Policy Modeling: Notes and Examples

Edward H. Kaplan, Yale School of Management

January 2026

Contents

Preface

The notes and examples to follow are material I have used in nearly 40 years of teaching my Policy Modeling course at the Yale School of Management. Most, but not all, of the examples were my own concoctions. This was mainly an MBA course, but was also required of PhD students in our Operations program to provide motivation for problem formulation in societal applications of operations research. Graduate students from many other Yale programs including public health, international and economic development, engineering, statistics, political science, and even physics and computer science have taken this course over the years. Much of the action in this class took place on the blackboard, and there is a lot of material that did not make it into these notes. Consequently there are several gaps in explanation plus an occasional reference to something that happened in the classroom that is absent from the notes. Whatever, you should get the flavor of what the class was like from the notes and examples to follow. I hope that you find this material of interest. Ed Kaplan, New Haven, January 2026

Chapter 1

Resource Allocation Models

Suppose that in the evaluation of some program or policy, you are faced with choosing the value of a critical policy variable we will denote as π (for *policy*). In cost benefit analysis, we assume that the costs and benefits can be measured on the same scale, typically monetary. So, as functions of your choice of the value for the policy variable π, let $b(\pi)$ be the total benefits and $c(\pi)$ be the total costs. You want to maximize the *net benefits*, that is, the difference between $b(\pi)$ and $c(\pi)$. So the decision problem you face is

$$\max_{\pi} \left\{ b(\pi) - c(\pi) \right\}.$$

Often this problem can be solved via marginal analysis: assuming diminishing marginal benefits (that is, $b(\pi)$ is concave) and increasing marginal costs (that is, $c(\pi)$ is convex), differentiation yields π^*, the optimal value of the policy variable, as the solution to the equation

$$b'(\pi^*) = c'(\pi^*)$$

where the primes denote differentiation. In words, you want to set the marginal benefit equal to the marginal cost. Note that this problem is identical to the profit maximization problem of the firm: to make as much money as possible, set marginal revenue equal to marginal cost.

In general, you might have several policy variables $\boldsymbol{\pi} = (\pi_1, \pi_2, ...\pi_n)$. You might also be restricted to a set of specific options, which we can denote in general by some set $\mathbf{\Pi}$. The problem then becomes: choose the values of the decision variables $\boldsymbol{\pi}$ so as to

$$\max_{\boldsymbol{\pi} \in \mathbf{\Pi}} \left\{ b(\boldsymbol{\pi}) - c(\boldsymbol{\pi}) \right\}.$$

If you think about it, this is just a decision analysis problem: choose the best option to maximize net benefits.

In cost effectiveness analysis, the complicating feature is a budget constraint: you can't spend more than B. Of course, the uncomplicating feature is that benefits and costs no longer need to be in the same units. So, for example, in a cost benefit analysis of an HIV prevention program, you need to assign a dollar value to prevented infections to contrast against costs incurred. In a cost effectiveness analysis, you can simply focus on how many infections you prevent given your budget.

Using the same notation as above, the cost effectiveness problem becomes:

$$\max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} b(\boldsymbol{\pi})$$

subject to

$$c(\boldsymbol{\pi}) \leq B.$$

In some special cases, simple results fall out. For example, if the benefit and cost functions are linear in the policy variables, and the policy variables are all non-negative but bounded, then the problem is

$$\max \sum_{i=1}^{n} b_i \pi_i$$

subject to

$$\sum_{i=1}^{n} c_i \pi_i \leq B$$

$$0 \leq \pi_i \leq u_i \text{ for } i = 1, 2, ..., n$$

where the u_i's are upper bounds. This has a simple solution: rank order the policy variables (which in this case might be effort levels in different programs) by the ratio of b_i/c_i (the bang for the buck), and assign all of the budget by starting with the highest ranking alternative, giving as much dough as possible ($\min(u_{[1]}, B)$) to the first ranked alternative (denoted by [1]), then giving as much as possible to the second ranked alternative, and continuing until the budget is exhausted.

1.1 Nation of Shoplifters

The Nation of Shoplifters, as its name suggests, is beset with a minor crime problem. On average across the population, citizens are committing ten crimes over the course of their shoplifting careers, even after receiving disciplinary action. Now,there are two programs – I and II – that are under consideration to reduce crime in the Nation of Shoplifters. Program I can reduce the expected number of crimes per career from 10 to 5, which means that Program I prevents 5 crimes per offender. Program II reduces the expected number of crimes per career from 10 to 6.67, which means that Program II prevents 3.33 crimes per criminal.

1.2 Bang for the Buck: Cost-effectiveness ratios

For Program I, on a per criminal basis, spending $6,000 prevents 5 crimes, for a cost-effectiveness ratio of $\$6,000/5 = \$1,200$.

For Program II, spending $3,000 prevents 3.33 crimes, for a cost-effectiveness ratio of $\$3,000/3.33 = \$900$.

Many analyses would simply stop here, and conclude that Program II is more *cost-effective* than Program I.

1.3 "Profit Maximization": Cost-benefit analysis

Net benefits consider the difference between the benefits of crime prevention and the costs of doing so. If crimes cost society $2,000 each on average, then there is a $2,000 benefit from preventing each crime.

For Program I, which prevents 5 crimes per criminal but costs $6,000 per criminal to do so, the net benefits are given by $\$2,000 \times 5 - \$6,000 = \$4,000$.

For Program II, which prevents 3.33 crimes per criminal at a cost of $3,000, the net benefits equal $\$2,000 \times 3.33 - \$3,000 = \$3,666.66$.

Many analyses would simply stop here, and conclude that Program I is more *cost-beneficial* than Program II.

1.4 A Budget Constraint

What if you can only spend \$2,000 per criminal? Since Program I costs \$6,000 per criminal, you could only reach \$2,000/\$6,000 = 1/3 of all those you would like to reach. Since you would gain no benefits (and suffer no costs) for those you do not reach, the budget constraint of \$2,000 per criminal reduces your net benefit for Program I from \$ 4,000 to $(1/3) \times \$4,000 + (2/3) \times 0 = \$1,333.33$.

For Program II, which costs only \$3,000 per criminal, you can reach \$2,000/\$3,000 = 2/3 of those you wish to reach. So the budget constraint changes the net benefit from \$3,666.66 to $(2/3) \times \$3,666.66 + (1/3) \times 0 = \$2,444.44$.

With this budget constraint, it is clear that you would pump everything you had into Program II.

1.5 A Linear Program

Let's pull this all together. Suppose we have 1,000 criminals in the population. Let X_1 and X_2 be the number of these we send to Programs I and II respectively. Suppose that we also have a budget constraint of $\$B$ in total to spend. The objective is to prevent as many crimes as possible subject to the budget constraint. Mathematically, this can be stated as:

$$\max 5X_1 + 3.33X_2 \text{ (prevent as many crimes as possible)}$$

subject to

$$X_1 + X_2 \leq 1,000 \text{ (1,000 criminals to allocate)}$$

$$\$6,000X_1 + \$3,000X_2 \leq B \text{ (can't spend more than the budget)}$$

$$X_1 \geq 0,\ X_2 \geq 0 \text{ (\# of criminals to either program must be non-negative)}$$

You can solve this graphically, or on the computer using Excel, or mathematically. The solution is:

1.5.1 For $0 \leq B < \$3$ million:

$X_1^* = 0$; $X_2^* = B/\$3,000$; avert $3.33 \times B/\$3,000$ crimes at a cost of \$900 per crime averted.

1.5.2 For $3 million $\leq B <$ $6 million:

$X_1^* = B/3,000 - 1000$; $X_2^* = 2,000 - B/3,000$; avert $1,666.66 + 0.00055B$ crimes at a *marginal* cost of $1,800 per crime averted (i.e. crimes averted goes from 3,333.33 to 5,000 as the budget B goes from $3 million to $6 million, thus the marginal cost per crime averted equals \$3 million$/(5,000 - 3,333.33) =$ $\$1,800$).

1.5.3 For $B \geq$ $6 million:

$X_1^* = 1,000$; $X_2^* = 0$; avert 5,000 crimes, dumb to spend more than $6 million.

1.6 A Tabular View:

Program	Averted Crimes/Criminal	Cost/Criminal
I	5	$6,000
II	3.33	$3,000

Arrange the programs in order of increasing benefits (averted crimes) per criminal:

Program	Benefit	ΔBenefit	Cost	ΔCost
Do Nothing	0	-	0	-
II	3.33	3.33	$3,000	$3,000
I	5	1.67	$6,000	$3,000

Now look at the incremental cost-effectiveness ratios ΔCost/ΔBenefit:

Program	ΔCost/ΔBenefit
Do Nothing	-
II	$\$3,000/3.33 = \$900$
I	$\$3,000/1.67 = \$1,800$

Note that these ratios are exactly the marginal costs per crime averted as computed from the linear program. Note that for Program I, this ratio differs from the $1,200 that was reported under “Bang for the Buck” earlier. That is because Program I there was compared directly to the do nothing option, whereas above we are learning the *marginal* cost-effectiveness of Program I relative to Program II.

1.7 What Happened to the Benefit Valuation?

When we face a binding budget constraint, it doesn’t matter what the benefit valuation is! That is because when the budget constraint is binding, you always spend everything you have. To see this, note that

$$\text{Net Benefits} = \$2000[5X_1 + 3.33X_2] - [\$6,000X_1 + \$3,000X_2]$$

but when the budget constraint is binding, we have $\$6,000X_1 + \$3,000X_2 = B$, so for a fixed budget constraint, maximizing Net Benefits is equivalent to

$$\max \$2000[5X_1 + 3.33X_2] - B$$

and, since the budget B is fixed, maximizing the above is just equivalent to maximizing $5X_1 + 3.33X_2$, the total number of crimes averted, which is what we have been doing!

1.8 Suppose There Is No Budget Constraint, and Benefit Valuation is Unknown!

In other words, suppose we have lots of money, but don’t know the benefit of preventing a crime. Well, let’s call b the benefit of preventing a crime. Then you will prefer Program I to Program II if it has larger net benefits per criminal, that is, if

$$5b - \$6,000 > 3.33b - \$3,000.$$

Guess what? When you solve this inequality, you find out that you prefer Program I to Program II, in a world without a budget constraint, if $b > \$1,800$.

Is this amazing or what – after all, \$1,800 is exactly the cost-effectiveness ratio for Program I relative to Program II! So if money was no obstacle, you would prefer I to II providing the value of preventing a crime is worth more than \$1,800.

1.9 Examples

1.9.1 No Free Riders!

The Town of No Haven is contemplating raising its bus fare. Currently, riders pay \$1.20 per trip. The town figures that on average they incur a \$1.00 per passenger expense. Since there are currently 100,000 riders per week, the town is making \$20,000 weekly as is.

Now, a consultant to the town has somehow prepared a demand curve that relates weekly ridership to the fare charged per trip. The resulting demand formula is given by:

$$q(r) \;= 100000e^{-2(r-1.2)}$$

where $q(r)$ is the weekly ridership that can be expected if a fare of \$$r$ is charged. Note that if $r = 1.2$, the weekly ridership equals 100,000.

(a) The consultant argues that the fares can be raised. If c, the cost per trip to the town, is \$1, while the income to the town per trip equals r, the consultant argues that the net benefit to the town as a function of the fare, $NB(r)$, is given by

$$\begin{aligned} NB(r) &= r \times q(r) - c \times q(r) \\ &= (r-1) \times 100000e^{-2(r-1.2)}. \end{aligned}$$

Following this logic, what should the town charge riders to maximize their net benefit? What would the resulting ridership, revenues, and costs equal? (Note: you can solve this numerically on a spreadsheet by experimenting with different fares, or for those of you who know some calculus, you can solve this analytically; either approach is fine).

Attached is a graph showing the net benefits (i.e. profit) that result as a function of the fare per ride r. I did this using Excel; you could have done this with any spreadsheet package, or you could have written your own little

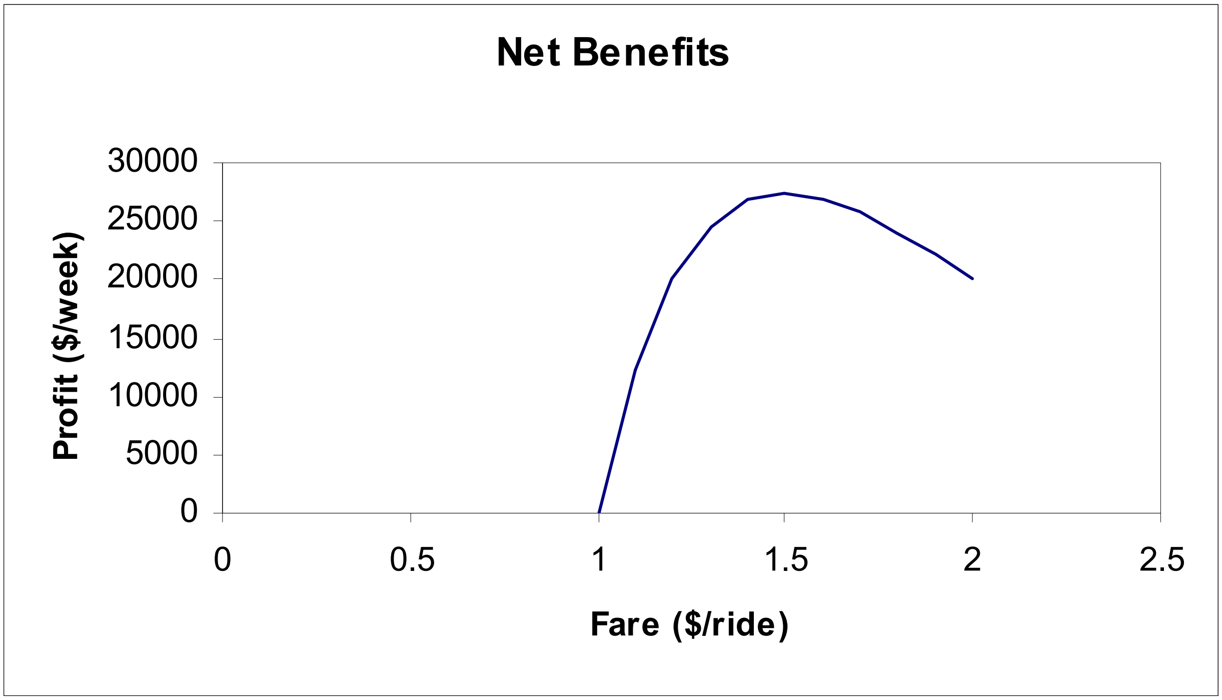


Figure 1.1:

computer program. Either way, the curve reaches a maximum when the fare r is set equal to \$1.50, so this formulation supports the consultant's view: raise the fare from \$1.20 to \$1.50.

A mathematical solution can be found applying calculus. We need to differentiate the net benefits with respect to r and equate the result to zero. Doing so we find that

$$\frac{dNB(r)}{dr} = [1 - 2(r - 1)]q(r)$$

so setting this equal to zero yields

$$r^* = 1 + \frac{1}{2} = \$1.50$$

as found numerically above.

At a fare of $1.50, the ridership would fall from 100,000 per week to 54,881 per week. This would bring in revenues of $\$1.50 \times 54,881 = \$82,321$, at a cost of $54,881, for a net benefit (profit) of $27,440 per week, which is $7,440 per week better than the existing situation.

(b) The town's new city planner says "Hold Everything!" Something is wrong. The average costs per rider were computed *given a ridership of 100,000!* The planner points out that the real costs of running the system are fixed with respect to the number of riders: the town's expenses are the costs of paying drivers, buying and maintaining buses, fuel, etc. Indeed, it turns out that the value of $c = \$1$ was computed by adding up the fixed costs and dividing; it really costs the town $100,000 per week to run the buses *without regard* to the number of passengers. If the city planner is right, how should the town set the fare? What consequences would follow?

If the city planner is right, then the net benefit function is given by

$$NB(r) = rq(r) - \$100,000.$$

This function has been plotted on the attached graph (Excel again). The function is maximized at a fare of only 50 cents (calculus would yield the same answer). The ridership would roughly quadruple (it would increase to 405,520 riders weekly), yielding revenues of $202,760, and a profit of $102,760, *much* better than the current situation.

(c) A recent SOM grad points out that, actually, the total costs depend on the number of buses deployed. Presumably this depends on the number of riders, as buses are capacitated: you can't put more than 50 people on a bus! From this point of view, which of the solutions above makes more sense?

According to our student, if the capacity of a bus is K, then the number of busloads per week is given by

$$\text{Busloads} = \left\lceil \frac{q(r)}{K} \right\rceil$$

where the notation $\lceil x \rceil$ refers to the smallest integer greater than or equal to x (and for obvious reasons is referred to as the *ceiling function*). If the cost per busload is given by k, then the total cost to the town as a function of the fare would equal

$$\text{Total Cost}(r) = k \times \left\lceil \frac{q(r)}{K} \right\rceil \approx \frac{k}{K} \times q(r)$$

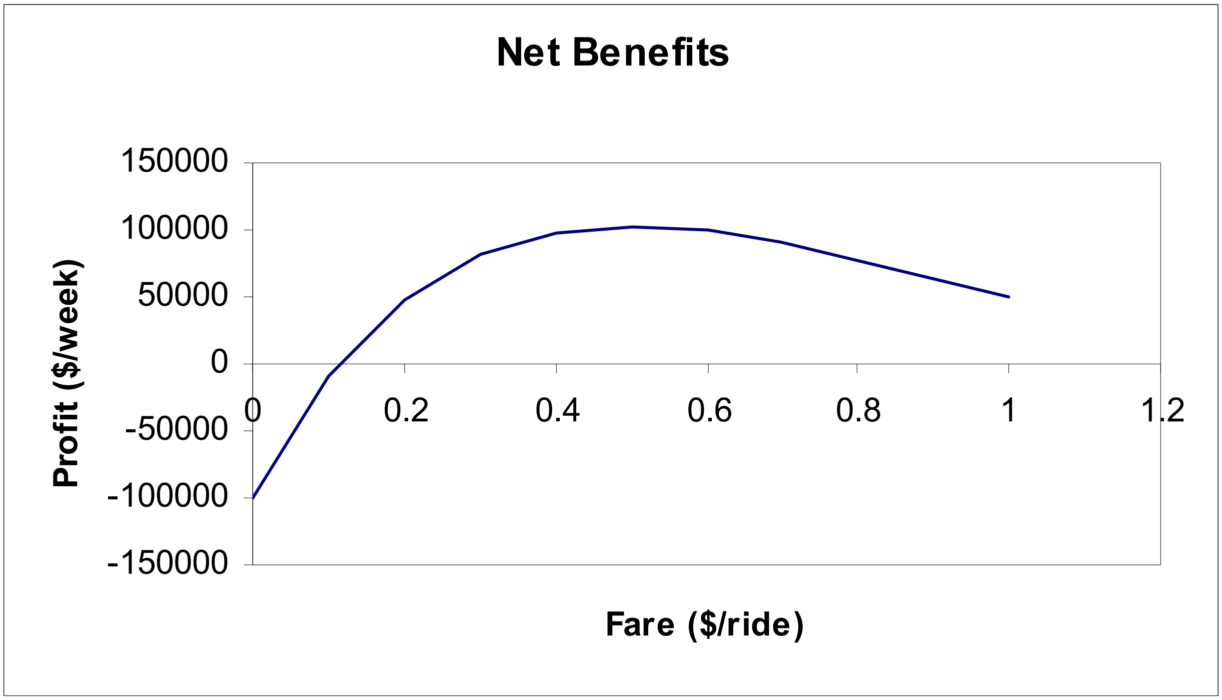

Net Benefits
150000
100000
50000
0
-50000
-100000
-150000
Profit ($/week)
0
0.2
0.4
0.6
0.8
1
1.2
Fare ($/ride)

Figure 1.2:

so the total cost is approximately proportional to the number of riders. This would support the formulation in (a) rather than the formulation in (b). Indeed, is it sensible to believe that the ridership could quadruple (from 100,000 to 405,520) without the costs changing (which is what you have to believe if you accept the analysis in (b))? So the moral of the story is: choose a consultant over a city planner!

1.9.2 Running For President

In the United States, presidents are elected in accord with a two-tiered process. First, there is a popular election held simultaneously in all states to see which presidential candidate gets the most votes. Then the election proceeds to the Electoral College where each state is represented by *electors* whose number in each state equals the total congressional representation of that state across the House of Representatives and the Senate. These electors cast votes for the president, and with few exceptions in certain states (e.g. Maine, Nebraska), whichever candidate won the *majority* of the popular vote in a given statewide election is awarded *all* of the electoral college votes in that state. To be elected president of the United States requires receiving at least 270 of the 538 votes cast in the electoral college.

Given these (approximate) rules for electing the president, consider the following question: conditional on the number of people voting in each state, what is the smallest total number of popular votes received with which a candidate could still be elected president by winning the electoral college vote? Since a candidate can win a state with (just over) half of the popular votes, an immediate back-of-the-envelope guesstimate follows from assuming that all states have the same number of voters, for in this case a candidate who wins (just over) half of the votes in half of the states could win the electoral college. Our simplest model thus suggests that a candidate could be president with only 50% × 50% = 25% of the total popular vote (as *no* votes need be won in states the candidate loses!). This result is not a forecast or prediction of any actual presidential election; rather it expresses a property of the electoral college system: under this approach to choosing the president, it is theoretically possible to be elected with as little as 25% of the popular vote. An immediate corollary is that a candidate could win up to 75% of the popular vote and still lose the election!

Of course, in reality all states do not have the same number of popular voters, and consequently do not share the same number of electoral college delegates. This follows from the simple fact that the population sizes of (and hence the number of voters in) the different states are not equal. In the 2020 election for example, the number of voters ranged from 276,765 in Wyoming to 17,511,515 in California, What happens if we allow both the number of voters and electoral college delegates to vary by state? To find out, define v_i to equal the number of voters in the popular election held in state i, and let e_i be the number of electoral college votes at stake in state i. Further, let x_i take on the value 1 if the candidate in question wins state i and zero otherwise. Again presuming that winning in a state requires half of the votes case while losing requires zero votes, the smallest number of popular votes that in principle could elect a president is given by the solution to

$$\min \sum_i \frac{v_i}{2} x_i$$

subject to

$$\begin{aligned} \sum_i e_i x_i &\geq 270 \\ x_i &= 0 \text{ or } 1 \text{ for all states } i \end{aligned}$$

This model succinctly seeks the smallest number of possible votes for a candidate possible while ensuring that the total number of electoral college votes for this same candidate is at least 270 (with 538 electoral college votes, a literal tie resulting from receipt of 269 votes is possible, and in such an event a bizarre set of tie-breaking rules would be called into effect, but those are not important for our present discussion).

Before trying to solve this exactly, think of this in cost-effectiveness terms: what is the “price” per electoral college vote in units of popular votes in each state? The answer is given by the ratio of v_i/e_i, the number of voters per electoral college vote in each state. It is “cheaper” to “buy” electoral college votes in states where this ratio is smaller compared to states where this ratio is larger. This suggests a simple ranking solution to the problem posed: rank all of the states in order of smallest to largest v_i/e_i ratio, start at the top of the list, and assign states to the candidate while recording the total electoral votes received, and keep going until the number of electoral college votes received equals or passes 270. In the 2020 election, the 270 electoral vote threshold is crossed when the state of Maryland is awarded to our candidate (resulting in 272 electoral votes). Under our rule of half the popular votes in winning states and zero popular votes in losing states, this assignment of states would deliver a presidential victory with 34,323,233 popular votes out of 158,537,765 in total, or 21.6%. Going through all this work of ranking the states from smallest to largest v_i/e_i has reduced the popular vote required to win the presidency from 25% to 21.6%. This seems like a lot of work for a small gain in precision!

But more precision is indeed possible. Note that the solution just suggested requires 272 electoral college votes. Is it possible to find an allocation that requires only 270 electoral college votes and uses a smaller number of popular votes in total? The answer is yes, and it comes from exactly solving the integer program described earlier. It turns out that the v_i/e_i ranking rule, while certainly efficient, is not guaranteed to be optimal due to the "all or nothing" allocation of winning states to the presidential candidate. With the help of the Solver in Excel, the solution to the integer program does not strictly follow the ranking rule – Oregon and Missouri are included as winning states while Missouri and Maryland, ranking rule winners, are out. Nonetheless this solution achieves exactly 270 electoral college votes, and does so with only 34,177,592 popular votes, or 21.5% of the total popular votes cast. Compared to the ranking rule, this last result required a *lot* more work just to reduce the fraction of the vote required by one tenth of a percentage point!

Summing up – a simple back-of-the-envelope model suggesting that a candidate could win the electoral college with half the votes in half of the states gave an immediate approximation that the presidency could be won with only 25% of the popular vote. An intuitively pleasing "price per electoral vote" ranking using actual voting data from the 2020 election required more work, but reduced the required popular vote percentage from 25% to 21.6%. And, tossing out the ranking rule and using the direct solution of the problem-defining integer optimization model lowered the result from 21.6% to 21.5%.

Diminishing returns is a well-known economic feature of many investment or production problems: doubling the amount of resources devoted to some task less than doubles the output. This example illustrates that diminishing returns can also apply to model formulation: we witnessed diminishing returns to insight in terms of effort in this presidential election example. It is often the case in policy modeling that the simplest, "first strike" model provides the greatest incremental insight, after which model refinements to improve realism are not always rewarded with more compelling or meaningful results. This is why with policy modeling, we try to analyze problems beginning with the simplest possible model with the hope of gaining great intuition before proceeding to more complicated models that capture additional detail.

1.9.3 Job Creation Programs

You are on the research staff of the United States Department of Labor. Your job for the past few years has been to review the evidence on different approaches to job creation, emphasizing the expected number of jobs that can be created by different programs at various levels of government funding.

A small sum, $10 million, has been made available annually to sponsor "showcase" job creation programs. You have been asked to review three programs that are of interest for various reasons. You summarize the data in the table below, which reports the number of jobs that could be created for investments of $0 through $5 million per year for three different non-overlapping programs:

$ (millions)	Program A	Program B	Program C
0	0	0	0
1	30	30	30
2	65	60	70
3	85	90	105
4	100	120	120
5	110	150	140

Suppose you are limited to allocating money in even increments of $1 million. Eager to impress, you decide to figure out the *optimal* allocations for all budgets ranging from $0 to $15 million in increments of $1 million. That is, you decide to figure out, for each budget: how much money each program should get; and the maximal number of jobs that can be created for each budget. Even though your available budget is only $10 million, you want to show how many additional jobs could be created were the budget expanded beyond $10 million (or how many new jobs would be foregone should the budget contract below $10 million).

