 - 15)} = 2.25 \text{ cust}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{2.25}{15} = .15 \text{ hour} = 9 \text{ min}$$

If they add a second person to the existing single checkout

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{15^2}{30(30 - 15)} = 0.5 \text{ cust}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{0.5}{15} = 0.0333 \text{ hour} = 2 \text{ min}$$

and if they add a completely separate second checkout then we need to use the formulas from section 11.3 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 $P_0 = 0.4545$, $L_q = 0.123$, and $W_q = 0.0082$ hours or 0.5 minutes. Considering the relative improvement and the cost of adding a second checkout, I would recommend option a). That is, to add a second person to the existing checkout. The average number in queue and the average time in queue appear to be acceptable for a grocery store.

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 Exercises

1. *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.

2. *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

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

3. *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.

4. *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.

5. *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 75% 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.

6. *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 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.

7. *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 relatively constant, so a gain in sales for Doodle's is a loss of sales for Lola's. Each company has their standard packaging for peanut butter and a new innovative packaging for peanut butter. Both companies may produce and sell both types of packaging. If both companies choose to market only their innovative 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 innovative product and Lola's markets their standard product, Doodle's will gain 10% of the market. If Lola's markets their innovative 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 advertising throughout the year, so a mixed strategy is possible.

   (b) Compute the value of the game.

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

Solution. The game is

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

Note that there is no saddle point, so the best strategy is mixed. Doodle's should mix the strategies standard and innovative in the ratio 4 to 5, while Lola's should

mix their strategies of standard to innovative 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 the Doodle's markets the standard product, they receive a payoff of -2 with probability 1/3 and payoff of 8 with probability 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 7c), 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.

8. *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).

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)`?

4.2 Descriptive Graphics

4.3 Exercises

Descriptive Statistics

1. *Summarizing a data set with statistics.* Each time a particular supermarket receives a shipment of peaches, the manager chooses one box at random and counts the number of spoiled peaches in the box. (Each box contains 48 peaches.) The file `fruit-spoilage.txt` contains data on the number of spoiled peaches for 55 such boxes.

   (a) From the data set of 55 observations, determine the following summary statistics: minimum, maximum, mean, median, mode, first quartile, third quartile, and standard deviation. To be clear, from this data set you are computing sample statistics; the true underlying distribution of the number of spoiled peaches is unknown.

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

2. *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`

3. *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.

4. *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.

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

6. *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?

7. *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. Seven independent replications of the simulation model were run, each for 30 days of simulated time, with the following results for average number of (simulated) jobs in the shop

$$18.9 \quad 22.0 \quad 19.4 \quad 22.1 \quad 19.8 \quad 21.9 \quad 20.2.$$

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 (simulated) 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.

Solution. Using the simulation output, we can construct a 95% confidence interval for the mean number of jobs in the shop. If the the CI contains 22.5, then we will conclude that the simulation output is consistent with the real system behavior (with a confidence level of 95%). Note that we have $n = 7$ observations, so we should use the t distribution. The confidence interval is

$$\bar{X} \pm t_{1-\alpha/a,n-1} \times \frac{S}{\sqrt{n}}$$

Using the simulation output and the t distribution we have

$$20.614 \pm 2.447 \times \frac{1.356}{\sqrt{7}}$$

The resulting CI is (19.36, 21.87), and we conclude that the simulation output is not consistent with system behavior.

8. *Confidence interval for a proportion.*

9. *Comparison of system configurations.*

10. *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

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

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.

11. *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 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

12. *College students and driving speed.* The file `speed_gender_height.csv` contains 1,325 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 Modeling

5.1 Regression

Statistical Models

Statistical models are simplified descriptions of data that involve mathematical relationships. In addition to the data itself, a statistical 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 speed, gender, and height 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? The model formula is

```
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 "∼" is the response or dependent variable and the terms to the right of the "∼" that are separated by a "+" are the predictors or independent variables. Mathematically, the formula above specifies the model

$$speed_i = \beta_0 + \beta_1 height_i + \beta_2 gender_i + \epsilon_i$$

where the coefficients $\beta_0,\ \beta_1,\ \beta_2$ 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. 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 formula above is equivalent to

```
speed ~ 1 + height + gender
```

If we do not want an intercept term in our model, say 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
```

Statistical modeling is a process that almost always involves iterating over the following steps [1].

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

Transformations on Data

Suppose that a variable u is related to a 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.

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,

$$\hat{\gamma_0} = e^{\hat{\beta_0}}, \qquad \hat{\gamma_1} = \hat{\beta_1}$$


```
## example of linear regression on the msleep data to
## notice that collinearity is present,
## choose variable transformation,
## understand the reported p-values,
## and understand the reported R-squared.
attach(msleep)

fm <- lm(sleep_total ~ brainwt + bodywt)
summary(fm)
## collinearity is present
plot(brainwt, bodywt)
plot(log(brainwt), log(bodywt))
fm <- lm(sleep_total ~ log(brainwt))
summary(fm)

## a one-unit increase in log(brainwt) decreases
## sleep time by about one hour (on avg).

## for interpretation, plug in a few values to get an idea
(x <- seq(from=0.5, to=5, by=0.5))
```

```
y.hat <- predict(fm, newdata=data.frame(brainwt=x))
plot(x, y.hat)

## the hypothesis test for b1

(t0 <- -1.0397/0.1914) # the t-statistic

pt(-5.432, 54) * 2  # the p-value
## should match the summary output

## compute R-squared
y.hat <- predict(fm)
y <- sleep_total[ !is.na(sleep_total) & !is.na(brainwt) ]

TSS <- sum( (y - mean(y))^2 )
RSS <- sum( (y - y.hat)^2 )
1 - RSS/TSS   # R-squared

(F <- ((TSS-RSS)/1) / (RSS/(56-1-1))) # the F-statistic
1 - pf(29.51, 1, 54)   # p-value for the F-test
```

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. *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.

Classification

Time Series

6. *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

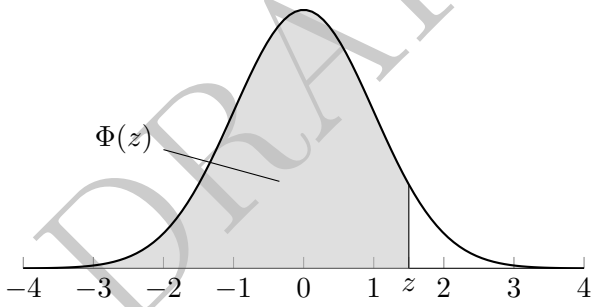
Standard Normal Distribution

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] John Chambers and Trevor Hastie. *Statistical Models in S*. Wadsworth and Brooks/-Cole, 1992.

[2] 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.

[3] GAMS. *Quick Start Tutorial*. (accessed June 5, 2020). URL: `https://www.gams.com/latest/docs/UG_TutorialQuickstart.html`.

[4] *GLPK (GNU Linear Programming Kit)*. `www.gnu.org/software/glpk`.

[5] George L. Nemhauser and Laurence A. Wolsey. *Integer and Combinatorial Optimization*. John Wiley and Sons, 1988.

[6] Richard E. Rosenthal. *A GAMS Tutorial*. (accessed June 5, 2020). URL: `https://www.gams.com/latest/docs/UG_Tutorial.html`.

[7] Paul Williams. *Model Building in Mathematical Programming*. Fourth. John Wiley and Sons, 1999.