(a) Prepare a graph showing the maximum number of jobs that can be created as a function of the budget.

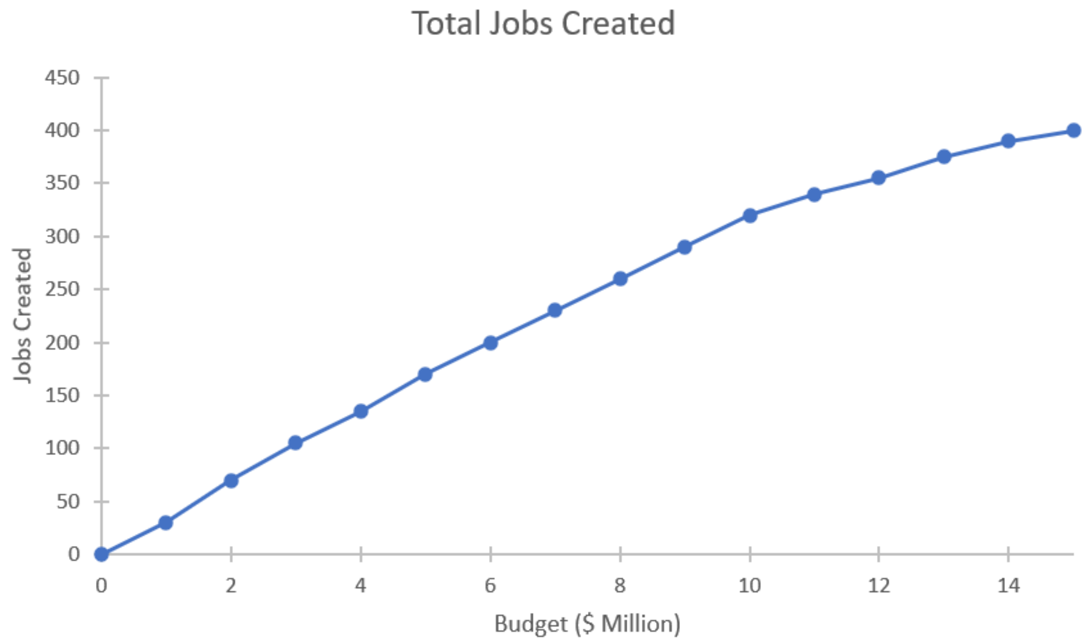
Let b_{ij} equal the benefits, measured in new jobs created, that result from spending \$$j$ million dollars on program i; $i = A, B, C$' $j = 1, 2, 3, ..., 10$. Let $x_{ij} = 1$ if you allocate \$$j$ million dollars to program i and 0 otherwise, again for $i = A, B, C$' $j = 1, 2, 3, ..., 10$. Let B denote the total budget. Then the resource allocation problem can be formulated as

$$\max \sum_{i=A}^{C} \sum_{j=0}^{5} b_{ij} x_{ij} \quad \text{(create as many jobs as possible!)}$$

subject to

$$\begin{aligned} \sum_{j=0}^{5} x_{ij} &= 1 \text{ for } i = A, B, C \text{ (each program receives some level of funding)} \\ \sum_{i=A}^{C} \sum_{j=0}^{5} j x_{ij} &\leq B \quad \text{(cannot spend more money in total than the budget)} \\ x_{ij} &= 0 \text{ or } 1 \text{ for } i = A, B, C; \ j = 1, 2, 3, ..., 10 \end{aligned}$$

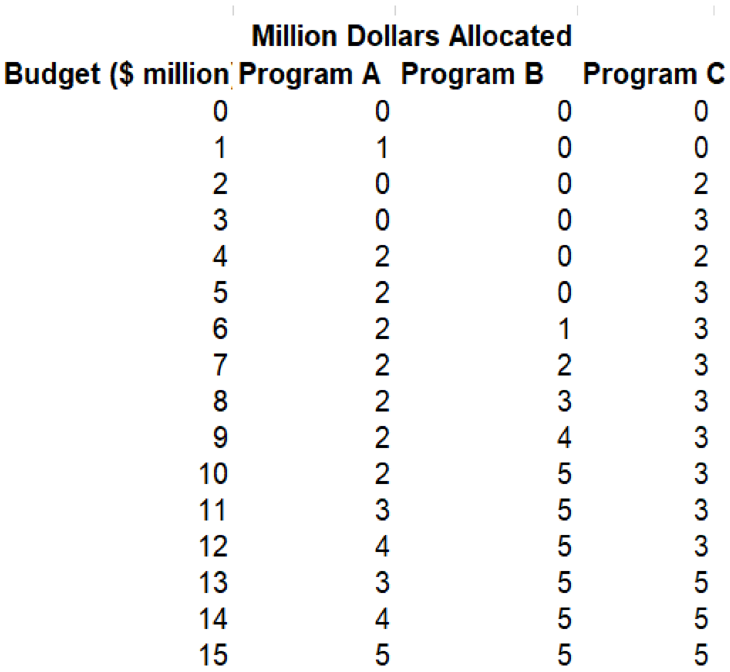
The data provided are the jobs created (b_{ij}) data. Using those values and solving repeatedly with the Solver in Excel for budgets running from \$0 to \$15 million yields the following graph of the maximum number of jobs created as a function of the budget:



For $10 million, you can get 320 jobs by investing $2, $5 and $3 million dollars in programs A, B and C yielding 65, 150, and 105 jobs respectively in each program.

(b) Prepare a table reporting the optimal amount of money given to each program at each budget level.

Repeated runs of the optimal resource allocation model above yields:



	Million Dollars Allocated		
Budget ($ million	Program A	Program B	Program C
0	0	0	0
1	1	0	0
2	0	0	2
3	0	0	3
4	2	0	2
5	2	0	3
6	2	1	3
7	2	2	3
8	2	3	3
9	2	4	3
10	2	5	3
11	3	5	3
12	4	5	3
13	3	5	5
14	4	5	5
15	5	5	5

Note how the optimal allocations jump around across the programs as the budget is increased; why do you think this is the case?

1.9.4 Screening for Learning Disabilities

It has been estimated that as many as 1 in 5 school children have learning disabilities or attention deficit issues. Left untreated, children with learning disabilities can quickly fall behind in their classes with potentially devastating lifelong consequences. Among young children (grades 1 through 5), it is often hard to distinguish a child with a learning disability from the variation in learning exhibited by young children without such disabilities. The good news is that there are interventions that, if implemented early, can essentially bring learning disabled kids back to grade level. The problem is that identifying kids with learning disabilities is far from straightforward so it is not obvious which students should receive such interventions. Further, these interventions are expensive, and schools simply cannot afford to offer interventions to all who might need them.

Given that interventions are only beneficial for truly learning disabled students, schools use very inexpensive pencil-and-paper screening tests to try and identify children with learning disabilities. The tests, while inexpensive, are also not completely accurate, though test accuracy does improve if offered at later rather than earlier grades. Suppose that for one such screening test, the true positive probability (that is, the test sensitivity) of identifying a truly learning disabled child is given by $p(t)$ if the test is administered during grade t. Suppose also that the false positive probability (that is, one minus the test specificity) that the screening test falsely identifies a child with no learning issues as disabled when the test is administered in grade t is given by $q(t)$. Also, suppose that the true prevalence of learning disability among incoming school students stays constant and equal to π (for πrevalence). Assume that entering students are either learning disabled or not, and that no changes to a child's true learning disability status occur over time.

(a) Suppose that the school has decided to administer a screening test for learning disabilities to all students in grade t, where t runs from 1 to 5 (so the first five grades of elementary school). What is the conditional probability that a student who tests positive on a screening test administered in grade t is truly learning disabled (that is, what is the positive predicted value of a test offered in grade t)?

The positive predictive value of a test administered in grade t, denoted here by $PPV(t)$, is the same thing as Pr{Student in grade t is learning disabled | student tests positive on the screening test in grade t}. This is a straightforward calculation using either the "never fail method" or Bayes' Rule, and is given by

$$PPV(t) = \frac{\pi p(t)}{\pi p(t) + (1 - \pi) q(t)}$$

Note that the numerator of this expression is just the probability that a kid is both learning disabled and tests positive on the screening test, while the denominator is the probability that a kid tests positive on the screening test (accounting for all children).

(b) Looking at your formula from part (a), describe how it behaves as a function of time t given that $p(t)$ is increasing with time while $q(t)$ is decreasing with time. That is to say, does the conditional probability that a child who tests positive is in fact learning disabled get larger, smaller, or stay the same over time given the assumptions governing the accuracy of the screening test?

Clearly $PPV(t)$ must be increasing with time. Intuitively, as time goes by, the screening test becomes more accurate at identifying both children with and without disabilities, thus the fraction of all students identified via the screening test who are truly disabled has to go up. A formal way to prove this is to show that the derivative of $PPV(t)$ with respect to time must be positive (you did not have to do this to gain full marks; the earlier intuition suffices). In simple words, the later you test, the more likely that a child who tests positive is truly learning disabled!

(c) Again looking at your formula from part (a), describe how it behaves as a function of π, the underlying prevalence of true learning disability in the student population. Does the conditional probability a child who tests positive is in fact learning disabled get larger, smaller, or stay the same as π varies?

Similarly, at any grade, so long as $p(t) > q(t)$ (which must be true for any reasonable screening test – it must be true that truly learning disabled students have a higher chance of testing positive than students without learning disabilities), the positive predictive value must increase with the prevalence π as the PPV becomes more heavily weighted towards $p(t)$ than $q(t)$ as π increases, and for any reasonable screening test $p(t) > q(t)$. You could also take the derivative of PPV with respect to π and show that it must be positive, but you didn't have to do that.

(d) Define $b(t)$ as the incremental benefit a *truly* learning disabled student would receive from an intervention administered in grade t (if the intervention is offered to a non-learning-disabled student, the incremental benefit is zero). Since interventions are so expensive, only students who test positive on the screening test can be offered an intervention, and indeed there might not be enough interventions for all students who would screen positive. Any intervention offered is therefore given to a randomly selected student from among those who tested positive on the screening test. Suppose that the screening test is offered to all students in grade t. Produce a formula that states the expected student benefit derived from a randomly selected intervention offered in grade t.

Let's let $\beta(t)$ represent the expected benefit from a randomly selected intervention offered in grade t. Since the intervention is only offered to those who test positive on the screening test, and since the benefit equals $b(t)$ for students who are truly learning disabled and 0 for students who are not learning disabled, and since the probability that a student who tests positive on the screening test is truly learning disabled is given by $PPV(t)$, it must be that:

$$\begin{aligned}\beta(t) &= PPV(t) \times b(t) + (1 - PPV(t)) \times 0 \\ &= PPV(t) \times b(t) \\ &= \frac{\pi p(t)}{\pi p(t) + (1-\pi)q(t)} \times b(t).\end{aligned}$$

(e) Now suppose that the school is able to administer the screening test to all students in a single grade t, but that the number of interventions that can be offered is, unfortunately, much less than the number of students who screen positive. Suppose that:

$\pi = 0.2$ (prevalence of learning disability in the student population)

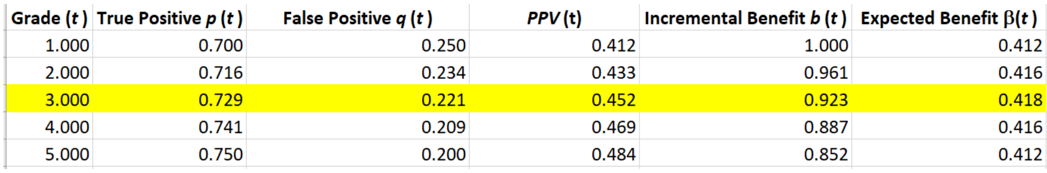
$b(t) = e^{-0.04(t-1)}$ for $t = 1, 2, 3, 4, 5$ (the incremental benefit from intervening with a truly learning disabled child in grade t)

$p(t) = 0.8 - 0.1e^{-0.1733(t-1)}$ for $t = 1, 2, 3, 4, 5$ (true positive rate or sensitivity)

$q(t) = 0.15 + 0.1e^{-0.1733(t-1)}$ for $t = 1, 2, 3, 4, 5$ (false positive rate or one minus specificity)

Produce a table that shows, as a function of grade (i.e. t), the numerical values of $p(t), q(t)$,the positive predictive value of the screening test (from part (a)), the incremental benefits $b(t)$, and the expected benefit derived from a randomly selected intervention in grade t (from part (d)). Based on this table, in which grade should the school administer screening tests if the goal is to maximize the expected benefit from offering interventions for learning disabilities?

This is effectively plug-and-play, and easy to do in Excel. Simply applying the stated formulas for $b(t), p(t)$ and $q(t)$, recognizing that $\pi = 0.2$, and using the result of part (a) for $PPV(t)$ and the result of part (d) for $\beta(t)$ we obtain:



Grade (*t*)	True Positive *p* (*t*)	False Positive *q* (*t*)	*PPV* (t)	Incremental Benefit *b* (*t*)	Expected Benefit β(*t*)
1.000	0.700	0.250	0.412	1.000	0.412
2.000	0.716	0.234	0.433	0.961	0.416
3.000	0.729	0.221	0.452	0.923	0.418
4.000	0.741	0.209	0.469	0.887	0.416
5.000	0.750	0.200	0.484	0.852	0.412

(f) How would your recommendation change (if at all) if, everything else remaining the same, the prevalence of learning disabilities in the population was as low as 10%? As high as 30%? Explain.

When $\pi = 0.1$, it makes sense to wait until grade 5 to administer the screening tests as that is when the highest expected benefit occurs ($\beta(5) = 0.2506$), whereas when $\pi = 0.3$, the highest expected benefit occurs when screening takes place in grade 1 ($\beta(1) = 0.5455$). The explanation lies with the tradeoff between increasing $PPV(t)$ and decreasing $b(t)$. Note that $b(t)$ falls from 1 to 0.852 from first through fifth grade, a decline of 14.8%. When $\pi = 0.1$, the PPV increases by 24% over the same grade range so it makes sense to delay screening given the large gain in PPV relative to $b(t)$. When $\pi = 0.3$, PPV only increases by 13%, which is not enough to overcome the drop in $b(t)$ so it makes sense to screen children as early as possible.

1.9.5 Allocating COVID Vaccine

The Advisory Committee on Immunization Practice (ACIP) issued recommendations for allocating the initial supplies of COVID-19 vaccine – see: https://www.cdc.gov/mmwr/volumes/69/wr/mm6949e1.htm. These recommendations stated that in the initial phase of the vaccination program, COVID-19 vaccine should be offered to health care personnel and residents of long-term care facilities (also known as nursing homes).

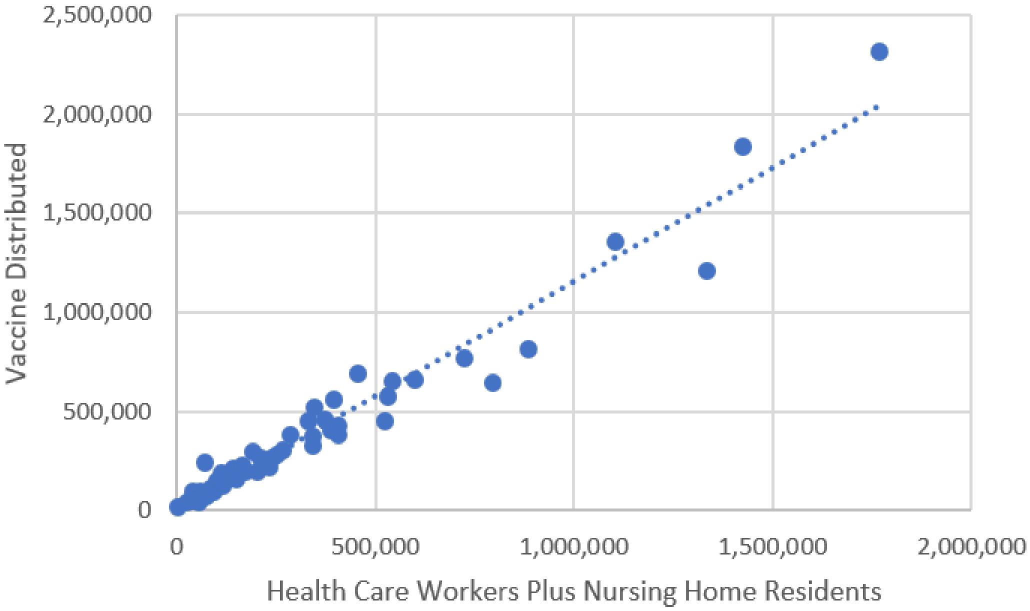
For each state in the United States plus Guam, the District of Columbia, and Puerto Rico, the Excel file titled Covid_Vaccine_Allocation contains the following data: total population, the sum of the number of health care personnel and nursing home residents, and number of vaccine doses distributed as of January 11, 2021. Please download this file.

(a) Plot the number of vaccine doses allocated (Y-axis) versus the sum of the number of health care personnel and nursing home residents (X-axis). On this same graph, also plot the number of vaccine doses that would be allocated to each jurisdiction if the vaccine was allocated in proportion to the sum of the number of health care personnel and nursing home residents, as suggested by the initial ACIP guidance. Let y_i (and $\widehat{y_i}$) denote the observed vaccine doses allocated to state i (and the number of vaccine doses if vaccine was distributed in proportion to the sum of the number of health care workers and nursing home residents in state i). The "root mean squared error" (or RMSE) is defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{53}(y_i - \widehat{y_i})^2}{53}}$$

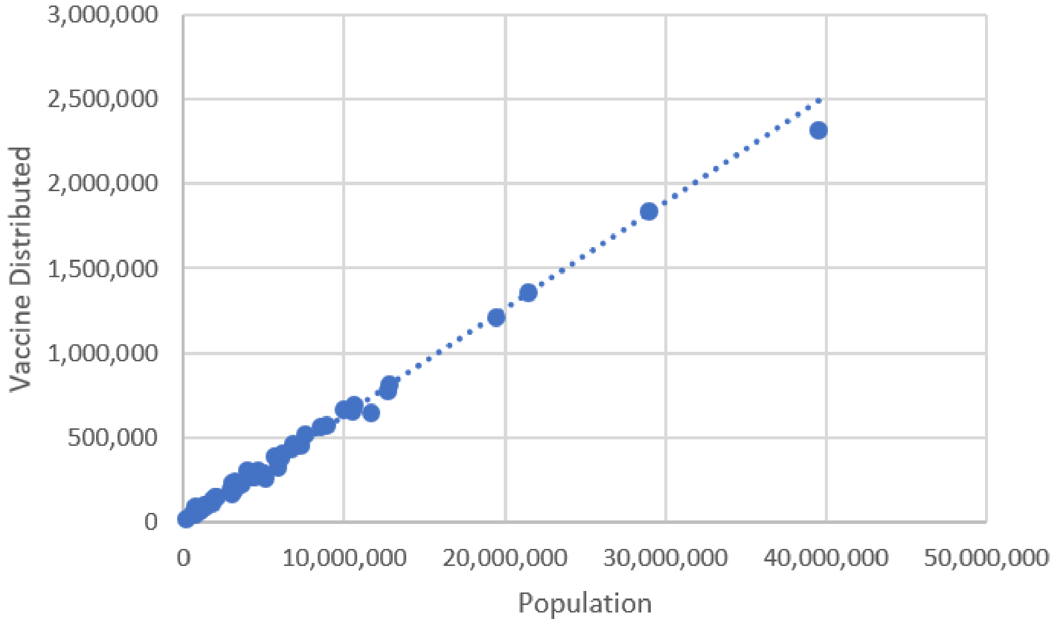
where the "53" arises from the 50 states plus Guam, DC and Puerto Rico. Calculate the RMSE for vaccine allocation in proportion to the sum of health care workers and nursing home residents, which we take as representing the ACIP recommendation.

The requested plot appears below (the dots show the actual allocations, while the dotted line shows allocation proportional to the sum of health care workers and nursing home recipients). The RMSE equals 96,930. While the amount of vaccine distributed roughly adheres to the sum of health care workers and nursing home residents in each jurisdiction, the pattern seems to break up once the size of the Phase 1a group exceeds 500,000.



(b) Similarly to part (a), on a new graph plot the number of vaccine doses allocated (Y-axis) versus the population (X-axis). On this same graph, also plot the number of vaccine doses that would be allocated to each jurisdiction if the vaccine was allocated in proportion to population, which is not the ACIP recommendation. Calculate the RMSE for vaccine allocation in proportion to the population.

Here is the plot comparing actual vaccine allocations to what would be expected if vaccine doses were distributed proportional to the population. With the exception of the largest jurisdiction (California), population alone does a much better job predicting the actual doses distributed as of the date these data were collected. The RMSE equals 34,605, which makes the "typical" error modeling allocation by the size of the Phase 1a numbers about 2.8 times larger than the error from modeling by population alone (that is, the RMSE for Phase 1a from part (a) above, 96,930, is 2.8 times larger than the RMSE of 34,605 calculated for allocating in proportion to population.



(c) Based on (a) and (b), what looks like a better predictor of how (at least as of January 11, 2021) COVID-19 vaccine has been distributed to individual states/jurisdictions: allocation based on the ACIP recommendations, or allocation based on population?

This is interesting – it looks like just allocating vaccine in proportion to population is a better back-of-the-envelope model for how vaccine has been distributed to the date of the data compared to allocating in proportion to the supposed Phase 1a numbers! This is clear from the two graphs, and also from the RMSE values calculated. There are many, many federal programs that allocate funds or other resources to states in proportion to population; that was not stated as the approach for Operation Warp Speed, but the data suggest allocations proceeded *as if* allocating in proportion to population was the real goal.

(d) Presumably vaccine is being allocated to save lives. Suppose those in charge of Operation Warp Speed (the program in charge of federal vaccine distribution to individual states/jurisdictions) believe that if y_i doses of vaccine are distributed to jurisdiction i with population p_i, then the expected number of lives saved in jurisdiction i, call this ℓ_i, would be given by

$$\ell_i = 0.0355 \times p_i \ln(y_i)$$

where $\ln(y_i)$ is the natural (base e) logarithm of y_i. Note that this formulation implies a marginally decreasing return in lives saved as the amount of vaccine allocated increases (that is, ℓ_i exhibits diminishing returns in y_i). Suppose that the total amount of vaccine available for distribution is given by v (as of January 11, 2021, $v = 20,908,375$ for the 53 jurisdictions considered in this problem). Also assume that political constraints dictate that at least 10,000 vaccine doses will be allocated to each jurisdiction. Formulate the mathematical resource allocation problem that, in words, says: find the non-negative vaccine allocations $\{y_i, i = 1, 2, ..., 53\}$ that maximize the expected total lives saved by vaccination subject to the constraints that each jurisdiction receives at least 10,000 vaccine doses, and the total number of vaccine doses allocated is equal to $v = 20,908,375$.

This just involves translating words into math. Succinctly, the problem is:

$$\max_{\{y_1, y_2, ..., y_{53}\}} \sum_{i=1}^{53} 0.0355 \times p_i \ln(y_i)$$

subject to the constraints

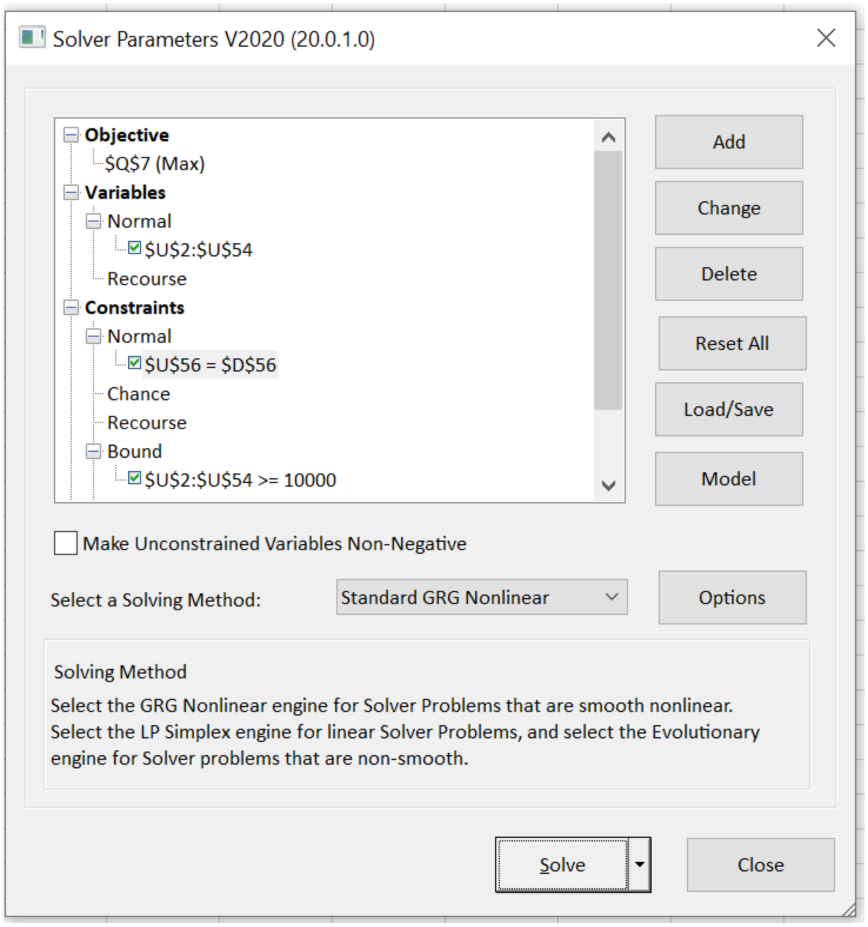
$$\sum_{i=1}^{53} y_i = v = 20,908,375$$

and

$$y_i \geq 10,000 \text{ for } i = 1, 2, ..., 53$$

(e) Either using the Solver in Excel, or using your mathematical skills, determine the vaccine allocations that solve the problem you defined in part (d). How do these "optimal" allocations compare to your answers to parts (a) and (b) above?

Using the Solver is pretty easy for this problem. The "changing cells" (or "decision variables" in operations research lingo) represent the number of vaccine doses (the $y_i's$), so one takes a column in Excel of length 53, and puts in initial numbers that represent your first guess at the answer (e.g. set the allocation for every state equal to $v/53 = 20,908,375/53 = 394,498$). The population values are given in the problem data, so you have those in a column that also has 53 rows. Now create a column that represents expected lives saved in each state, and each row value is the result of the Excel formula $= 0.0355 * p_i * \ln(y_i)$ where p_i is the population in state i, and y_i is your initial proposed vaccine allocation. The objective function to maximize is then just the sum of the rows in this column. Using the Solver to specify the "changing cells" and also to define the constraints (sum of the values in the number of doses allocated column equals 20,908,375, and the number of doses in each row is at least 10,000), you set the objective function equal to the sum of lives saved calculated earlier using the logarithmic function shown, and find the maximum. Here's what a sample Solver setup looks like when the sum of lives saved is in cell Q7, the vaccine allocations are in cells U2:U54 (note: 53 rows!); the sum of the vaccine allocations are in cell U56, and these are set equal to the total vaccines to be allocated which in this case was stuck in cell D56 (and equal to 20,908,375), and there is a bound stating that vaccine allocations must be at least 10,000 in each jurisdiction.



Solver Parameters V2020 (20.0.1.0)

- Objective
  - $Q$7 (Max)
- Variables
  - Normal
    - ☑ $U$2:$U$54
  - Recourse
- Constraints
  - Normal
    - ☑ $U$56 = $D$56
  - Chance
  - Recourse
  - Bound
    - ☑ $U$2:$U$54 >= 10000

□ Make Unconstrained Variables Non-Negative

Select a Solving Method: Standard GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

When you solve this problem, you should discover that vaccine allocations from this model turn out to be proportional to the population in each state, which as we saw in (b) above is very close to what seems to have happened. Now the question is, why?

This next part of the solution is a bit mathy, so for those of you who have not seen optimization problems like this, what follows is "music appreciation." Anyway, here goes: ignoring for a moment the minimum constraints $y_i \geq 10,000$ for $i = 1, 2, ..., 53$, we can reformulate the problem as

$$\max_{\{y_1, y_2, \ldots, y_{53}\}} \sum_{i=1}^{53} 0.0355 \times p_i \ln(y_i) - \lambda(\sum_{i=1}^{53} y_i - v)$$

where we have introduced a new variable, λ, referred to as a Lagrange multiplier. We have 54 variables in this problem: the 53 values y_i (one per jurisdiction) plus the new variable λ. Applying standard calculus to get the first order condition for a maximum requires differentiating the function above with respect to each variable and setting the result equal to zero. So, let's do that!

$$\frac{\partial}{\partial y_i}\left(\sum_{i=1}^{53} 0.0355 \times p_i \ln(y_i) - \lambda(\sum_{i=1}^{53} y_i - v)\right) = 0.0355 \times \frac{p_i}{y_i} - \lambda \to 0 \text{ for } i = 1, 2, ..., 53$$

and

$$\frac{\partial}{\partial \lambda}\left(\sum_{i=1}^{53} 0.0355 \times p_i \ln(y_i) - \lambda(\sum_{i=1}^{53} y_i - v)\right) = \sum_{i=1}^{53} y_i - v \to 0.$$

From the first equation, we see that

$$y_i = 0.0355 \times \frac{p_i}{\lambda}$$

while from the second equation we see that

$$\sum_{i=1}^{53} y_i = v.$$

Substituting the results from the first equation in to the second yields

$$\sum_{i=1}^{53} y_i = \sum_{i=1}^{53} 0.0355 \times \frac{p_i}{\lambda} = v$$

which means that

$$\lambda = \frac{0.0355}{v} \sum_{i=1}^{53} p_i.$$

Now that we know what λ is, we plug it back into our solution for y_i and get the final result

$$y_i = 0.0355 \times \frac{p_i}{\lambda} = 0.0355 \times \frac{p_i}{\frac{0.0355}{v} \sum_{i=1}^{53} p_i} = \frac{p_i}{\sum_{i=1}^{53} p_i} \times v \text{ (!!)}$$

We have just shown that if you maximize expected lives saved under the presumption that they follow from the logarithmic model specified, ignoring the minimum allocations in each jurisdiction you get vaccine allocation exactly in proportion to population! So what about the minimum allocation condition? Well, guess what – in this problem, if you allocate strictly according to population, every jurisdiction gets at least 10,000 doses (though Guam is close at only 10,452), so the minimum constraints turn out not to matter.

Now, you might be wondering why this is such a big deal when at no point in this problem did we attempt to justify the logarithmic expected lives saved formula. But that turns out to be exactly the point – the feds appear to have allocated vaccine *as if* they believed that the expected lives saved follow the logarithmic formula shown. And ladies and gentlemen, *that* is a head scratcher!

1.9.6 Allocating HIV Prevention Resources

In this question, you get to replicate some of the analysis contained in the report of the Institute of Medicine's Committee on HIV Prevention Strategies in the United States. To prepare, first read carefully Chapter 3 ("Allocating Resources") and Appendix D ("Description and Mathematical Statement of the HIV Prevention Resource Allocation Model"), both contained in your coursepack. Then, download the Excel file titled Holmberg_Data from the Policy Modeling Web site. This file contains the state-by-state data reporting total population; and risk group population, number of current HIV infected persons in the risk group, and annual number of new HIV infections in the risk group for the three risk groups considered: injecting drug users (IDUs), men who have sex with men (MSMs), and heterosexuals at high risk (HETs). Note that all of these numbers were estimated by Dr. Scott Holmberg of the Centers for Disease Control.

Define:

$$
\begin{aligned}
x_{ij} &= \text{dollars for programs serving group } j \text{ in state } i \\
c_j &= \text{cost per person in programs in group } j \\
e_j &= \text{relative reduction in HIV incidence for programs in group } j \\
f_j &= \text{maximum fraction of the population reachable in group } j \\
r_{ij} &= \text{total number of new HIV infections per year in state } i\text{, group } j \\
n_{ij} &= \text{number of persons at risk in state } i\text{, group } j \\
B &= \text{HIV prevention budget}
\end{aligned}
$$

Then the number of persons who can be reached with programs in state i that serve persons in group j equals x_{ij}/c_j (note the units: dollars divided by dollars per person equals persons). Also, the number of infections that can be prevented with programs in state i that serve persons in group j is given by $(x_{ij}/c_j) \times (r_{ij}/n_{ij}) \times e_j$. Focusing on group j in state i, this product can be understood as the number of persons reached with prevention programs (x_{ij}/c_j) times the per capita rate of new infections per person $(r_{ij}/n_{ij}$, times the relative reduction in the infection rate due to the prevention programs (e_j).

A linear program for optimally allocating HIV prevention resources is then given by:

$$\max_{\{x_{ij}\text{'s}\}} \sum_i \sum_j \frac{x_{ij}}{c_j} \times \frac{r_{ij}}{n_{ij}} \times e_j \text{ (Prevent as many infections as possible!)}$$

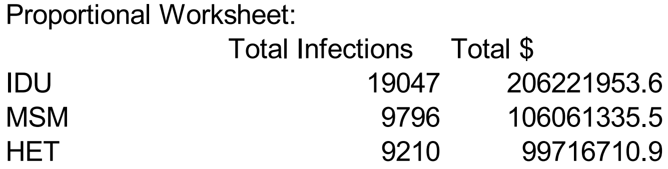
subject to

$$
\begin{aligned}
x_{ij} &\leq f_j c_j n_{ij} \; \forall i,j \;\; \text{(Can't reach more than } 100f\text{\% of population)} \\
\sum_i \sum_j x_{ij} &\leq B \text{ (Can't spend more than the budget)} \\
x_{ij} &\geq 0 \; \forall i,j \text{ (non-negative resource allocation)}
\end{aligned}
$$

Focus on the base case parameters as reported in Table 3-1 of Chapter 3. Also presume a budget constraint of \$412 million dollars, which is roughly what the CDC spent on prevention programs in Fiscal Year 1999 (see p. 29, “Allocating Resources at the National Level”). Then, using the data in the spreadsheet and the following the formulation in Appendix D, answer the following questions:

(a) Under *proportional allocation to new HIV infections,* what is the total amount of money that would be allocated to programs for IDUs, MSMs, and HETs across all states? (You need to figure out the proportional allocation of funds to programs for each risk group in each state, and then add across states for each risk group to obtain the three numbers sought).

This is straightforward – there is a $412 million budget that is split in proportion to the number of new infections in each risk group/location combination (i.e. to the values of the r_{ij}'s in Holmberg's data). Directly from these data, one obtains the following:



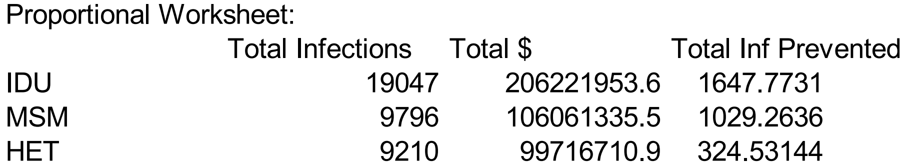
Proportional Worksheet:

	Total Infections	Total $
IDU	19047	206221953.6
MSM	9796	106061335.5
HET	9210	99716710.9

So IDUs get about 50% of the money, while MSM's and HET's each get about 25%. Not a huge surprise when you consider that these are just the relative proportions of new infections in each group!

(b) Under *proportional allocation to new HIV infections*, what is the total annual number of infections prevented in each of the three risk groups, and hence what is the total annual number of infections prevented overall? (Don't forget to incorporate the assumption that, in the base case from Table 3-1, only 50% of risk group members in any location can be reached!)

This is a little trickier. For each risk group/location combination, you need to look at how much money was allocated, and then how many people could be reached in principle. However, this number cannot be larger than 50% of the risk group population (by assumption of the 50% reach parameter). Once you have figured out the number of persons reached in each group, you multiply through by the ratio of new infections to the population risk group size, and then by the presumed efficacy estimate to obtain infections averted. This calculuation is detailed at the top of p. 136 in Appendix D. Finally, you just add within the different risk groups. I obtain:



Proportional Worksheet:

	Total Infections	Total $	Total Inf Prevented
IDU	19047	206221953.6	1647.7731
MSM	9796	106061335.5	1029.2636
HET	9210	99716710.9	324.53144

In total, proportional allocation of $412 million looks like it would avert about 3,000 infections.

(c) Now, determine the amount of money that would be allocated to programs for each risk group in each state in order to *prevent as many infections as possible*, and then report the total amount of money allocated nationwide to programs for IDUs, MSMs, and HETs.

There are two different ways you can proceed. One approach is just to figure out the cost-effectiveness ratios for all risk groups in all states: how many infections can you prevent per dollar in each case? Then, once you have done this, allocate the money from most to least cost-effective location/risk-group combinations, noting that for each such combination you will either run out of people (because of the 50% maximum reach assumption) or, eventually, run out of money (because of the $412 million budget constraint).

The other way to proceed, for those who are familiar with Excel's *Solver* tool, is to exactly formulate the linear program presented in Appendix D inside your spreadsheet and solve using the *Solver*.

Using the first approach, you just compute for each risk group/location combination the cost-effectiveness ratio (really infections averted per dollar) given by

$$ratio = \frac{1}{c_j} \times \frac{r_{ij}}{n_{ij}} \times e_j$$

where c and e are the cost per person, and fractional reduction in incidence (i.e. the efficacy), and r and n are the rate of new infections (new infections per year) and the size of the risk group. Then you sort all combinations from highest to lowest ratios. Then you figure out the maximum number of persons you could possibly place in programs (given by 50% of the population size), and the cost of doing so (given by c times the reachable population). The infections prevented will then just equal the money spent times the cost-effectiveness ratio above. Add up the money until you run out. It turns out that you run out of money when considering programs for HETs in New York. Up until that group, you would have spent a total of $391,207,500 on group/location combos up to and including IDUs in Texas. To fund programs to cover all HETs in New York would require cumulative spending of $439,762,500. But you only have $412 million, so you can only spend $412 million - $391,207,500 million = $20,792,500 on HETs in New York. That's when you run out of money.

Summing the amounts allocated over the different risk groups, I obtain the following results:
IDUs: \$135,030,000
MSMs: \$214,837,500
HETs: \$62,132,500

It adds to \$412 million. Hooray!

(d) Having determined the amount of money that would be allocated to programs to *prevent as many infections as possible*, report the annual number of infections prevented in each of the three risk groups, and hence the total annual number of infections prevented overall. What percentage improvement does this offer over the policy of proportional allocation?

In figuring out the sums spent on each risk group in part (c), I noted that for all risk group/location combinations more efficient than New York HETs, one reaches 50% of the population at risk in the group. One then prevents 100e% of the new infections that would have occurred in that group. For New York HETs, you can reach only \$20,792,500 / \$300 = 69,308 of the 323,700 at risk there. You then figure out infections prevented as before after adding in the partial result for New York HETs. I obtain:
IDUs: 1648.1
MSMs: 1959.2
HETs: 289.0

This adds to 3896.3, or about 3,900. Recall that with proportional allocation, the number of infections prevented per year was 3,000. So, the improvement is given by roughly $(3,900 - 3,000)/3,000 = 30\%$. Also, note that the number of infections prevented among IDUs is virtually identical under the proportional and cost-effective approaches, but much cheaper in the second case. More infections are prevented among MSMs under the cost-effective allocations (1959 versus 1029), but fewer among HETs (289 versus 325). The gain in efficiency really comes from targeting more resources to programs for MSMs, which (by assumption) are more efficacious.

1.9.7 Optimizing Needle Exchange

In a needle exchange program, participating drug injectors are legally allowed to exchange used needles for clean ones with the goal of preventing HIV and other infections. Define λ as the needle sharing rate per injector in the program, v as the number of needles exchanged per unit time per injector in the program, and I_0 as the number of new HIV infections per injector per unit time in the absence of needle exchange (the HIV *incidence rate*). A simplified model of needle exchange equates the number of HIV infections averted per drug injector, denoted by $\Delta I(v)$, to

$$\Delta I(v) = I_0 \frac{v}{\lambda + v}.$$

(a) Let b = the benefit of each infection averted, and c = the cost of each needle exchanged. Explain why the net benefit from operating a needle exchange program, denoted by $\beta(v)$, is given by

$$\beta(v) = b\Delta I(v) - cv.$$

The program prevents $\Delta I(v)$ infections per participant per unit time, each of which provides a benefit of b, while it costs c per needle exchanged, and v needles are exchanged per participant per unit time. So, the net benefit of operating the program is given by $\beta(v) = b\Delta I(v) - cv$ per participating drug injector per unit time.

(b) What value of the needle exchange rate v maximizes the net benefit of operating the program? Provide an explicit formula

We seek to maximize $\beta(v)$, which we will accomplish via differentiation. We find that

$$\frac{d}{dv}\beta(v) = \frac{d}{dv}\left(bI_0 \frac{v}{\lambda + v} - cv\right) = \frac{1}{(v+\lambda)^2}\left(bI_0\lambda - c(v+\lambda)^2\right).$$

Equating this to zero and solving for v yields

$$v^* = \sqrt{\frac{b\lambda I_0}{c}} - \lambda$$

where we have been careful to take the positive solution from the quadratic above.

(c) Suppose that very conservatively, preventing an HIV infection averts \$56,000 in healthcare costs (b = \$56,000), that needles cost 15 cents each

($c = \$0.15$), that HIV incidence equals 2% ($I_0 = 0.02$), and that injectors each share needles 246 times per year on average ($\lambda = 246/\text{yr}$). What is the optimal needle exchange rate?

Plug and play! We have

$$v^* = \sqrt{\frac{56,000 \times 246 \times 0.02}{0.15}} - 246 \approx 1,109$$

(d) Suppose that the needle exchange rate v is sufficiently small relative to the needle sharing rate λ that

$$bI_0 \frac{v}{\lambda + v} \approx \frac{bI_0 v}{\lambda}.$$

What then is the marginal benefit per needle exchanged per injector, what is the marginal cost per needle exchanged, and consequently when does it make sense to even establish a needle exchange program?

In this case, the net benefit of the exchange program per injector is given by $(bI_0 v/\lambda - cv) = (bI_0/\lambda - c)v$ and consequently the marginal benefit per needle exchanged is just given by $bI_0/\lambda - c$. It would only make sense to establish a program if the marginal benefit is positive, that is, if

$$\frac{bI_0}{\lambda} > c$$

which can also be written as

$$I_0 > \frac{c\lambda}{b}$$

which means that if the HIV incidence rate I_0 is too small (less than $c\lambda/b$), it does not even make sense to establish the program.

Another way to see this condition is to revisit your solutions to part (b) for the optimal needle exchange rate, and ask when that optimum exchange rate is positive. You will discover that

$$v^* = \sqrt{\frac{b\lambda I_0}{c}} - \lambda > 0$$

iff

$$I_0 > \frac{c\lambda}{b}.$$

(e) Now suppose that $x =$ number of clients in the program, and consequently the *number* of infections prevented is given by $x\Delta I(v)$. The benefit and cost of preventing infections are given by $bx\Delta I(v)$ and cxv respectively. Also, suppose that the *marginal* cost of recruiting new participants is linear in the number of participants because participants are found in order of easiest to hardest to reach. It then follows that $kx =$ *marginal* cost of x^{th} client, and consequently the total cost of x clients is given by $kx^2/2$. Argue that the net benefits of operating a program with x participants and an exchange rate of v needles per participant, $\beta(x, v)$ is given by

$$\beta(x, v) = bx\Delta I(v) - cxv - k\frac{x^2}{2}\cdot$$

The new net benefits function is now expressed in dollars per unit time instead of dollars per injector per unit time. As before, we have that $b\Delta I(v) - cv$ represents the net benefits of exchanging needles per participant, so we need to multiply this by the number of participants (x) to get the total net benefits from exchanging needles. But, we now also need to include the recruiting costs of getting x clients into the program, which requires subtracting off $kx^2/2$. This gives us the stated formula for $\beta(x, v)$.

(f) What is the optimal number of participants in the program, and what is the optimal needle exchange rate?

Now we need to optimize over both the needle exchange rate v and the number of clients x. Again we proceed via differentiation. Note that

$$\frac{\partial}{\partial v}\beta(x, v) = x\frac{d}{dv}\beta(v)$$

where $\beta(v)$ was defined in part (a). Setting this result equal to zero yields the same equation $\frac{d}{dv}\beta(v) = 0$ as in part (b) so we see immediately that the optimal exchange rate remains

$$v^* = \sqrt{\frac{b\lambda I_0}{c}} - \lambda$$

as before. Continuing, we now set

$$\frac{\partial}{\partial x}\beta(x, v) = 0$$

which yields

$$x^*(v) = \frac{bI_0\frac{v}{\lambda+v} - cv}{k}.$$

Substituting the result for the optimal needle exchange rate and simplifying yields

$$x^* = \frac{cv^{*2}}{k\lambda} = \frac{(\sqrt{bI_0} - \sqrt{c\lambda})^2}{k}$$

(g) The program also has fixed costs – e.g. the salaries for the outreach workers and the cost of upkeep for the building (or van) where the program is operating. Denote these fixed costs by f. Taking such fixed costs into account, when does it make sense to establish a needle exchange?

Now we need to subtract off the fixed costs from the net benefits of exchanging needles and recruiting clients, so we need to ask when

$$bx^*\Delta I(v^*) - cx^*v^* - k\frac{x^{*2}}{2} - f > 0,$$

that is, when we still get positive net benefits from operating the program at the optimal exchange rate and client population level (which would stay the same given the fixed costs). Substituting in the results from part (f) you will discover that the function above remains positive so long as

$$I_0 > \frac{(\sqrt[4]{2kf} + \sqrt{c\lambda})^2}{b}.$$

Note that if $f = 0$, this condition reduces to what we found in part (d) when there were no fixed costs.

(h) Suppose that the health department has a budget of B dollars. Taking into account fixed costs, the cost of exchanging needles, and the cost of recruiting participating drug injectors, how would you determine the number of program participants and needle exchange rate to prevent as many infections as possible while satisfying the budget constraint?

You need to solve the following optimization problem:

$$\max_{v,x} I_0\frac{v}{\lambda+v}x$$

subject to

$$cxv + k\frac{x^2}{2} \leq B - f$$

and of course v and x both non-negative. This can be solved using a Lagrange multiplier and recognizing that since the objective function is increasing in both v and x, one would always exhaust the budget. If $B < f$ one cannot establish a program (and $v^* = x^* = 0$). Otherwise the solution follows from (letting γ be the Lagrange multiplier)

$$\max_{v,x,\gamma} \left\{ I_0 \frac{v}{\lambda + v} x - \gamma(cxv + k\frac{x^2}{2} + f - B) \right\}.$$

Differentiating with respect to $v, x,$ and γ and setting to zero yields three equations:

$$\frac{d}{dv}\left(I_0 \frac{v}{\lambda + v} x - \gamma(cxv + k\frac{x^2}{2} + f - B) \right) = -\frac{x}{(v+\lambda)^2}\left(c\gamma v^2 + 2c\gamma v\lambda + c\gamma\lambda^2 - I_0\lambda \right) = 0$$

$$\frac{d}{dx}\left(I_0 \frac{v}{\lambda + v} x - \gamma(cxv + k\frac{x^2}{2} + f - B) \right) = -\frac{1}{v+\lambda}\left(cv^2\gamma - vI_0 + kvx\gamma + cv\lambda\gamma + kx\lambda\gamma \right) = 0$$

$$\frac{d}{d\gamma}\left(I_0 \frac{v}{\lambda + v} x - \gamma(cxv + k\frac{x^2}{2} + f - B) \right) = -\frac{1}{2}kx^2 - cvx + B - f = 0$$

Solving each of the first two equations for the Lagrange multiplier γ yields two different expressions for γ, namely

$$\gamma = \frac{\lambda I_0}{cv^2 + c\lambda^2 + 2cv\lambda} \text{ (from differentiating by } v\text{)}$$

and

$$\gamma = \frac{vI_0}{cv^2 + cv\lambda + kx\lambda + kvx} \text{ (from differentiating by } x\text{).}$$

Equating these two expressions for γ and solving for x yields

$$x = \frac{cv^2}{\lambda k}.$$

Substituting this expression for x into the budget constraint $cxv + k\frac{x^2}{2} + f = B$ yields the following equation for v:

$$\frac{c^2v^3}{\lambda k} + \frac{c^2v^4}{2\lambda^2 k} + f = B.$$

So, in any instance of this problem, you would solve the equation above for v to find the optimal needle exchange rate, and substitute the resulting value into the equation for x to obtain the optimal number of participants in the program.

Chapter 2

Bernoulli Process Models

In discrete time, an “event” (or “arrival” or “success”) occurs at any point in time with probability p, and does not occur with probability $1 - p$. The outcome of any trial at any point in time is independent of the outcomes of all other trials. This scenario gives rise to the *Bernoulli process*. Examples include the successive flips of coins, the reliability of components or devices over time (does the power plant blow up each week - yes or no?), the duration of stay in hospitals, housing projects, etc. (each week there is a chance of leaving), and so forth. Although the assumption of a constant success probability p may seem unrealistic in some cases, this assumption yields the simplest modeling approach available that retains the major features of these phenomena, often leads to key insights which do not change when more accurate assumptions are made, and is surprisingly accurate in some cases.

2.1 The Bernoulli Distribution

We say that the random variable X has the *Bernoulli distribution* if

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

Thus, $X = 1$ with probability p, and $X = 0$ with probability $1 - p$. Note that the mean of the Bernoulli random variable is given by

$$E(X) = p \times 1 + (1-p) \times 0 = p$$

while the mean squared Bernoulli equals

$$E(X^2) = p \times 1^2 + (1-p) \times 0^2 = p$$

as well. The variance of the Bernoulli random variable thus equals

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p).$$

Note that the variance reaches a maximum at $p = 1/2$, reflecting the notion that maximum uncertainty is "like a coin toss."

2.2 The Binomial Distribution

Let X_n denote the number of successes in n Bernoulli trials. X_n follows the *Binomial* distribution:

$$\Pr\{X_n = x\} = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, ..., n.$$

The first term represents the number of ways to obtain x "successes" in n "trials," while the second term represents the chance of obtaining a particular path (or sequence) of x successes and $n - x$ failures. Note that X_1 is the Bernoulli random variable.

The Binomial random variable has a physical interpretation. X_n is the sum of n independent, identically distributed Bernoulli variables, that is,

$$X_n = X_1^{(1)} + X_1^{(2)} + ... + X_1^{(n)}$$

where $X_1^{(i)}$ is a Bernoulli random variable, $i = 1, 2, ..., n$. Thus, the mean and variance of the Binomial are easily found from the underlying Bernoulli variables:

$$\begin{aligned} E(X_n) &= E[X_1^{(1)} + X_1^{(2)} + ... + X_1^{(n)}] \\ &= \underbrace{p + p + ... + p}_{n \text{ times}} \\ &= np. \end{aligned}$$

$$\begin{aligned} Var(X_n) &= Var[X_1^{(1)} + X_1^{(2)} + ... + X_1^{(n)}] \\ &= \underbrace{p(1-p) + p(1-p) + ...p(1-p)}_{n \text{ times}} \\ &= np(1-p). \end{aligned}$$

As with the Bernoulli, the variance of X_n is largest when $p = 1/2$. Note that at $p = 0$ or 1, the variance of X_n equals zero, reflecting the fact that failure or success, respectively, always occurs.

2.3 The Geometric Distribution

Random variable T has the *Geometric* distribution if

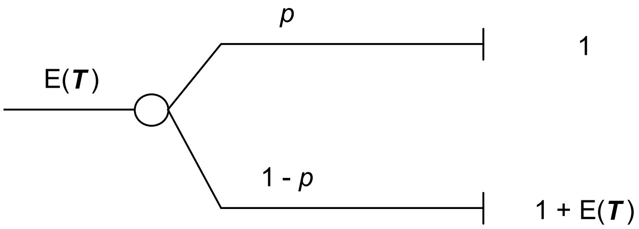
$$\Pr\{T = t\} = p(1-p)^{t-1} \text{ for } t = 1, 2, 3, ...$$

T represents the number of trials necessary to obtain the first success in a Bernoulli process. To obtain the first success on trial t, the first $t-1$ trials must result in failure; this occurs with probability $(1-p)^{t-1}$. On trial t, success must occur; this has probability p.

To find the expected value of T, we could evaluate

$$E(T) = \sum_{t=1}^{\infty} t \times p \times (1-p)^{t-1}$$

but this looks ugly. Rather, we will use the "repetition method." Note that



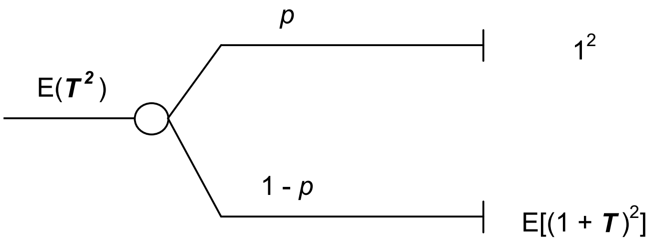
Resolving the lottery above leads to the equation

$$E(T) = p + (1-p) \times (1 + E(T))$$

which has the solution

$$E(T) = \frac{1}{p} \cdot$$

We have used the principle that if a failure occurs on the first trial, the remaining future is probabilistically identical to the future as viewed before the first trial takes place. We can use this method to find $Var(T)$ as well. First, we need to obtain $E(T^2)$. Proceeding as above, we establish the following lottery:



Note that

$$\begin{aligned} E[(1+T)^2] &= E[1^2 + 2T + T^2] \\ &= 1 + 2E(T) + E(T^2) \\ &= 1 + \frac{2}{p} + E(T^2). \end{aligned}$$

Solving for $E(T^2)$ yields

$$E(T^2) = \frac{2}{p^2} - \frac{1}{p}\cdot$$

Finally, noting that $Var(T) = E(T^2) - E(T)^2$ we see that

$$Var(T) = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}\cdot$$

The *smaller* the value of p, the larger the variance in T, the number of trials until the first success.

2.4 Examples

2.4.1 Nation of Shoplifters Revisited

Recall the Nation of Shoplifters example where you were told how many crimes were being committed over the course of a criminal career with and without intervention. Now let's model how such statistics might come about. Suppose that shoplifters *recidivate* (commit another crime) after each crime committed with probability 9/10, but with probability 1/10 they retire from crime forever. This creates a Bernoulli process model for the occurrence of shoplifting crimes over time.

The Nation's Elders are entertaining a proposal for a new intervention that, in theory, should reduce the recidivism probability from 9/10 to 8/10. "Why should we be excited by this," an Elder exclaims, "when instead of continuing to commit crimes 90% of the time, crimes would be committed 80% of the time? That's only an 11% reduction in the recidivism rate!" And truth be told, $(90\% - 80\%)/90\% = 1/9 = 11\%$.

But wait – the repetition method introduced above suggests that the expected number of crimes committed over an entire shoplifting career, call this τ, when the recidivism rate is given by r, is the solution to

$$\tau = 1 - r + r \times (1 + \tau) = 1 + r\tau$$

and consequently the expected number of crimes per career is given by

$$\tau = \frac{1}{1 - r}.$$

There is another way to explain this result. Imagine some large number of crimes, say n. Under the Bernoulli process model, with probability $1 - r$ any crime is a career-ending crime, that is, the offender quits shoplifting for all time with probability $1 - r$. The expected number of career-ending crimes out of the n crimes considered then equals $n \times (1 - r)$, and consequently the number of crimes per career, which is the same as the number of crimes per career-ending crime, is given by $n/(n \times (1 - r)) = 1/(1 - r) = \tau$ as before.

Substituting $r = 0.9$ shows that the current situation ($r = 0.9$) results in an expected 10 crimes per criminal career, while substituting $r = 0.8$ yields 5 crimes per career. The 11% reduction in recidivism rates equates to cutting all future crime in half! There is a lesson here – reporting an 11% reduction in recidivism and a 50% reduction in crime are both correct statements, but the second is much more powerful than the first (and could convince the Elders to adopt the new intervention when the simple reduction in recidivism rates might fail to do so).

2.4.2 Bernoulli Hiring and the 4/5ths Rule

Consider a firm that is trying to fill n available positions. The population of qualified applicants is much greater than the number of slots to fill, and for our purposes this population can be considered infinitely large. However there are two different but equal sized groups of applicants, say group 1 and group 2, whose true job qualifications are uncorrelated with group identity (e.g. men and women, young and old, whites and persons of color). The hiring process works as follows: an applicant is selected at random for an interview (implying that the applicant so selected is a member of group 1 with probability 1/2). Interviewed applicants from group i are hired with probability p_i ($i = 1, 2$). If an interviewed applicant is not hired, however, then the firm randomly picks the next applicant to interview from the pool. This process continues until all n positions are filled.

Here are several questions regarding this Bernoulli hiring process:

1. What is the probability of hiring an applicant who has just been interviewed?

2. Suppose that m interviews have been completed. What is the probability that exactly x persons were hired as a result of these m interviews? What is the mean and variance of the number of persons hired from m interviews?

3. The interview and hiring process has just begun! What is the probability distribution for the number of persons interviewed until the first position is filled? What is the mean and variance of the number of persons interviewed to fill the first position?

4. Recall that there are a total of n positions to fill. What is the probability distribution for the number of persons interviewed until all n positions are filled? What is the mean and variance of the total number of persons interviewed to fill all n positions?

5. What is the conditional probability that when a person is hired, that person is a member of group 1?

6. After all n positions have been filled, what is the probability distribution of the number of persons hired from group 1? What is the mean and variance of the number of persons from group 1 who are hired?

7. Federal law states that "A selection rate for any race, sex, or ethnic group which is less than four-fights (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact." (REFERENCE: Section 1607.4 of Equal Employment Opportunity Commission Act). What is the probability that the results of this firm's interview and hiring process would trigger the "4/5ths

rule" of adverse impact? In particular, suppose that there is equal hiring opportunity, meaning that $p_1 = p_2 = p$; what is the probability of triggering the 4/5ths rule and generating a "false positive" result? Alternatively, suppose that there is hiring discrimination (meaning that $p_1 = p_2$); what is the probability of failing to trigger the 4/5ths rule and generating a "false negative" result? What do these calculations suggest regarding the 4/5ths rule?

Now let's answer these questions:

1. The probability of hiring an applicant who has just been interviewed

Since the two applicant groups are equally numerous, both the probability of interviewing an applicant from group 1 and the probability of interviewing an applicant from group 2 equal 1/2. Since the conditional probability of hiring a group i applicant equals p_i, the unconditional probability of hiring a person who was just interviewed, call this q, is given by

$$q = \frac{1}{2}p_1 + \frac{1}{2}p_2$$

which is the simple average of the conditional hiring probabilities. Suppose that instead of being equally numerous, the fraction of all applicants from group 1 was equal to f. In this case the unconditional probability of hiring a person just interviewed would equal

$$q = f \times p_1 + (1 - f) \times p_2.$$

2. The number of applicants hired in m interviews

Let X denote the number of applicants hired in m interviews. We just learned that the probability any person interviewed is hired is given by q as defined above, thus the probability distribution of the number of positions filled from m interviews follows the Binomial distribution

$$\Pr\{X = x\} = \binom{m}{x} q^x (1 - q)^{m-x},\ x = 0, 1, ..., m.$$

The mean and variance of the number of applicants hired just follows the Binomial results as

$$E(X) = mq$$

and

$$Var(X) = mq(1-q).$$

3. The number of interviews until the first applicant is hired

Recalling again that any interview results in a filled position with probability q, the number of interviews required to hire the first applicant, call this T, follows the Geometric distribution

$$\Pr\{T = t\} = q(1-q)^{t-1},\ t = 1, 2, 3, ...$$

with mean and variance given by

$$E(T) = \frac{1}{q}$$

and

$$Var(T) = \frac{1-q}{q^2}.$$

4. The number of interviews until all n positions are filled

Let's define T_n as the time needed to fill all n positions. To determine the probability that $T_n = t$ for $t = n,\ n+1,\ n+2...$ (as clearly there must be at least n interviews to hire n applicants!), note that the only way the n^{th} position is filled on the t^{th} interview is if exactly $n-1$ persons have been hired as a result of the first $t-1$ interviews, and a person is hired as a result of the t^{th} interview. Consequently

$$\begin{aligned}\Pr\{T_n = t\} &= \underbrace{\binom{t-1}{n-1} q^{n-1}(1-q)^{(t-1-(n-1))}}_{\Pr\{\text{hire } n-1 \text{ persons in } t-1 \text{ interviews}\}} \times \underbrace{q}_{\Pr\{\text{hire}\}} \\ &= \binom{t-1}{n-1} q^{n}(1-q)^{(t-n)},\ t = n,\ n+1,\ n+2...\end{aligned}$$

This is known as the Pascal probability distribution. To determine the mean of the number of interviews required to hire n people, let $T_{i-1,i}$ denote the number of interviews required to fill the i^{th} position beyond the number of interviews required to fill the $(i-1)^{st}$ position. With this notation, note that $T_{0,1}$ corresponds to the number of interviews to fill the first position from the start of the hiring process, $T_{n-1,n}$ is the number of interviews required to fill the last position counting from when the $(n-1)^{st}$ person was hired, and the total number of interviews T_n required to hire n applicants is given by

$$T_n = \sum_{i=1}^{n} T_{i-1,i}.$$

Now, each of the random variables $T_{i-1,i}$ corresponds to the number of interviews required to fill a single position, and consequently $T_{i-1,i}$ follows the Geometric distribution discussed earlier. Since $E(T_{i-1,i}) = E(T) = 1/q$, we have just discovered that

$$E(T_n) = E(\sum_{i=1}^{n} T_{i-1,i}) = \sum_{i=1}^{n} E(T_{i-1,i}) = \frac{n}{q}.$$

Similarly, since the random variables $T_{i-1,i}$ are mutually independent (as the number of interviews from a given hire to fill the next position is in no way is affected by the number of interviews needed to fill other positions), the variance of the total number of interviews needed to fill all n positions may be written as

$$Var(T_n) = Var(\sum_{i=1}^{n} T_{i-1,i}) = \sum_{i=1}^{n} Var(T_{i-1,i}) = n \times \frac{1-q}{q^2}.$$

.

5. The probability a position is filled with an applicant from group 1

We have already deduced that the likelihood a position is filled is given by $q = (p_1+p_2)/2$ (assuming our two groups are equally numerous). Conditional upon filling a position, the probability that the applicant came from group 1, call this π, is found from

$$\pi = \frac{\Pr\{\text{Applicant interviewed is from group 1 and is hired}\}}{\Pr\{\text{Applicant interviewed is hired}\}} = \frac{\frac{1}{2} \times p_1}{\frac{1}{2} \times (p_1 + p_2)} = \frac{p_1}{p_1 + p_2}.$$

Before continuing, ask yourself how π would change if the fraction of all applicants emanating from group 1 was equal to f instead of $1/2$.

6. The number of positions filled by applicants from group 1

Let H_i denote the number of applicants hired from group i. Since the probability that any position is filled by an applicant from group 1 is given by π, and since all positions are filled independently of each other in this Bernoulli hiring process, the probability distribution of H_1 is also Binomial with n trials and "success probability" π, that is

$$\Pr\{H_1 = h\} = \binom{n}{h} \pi^h (1-\pi)^{n-h},\ h = 0, 1, ..., n$$

with mean and variance given by

$$E(H_1) = n\pi = n \times \frac{p_1}{p_1 + p_2}$$

and

$$Var(\pi) = n\pi(1-\pi) = n \times \frac{p_1}{p_1 + p_2} \times \frac{p_2}{p_1 + p_2}.$$

7. Triggering the 4/5ths rule

Recall that the EEOC "$4/5^{ths}$ rule" states that if one group gets less than 80% (or 4/5) of the number of jobs given the other, then there is evidence of adverse impact. Focus on the original assumption that groups 1 and 2 are equally numerous. In this case, the firm's hiring process will trigger the $4/5^{ths}$ rule if either

$$\begin{aligned} H_1 &< \frac{4}{5} \times H_2 \\ \text{or } H_2 &< \frac{4}{5} \times H_1. \end{aligned}$$

Now since there are n positions filled, we have

$$H_1 + H_2 = n.$$

Recognizing that $H_2 = n - H_1$, we can re-write the $4/5^{ths}$ rule triggers as

$$\begin{aligned} H_1 &< \frac{4}{5} \times (n - H_1) \implies H_1 < \frac{4}{9}n \\ \text{or } n - H_1 &< \frac{4}{5} \times H_1 \implies H_1 > \frac{5}{9}n. \end{aligned}$$

The probability that the hiring results would *not* trigger the $4/5^{ths}$ rule (and thus Federal agencies would conclude there is no adverse impact) is thus given by

$$\Pr\{\text{Conclude no adverse impact}\} = \Pr\{\frac{4}{9}n \leq H_1 \leq \frac{5}{9}n\}$$

where H_1 follows the Binomial distribution deduced above.

So, let's get specific. Suppose that in truth there is no discrimination ($\pi = \frac{1}{2}$), and that there are 10 positions to be filled (i.e. $n = 10$). Then the EEOC would conclude there is no discrimination with probability given by

$$\Pr\{\frac{4}{9} \times 10 \leq H_1 \leq \frac{5}{9} \times 10\}$$

$$= \Pr\{4.44 \leq H_1 \leq 5.55\}$$

$$= \Pr\{H_1 = 5\} = 0.2461.$$

Thus, with "false positive" probability $1 - 0.2461 = 0.7539$, the EEOC would conclude that there is adverse impact when in truth there is none!

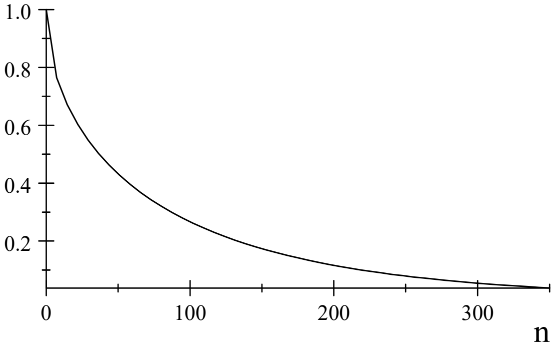
Now, let's suppose that the number of positions n gets large. Providing that $\min(n\pi, n(1-\pi)) \geq 5$, we can employ a normal distribution approximation to the Binomial as found in any introductory probability text. In particular, in the case where there is no adverse impact in truth, that is, when $\pi = 1/2$, then the mean μ and variance σ^2 of the the approximating normal are given by $\mu = n/2$ and $\sigma^2 = n/4$, and as a consequence we would find that

$$\Pr\{H_1 < \frac{4}{9}n\} \approx \Pr\{Z < \frac{\frac{4}{9}n - \frac{1}{2}n - \frac{1}{2}}{\sqrt{\frac{n}{4}}}\} \approx \Pr\{Z < -\frac{\sqrt{n}}{9}\}$$

where Z is the standard normal random variable (mean 0, variance 1), and assuming that n is big enough to ignore the continuity-correction (the 1/2 subtracted in the numerator). By the same reasoning, note that $\Pr\{H_1 > \frac{5}{9}n\} \approx \Pr\{Z > \frac{\sqrt{n}}{9}\} = \Pr\{Z < -\frac{\sqrt{n}}{9}\}$ (with the last equality due to the symmetry of the normal distribution). Thus, the probability of concluding discrimination when there is none is approximately equal to $2\Pr\{Z < -\frac{\sqrt{n}}{9}\}$.

To compare to our earlier calculation, when $n = 10$, $\Pr\{Z < -\frac{\sqrt{10}}{9}\} = \Pr\{Z < -0.35\} = 0.3632$, and doubling this gives 0.7264 which is very close to the "exact" answer given earlier.

How large does the number of jobs n have to be such that the probability of erroneously concluding discrimination when there is none is no larger than, say, 5%? Well, since $\Pr\{Z < 1.96\} = 0.025$, set $\frac{\sqrt{n}}{9} = 1.96$ which makes $2\Pr\{Z < -\frac{\sqrt{n}}{9}\} = 5\%$, and results in $n \geq 311.17$ or 312. It seems that even for reasonably large values of n, the EEOC rule can yield many false positive results! The graph below reports the (approximate) probability of triggering the $4/5^{ths}$ rule and concluding adverse impact when there is none.

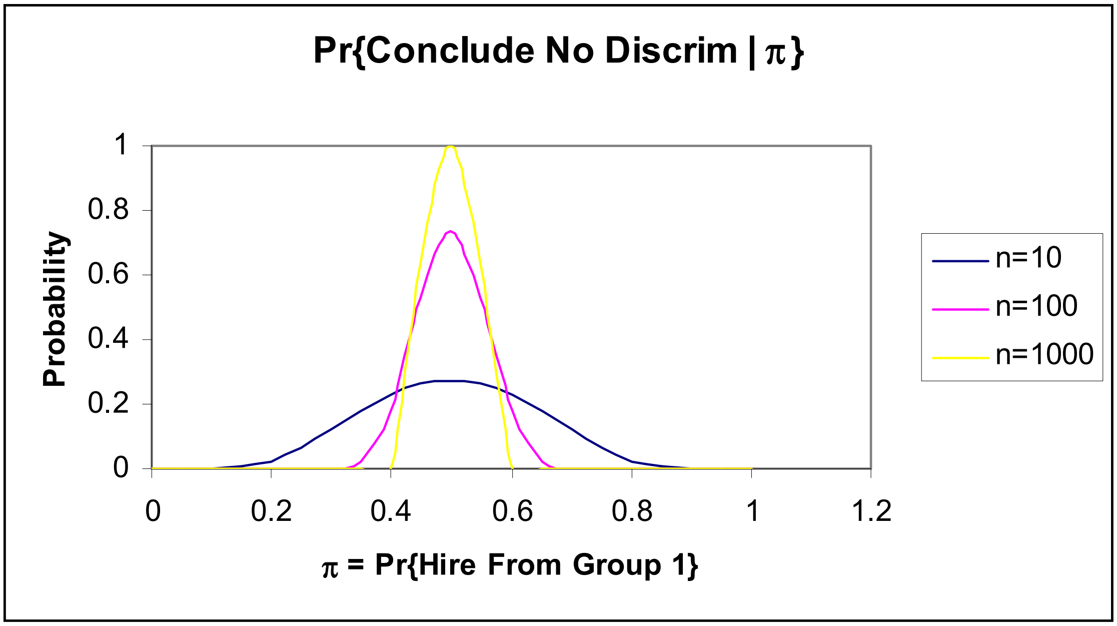


$4/5^{ths}$ Rule:Pr{Adverse Impact | None}

So far we have focused on the false positive probability of concluding adverse impact when there is none. Alternatively, one can compute the probability of concluding no adverse impact when discrimination does exist (i.e. when $\pi \neq 1/2$). The exact calculation would follow from the binomial distribution for H_1 mentioned earlier, but we will again use the normal approximation (skipping the continuity-correction term) to arrive at the approximation

$$\Pr\{\text{Conclude No Adverse Impact} \mid \pi\} = \Pr\{\frac{(\frac{4}{9} - \pi)\sqrt{n}}{\sqrt{\pi(1-\pi)}} \leq Z \leq \frac{(\frac{5}{9} - \pi)\sqrt{n}}{\sqrt{\pi(1-\pi)}}\}.$$

A graph of this appears below for $n = 10$, 100 and 1,000. What does this figure state about the probability of concluding there is no discrimination when there actually might be some (that is, when $\pi \neq 1/2$)?



2.4.3 Restricting the Size of University Events due to Covid-19

Yale's COVID-19 public health committee was asked to consider restricting the size of university events in response to the possible spread of SARS-CoV-2 infection. The quickly agreed-to goal of restricting attendance was that no new transmissions of infection should occur as a result of an event. Now, there are two ways that no new infections would be transmitted at an event in an already sterile environment: either no infected persons attend the event, or some infected persons enter but all fail to transmit infection to others. If X is the random number of infected persons entering an event with n attendees in total (each of whom has probability p of being infected), then

$$\begin{aligned}\Pr\{\text{No Transmission}\} &= \Pr\{X=0\} + \Pr\{X>0 \;\cap\; \text{all } X \text{ fail to transmit}\} \\ &\geq \Pr\{X=0\} = (1-p)^n. \qquad (2.1)\end{aligned}$$

Framing the issue this way changed the focus of discussion from infection control principles such as spacing between event attendees to the realization that *the best way to prevent transmission of infection is to ensure no infected persons enter the event.*

Producing a recommended crowd limit required two additional inputs: a comfortable lower bound for the probability that no transmission would ensue (the Chair of Yale's Epidemiology of Microbial Diseases department, Prof. Albert Ko, suggested 99%; this became known as the Ko Kriterion), and an estimate of p, the prevalence of infection among event attendees.

Of course, with no COVID-19 cases yet reported in Connecticut, nobody knew the value of p. Still what really mattered was the largest value of p against which Yale wished to defend. The public health committee believed that the underlying prevalence of infection was very small and likely no larger than 1 in 100,000 to 1 in 50,000. Given such small prevalence estimates, one could safely approximate $(1-p)^n \approx 1 - np$, and via the Ko Kriterion of 99%, the following very easily understood formulas took shape:

$$\Pr\{\text{No Transmission}\} \geq 1 - np \geq 0.99 \tag{2.2}$$

and thus

$$n \leq \frac{0.01}{p} \tag{2.3}$$

Thinking of *defending* against a prevalence of infection versus *estimating* the actual level, the public health committee felt comfortable choosing a value of p that was five- to ten-times higher than *a priori* beliefs, thus p was set to 1 in 10,000, and the university administration accepted our recommendation to restrict events to at most $n = 100$ attendees. The university community was notified of this decision in an e-mail message on March 7, 2020 that was also posted to Yale's COVID-19 communications website (https://tinyurl.com/v7af7bd).

Fast forward to March 14, the date a Yale community member was first diagnosed with COVID-19. By then 126 cases had been diagnosed in Connecticut, 40 of whom had been admitted to the hospital, but as yet no COVID-19 deaths had been recorded. The public health committee advised the university that given the rapid rise in cases over the previous two weeks, reducing event sizes further was required at a minimum, while many felt all group meetings should be abandoned. An argument to further restrict events to 20 or fewer participants went like this: early evidence suggested that SARS-CoV-2, the underlying infection responsible for COVID-19 disease, grew exponentially at a rate of 10% *per day* (Li et al, 2020). This rate implies a *quintupling* of the underlying incidence of infection over a 16 day period. Since we defended against a prevalence as large as 1 in 10,000 two short weeks ago, we should now defend against a prevalence as large as 5 in 10,000; setting p to 0.0005 in equation (2.3) led to a new maximum event size of 20.

Even with such a simple model, some misinterpreted the result as meaning that one event of 100 participants could be replaced by five events of size 20. While the likelihood that no infected person enters a group of 20 is greater than the same for a group of 100, one also must account for the factor-of-five increase in the number of events. Mathematically,

$$\Pr\{\text{No Transmission in 5 Groups of 20}\} \geq \left[(1-p)^{20}\right]^{5} = (1-p)^{100} \quad (2.4)$$

so splitting the large event into smaller events does not help if one is only considering whether any infected persons participate.

There was some worry that the rule proposed was too conservative; suppose one knew the transmission probability q from an infected to an uninfected person given exposure. In an event with n people, X of whom are infected, a more complete model for the probability that no infections are transmitted is given by

$$\begin{aligned} \Pr\{\text{No Transmission}\} &= E_X[\{(1-q)^{n-X}\}^X] \qquad (2.5)\\ &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \{(1-q)^{n-x}\}^x \end{aligned}$$

assuming that both already infected persons and transmission from them to others follow Bernoulli processes. While one estimate of $q \approx 1/200$ based on two infections among 445 home or hospital contacts of 10 patients who were infected outside of the United States was known (Burke et al, 2020), the public health committee did not feel comfortable relying on this number to increase n by using equation (2.5). In the end it did not matter, because on March 16, then President Trump issued a national guideline asking Americans to avoid gatherings of more than 10 people (https://tinyurl.com/reg85wo), and the university immediately followed suit.

2.4.4 Notes on the Two Suspect Scenario (after a paper by Brian Netter)

Two suspects have been accused of a serious crime (e.g. murder). One is guilty and one is innocent, but the prosecutor does not know which is which

– from her point of view, there is a 50/50 chance that either is the guilty party. The prosecutor also knows that juries make mistakes. Of import are the *error probabilities* given by

$$\begin{aligned}\alpha &= \Pr\{\text{Convict} \mid \text{Innocent}\}\\ \beta &= \Pr\{\text{Acquit} \mid \text{Guilty}\}\end{aligned}$$

and note also that

$$1 - \beta = \Pr\{\text{Convict} \mid \text{Guilty}\}.$$

There are three possible trial regimes under consideration:

Independent Trials

Under this regime, both suspects are tried, and all possible outcomes (dual acquits, dual convictions, the correct result of convicting the guilty while acquitting the innocent, and the worst-case result of convicting the innocent and releasing the guilty) are allowed.

Judicial Estoppel

This regime says pick one of the suspects at random, try him, stop if a conviction is reached, or if the decision is to acquit, then try the second suspect. Dual convictions are not possible here as the process stops with any conviction, but dual acquits, the correct result, and the worst-case are all possible.

Offsetting Convictions

Under this regime, both suspects are tried. If anything other than dual convictions results, the process halts after two trials and whatever outcome resulted (dual acquit, the correct result, or the worst-case) stands. However, if both suspects are found guilty, the process repeats until there are two trials that do not both result in convictions.

Performance Measures

For each of these three regimes, we wish to compute the following performance measures:

$$\begin{aligned} q &= \Pr\{\text{Convict the innocent suspect}\} \\ r &= \Pr\{\text{Correct result}\} = \Pr\{\text{Acquit innocent and convict guilty}\} \\ E(T) &= \text{expected number of trials until resolution} \\ \Pr\{T &= t\} = \text{probability distribution of } T\text{, the number of trials until resolution} \end{aligned}$$

The idea is to see which regime has the more desirable properties. Clearly we would like q to be as low as possible (don't want to convict the innocent), r to be as large as possible (want justice to be served), $E(T)$ to be small (would like the shortest process possible), and $\Pr\{T = t\}$ to favor (have higher probabilities) for smaller versus larger values of t, the number of trials.

Analyzing The Trial Regimes

Our appropriate sample space for each regime, figure out the formulas for the performance measures for each, and then ask which regime has the best properties. More formally, we seek the conditions under which some regimes are better than others.

Analyzing Independent Trials

The sample space for this experiment is easy – just like tossing a red coin and a blue coin with color-specific probabilities of gaining heads. We can nicely summarize everything in a table (we could also use a tree like I did in class, but to keep you on your toes – I mean, versatile – let's use a table) – the interior cells report the joint probabilities of the events in the header row/column, while the marginal cells report the marginal probabilities.

Joint Probabilities: Independent Trials

	Acquit | Innocent	Convict | Innocent	
Acquit | Guilty	$(1-\alpha)\times\beta$	$\alpha\times\beta$	β
Convict | Guilty	$(1-\alpha)\times(1-\beta)$	$\alpha\times(1-\beta)$	$1-\beta$
	$1-\alpha$	α	1

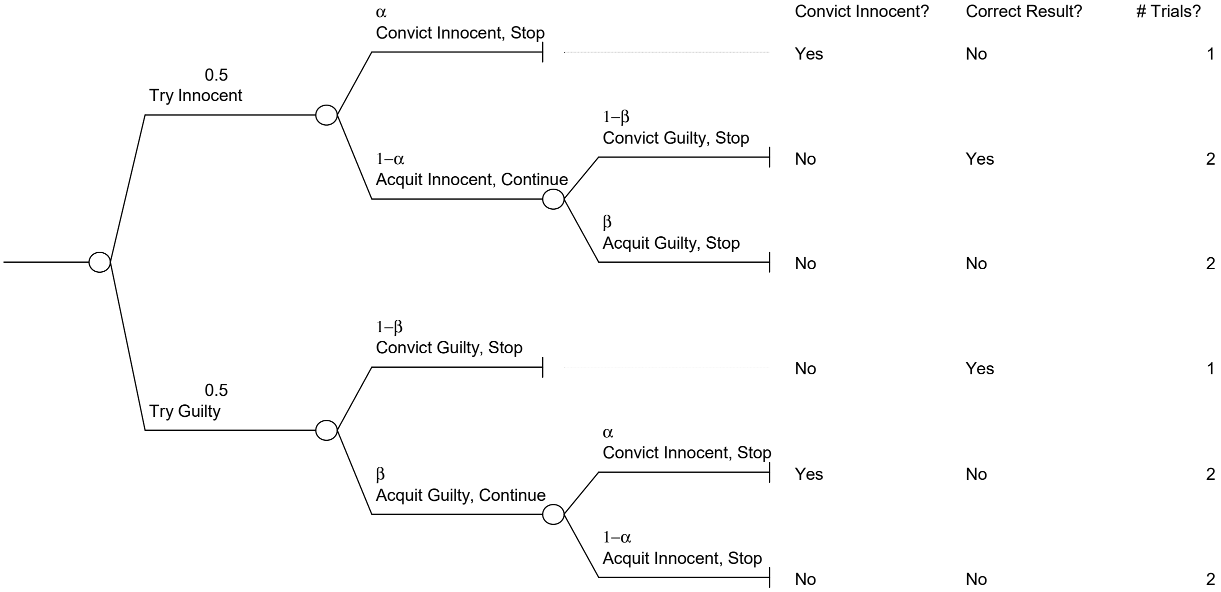
We can simply read the values of q and r right off this table. We see that

$$\begin{aligned} q &= \Pr\{\text{Convict Innocent}\} = \alpha\times\beta+\alpha\times(1-\beta) = \alpha \text{ (duhhhh.... we knew that already)} \\ r &= \Pr\{\text{Correct Result}\} = (1-\alpha)\times(1-\beta) \end{aligned}$$

We also know that there will be exactly 2 trials under this regime – one for each suspect – so we have immediately that $E(T) = 2$, and in fact $T = 2$ with probability 1 (i.e. $\Pr\{T = 2\} = 1; \Pr\{T = t\} = 0$ for all $t \neq 2$).

Analyzing Judicial Estoppel

Judicial estoppel is more complicated as the order matters (because the conviction probabilities differ by suspect). So, here we will revert to the probability tree used in class. The tree is shown below:

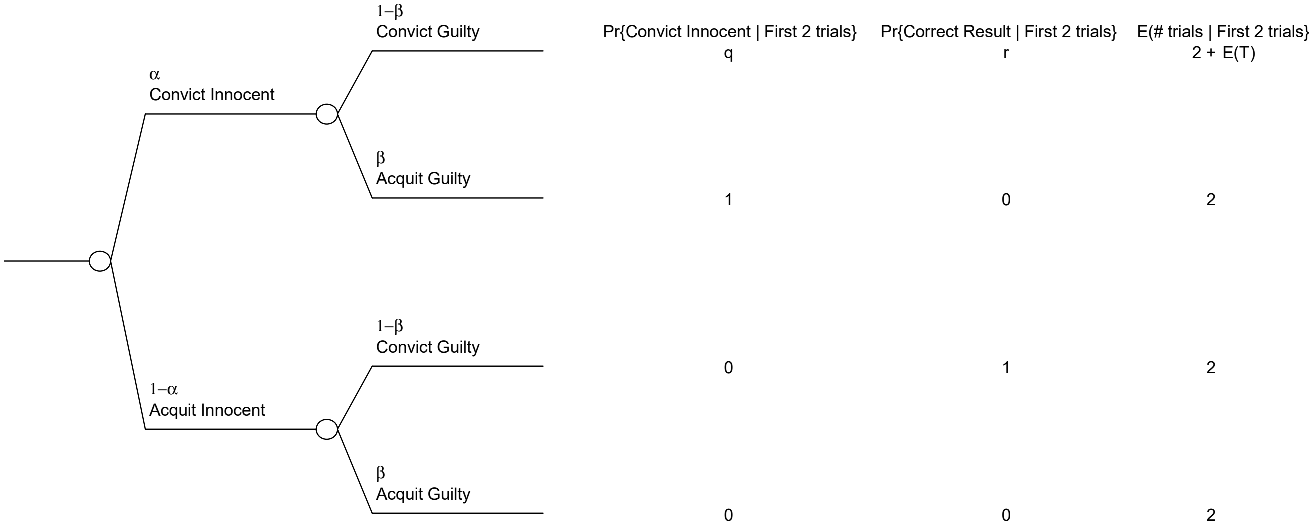


We can just read the values of the performance measures off of the tree. We have:

$$\begin{aligned}
q &= \Pr\{\text{Convict Innocent}\} = \frac{1}{2} \times \alpha + \frac{1}{2} \times \beta \times \alpha = \frac{1}{2}\alpha(1+\beta) \\
r &= \Pr\{\text{Correct Result}\} = \frac{1}{2} \times (1-\alpha) \times (1-\beta) + \frac{1}{2} \times (1-\beta) = \frac{1}{2}(1-\beta)(2-\alpha) \\
\Pr\{T &= 1\} = \frac{1}{2} \times \alpha + \frac{1}{2} \times (1-\beta) \\
\Pr\{T &= 2\} = \frac{1}{2} \times (1-\alpha) + \frac{1}{2} \times \beta \\
\Pr\{T &= t\} = 0 \text{ for } t \neq 1, 2 \\
E(T) &= 1 + \frac{1}{2} \times (1-\alpha) + \frac{1}{2} \times \beta
\end{aligned}$$

Analyzing Offsetting Convictions

Here we will work with the tree version of the independent trials scenario, but modified to take account of the possibility of offsetting convictions which necessitate repeating the entire process. Note how we define q, r, and $E(T)$ recursively via the tree – the outcomes at the end of the branches represent conditional expectations given the results of the first two trials. Here's the tree:



Let's solve for q, the probability of convicting the innocent. From the tree above we see that

$$\begin{aligned} q &= \alpha(1-\beta)\times q + \alpha\beta \times 1 + (1-\alpha)(1-\beta)\times 0 + (1-\alpha)\times\beta\times 0 \\ &= \alpha(1-\beta)q + \alpha\beta. \end{aligned}$$

Solving for q we obtain

$$q(1-\alpha(1-\beta)) = \alpha\beta$$

and hence

$$q = \frac{\alpha\beta}{1-\alpha(1-\beta)}.$$

Similarly, you can find r from the tree:

$$r = \alpha(1-\beta)\times r + (1-\alpha)(1-\beta)$$

which leads to

$$r(1-\alpha(1-\beta)) = (1-\alpha)(1-\beta)$$

and hence

$$r = \frac{(1-\alpha)(1-\beta)}{1-\alpha(1-\beta)}.$$

Finally, for $E(T)$ we get

$$\begin{aligned} E(T) &= \alpha(1-\beta)(2+E(T)) + (1-\alpha(1-\beta)) \times 2 \\ &= 2 + \alpha(1-\beta)E(T) \end{aligned}$$

which leads to

$$E(T)(1-\alpha(1-\beta)) = 2$$

and hence

$$E(T) = \frac{2}{1-\alpha(1-\beta)}.$$

The probability distribution of T for the offsetting convictions regime looks like a geometric, but each repetition of the process requires two real trials. The "success" probability of reaching a conclusion on any pair of trials is equal to $1-\alpha(1-\beta)$, which is just the probability of *not* obtaining offsetting convictions. So, the probability distribution for the total number of trials under the offsetting convictions regime is given by

$$\Pr\{T=t\} = [\alpha(1-\beta)]^{t/2-1} \times (1-\alpha(1-\beta)) \text{ for } t = 2,4,6,8,...$$

Comparing Trial Regimes

Now we wish to compare the different trial regimes. Let's start by comparing the probability of convicting the innocent suspect, q, for the various regimes. We have:

1. Independent trials: $q = \alpha$

2. Judicial estoppel: $q = \frac{1}{2}\alpha(1+\beta) \leq \frac{1}{2}\alpha(1+1) = \alpha$ (since $\beta \leq 1$). This means that with respect to q, judicial estoppel is better than independent trials.

3. Offsetting convictions: $q = \frac{\alpha\beta}{1-\alpha(1-\beta)} \leq \alpha$ (follows from algebra – note that $\alpha+\beta-\alpha\beta \leq 1$). This means that with respect to q, offsetting convictions always beats independent trials. But what about comparing

offsetting convictions with judicial estoppel. With respect to q, offsetting trials is always better if

$$\frac{\alpha\beta}{1-\alpha(1-\beta)} \leq \frac{1}{2}\alpha(1+\beta).$$

It is a matter of algebra to show that this condition is true if $1-\beta \geq 2-1/\alpha$. Note that if juries add value to the process, then it must be that $\Pr\{\text{Convict} \mid \text{Guilty}\} > \Pr\{\text{Convict} \mid \text{Innocent}\}$, that is, $1-\beta > \alpha$. If this condition is true, then offsetting convictions beats judicial estoppel. In words, for any reasonable jury, the chance of convicting the innocent is less for offsetting convictions than for judicial estoppel (which in turn does better than independent trials).

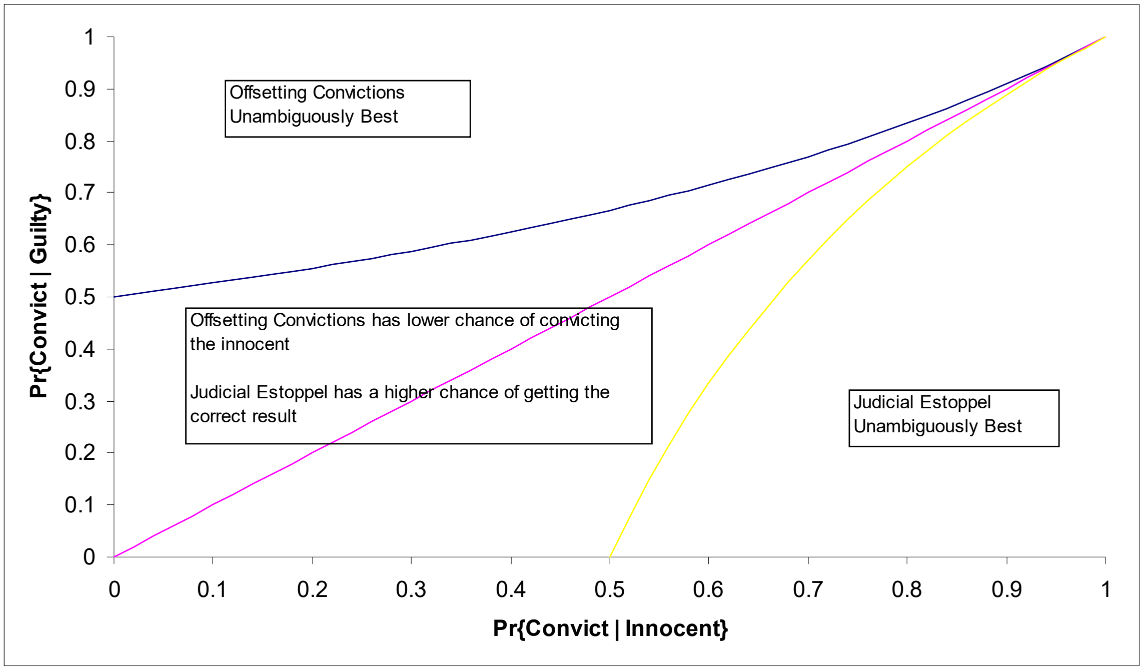
Similar analysis yields results with respect to r, the probability of getting the correct result. Recall that:

1. Independent trials: $r = (1-\alpha) \times (1-\beta)$.

2. Judicial estoppel: $r = \frac{1}{2}(1-\beta)(2-\alpha) \geq \frac{1}{2}(1-\beta)(2-2\alpha) = (1-\alpha) \times (1-\beta)$, so judicial estoppel beats independent trials in terms of getting the correct result too.

3. Offsetting convictions: $r = \frac{(1-\alpha)(1-\beta)}{1-\alpha(1-\beta)}\cdot$ Algebra shows that $\frac{(1-\alpha)(1-\beta)}{1-\alpha(1-\beta)} \geq \frac{1}{2}(1-\beta)(2-\alpha)$ if $1-\beta \geq \frac{1}{2-\alpha}\cdot$ So sometimes offsetting convictions clearly wins, and sometimes it doesn't.

These results are summarized in the graph below which plots $1-\beta$ versus α. There are combinations of the conviction probabilities for which offsetting convictions unambiguously dominates judicial estoppel, combinations for which offsetting convictions has a lower chance of convicting the innocent but also a lower chance of getting the correct result, and (unlikely) combinations where judicial estoppel is unambiguously the best – considering only q and r.



What about $E(T)$? Ah – here we see that judicial estoppel wins – it is the only approach that could end in only one trial, and requires at most two. Independent trials require exactly two trials, while offsetting convictions require *at least* two trials. If you care more about the cost of trials than the fairness of the outcomes, judicial estoppel is your best bet. But while trials are costly, I think we would all agree that the cost pales to that of sentencing the innocent in a serious crime.

2.4.5 Buying a Car in Rookesville

You are a staffer at the US Embassy in Rookesville, a small city-state with exactly one car dealership. You have been asked to purchase a new car for the embassy. All cars at the dealership look the same ("sandy"), and each costs \$8,000 (Rookesville dollars). However, 70% of these cars are "lemons". Lemons each require an additional expenditure of \$2,000 in maintenance fees within a short time period after purchase. The other 30% of the cars are "peaches." Peaches require no further expenditures, as they offer trouble free motoring for life. The dealership maintains an infinite supply of cars (low inventory costs apparently).

(a) Suppose you randomly select a car and buy it. What are your expected total costs (including purchase and maintenance fees if any)?

Well, with probability .3 you pick a peach and pay $8,000, while with probability .7 you pick a lemon, pay the $8,000 for the car but also pay $2,000 in maintenance costs. So, your expected costs from picking a car at random are

$$E(\text{cost}) = .3 \times \$8,000 + .7 \times (\$8,000 + \$2,000) = \$9,400$$

Now wasn't that easy?

(b) You have the option to hire a perfect evaluator. The evaluator, for the fee of $400 per car inspected, will declare "this car is a peach" or "this car is a lemon" (depending, of course, on the true nature of the car - the evaluator is perfect after all). If the evaluator says that a car is a lemon, you cough up another $400 and choose another car for the evaluator to consider. You continue in this fashion until the evaluator declares "this car is a peach." Would you hire the evaluator? Explain.

Well let's see. The evaluator will tell you if you have picked a lemon, so you will keep inspecting cars until the evaluator tells you that you have picked a peach. Since 30% of the cars are peaches, the number of cars you will inspect up to and including the one you purchase is the same as the number of times you toss a coin until you get heads for the first time when there is a 30% chance of getting a head! So, recognizing that the probability distribution of the number of cars inspected until purchase is geometric with success probability .3, the expected number of cars inspected just equals 1/.3 or 3.3333... Your total purchase price using the evaluator would then equal $8,000 (the price of buying a peach) plus 3.3333 ×$400 (since you pay $400 each time your evaluator checks out a car). The expected total cost of doing this is thus

$$3.3333 \times \$400 + \$8,000 = \$9,333.33.$$

Now, if you don't use the evaluator, your expected costs from (a) above equal $9,400, which is higher than $9,333.33. So, you actually would save a little (on average) by using the evaluator.

Here's another way to see this. Let t be the total costs from using the evaluator and consider the first car. Either the car is a peach (with probability .3, in which case you pay the evaluator $400 and the dealer $8,000 and you're done), or the car is a lemon (with probability .7, in which case you pay the evaluator $400 and, looking into the future, you expect to pay another, well, t dollars – since the future looks exactly the same). So you end up with the equation

$$t = .3 \times (\$8,000 + \$400) + .7 \times (\$400 + t)$$

which solves to yield

$$t \times (1 - .7) = .3 \times \$8,400 + .7 \times \$400 = \$2,800$$

or

$$t = \frac{\$2,800}{1 - .7} = \frac{\$2,800}{.3} = \$9.333.33$$

as before.

Using this same logic, suppose that the evaluator charged c dollars per evaluation instead of $400. Simply substituting c for $400 in the approach above yields the equation (after simplifying)

$$t = \frac{10}{3} \times \$c + \$8,000.$$

Comparing this to $9,400 (your expected purchase price if you select a car at random), you see that you can do better with the evaluator as long as her fee of $\$c < \$420$. In the problem I gave you, $c = 400$ so you should hire the evaluator, but the little calculation above shows that you could go up to $420 and still be better off with the evaluator than going it alone.

(c) You learn that there is a cheaper evaluator available who charges only $310 per car inspected. However, this evaluator is imperfect: 90% of all cars declared to be peaches are actually peaches, while 90% of all cars declared to be lemons are actually lemons. You will still behave in accord with the judgement of the evaluator (so if you hire this wire, you behave as in part (b)). First, having examined a randomly chosen car, what is the probability that the (imperfect) evaluator declares "This car is a lemon?" (Incidentally, how does this compare to the probability that the perfect evaluator of part (a) declares that a randomly inspected car is a lemon?)

Hmmmm - well, we know that the probability the evaluator inspects a lemon equals .7 (for you choose the car, and the cars are indistinguishable).

We also know that the probability the car is a lemon given that the evaluator says the car is a lemon equals .9. Come to think of it, we know that the probability the car is a lemon, given that evaluator says a *peach* has been inspected, equals .1 (for given that the evaluator says "peach" there is a .9 chance that the car is a peach and hence a .1 chance that the car is a lemon). So, let p be the probability that the evaluator says lemon (and $1-p$ be the probability the evaluator says peach). Consistency requires that

$$p \times .9 + (1-p) \times .1 = .7$$

(since 70% of the cars are really lemons), and thus $p = 3/4$. So, this evaluator will state that cars are lemons 75% of the time, though we know that in fact only 70% of all cars are lemons! And of course, the perfect evaluator from part (b) would declare that a randomly inspected car is a lemon with probability 0.7.

(d) Would you hire the imperfect evaluator (as compared to taking your chances alone)? Explain.

Well, again let's recognize that you keep looking at cars until the evaluator says *peach!* Since there is a 25% chance that the evaluator will say a car is a peach, we can expect to look at 4 cars. So our total expected costs will be the sum of the purchase price ($8,000), the expected evaluator fees ($4 \times \$310 = \$1,240$), and the expected maintenance fees ($.1 \times \$2,000 = \$200$, since when the evaluator tells us we have a peach, there is a 10% chance she has made a mistake and we end up with a lemon). Adding everything up we get

$$\$8,000 + \$1,240 + \$200 = \$9,440$$

which is greater than $9,400, so would be making a mistake to choose this evaluator, even though her fee is less. More generally, if this evaluator charged $\$c$, your total cost from using her would equal

$$\$8,200 + \$4c$$

and comparing to $9,400, it would only make sense to hire her if $c < \$300$.

2.4.6 Condom Failures (circa 1998)

The Food and Drug Administration tested condoms by submitting them to federal water-leakage standards in the lab. If *more* than 4 condoms out of 1000 were found defective, the batch from which the 1000 condoms were drawn is rejected. 12% of all batches tested in this manner were rejected. What then is the probability that a randomly selected condom would be found defective?

If condoms fail with probability p, and if X is the number of condoms that fail from a batch of size 1000, then X has a binomial distribution with $n = 1000$ and success probability p (here success means condom failure). Now, a batch fails if more than 4 condoms fail in the 1000 tested. Thus,

$$\Pr\{\text{Batch Fails}\} = \Pr\{X > 4\} = 1 - \Pr\{X \le 4\} = 1 - \sum_{x=0}^{4} \binom{1000}{x} p^x (1-p)^{1000-x}.$$

We also know that the probability that a batch fails is given by 12%. Thus, one needs to find a value of p such that the expression above equals 12%; this is equivalent to solving

$$\sum_{x=0}^{4} \binom{1000}{x} p^x (1-p)^{1000-x} = .88$$

for p. Doing so (for example, in Excel) yields $p = 0.00258$.

2.4.7 To Catch A Thief

That notorious crook, The Joker, is at it again, but this time the police have been tipped off. Rather than run around town looking for him, the police have stationed undercover agents at 10% of the intersections downtown. If The Joker passes through an intersection where there is an undercover agent, he will be caught. Now, The Joker magically arrives at a random intersection downtown. If he is not caught (that is, if he doesn't show up at an intersection with an undercover agent), he strolls down one of the streets. Each street has 10 shops. The Joker will target each shop he passes with probability 1/3, and once a shop is targeted, there is no hope – The Joker is very good. However, if The Joker reaches the end of a block without having targeted any shops, he randomly picks a new direction to walk in. Note that there is an intersection at the end of the block too! And, for purposes of this problem, you may assume that for all practical purposes, there are an infinite number of intersections downtown.

(a) What is the probability that The Joker is caught at the first intersection visited?

There are agents at 10% of the intersections, so there is a 10% chance The Joker is caught at the first intersection visited.

(b) Suppose The Joker is not caught at this first intersection. What is the probability that he targets at least one of the ten shops on the first street he visits?

There is a 1/3 chance of targeting any shop. There are ten shops, so the chance he targets at least one of them is just 1 minus the chance he targets none of them. The chance he targets none of the 10 shops equals $(2/3)^{10} = 0.0173$, so the chance he hits one of them equals $1 - .0173 = 0.982\,7$.

(c) What is the probability that The Joker is caught before he pilfers a shop?

Here is the easy way to do this. First, suppose The Joker is just getting started. Either he is caught right away in the first intersection (with probability 10%), or with probability $.9 \times .9827$, he is not caught at the first intersection and he pilfers a shop on the block, or with probability $.9 \times 0.0173$, he is not caught at the first intersection and he does not target a shop on the block, in which case the entire situation repeats starting with a new intersection. So, the process ends on a given block if The Joker is caught at the initial intersection (with probability .1), of if he is not caught and targets a shop on the block (with probability $.9 \times .9827$). *Given* that the process ends on a particular block, the conditional probability that *the process ends because The Joker was caught* just equals $.1/(.1 + .9 \times .9827) = 0.101\,6$

You could also use the repetition method: let q be the chance that The Joker is caught before pilfering a store. With probability .1, The Joker is caught at the first intersection. With probability $.9 \times .9827$, The Joker pilfers a store on the first block visited. With probability $.9 \times 0.0173$, the process repeats. This leads to the following equation for q:

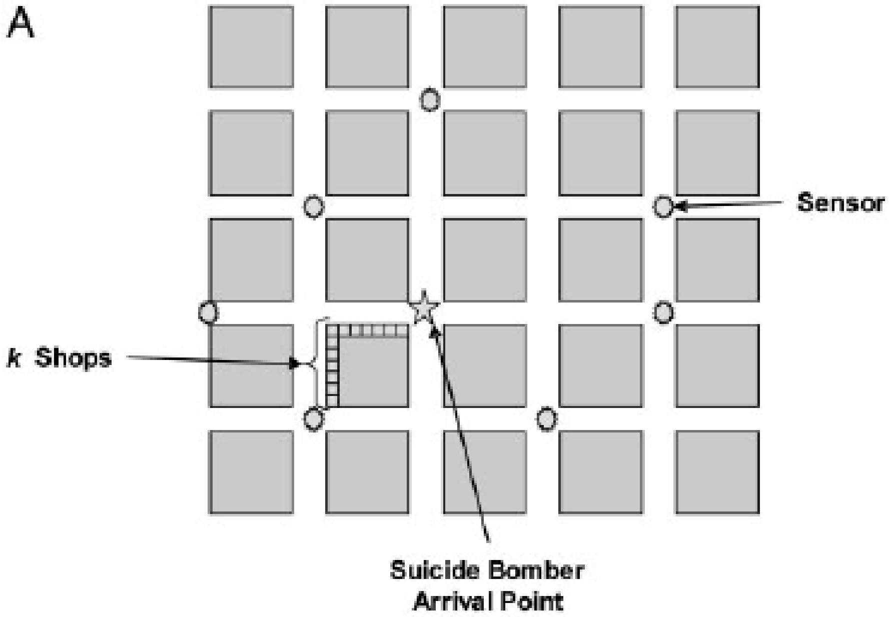
$$q = .1 \times 1 + .9 \times .9827 \times 0 + .9 \times .0173 \times q$$

which leads to the result

$$q = \frac{.1}{1 - .9 \times .0173} = 0.1016$$

as before.

Now, here is a different version of this same question: change “The Joker” to “suicide bomber” and change “undercover agent” to “suicide bomber detector” (i.e. a sensor that could detect explosives by seeing through clothes, or an undercover intelligence agent, etc.). Same problem holds – here is a picture to show you the geometry of what’s going on:



See the *PNAS* article “Operational effectiveness of suicide-bomber-detector schemes” in your coursepack for a more elaborate analysis of the suicide bomber detection problem.

2.4.8 Cyber SCADAddle!

The power grid in the United States is controlled through a large number of decentralized *S*upervisory *C*ontrol *A*nd *D*ata *A*cquisition (or SCADA) systems that we will refer to as SCADAs for short. SCADAs are extremely vulnerable to cyberterrorism attacks. Furthermore, due to the interconnected nature of the power grid, an attack on a single SCADA can have devastating consequences for the rest of the grid.

Suppose that SCADAs are aligned in series so that each SCADA (represented as a $\Diamond$) is linked to two other SCADAs, one to the "left" and one to the "right" as shown below:

————◊————◊————◊————◊————

For our purposes, the number of SCADAs is sufficiently large to be considered infinite.

Initially, all SCADAs are vulnerable to a cyber attack with a computer virus, that is, they are uninfected. If a terrorist attacks a single SCADA using a computer virus, the attack is successful with probability p. Given a successful initial attack, the virus spreads from infected SCADAs to "next door" neighboring uninfected SCADAs with probability q in mutually independent fashion . For example, suppose that the second SCADA in the figure above is attacked. It is successfully infected with probability p. If the attack infects the second SCADA, then the first SCADA becomes infected with probability q, as does the third SCADA; owing to independence, *both* the first and third SCADA become infected with probability q^2, while the probability that the first SCADA escapes infection *and* the third becomes infected equals $(1-q)q$. If the third SCADA becomes infected, it would infect the fourth with probability q, and so on *ad infinitum*.

(a) We start with a completely vulnerable system with an infinite number of uninfected SCADAs as described above. A terrorist attacks a single SCADA. What is the probability that *at least two* SCADAs are infected as a result of this attack?

As is so often the case in probability, it is easiest to work with complementary events. So, rather than figuring out the chance that at least two SCADAs are infected directly, we write

$$\begin{aligned}\Pr\{\text{At least two SCADAs infected}\} &= 1-\Pr\{\text{no SCADAs infected}\}\\ &\qquad -\Pr\{\text{exactly 1 SCADA infected}\}.\end{aligned}$$

Now, for no SCADAs to be infected, the initial terror attack must fail, and as the initial attack succeeds with probability p, we immediately have

$$\Pr\{\text{no SCADA infected}\}=1-p.$$

As for exactly one SCADA getting infected, there is only one way that this can happen: the initial attack succeeds, but the cybervirus fails to spread in either direction. Given a successful initial attack, additional infections to neighboring SCADAs take place independently of each other, and each such viral infection occurs with probability q, which means that the conditional probability that the virus *fails* to spread given a successful initial attack equals $(1-q)^2$. Unconditioning we see that

$$\Pr\{\text{exactly 1 SCADA infected}\} = p(1-q)^2.$$

Thus,

$$\begin{aligned}\Pr\{\text{At least two SCADAs infected}\} &= 1-(1-p)-p(1-q)^2 \\ &= p(1-(1-q)^2)\end{aligned}$$

which shows how you could have derived this result directly: for at least two SCADAs to get infected, there must be a successful terror attack (with probability p) and at least one of the neighboring SCADAs must be infected cybervirally (with conditional probability $1-(1-q)^2$).

(b) We start with a completely vulnerable system with an infinite number of uninfected SCADAs as described above. A terrorist attacks a single SCADA. What is the expected *total* number of SCADAs infected?

Let's see. Suppose the SCADA attacked gets infected. Then the virus can spread both to the "right" and to the "left." Define τ_L as the conditional expected total number of additional SCADAs infected moving to the "left," given that the initial SCADA attacked gets infected (and let τ_R denote the same but for cyberspread to the "right"). Let's apply the repetition method to compute τ_L. With probability $1-q$, we have no additional infections, while with probability q, we get $1+\tau_L$ additional infections in expectation. Thus, we have

$$\tau_L = (1-q)\times 0 + q \times (1+\tau_L)$$

which solves to yield

$$\tau_L = \frac{q}{1-q}.$$

This works because there are an infinite number of SCADAs, and the infection probability is constant at q. In addition, symmetry tells that infection to the “right” works exactly the same as infection to the “left” so it is also the case that

$$\tau_R = \frac{q}{1-q}.$$

Now, these are conditional expectations given that the SCADA attacked gets infected. Of course, the chance that the initial SCADA is infected equals p, and if the attack is successful, then we have an infected SCADA! Putting all this together, we obtain

$$\begin{aligned} E(\text{total number of SCADAs infected}) &= p(1+\tau_L+\tau_R) \\ &= p(1+\frac{2q}{1-q}) \\ &= p\frac{1+q}{1-q}. \end{aligned}$$

(c) Now we will slightly modify the situation to a one-sided infinite SCADA network: suppose there is a "head-of-the-line" SCADA (henceforth the HOL) which then feeds into successive SCADAs in linear fashion. The HOL only connects to a single neighbor, after which all successive SCADAs have two neighbors as before. Suppose that a terrorist has successfully infected the HOL (all other SCADAs are vulnerable), but that now, the spread of the virus takes time. Specifically, suppose that starting from time zero when the HOL is infected, in any time period t there is a probability q that the furthest "downstream" infected SCADA (as of time t) will infect its immediate uninfected neighbor in the next time period, and probability $1-q$ that the situation remains as is and repeats in the next time period.

(i) What is the probability that at the end of the first 5 time periods following infection of the HOL, exactly 3 additional SCADAs have become infected?

This is easier than it looks: since there is always a probability q of infecting a SCADA in any time period (since if the cybervirus does not transmit in a given time period, the situation stays as is and repeats in the next time period, and there is an infinite number of SCADAs), this problem is *exactly* the same as flipping a coin 5 times and asking for the chance of getting exactly 3 heads when the chance of getting a head equals q. This of course is found from the binomial distribution, and the answer is

$$\begin{aligned}\Pr\{\text{infect 3 SCADAs in 5 time periods}\} &= \binom{5}{3} q^3(1-q)^{5-3} \\ &= \frac{5!}{3!2!} q^3(1-q)^2 \\ &= 10q^3(1-q)^2.\end{aligned}$$

(ii) What is the probability that exactly 5 time periods following infection of the HOL are required to infect exactly 3 additional SCADAs?

This is a bit trickier, but still requires application of Bernoulli process reasoning. Exactly 5 time periods are required to infect exactly 3 SCADAs if the 3^{rd} SCADA is infected in the 5^{th} time period. If the 3^{rd} SCADA was infected in the 4^{th} time period, then it would not be true that exactly 5 time periods are *required* to infect 3 SCADAs, right? So, the problem reduces to finding the probability that the 3^{rd} SCADA is infected in the 5^{th} time period. For this to happen, it must be that after the 4^{th} time period, exactly 2 SCADAs have been infected, and then the 3^{rd} SCADA gets infected in the 5^{th} time period. As in part (c-i), we have

$$\begin{aligned}\Pr\{\text{infect 2 SCADAs in 4 time periods}\} &= \binom{4}{2} q^2(1-q)^{4-2} \\ &= 6q^2(1-q)^2.\end{aligned}$$

And, given that 2 SCADAs have been infected in the first 4 time periods, the chance of infecting the 3^{rd} SCADA in the 5^{th} time period is the same as the chance of infecting any SCADA in any time period, namely q. So, the solution is

$$\begin{aligned}\Pr\{3^{rd}\text{ SCADA infected in the }5^{th}\text{ time period}\} &= 6q^2(1-q)^2 \times q \\ &= 6q^3(1-q)^2.\end{aligned}$$

Note that this is smaller than your answer in part (c-i), as it should be: part (c-i) asked for the chance of infecting 3 SCADAs in 5 time periods, while part (c-ii) asked for the chance that the 3^{rd} SCADA is infected in the 5^{th} time period, which is only one way that 3 SCADAs could be infected in 5 time periods (so the event in part (c-ii) is included in the event in part (c-i)).

2.4.9 One More River To Cross...

In a given year, 100 people try to cross the border illegally. The probability that an individual attempting to cross is apprehended is 30%. An apprehended individual is returned to the original (foreign) side of the border, where (s)he tries again to cross, again facing a 30% probability of being apprehended. Individuals continue to attempt to cross the border until they successfully make it across.

(a) On average, how many attempts must a random crosser make before crossing the border?

This is a Bernoulli process with success probability 0.7, which means that the number of attempts to cross will follow a geometric distribution, just like the nation of shoplifters. If p is the probability of being apprehended, then $1 - p$ is the probability of crossing successfully, and the expected number of attempts required to cross successfully equals $1/(1 - p)$. Specializing to the case of $p = 0.3$, we have

$$E(\text{number of attempts}) = \frac{1}{(1 - 0.3)} = \frac{1}{0.7} = 1.429$$

(b) What is the chance that a random crosser will need more than 3 tries to get across?

It will take more than three tries to cross only if the random crosser is caught on each of the first three times. If p is the apprehension probability, then the chance of being caught three times in a row is just p^3, and specializing to the case of $p = 0.3$, we have

$$\Pr\{\text{number of attempts} > 3\} = 0.3^3 = 0.027$$

(c) The Border Patrol is planning to hire 10 new agents. This would increase the probability of apprehension per crossing attempt to 40%. On average, how many more arrests would result from hiring the new agents? Is this an effective strategy to reduce the flow of undocumented immigrants?

It is easy to see that

$$\begin{aligned} E(\text{number of apprehensions}) &= \text{number of crossers} \\ &\quad \times E(\text{number of attempts}) \\ &\quad \times \text{probability of apprehension per attempt} \end{aligned}$$

Since the expected number of attempts per crosser when the probability of apprehension is p equals $1/(1-p)$, and there are 100 crossers, this simplifies to

$$E(\text{number of apprehensions}) = 100 \times \frac{1}{1-p} \times p$$

so substituting for the relevant apprehension probabilities and subtracting, we get the increase in the expected number of apprehensions which is

$$E(\text{increase in arrests}) = 100 \times \frac{1}{1-0.4} \times 0.4 - 100 \times \frac{1}{1-0.3} \times 0.3 = 23.81$$

You could equivalently note that since the expected number of attempts until a successful crossing equals $1/(1-p)$ when the probability of apprehension equals p, the expected number of arrests is one less than the expected number of attempts (since the last attempt is a successful crossing), that is,

$$E(\text{number of apprehensions}) = 100 \times (\frac{1}{1-p} - 1).$$

Substituting and subtracting yields

$$\begin{aligned} E(\text{increase in arrests}) &= 100 \times \left(\frac{1}{1-0.4} - 1\right) - 100 \times \left(\frac{1}{1-0.3} - 1\right) \\ &= 100 \times \left(\frac{1}{0.6} - \frac{1}{0.7}\right) = 23.81 \end{aligned}$$

This is of course *not* a successful strategy to reduce the flow of undocumented immigrants, because while there are more arrests, all 100 crossers still make it through. All increasing the number of border agents does is increase the number of *attempts* to cross without decreasing the number of successful crossers. A more successful policy would serve to deter entrants from crossing. Of course, that agents do not deter crossers is an assumption of the model in this problem, and it's a pretty questionable assumption. More realistic alternatives were discussed in class.

2.4.10 How Many Border Crossers Are There?

Fazel-Zarandi, Feinstein and Kaplan (2018) estimated the annual number of undocumented immigrants crossing the southern border using a Bernoulli model that relied on only three pieces of data: the observed number of attempted border crossings that were apprehended (call this A, a statistic reported by the Department of Homeland Security), of these A apprehensions the number that involved the same person at least two times (call this A^+ and note that $A - A^+$ accounts for the number of different individual border crossers who were apprehended; this is also reported by the Department of Homeland Security), and the fraction of apprehended border crossers who were deterred from further border crossing attempts (call this d; this was estimated based on Mexican border guard interviews of undocumented immigrants apprehended at the border and returned to Mexico). How did they do this?

First, define c as the number of undocumented persons trying to cross the border in a given year, and q as the probability that any person attempting to cross the border is ultimately successful. Then the expected number of successful border crossers just equals $c \times q$. So, if one knows c and q, one can estimate the number of border crossers.

To model q, define p as the probability that any given border crossing attempt results in the apprehension of that border crosser. There are then three possible outcomes to any single border crossing attempt:

*the person is not apprehended and successfully crosses the border, which occurs with probability $1 - p$

*the person is apprehended and deterred from future border crossing attempts, which occurs with probability $p \times d$

*the person is apprehended but not deterred and thus attempts another border crossing in the future; this occurs with probability $p \times (1 - d)$
Note that the deterrence probability d is *conditional* upon a border crosser being apprehended.

Now ask for the conditional probability that a given individual eventually crosses the border given each of the three outcomes above. In the first case where there is no apprehension, the conditional probability of eventually crossing the border is 100%! In the second case where the individual is apprehended and deterred, the conditional probability of eventually crossing the border is equal to 0. In the third case where the person is apprehended but not deterred, the probability of ultimately succeeding in crossing the border is equal to q by definition. We conclude that

$$q = (1 - p) \times 1 + p \times d \times 0 + p \times (1 - d) \times q$$

and solving for q gives

$$q = \frac{1-p}{1-p\times(1-d)}$$

as a model for the probability that any border crosser ultimately succeeds.

Unfortunately we don't know what p is, we only know A, A^+ and d. We seem to have only replaced one unknown parameter (q) with another (p). But we're not done yet! Suppose we knew c, the number of persons trying to cross the border (we don't...yet). Then it must be that the average number of apprehensions across all individual border crossers, call this a, is defined by

$$a = \frac{A}{c}\cdot$$

Is there a way to estimate a from the Bernoulli border crossing process? Consider the same three outcomes of a single border crossing attempt we utilized earlier in modeling q, and ask for the conditional expected number of apprehensions over all time for a single individual given each of these three outcomes. In the first case where there is no apprehension, the conditional expected number of apprehensions is zero! In the second case where the individual is apprehended and deterred, there is exactly one apprehension due to deterrence. In the third case where the person is apprehended but not deterred, there is one apprehension corresponding to this first crossing attempt, plus by definition an additional a apprehensions resulting from all future border crossing attempts! Applying the repetition method then yields the equation

$$a = (1-p)\times 0 + p\times d\times 1 + p\times(1-d)\times(1+a)$$

which solves to yield

$$a = \frac{p}{1-p\times(1-d)}\cdot$$

This is a very helpful result as it pins down a model for the population of border crossers c, namely

$$c = \frac{A}{a} = \frac{A}{\frac{p}{1-p\times(1-d)}}\cdot$$

Combining with our model for the probability that any individual border crosser eventually makes it across the border (q), the expected number of successful border crossers is given by

$$cq = \frac{A}{\frac{p}{1-p\times(1-d)}} \times \frac{1-p}{1-p\times(1-d)} = A \times \frac{1-p}{p}.$$

It looks like we are getting somewhere since we do know the reported value of A. We still don't know the apprehension probability p. But we do know the number of "recidivist" apprehensions A^+, which means we also know the number of first-time apprehensions $A - A^+$. And, since there are a total of c persons trying to cross the border, the expected number of first-time apprehensions must equal $c \times p$ since by the Bernoulli model, any border crossing attempt is apprehended with probability p. This leads to an identifying equation for the apprehension probability p by solving

$$\begin{aligned} p &= \frac{A - A^+}{c} \\ &= \frac{A - A^+}{\frac{A}{\frac{p}{1-p\times(1-d)}}} \end{aligned}$$

which leads to

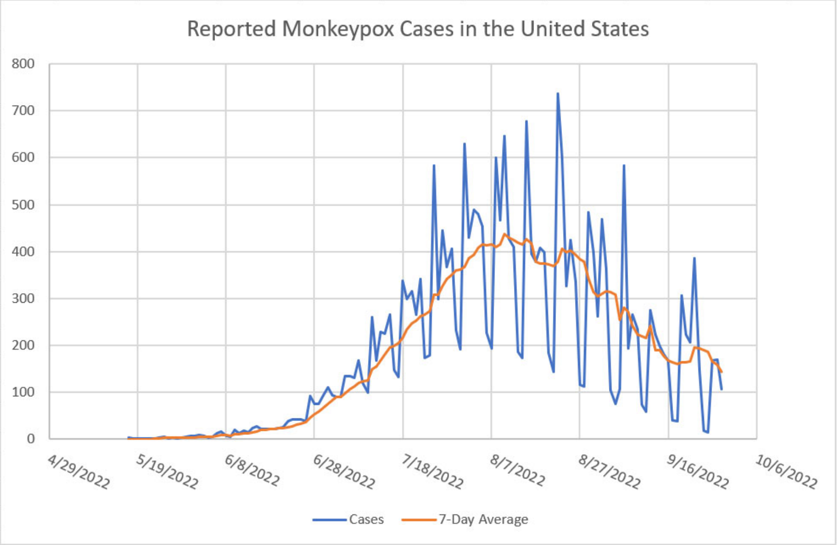
$$p = \frac{A^+/A}{1-d}.$$

This closes the loop: A, A^+ and d determine p; A, p and d determine c; and p and d determine q. The estimated number of undocumented border crossers in a year is thus given by

$$\begin{aligned} E(\text{Number of Border Crossers}) &= c \times q \\ &= A \times \frac{1-p}{p} \\ &= A \times \frac{1 - \frac{A^+/A}{1-d}}{\frac{A^+/A}{1-d}} \\ &= \frac{A}{A^+} \times \left(A - Ad - A^+\right). \end{aligned}$$

2.4.11 MPOX: Major or Minor?

Starting in the late spring of 2022, the United States experienced an outbreak of MPOX (formerly monkeypox) virus. The graph below shows the daily reported cases of monkeypox from May 17 through September 28 of 2022. The weekly cyclicity in these data is clearly a reporting artifact; the 7 day moving average is thus more suggestive of the true shape of the outbreak.



(a) The reproductive number R for monkeypox was estimated in https://www.medrxiv.org/content/10.1101/2022.07.26.22278042v1 as equal to 1.29 (with a 95% confidence interval running from 1.26 to 1.33). Assume that $R = 1.29$, and that the number of infections transmitted per newly infected man in the monkeypox outbreak, X, followed the shifted geometric distribution given by

$$\Pr\{X = j\} = q(1-q)^j \text{ for } j = 0, 1, 2, ...$$

where $q = 1/(R+1)$. According to this model, what is the probability π that the resulting pattern of infections would constitute a *minor* outbreak?

As discussed in class, the probability of a minor outbreak π is given by the solution to the equation

$$\pi = g(\pi) = \sum_{j=0}^{\infty} \Pr\{X = j\} \times \pi^j.$$

When X follows the shifted geometric, the parameters q and R are related as $q = 1/(R+1)$, and consequently $g(\pi)$ simplifies to $1/(1 + R - R\pi)$, which admits two solutions to $\pi = g(\pi)$:

$$\pi = \begin{cases} 1 & \text{if } R \leq 1 \\ \\ 1/R & \text{if } R > 1 \end{cases}$$

Since we are told that $R = 1.29$, we see that $\pi = 1/1.29 = 0.78$.

2.4.12 Time to Detect a Bioterror Attack

This problem provides a model of the time required to detect a bioterror attack from symptomatic victims. Suppose that at time 0, a covert (i.e. undetected) attack occurs that infects n people with a noncontagious agent such as anthrax. Suppose also that the *incubation time* from infection through symptoms associated with this agent can be well-approximated by a geometric distribution with mean $1/p$ days. So, if T is the incubation time, then

$$\Pr\{T = t\} = p(1-p)^{t-1} \text{ for } t = 1, 2, 3, ...$$

All persons infected in the attack progress to symptoms independently of each other in accord with the incubation distribution above. Finally, assume (optimistically) that the attack is detected when the first symptomatic patient(s) are observed.

To make the problem more concrete, suppose that $p = 1/10$ (so on average it takes 10 days to progress from infection to symptoms), and also that $n = 10$ persons are infected in the initial attack. So, for each of the 10 persons infected in the attack, the probability that a person develops symptoms after 1 day equals 1/10, the probability that it takes 2 days to develop symptoms equals $9/10 \times 1/10 = 9/100$, and in general the probability that it takes t days to develop symptoms equals $(9/10)^{t-1} \times 1/10$. Remember, the attack is with a noncontagious agent, so the 10 persons initially infected in the attack will not transmit infection to anyone else.

(a) What is the probability that an individual infected at time 0 requires *more* than t days to progress to symptoms?

I'll provide answers in general terms using p and n instead of $p = 1/10$ and $n = 10$; the specific results can then be found by just plugging in for p and n.

For an individual to require *more* than t days to progress to symptoms, the individual must fail to progress on each of the first t days after the attack, and given that such failure occurs with probability $1-p$ on each day (T follows a geometric distribution, so right away we know we are dealing with a Bernoulli process model), independently of whatever happened on prior days, the answer is simply given by $(1-p)^t = (9/10)^t$.

(b) What is the probability that *all* of the 10 persons infected in the attack require *more* that t days to progress to symptoms? What is the probability that the time required to detect the attack exceeds t days? Assume that the time required for any one person to progress to symptoms is independent of the time required for any other person to progress to symptoms.

Assuming that the time required for any one person to progress to symptoms is independent of the time required for any other person to progress to symptoms, the chance that all n people infected in the attack require more than t days to progress to symptoms equals $[(1-p)^t]^n = (1-p)^{nt} = (9/10)^{10t}$. And, since the attack will be detected when the first person(s) progress to symptoms, the probability that the time to detect the attack exceeds t days is equal to the time that everyone infected requires more than t days to progress to symptoms – which we just showed is equal to $(1-p)^{nt} = (9/10)^{10t}$.

(c) What is the probability that the *first person(s) to progress to symptoms* do so on day t after the attack? What is the probability that the attack is detected t days following the attack?

Again, note that both questions are the same question! Now, for the attack to be detected on day t, it must be that the time to detect the attack *exceeds* $t-1$ days, and given this that the attack is detected on day t. From part (b) we know that the probability that the time to detect the attack exceeds $t-1$ days is equal to $(1-p)^{n(t-1)} = (9/10)^{10(t-1)}$. Now, given that the attack has not yet been detected, the probability that the attack is detected on the next day is equal to the probability that *at least one* person infected in the attack develops symptoms on that day. But since we are dealing with a Bernoulli process, this probability simply equals $1-(1-p)^n = 1-(9/10)^{10} = 0.6513$. Why? Because if p is the probability that any one person progresses to symptoms on a given day, then $(1-p)$ is the probability such a person does not progress; $(1-p)^n$ is the probability that none of the n infected persons progress; and hence $1-(1-p)^n$ is the probability that at least one person progresses, resulting in the detection of the attack. Combining results, we have that the probability that the attack is detected on day t after the attack is given by

$$\begin{aligned}\Pr\{\text{Detect attack on day } t\} &= (1-p)^{n(t-1)}[1-(1-p)^n] \\ &= (9/10)^{10(t-1)}[1-(9/10)^{10}] \text{ for } t = 1, 2, 3, ...\end{aligned}$$

(d) What is the expected time from the attack until the first person(s) are observed to progress to symptoms; that is, what is the expected time required to detect that an attack has occurred?

Look carefully at the formula for the probability distribution of the time required to detect the attack from part (c). Define θ as the daily probability of detecting the attack, conditional on it not having yet been detected, and note that

$$\theta = 1-(1-p)^n = 1-(9/10)^{10} = 0.6513.$$

Note also that $1-\theta = (1-p)^n = (9/10)^{10} = 0.3487$, and thus

$$(1-p)^{n(t-1)} = [(1-p)^n]^{t-1} = (1-\theta)^{t-1} = (0.3487)^{t-1}.$$

In terms of θ, the probability distribution of the time to detect the attack is given by

$$\begin{aligned}\Pr\{\text{Detect attack on day } t\} &= (1-\theta)^{t-1} \times \theta \\ &= (0.3487)^{t-1} \times 0.6513 \text{ for } t = 1, 2, 3, ...\end{aligned}$$

This is also a geometric distribution, but with success probability $\theta = 0.6513$ instead of $p = 1/10$. So, the expected time to detect the attack immediately follows from the geometric as

$$E[\text{Time to detect attack}] = \frac{1}{\theta} = \frac{1}{1-(1-p)^n} = \frac{1}{0.6513} = 1.54 \text{ days.}$$

Chapter 3

Hazard Function Models

The Bernoulli process provides a simple model for duration problems where the random variable T describes the time (or number of Bernoulli trials) required until an event-of-interest occurs. The resulting Geometric distribution for T is particularly easy to work with, but it does rely on the fundamental Bernoulli assumption that the event-of-interest has a constant probability of occurring on any given trial. In this section we relax that assumption by allowing the conditional probability of a process-ending event occurring to depend upon the age (or number of elapsed trials) of the process.

3.1 Discrete Hazard Functions

Recall the Bernoulli Nation of Shoplifters. As before, offenders will continue to recidivate and commit crimes until they quit. However, instead of presuming that an offender retires from crime with probability $1 - r$ after any offense, we now assume that the offender quits with conditional probability $h(t)$ following the t^{th} offense (given that the offender has already committed $t - 1$ crimes). For reasons to be explained later, we refer to $h(t)$ as the *hazard function*. Letting T denote the number of crimes in (equivalently the duration of) a criminal career, we see that

$$\Pr\{T = 1\} = h(1)$$

since $h(1)$ by definition is the probability of quitting after the first offense. Continuing, the probability that exactly two crimes are committed in a career equals

$$\Pr\{T = 2\} = h(2) \times (1 - h(1))$$

since in order to quit after the second crime, the offender first has to commit a first crime and *not* quit (this happens with probability $1-h(1)$), and given this the offender quits after the second crime (with conditional probability $h(2)$). Generalizing, we see that the probability a criminal career consists of exactly t crimes, $\Pr\{T=t\}$, is given by

$$\Pr\{T=t\} = h(t) \times \prod_{j=1}^{t-1}(1-h(j)),\ t=1,2,3,...$$

This expression can be easily understood as

$$\Pr\{T=t\} = h(t) \times \Pr\{T \geq t\},\ t=1,2,3,...$$

where

$$\Pr\{T \geq t\} = \prod_{j=1}^{t-1}(1-h(j)),\ t=1,2,3,...$$

is the probability that the offender commits *at least* t crimes, for given that $T \geq t$, by definition the probability an offender retires following the t^{th} offense equals $h(t)$.

The discussion above has shown how to determine the probability distribution of T, the duration of the process, given the hazard function $h(t)$, $t=1,2,3,....$ However, equation () shows how to go in the other direction and determine the hazard function from the probability distribution of $T:$

$$h(t) = \frac{\Pr\{T=t\}}{\Pr\{T \geq t\}},\ t=1,2,3,...$$

Suppose that $h(t)=p$, a constant that does not depend upon t. Substituting into equation () recovers the Geometric distribution

$$\Pr\{T=t\} = p \times \prod_{j=1}^{t-1}(1-p) = p \times (1-p)^{t-1},\ t=1,2,3,...$$

Alternatively, computing $h(t)$ from the Geometric distribution yields

$$h(t) = \frac{\Pr\{T=t\}}{\Pr\{T \geq t\}} = \frac{p \times (1-p)^{t-1}}{(1-p)^{t-1}} = p,\ t=1,2,3,...$$

which shows that the Geometric distribution has a constant hazard function.

The hazard function $h(t)$ gets its name from reliability theory where the random variable T is the time until some system fails. As failures are typically due to some hazard – temperature, pressure, mechanical wear and tear for examples – the conditional failure probability became known as the hazard function. The term *failure rate* is also used to describe hazard functions, while in actuarial science, demography and epidemiology, the term *force of mortality* is used to define the conditional probability of dying given survival to a given age, which of course is also a hazard function.

Finding the expected value or variance of a random duration with an arbitrary hazard function does not admit simpler formulas than those generally used to define the expected value and variance of a random variable, that is:

$$\begin{aligned} E(T) &= \sum_{t=1}^{\infty} t \times \Pr\{T = t\} \\ &= \sum_{t=1}^{\infty} t \times h(t) \times \prod_{j=1}^{t-1}(1 - h(j)) \end{aligned}$$

and

$$\begin{aligned} Var(T) &= \sum_{t=1}^{\infty}(t - E(T))^2 \times \Pr\{T = t\} \\ &= \sum_{t=1}^{\infty}(t - E(T))^2 \times h(t) \times \prod_{j=1}^{t-1}(1 - h(j)) \end{aligned}$$

though of course one can always use the identity

$$Var(T) = E(T^2) - [E(T)]^2$$

where

$$E(T^2) = \sum_{t=1}^{\infty} t^2 \times \Pr\{T = t\} = \sum_{t=1}^{\infty} t^2 \times h(t) \times \prod_{j=1}^{t-1}(1 - h(j)).$$

Here are some examples.

3.2 Example

3.2.1 A Drinking Game

Friends at a party decide to play a drinking game, in which a contestant consumes one alcoholic beverage after another until they can drink no more. Suppose that T denotes the number of drinks a randomly selected friend will consume before stopping. Suppose further that the hazard function $h(t)$ for this game, which describes the conditional probability of stopping after consuming t drinks, is given by the formula

$$h(t) = \frac{1}{11-t}, \, t = 1, 2, 3, ..., 10.$$

What is the probability distribution of T, and what is the expected number of drinks per (randomly selected) friend playing this game?

First, note that

$$\Pr\{T = 1\} = h(1) = \frac{1}{11-1} = \frac{1}{10}$$

so there is a 10% chance that a given player quits following one drink. How about quitting after two drinks? Well,

$$\begin{aligned} \Pr\{T = 2\} &= h(2) \times (1 - h(1)) = \frac{1}{11-2} \times (1 - \frac{1}{10}) \\ &= \frac{1}{9} \times \frac{9}{10} = \frac{1}{10}(!) \end{aligned}$$

This is interesting – the chance a player quits after a second drink also equals 10%. And after a third,

$$\begin{aligned} \Pr\{T = 3\} &= h(3) \times (1 - h(1)) \times (1 - h(2)) = \frac{1}{11-3} \times (1 - \frac{1}{10}) \times (1 - \frac{1}{9}) \\ &= \frac{1}{8} \times \frac{9}{10} \times \frac{8}{9} = \frac{1}{10}(!!) \end{aligned}$$

Let's skip ahead to stopping after a tenth drink – here we see that $h(10) = 1/(11 - 10) = 1$ so the conditional probability of stopping after a tenth drink is 100%. But unconditionally, the probability a a player has 10 drinks before stopping is given by

$$\begin{aligned}\Pr\{T=10\} &= h(10)\times\prod_{j=1}^{10-1}(1-h(j)) = 1\times(1-\frac{1}{10})\times(1-\frac{1}{9})\times(1-\frac{1}{9})...\times(1-\frac{1}{2})\\ &= 1\times\frac{9}{10}\times\frac{8}{9}\times...\times\frac{1}{2}=\frac{1}{10}(!!!)\end{aligned}$$

Indeed, what we have just discovered is that $\Pr\{T=t\}=1/10$ for $t=1,2,3,...,10$ (and $\Pr\{T=t\}=0$ for all other values of t). The hazard function $1/(11-t)$ thus induces a *uniform distribution* over the duration of the drinking game. A randomly chosen friend is *equally likely* to stop after any number of drinks between 1 and 10 inclusive. This being the case, the expected number of drinks consumed for a randomly selected player equals the average of the two endpoints of the uniform distribution or $(10+1)/2=5.5$. With ten friends that's an expected 55 drinks in total; better have a good liquor supply!

3.3 Different Hazard Functions

The drinking game hazard function, $h(t)=1/(11-t)$, was an example of an increasing (really non-decreasing) hazard in that $h(t+1)\geq h(t)$ for $t=1,2,...,9$ (the process must end at time 10 since $h(10)=1$ so there is no point defining $h(t)$ for $t>10$). Increasing hazard functions typify processes which are more likely to end as they age. While a machine or mechanical/electrical part thereof is a typical example (think of the lifetime of a car or computer), one also encounters increasing hazard functions in epidemiology. For example, the duration of infectiousness (the "infectious period") or the time from infection until symptoms develop (the "incubation time") for viruses like influenza or SARS-CoV-2 (the coronavirus responsible for COVID-19) both have increasing hazard functions, as do the infectious periods and incubation times for many other viral infections.

The next example features a decreasing hazard function.

3.3.1 To Be Or Not To Be: IVF

Women who have been unable to become pregnant absent reproductive assistance might seek to enroll in an *in vitro fertilization* (or IVF) program.

Unfortunately, IVF does not work for all women. Suppose that a fraction ϕ of those women attempting to become pregnant via IVF at a clinic can conceivably conceive (that is, in principle could become pregnant via IVF), and that among that subset of women who could conceive, the probability of achieving a viable pregnancy on any IVF attempt is equal to p. For those women who cannot conceive of course, the probability of achieving a viable pregnancy on any trial equals zero. Unfortunately, it is not possible to know ahead of time whether a new IVF patient is a member of the group that could conceive or not.

Let random variable T denote the number of IVF trials required for a new IVF patient to achieve a viable pregnancy. For women who can conceive, the number of trials to success just follows the Geometric distribution, and such women account for the fraction ϕ of the population of patients, thus

$$\Pr\{T = t\} = \phi \times p \times (1 - p)^{t-1},\ t = 1, 2, 3, ...$$

The probability that a patient would spend *at least* t trials in this program requires some thought. For women who can conceive, the chance of requiring at least t trials is the same as the chance of failing to conceive $t-1$ times in a row or $(1-p)^{t-1}$. However, for women who cannot conceive, ignoring drop out due to financial or other reasons unrelated to getting pregnant for the moment, the chance of spending at least t trials equals 100% as such women never achieve pregnancy. We thus see that

$$\Pr\{T \geq t\} = \phi \times (1 - p)^{t-1} + (1 - \phi) \times 1,\ t = 1, 2, 3...$$

The conditional probability of achieving a viable pregnancy on the t^{th} trial given failure on the first $t-1$ attempts, which of course is the hazard function, is then given by

$$h(t) = \frac{\Pr\{T = t\}}{\Pr\{T \geq t\}} = \frac{\phi \times p \times (1 - p)^{t-1}}{1 - \phi + \phi \times (1 - p)^{t-1}},\ t = 1, 2, 3, ...$$

The probability of achieving pregnancy on the first trial, $h(1)$, is just given by $\phi \times p$ – first the woman must be able to conceive (this has probability ϕ), and second the first trial must succeed given that success is possible (this has conditional probability p). As the number of trials increases beyond the first, however, the hazard function declines because $(1-p)^{t-1}$ is a decreasing function of the number of trials t. Indeed, as t becomes large, the conditional probability of success approaches zero.

Why does this model imply a declining hazard function? The answer is simply that those women who are capable of achieving pregnancy do so, albeit probabilistically with success probability p on each trial. As these women become pregnant, the fraction of the *remaining* women in the program who cannot ever conceive grows, which in turn lessens the chance at each trial that a randomly chosen IVF participant can achieve a viable pregnancy. This provides an example of how population heterogeneity (in this case in the likelihood of achieving a viable pregnancy) leads to a declining hazard function.

Another consequence of this "split-population" model is that the probability of never achieving pregnancy approaches 100% over time. To see this, let $\eta(t)$ be the conditional probability that a woman who has failed on t successive attempts will never achieve a viable pregnancy. This probability can be written as

$$\begin{aligned}\eta(t) &= \frac{\Pr\{\text{Woman cannot conceive}\}}{\Pr\{\text{Woman cannot conceive}\} + \Pr\{\text{Woman can conceive but has failed } t \text{ times in a row}\}} \\ &= \frac{1-\phi}{1-\phi+\phi\times(1-p)^t},\ \ t = 1, 2, 3...\end{aligned}$$

Suppose a woman has just entered the program and has yet to attempt IVF. This corresponds to zero failures, and

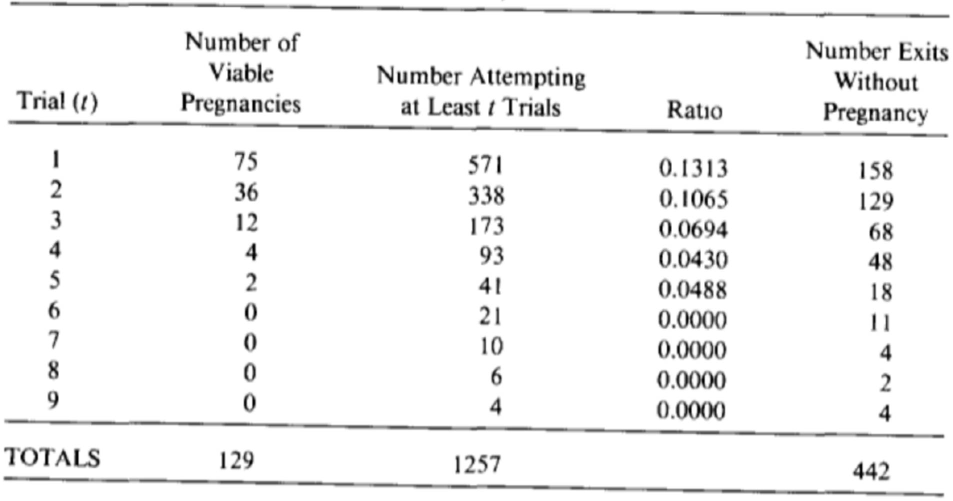
$$\eta(0) = \frac{1-\phi}{1-\phi+\phi\times(1-p)^0} = 1-\phi,$$

which simply says that the probability a new IVF program participant will never get pregnant is equal the fraction of women the program population who can never conceive. However, as the number of trials t grows, $(1-p)^t$ approaches zero and $\eta(t)$ approaches 100%. This is another reflection of the fact that women who conceive leave the program, leaving those who can never conceive to form a greater and greater fraction of the women who remain.

The models above are quite simple, yet they also have an empirical basis. Consider the data shown in the table below which reflect the experience of 571 women who attempted IVF at Yale between 1983-87.

TABLE 1

IVF-ET Data from Yale



Trial (t)	Number of Viable Pregnancies	Number Attempting at Least t Trials	Ratio	Number Exits Without Pregnancy
1	75	571	0.1313	158
2	36	338	0.1065	129
3	12	173	0.0694	68
4	4	93	0.0430	48
5	2	41	0.0488	18
6	0	21	0.0000	11
7	0	10	0.0000	4
8	0	6	0.0000	2
9	0	4	0.0000	4
TOTALS	129	1257		442

The fraction of women who achieved a viable pregnancy on a given trial out of the number of women who attempted to conceive on that trial falls from just over 13% on the first trial, to about 11% on the second trial, and continues to decline to zero by the sixth trial. While the overall likelihood of success on a randomly chosen trial from all IVF attempts equals $129/1257 = 0.102\,6$, these data demonstrate clearly that the conception probability declines as the number of trials increases, which would make the pure Bernoulli model of constant success probability per trial quite inaccurate.

However, the “split-population” model derived above provides an excellent fit to these data as seen from the excellent χ^2 goodness-of-fit statistics in the table below:

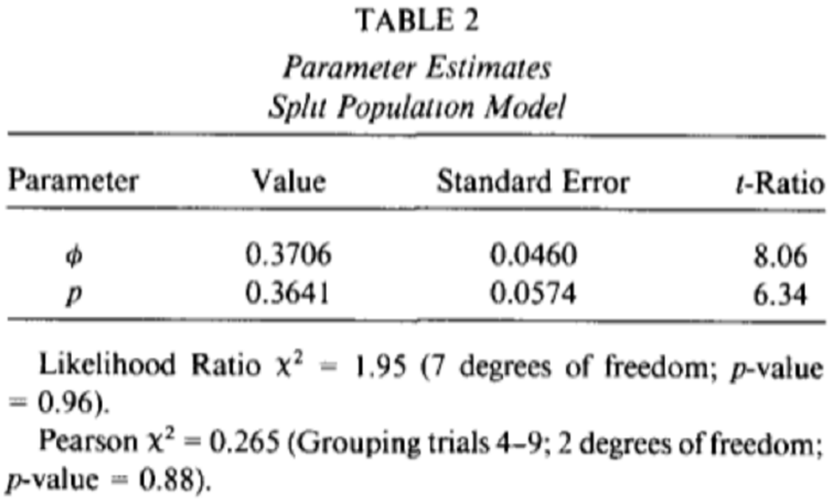


TABLE 2

Parameter Estimates
Split Population Model

Parameter	Value	Standard Error	*t*-Ratio
ϕ	0.3706	0.0460	8.06
p	0.3641	0.0574	6.34

Likelihood Ratio χ^2 = 1.95 (7 degrees of freedom; *p*-value = 0.96).

Pearson χ^2 = 0.265 (Grouping trials 4–9; 2 degrees of freedom; *p*-value = 0.88).

The estimated parameters suggest that a little over one third of all women entering this Yale IVF program could conceivably conceive ($\phi = 0.3706$), while among those women who are able to conceive, the chance of achieving a viable pregnancy is also just over one third ($p = 0.3641$). Indeed, the probability that a woman who failed to conceive after four attempts will *never* be able to conceive via IVF is equal to

$$\eta(4) = \frac{1 - 0.3706}{1 - 0.3706 + 0.3706 \times (1 - 0.3641)^4} = 0.9122.$$

Having failed four times in a row thus signals a greater than 90% chance of never being able to conceive via IVF. Indeed, a total of 41 women made at least five attempts each for a total of 82 attempts beyond four failures in a row. Only two of these attempts succeeded. One can only wonder what advice would have been given to these 41 women if it was known that the chance of ever conceiving following four consecutive failed attempts was so low.

3.3.2 Looking For A Job

You are looking for a job, and have decided to target your search to the policy modeling market. This is a tight market of course; others are looking too. You estimate that the chance you will get hired on your first interview, without regard to where that interview takes place, is equal to 1/4. However, since all those hiring policy modelers talk to each other all the time (who else would talk to them?), if you are not instantly hired by the first place you look, everyone else hiring will immediately know this, in turn lowering your chance of getting hired elsewhere. Indeed, suppose that you have interviewed and been turned down $t-1$ times. The conditional probability that you will be hired immediately following the t^{th} interview given failure on the first $t-1$ attempts is given by $1/(t+1)^2$. Note that, consistent with this formula, the chance that you will be hired on your first interview equals $1/(1+1)^2 = 1/4$ as stated earlier.

Now, you figure that you'll net \$98,000 if you get a policy modeling job. However, the cost for you to prepare for each interview is \$2,000.

(a) What is the *unconditional* probability that you will be hired immediately following t interviews for $t = 1, 2, 3, 4$? (You have to give me 4 numbers here!)

The probability of being hired after interview t is equal to the probability of being rejected on the first $t-1$ interviews times the conditional probability of being hired after interview t given rejection on the first $t-1$. Let $h(t)$ be the conditional probability of getting hired after interview t given failure until then. From the problem statement, we have $h(t) = 1/(t+1)^2$. Let T be the number of the interviews until hiring (if it occurs!). Then

$$\begin{aligned}\Pr\{T = t\} &= h(t)\prod_{j=1}^{t-1}(1 - h(j)) \\ &= \frac{1}{(t+1)^2}\prod_{j=1}^{t-1}(1 - \frac{1}{(j+1)^2}).\end{aligned}$$

Rolling through for $t = 1, 2, 3, 4$ we see that the unconditional probability of getting hired equals 0.25, 0.0833, 0.0417, and 0.025 respectively.

(b) What is the *unconditional* probability that the number of interviews required to get hired *exceeds* t (that is, is *more than* t) for $t = 1, 2, 3, 4$? (4 numbers again please!)

For the number of interviews to exceed t, it must be that all t interviews failed! We thus have

$$\begin{aligned}\Pr\{T > t\} &= \prod_{j=1}^{t}(1 - h(j)) \\ &= \prod_{j=1}^{t}(1 - \frac{1}{(j+1)^2}).\end{aligned}$$

Rolling through for $t = 1, 2, 3, 4$ we get 0.75, 0.6667, 0.625, and 0.6 respectively.

For the curious among you, you might note that even if you were to interview forever, you would not be guaranteed to land a job! In fact, the limit of the probability that the number of interviews required for a job exceeds t as t becomes very large is given by

$$\lim_{t\to\infty}\prod_{j=1}^{t}(1 - \frac{1}{(j+1)^2}) = \frac{1}{2}.$$

So even if you interviewed forever, there is only a 50% chance of landing a job in policy modeling! (This is an example of what is called a *defective distribution*; the sum of the unconditional probabilities of landing a job after t interviews over all t does not equal 1; it only equals 0.5!)

(c) What is the largest number of interviews you should attempt in the policy modeling job market?

Well, the results above should convince you that it doesn't make sense to stick around too long in any event. At the margin, you stand to gain \$98,000 if you get hired on the next interview, but it costs you \$2,000 at the margin to take another interview, *whether you are hired or not!* Since $h(t)$ is the marginal probability of success, it makes sense to keep interviewing as long as

$$\$98,000 \times h(t) \geq \$2,000$$

which is equivalent to

$$h(t) \geq \frac{1}{49}$$

or

$$\frac{1}{(t+1)^2} \geq \frac{1}{49}$$

or

$$(t+1)^2 \leq 49$$

which means that $t+1 \leq 7$, so $t \leq 6$. This means that you should attempt no more than 6 interviews!

3.3.3 Another Undocumented Immigration Problem

Imagine a scenario in which an average of 100 new undocumented immigrants arrive to the United States each year from the fictional country of Eyfo (so we are only considering a small portion of the total inflow of undocumented immigrants). 80% of these new arrivals are "movers" (short-term residents), while the remaining 20% of new arrivals are "stayers" (longer-term settlers). Suppose further that at the end of each year in the United States, a mover returns to Eyfo with probability 25%, while a stayer returns to Eyfo after each year with probability 2%.

(a) What is the average length of time spent in the United States per visit by a mover? What is the average for a stayer?

We are told that once in the United States, a mover departs with probability 0.25 each year, while a stayer departs with probability 2%. Both of these situations imply that the duration of time spent in the United States follows a geometric distribution, just like in the Nation of Shoplifters from the very first day of class! If p is the probability of returning to Eyfo each year, then the average time spent in the United States is just equal to $1/p$. Consequently, the average length of time spent in the US for a newly entering mover equals $1/0.25 = 4$ years, while for stayers the average duration of stay is given by $1/0.02 = 50$ years.

(b) A new undocumented immigrant from Eyfo has just arrived in the United States. What is the probability that this immigrant returns to Eyfo after one year?

Well, we know that if the newly arriving immigrant is a mover, there is 25% chance (s)he will return to Eyfo at the end of the year, while if the arrival is a stayer, there is only a 2% chance of leaving after a year. As 80% of new arrivals are movers and 20% are stayers, the probability that a newly arriving immigrant returns to Eyfo after one year is given by

$$0.8 \times 0.25 + 0.2 \times 0.02 = 0.204$$

or just over 20%.

(c) Now suppose that a new undocumented immigrant from Eyfo arrives to the United States at time 0. What is the probability that this new immigrant returns to Eyfo after spending exactly t years in the United States?

It seems that we answered this question in part (b) for the special case of $t = 1$. But what about for an arbitrary duration of time t? Again, the key is recognizing that, conditional on knowing whether the new immigrant is a mover or a stayer, the time spent in the United States follows a geometric distribution. For movers, the probability that an immigrant leaves after exactly t years is given by $0.25 \times (1-0.25)^{t-1} = 0.25 \times 0.75^{t-1}$. For stayers, the probability of leaving after exactly t years is given by $0.02 \times (1 - 0.02)^{t-1} = 0.02 \times .98^{t-1}$. Let T denote the time spent in the US by a newly arriving immigrant from Eyfo. Again recognizing that 80% of new arrivals from Eyfo are movers while the remaining 20% are stayers, we conclude that the probability a newly arriving immigrant returns to Eyfo after spending exactly t years in the United States is given by

$$\Pr\{T = t\} = 0.8 \times 0.25 \times 0.75^{t-1} + 0.2 \times 0.02 \times .98^{t-1}.$$

Note that if $t = 1$, we just get $0.8 \times 0.25 + 0.2 \times 0.02 = 0.204$ as in part (b).

(d) What is the conditional probability that a new undocumented immigrant from Eyfo who arrived to the US at time 0 returns to Eyfo after spending exactly t years in the United States, *given* that this immigrant spends *at least* t years in the United States?

Hmmmm – the probability we are being asked to find is

$$h(t) = \Pr\{T = t|T \geq t\} = \frac{\Pr\{T = t\}}{\Pr\{T \geq t\}}.$$

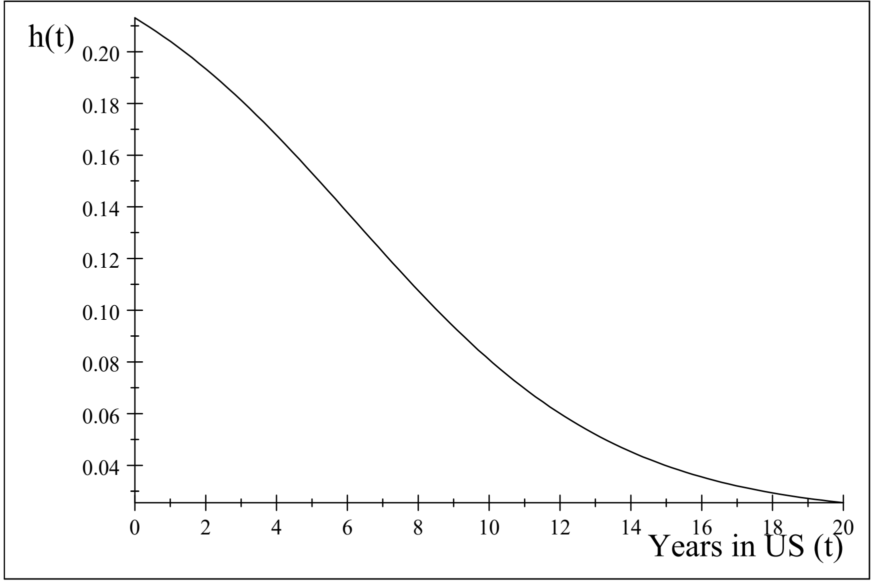
In other words, we are being asked to find the hazard function (which is equivalent to the *emigration rate*) for newly arriving undocumented immigrants from Eyfo. We already have the numerator for this hazard function from part (c) above. For the denominator, note again that from the geometric distribution, the likelihood that a newly arriving immigrant stays at least t years in the United States is the same as the probability that this immigrant does not leave after each of the $t - 1$ years after arrival. Again, if p is the probability of leaving at the end of any year in a Bernoulli process, then the probability of remaining in the US for at least t years equals $(1-p)^{t-1}$. We have two such processes happening, one for movers and one for stayers. So, taking into account the 80/20 mover/stayer split among new arrivals, we conclude that

$$\Pr\{T \geq t\} = 0.8 \times 0.75^{t-1} + 0.2 \times 0.98^{t-1},$$

which in turn implies that the hazard function we seek is given by

$$h(t) = \frac{0.8 \times 0.25 \times 0.75^{t-1} + 0.2 \times 0.02 \times .98^{t-1}}{0.8 \times 0.75^{t-1} + 0.2 \times 0.98^{t-1}}$$

A plot of this hazard function is shown below; note that the hazard declines, and asymptotes at 0.02, which is the migration rate for stayers.



(e) Suppose that this process has continued long enough to reach an equilibrium (steady state). Recall that 80% of new arrivals from Eyfo are movers. In equilibrium, what is the total number of undocumented immigrants in the United States who originated from Eyfo? And, of all of the undocumented immigrants from Eyfo residing in the United States, what fraction are movers?

A gift I tell you, a gift! From the problem statement, we know that an average of 80 movers and 20 stayers arrive each year. From our answers to part (a), we know that the average duration of stay in the US equals 4 years for movers and 50 years for stayers. Thus, the steady state total number of undocumented Eyfo immigrants in the US is given by

$$80 \times 4 + 20 \times 50 = 1,320.$$

Of these 1,320 immigrants, we expect that $80 \times 4 = 320$ are movers, thus the fraction of all undocumented Eyfo immigrants in the US that are movers is given by $320/1,320 = 0.242$ or just under a quarter; this in spite of the fact that 80% of newly arriving immigrants are movers! Also, note that the average duration of stay in the United States follows from $L = \lambda W$ setting $L = 1,320$ and $\lambda = 100$, which means that $W = 13.2$ years. Of course, we could also have obtained this result directly from part (a) by noting that the average stay for a newly arriving immigrant, taking into account the 80/20 split between movers and stayers, is given by $0.8 \times 4 + 0.2 \times 50 = 13.2$ years as claimed.

(f) Consider the first 20 years of the immigration from Eyfo. At the end of this 20 year period, researchers conduct a survey *in Eyfo* of undocumented immigrants *who have returned to Eyfo.* Note that only people who both left Eyfo for the United States and returned within these 20 years can possibly be included in the survey. Assuming that the researchers are able to obtain a random sample from the population of returnees to Eyfo over these first 20 years, what fraction of the survey respondents are movers? (BIG HINT: after the first year of immigration, only those who had arrived in the United States from Eyfo at the start of the first year could return to Eyfo at the end of the first year of immigration; after the second year of immigration, only those who had arrived in the United States at the start of the first year from Eyfo and spent two years in the US, plus those who arrived in the US at the start of the second year from Eyfo and spent one year in the US, could return to Eyfo at the end of the second year of immigration; ...)

This is a bit trickier, but the hint helps. Let's first think about the movers. There were 20 waves of immigration, each on average seeing 80 movers arrive in the US each year. Consider those in the most *recent* wave. 25% of those will return to Eyfo at the end of the year, and thus be back in Eyfo in time for the survey. Now consider those in the second most recent wave. Of those, $0.25 + 0.75 \times .25 = 0.437\,5$ would have been expected to return. Note that if T is the duration of stay in the US for movers, this is exactly the same as $\Pr\{T \leq 2\} = 1 - 0.75^2$. In general, the probability that a mover in the t^{th} *most recent* wave of immigration to the US returned to

Eyfo in time for the survey equals $\Pr\{T \leq t\} = 1 - 0.75^t$. Thus, the expected total number of movers who went the US and returned to Eyfo within the first 20 years of immigration is equal to

$$80 \times \sum_{t=1}^{20} \Pr\{T \leq t\} = 80 \times \sum_{t=1}^{20} (1 - 0.75^t) = 1,360.8.$$

We can do the same thing for stayers, and discover that the expected total number of stayers who went to the US and returned to Eyfo within the first 20 years of immigration is equal to

$$20 \times \sum_{t=1}^{20} (1 - 0.98^t) = 74.256.$$

Consequently, fraction of those in the survey who are movers is given by

$$\frac{1,360.8}{1,360.8 + 74.256} = 0.9483.$$

So, while movers make up 80% of the immigrants departing Eyfo, they would comprise almost 95% of the sample of migrants who returned to Eyfo within the first 20 years of immigration.

(g) Once again consider the survey described in (f) above. What should the average duration of stay in the United States equal *over all respondents to the survey*?

This is even trickier. Consider a mover who went from Eyfo to the US in the t^{th} most recent wave of immigration. Such a mover would have stayed exactly $j \leq t$ years with probability $\Pr\{T = j\} = 0.25 \times 0.75^{j-1}$, and would report having spent j years in the US in the survey (assuming truthful reporting). This means that the total person years spent in the US reported by movers in the survey in Eyfo accounting for all 20 years of migration is given by

$$80 \times \sum_{t=1}^{20} \sum_{j=1}^{t} j \times .25 \times .75^{j-1} = 4,501.3 \text{ years}$$

Similarly, the total person years spent in the US reported by *stayers* in the survey in Eyfo accounting for all 20 years of migration is given by

$$20 \times \sum_{t=1}^{20} \sum_{j=1}^{t} j \times .02 \times .98^{j-1} = 510.7 \text{ years.}$$

The average time spent in the US as computed from the survey would thus equal the total person years spent in the US divided by the total number of returning immigrants from Eyfo, and we know the latter from part (f) above. We therefore conclude that

$$\text{Average Time Spent in US per Immigrant Reported in Survey } = \frac{4,501.3 + 510.7}{1,360.8 + 74.256} = 3.49 \text{ years.}$$

What has happened? From the basic problem statement, we know that in truth, the average time spent in the US per immigrant from Eyfo equals 13.2 years (see part (e) above). But, by studying only those who had returned, even over a 20 year time span, the average time spent in the US *as computed from survey respondents* equals just under 3.5 years!

3.3.4 Covid Vaccine Effectiveness

In class we reviewed the data from the Pfizer Covid vaccine trial, and showed how to compute the hazard function for infection for both the placebo and vaccinated groups. We saw that as widely reported in the press, there was about a 95% reduction in the probability that a person who received both Pfizer doses was infected symptomatically compared to placebo.

(a) Let T denote the time from vaccination until infection for someone in the placebo group, and let

$$F(t) = \Pr\{T \leq t\}$$

be the cumulative probability distribution of the time to infection. Now define $F_V(t)$ as the probability that someone in the vaccinated group gets infected by time t. Suppose that with probability p, the vaccine prevents a person vaccinated at time 0 from ever becoming infected with SARS-CoV-2, but with probability $1 - p$, a vaccinated person has the same likelihood of getting infected by time t after vaccination as a person who received the placebo shot at time 0. Under this assumption, produce an equation for $F_V(t)$ in terms of $F(t)$.

With probability p, a vaccinated person is protected and will never be infected by time t, while with probability $1 - p$, a vaccinated person would be infected with the same chance as someone in the placebo group, namely $F(t)$. Consequently,

$$F_V(t) = p \times 0 + (1 - p) \times F(t) = (1 - p)F(t).$$

(b) Now think about the clinical trial conducted by Pfizer. From class you know how to use the data to estimate the hazard function for infection in both groups, and using those hazard functions, you know how to compute $F(t)$ and $F_V(t)$ directly (that is, without relying on the model for $F_V(t)$ in part (a)). Let "time 0" correspond to "week 5" in the Pfizer trial, which is the 1st week at risk one week following the second vaccine dose. Using the hazard functions for infection for both groups that we computed in class (and are contained in the Excel file Pfizer_Covid_Vaccine_Data), compute $F(10)$ and $F_V(10)$, that is, the probability that someone uninfected entering week 5 (time 0) has gotten infected by the end of week 15 (time 10) for the placebo and vaccinated groups respectively. What is the percentage reduction in the probability that a vaccinated person became infected compared to an unvaccinated person, that is, what is the numerical value of $(F(10) - F_V(10))/F(10)$?

Let $h(t)$ be the time t hazard function estimate for the placebo group. Remember that week 5 is the same as time 0. We know that the probability an unvaccinated person would not be infected in any of weeks 5 through 15, which corresponds to time 0 through time 10, is given by $\prod_{t=0}^{10}(1 - h(t))$, and thus the probability that a member of the placebo group would have been infected at some point during trial, which is the same as $F(10)$, is given by

$$F(10) = 1 - \prod_{t=0}^{10}(1 - h(t)).$$

From the placebo hazard function estimates in the spreadsheet, we obtain an estimate of $F(10) = 0.0219$. The same analysis applied to the vaccine group leads to an estimate of $F_V(10) = 0.0013$.The percentage reduction in the probability of getting infected when comparing the vaccine to the placebo group is then

$$\frac{F(10) - F_V(10)}{F(10)} = \frac{0.0219 - 0.0013}{0.0219} = 0.94$$

or 94%, pretty much what was reported in the press.

(c) Returning to the model of part (a), what is the percentage reduction in infection for vaccinated persons compared to unvaccinated persons at *any* time t, that is, what is $(F(t)-F_v(t))/F(t)$ for *any* value of t? With this insight and the numerical result of part (b), how do you feel about the performance of the Pfizer vaccine?

The model of part (a) suggests that since $F_V(t) = (1-p)F(t)$, the percentage reduction in the chance of being infected from time 0 to time t for any value of t is given by

$$\frac{F(t) - F_V(t)}{F(t)} = \frac{F(t) - (1-p)F(t)}{F(t)} = p.$$

What this says is that *if* the vaccine is perfectly protective with probability p while with probability $1-p$ the likelihood of a vaccinated person becoming infected follows the experience in the placebo group, *then* the percentage reduction in the probability of being infected within any fixed time interval from 0 to t is exactly equal to the vaccine success probability p. Wow! We should hope this model is true, because if so, our result from part (b) suggests that if you get vaccinated, with 94% probability you are forever protected from being infected with SARS-CoV-2!! Of course, this clinical trial took place before the Delta variant emerged, and also before it became clear that vaccine effectiveness wanes over time, so unfortunately what looked promising early on did not come to pass.

(d) Consider the following "split population" model: for those who are vaccinated, let T_V be the time until infection (again starting at time 0, which is week 5 in the Pfizer trial). The probability a vaccinated person gets infected during week t is presumed to follow

$$\Pr\{T_V = t\} = (1-p) \times \Pr\{T = t\}$$

where $\Pr\{T = t\}$ is the estimated probability that an *unvaccinated* person is infected during week t as calculated from the hazard function for the placebo group in the Pfizer_Covid_Vaccine_Data file. The *survivor function* $\Pr\{T_V \geq t\}$ for vaccinated individuals in this model is given by

$$\Pr\{T_V \geq t\} = p + (1-p) \times \Pr\{T \geq t\}$$

since $T_V \geq t$ for sure if the vaccine works (with probability p), or if the vaccine fails but a placebo subject takes at least t weeks to get infected (with probability $(1-p) \times \Pr\{T \geq t\}$). Note the similarity to the split population model in the IVF example discussed in class. The hazard function for infection for those who receive the vaccine is then given by

$$\begin{aligned} h_V(t) &= \frac{\Pr\{T_V = t\}}{\Pr\{T_V \geq t\}} \\ &= \frac{(1-p) \times \Pr\{T = t\}}{p + (1-p) \times \Pr\{T \geq t\}} \end{aligned}$$

where $\Pr\{T = t\}$ and $\Pr\{T \geq t\}$ are computed based on the hazard functions estimated for the *placebo* group (and again those hazard functions have been provided for you in Pfizer_Covid_Vaccine_Data). This makes the hazard for infection for *vaccinated* persons, $h_V(t)$, a function of just one variable – the vaccine success probability p – given the results for the placebo group. Suppose that $p = 0.95$, that is, the vaccine protects completely against infection 95% of the time. On one graph, plot the infection hazard $h_V(t)$ from the model above, and the direct empirically estimated hazard for infection *among those vaccinated* that we derived in class (and, once again, is available to you already in Pfizer_Covid_Vaccine_Data). What do the results suggest?

The first thing we need to do is figure out the probability distribution and survivor function for random variable T that describes the placebo group. This is easy, since we already have the hazard functions at our disposal. As argued in class, the survivor function $\Pr\{T \geq t\}$ is given by

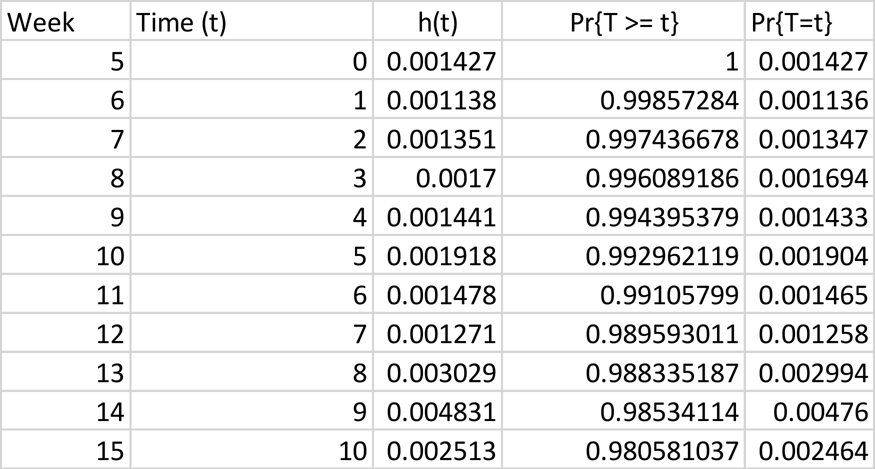
$$\Pr\{T \geq t\} = \prod_{j=0}^{t-1}(1 - h(j))$$

while the probability distribution $\Pr\{T = t\}$ is given by

$$\begin{aligned} \Pr\{T = t\} &= h(t)\prod_{j=0}^{t-1}(1 - h(j)) \\ &= h(t)\Pr\{T \geq t\}. \end{aligned}$$

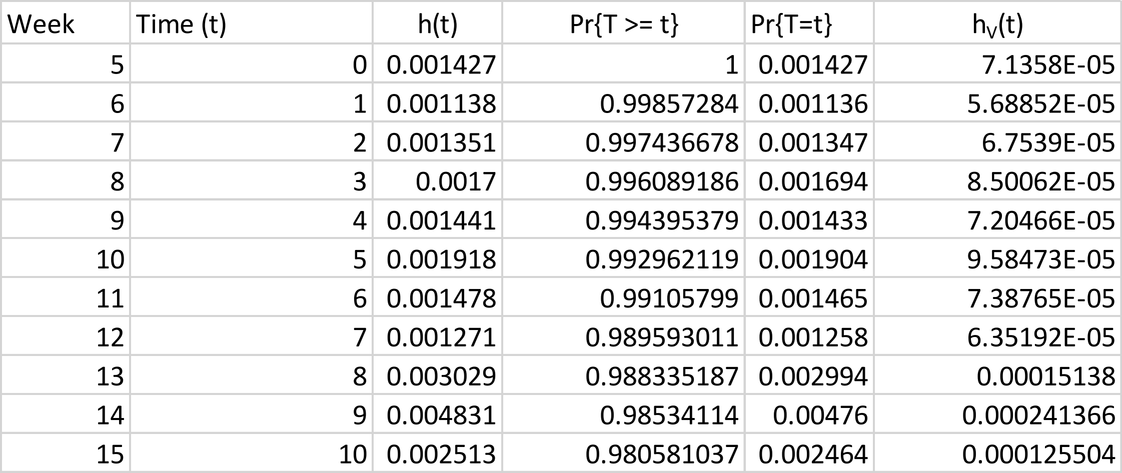
Note that these products run from $j = 0$ instead of $j = 1$ due to the timing convention in the data – one could become infected during week 5 = time 0. Note also that $\Pr\{T \geq 0\} = 1$.

Armed with these definitions, it is a simple matter to compute the corresponding numerical values using the hazard functions in the problem spreadsheet. Here are the results:



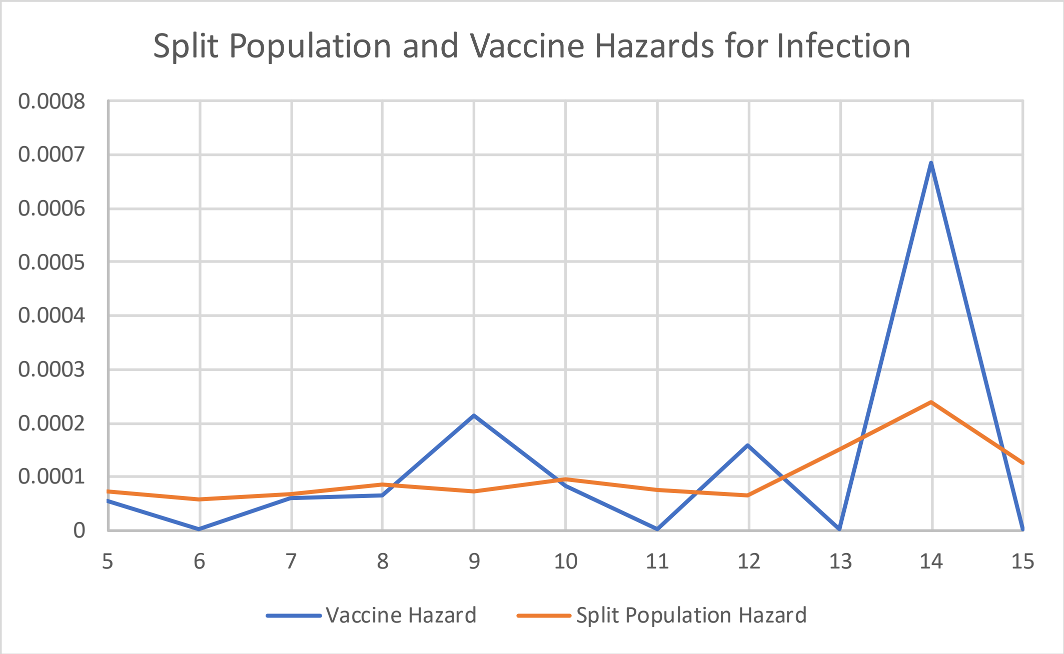
Week	Time (t)	h(t)	Pr{T >= t}	Pr{T=t}
5	0	0.001427	1	0.001427
6	1	0.001138	0.99857284	0.001136
7	2	0.001351	0.997436678	0.001347
8	3	0.0017	0.996089186	0.001694
9	4	0.001441	0.994395379	0.001433
10	5	0.001918	0.992962119	0.001904
11	6	0.001478	0.99105799	0.001465
12	7	0.001271	0.989593011	0.001258
13	8	0.003029	0.988335187	0.002994
14	9	0.004831	0.98534114	0.00476
15	10	0.002513	0.980581037	0.002464

From these numbers and the definition of $h_V(t)$ above, and using $p = 0.95$, we easily obtain $h_V(t)$ in the spreadsheet. Here are the results:



Week	Time (t)	h(t)	Pr{T >= t}	Pr{T=t}	h_V(t)
5	0	0.001427	1	0.001427	7.1358E-05
6	1	0.001138	0.99857284	0.001136	5.68852E-05
7	2	0.001351	0.997436678	0.001347	6.7539E-05
8	3	0.0017	0.996089186	0.001694	8.50062E-05
9	4	0.001441	0.994395379	0.001433	7.20466E-05
10	5	0.001918	0.992962119	0.001904	9.58473E-05
11	6	0.001478	0.99105799	0.001465	7.38765E-05
12	7	0.001271	0.989593011	0.001258	6.35192E-05
13	8	0.003029	0.988335187	0.002994	0.00015138
14	9	0.004831	0.98534114	0.00476	0.000241366
15	10	0.002513	0.980581037	0.002464	0.000125504

Finally, plotting the split population hazard on the same graph as the empirically estimated hazard function for the vaccinated group in the clinical trial (and those hazards were already provided to you in the problem file) yields:



This is pretty interesting – it suggests that the *modeled* hazard function you would get for the vaccinated group based on the *observed* hazard for the placebo group combined with the assumption of 95% vaccine success is very close to the *actual* hazard function *observed* for the vaccinated group. Very encouraging! Note – this does not prove that the Pfizer vaccine is 95% protective against infection over all time. What it does show is that *if* the vaccine offers true 95% protection, *then* the data seen in the trial in the vaccinated group based on the infection rates observed in the placebo group are just what we would expect! Of course, we need to temper our enthusiasm with facts – as we now know, vaccine effectiveness wanes with time and is also lower against the Delta and Omicron variants of SARS-CoV-2 compared to the virus that was circulating during the vaccine trials. On the bright side, the vaccine has proven to be very effective against severe illness and hospitalization, especially with a 3^{rd} (now n^{th}) booster shot. How well the newest bivalent vaccines perform remains to be seen.

Chapter 4

Poisson Process Models

4.1 When the Bernoulli grows up...

Suppose we observe a Bernoulli process with the following characteristics:

(i) A new Bernoulli trial occurs once every Δt time units (Δt is really pretty small!)

(ii) The success probability p is proportional to the length of the time period Δt. Let the proportionality constant equal λ. Thus, $p = \lambda \Delta t$ is the probability of an "arrival" on any Bernoulli trial (or in any time period of length Δt).

What happens?

4.2 When the Binomial grows up...

If we watch this process for n trials, this is equivalent to watching the process for a total length of time τ, where $\tau = n\Delta t$ (and thus $n = \tau/\Delta t$). Let $A(\tau)$ be the random variable representing the number of arrivals (or successes) over a time period of length τ (i.e. over the n trials). In this time period of length τ (or over these n trials), what is the expected number of arrivals (successes)? From the Bernoulli process, we know that:

$$\begin{aligned} E[A(\tau)] &= np \text{ (from the Binomial!)} \\ &= \frac{\tau}{\Delta t} \times \lambda \Delta t \\ &= \lambda \tau. \end{aligned}$$

The parameter λ thus has a physical interpretation: λ is the average arrival rate per unit time. In a period of length τ, $\lambda\tau$ arrivals are expected (as shown above), so the number of arrivals expected divided by the length of the time period (i.e. the average arrival rate per unit time) equals

$$\text{Avg Arrival Rate/Unit Time } = \frac{E[A(\tau)]}{\tau} = \frac{\lambda\tau}{\tau} = \lambda.$$

What about the variance of $A(\tau)$, the number of arrivals in a time period of length τ? Again, the Bernoulli process yields:

$$\begin{aligned} Var[A(\tau)] &= n \times p \times (1-p) \text{ (Binomial again!)} \\ &= \frac{\tau}{\Delta t} \times \lambda\Delta t \times (1 - \lambda\Delta t) \\ &= \lambda\tau \times (1 - \lambda\Delta t). \end{aligned}$$

Now we apply the squeeze! Let $\Delta t \rightarrow 0$. This implies that

$$Var[A(\tau)] \overset{\Delta t \rightarrow 0}{\longrightarrow} \lambda\tau.$$

Finally, what is the probability distribution of $A(\tau)$? Well, working with the Binomial again we have

$$\begin{aligned} \Pr\{A(\tau) = a\} &= \binom{n}{a} p^a (1-p)^{n-a} \\ &= \binom{\tau/\Delta t}{a} (\lambda\Delta t)^a (1 - \lambda\Delta t)^{\tau/\Delta t - a}. \end{aligned}$$

Once again, apply the squeeze! As $\Delta t \rightarrow 0$, the laws of calculus require that

$$\Pr\{A(\tau) = a\} = \frac{(\lambda\tau)^a e^{-\lambda\tau}}{a!} \text{ for } a = 0, 1, 2, ...$$

where $e = 2.718...$ is the base of the natural logarithms. This probability distribution is known as the *Poisson* distribution, and the underlying random process is called the *Poisson process.*

To summarize thus far: if "arrivals" occur at a constant average rate; the likelihood of an arrival in any very short time period (i.e. Δt) is proportional to the length of that time period; at most one arrival can occur in any very short time period; and whether an arrival occurs in any very short time period is independent of whether an arrival occurs in any other (i.e. "non-overlapping") very short time period; then we have a Poisson process. Letting $A(\tau)$ be the number of arrivals in a time period of length τ, $A(\tau)$ follows the Poisson distribution, with mean (and variance) equal to $\lambda\tau$ as shown above. λ is the average arrival rate of the Poisson process.

4.3 When the Geometric distribution grows up...

Suppose that we seek the probability that T, the time until the first arrival in a Poisson process, exceeds some time t. This is equivalent to stating that the first success exceeds the k^{th} trial if we set $k\Delta t = t$, or $k = t/\Delta t$. From the Geometric distribution, we have

$$\Pr\{T > t\} = (1-p)^k = (1-\lambda\Delta t)^{t/\Delta t}.$$

Again, apply the squeeze: as $\Delta t \to 0$, the laws of calculus assure us that

$$\Pr\{T > t\} = e^{-\lambda t} \text{ for } t \geq 0.$$

Here is another way to get the same result. For the time until the first arrival to exceed t, it must be true that *no* arrivals occur in the time period of length t preceding the first arrival. From the Poisson distribution noted earlier, we see the equivalence

$$\Pr\{T > t\} = \Pr\{A(t) = 0\} = e^{-\lambda t} \text{ !!!!}$$

So, the probability that the first arrival occurs at or before time t equals

$$\Pr\{T \leq t\} = 1 - e^{-\lambda t}.$$

This distribution is known as the *Exponential* distribution. It describes the behavior of random variable T, the time until the first arrival in a Poisson process. Without going into details (though you can derive what follows by first working with the Geometric distribution), we have:

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

and

$$Var(T) = \frac{1}{\lambda^2}.$$

Note that the mean equals the standard deviation for the Exponential distribution. Think about what this means for random incidence if "gap lengths" in a renewal process follow the Exponential distribution. By the way, note the reverse: if the gap lengths in a renewal process follow the Exponential distribution, then the renewal process is in fact the Poisson process!!

4.4 Is it Live, or is it Memoryless?

Suppose that the time in between buses follows an Exponential distribution (or equivalently, that the buses arrive in accord with a Poisson process). Having just missed the bus, you have been waiting for 5 minutes. You muse to yourself that the chance of waiting at least this long equals $e^{-5\lambda}$ where λ is the mean arrival rate of buses (in arrivals per minute). You ask yourself: what is the chance I have to wait at least another 5 minutes? You compute:

$$\begin{aligned} & \Pr\{\text{Wait additional 5} \mid \text{Waited 5 already}\} \\ = & \Pr\{T > 10 \mid T > 5\} \\ = & \frac{e^{-10\lambda}}{e^{-5\lambda}} = e^{-5\lambda} \text{ (!!)}. \end{aligned}$$

Slightly confused, you slowly realize that the chance of waiting at least another 5 minutes, having already waited 5 minutes, is unchanged from the likelihood of having to wait at least 5 minutes in the first place! In general:

$$\Pr\{T > t_1 + t_2 \mid T > t_1\} = \Pr\{T > t_2\}$$

if T follows the Exponential distribution. The chance of waiting at least another t_2 minutes is independent of t_1, the time spent already waiting! Similarly, the expected value of T given that you have already waited t time units is given by

$$E(T \mid T > t) = t + \frac{1}{\lambda}.$$

How unfair – your remaining expected waiting time exactly equals your original expected waiting time of $1/\lambda$! This is known as the *memoryless property* of the Exponential distribution. The random variable "forgets" how long you have been waiting for!

4.5 Examples

4.5.1 MPOX Revisited

(b) Suppose instead that the number of infections transmitted per newly infected man in the monkeypox outbreak followed a Poisson distribution with mean $R = 1.29$, that is,

$$\Pr\{X = j\} = \frac{R^j e^{-R}}{j!} \text{ for } r = 0, 1, 2, ...$$

Now what is your estimate of the minor outbreak probability π?

There are different ways to arrive at the single correct answer. You could do this numerically – in Excel, just evaluate

$$g(\pi) = \sum_{j=0}^{\infty} \frac{R^j e^{-R}}{j!} \times \pi^j$$

for different values of π and see when the value of π you enter into the right hand side of this formula equals $g(\pi)$, the result of the computation. Note that the formula is easily evaluated using the =sumproduct command. You simply fix a single cell to contain your trial value of π, then in one column enter the integers 0, 1, 2, ... (going only as far as 10 works fine for this problem), in the next column enter =poisson(j,1.29,0) and drag down (where j is the cell address for the integers 0, 1, 2... that you entered previously), and then in the next column enter $=\pi\hat{\ }j$ where π is your trial value on the right hand side (contained in a single cell and fixed in this computation using the absolute $ reference in Excel). Then just take the sumproduct of the second two columns – that evaluates $g(\pi)$ for the particular trial value. You can quickly get a match of the input and output at $\pi = 0.59$ (to two decimals). Or, you can use the Solver if you don't like trial and error – find the smallest value of π such that $\pi = g(\pi)$. You'll also get $\pi = 0.59$. Or, for those of you who like solving things mathematically, note that

$$g(\pi) = \sum_{j=0}^{\infty} \frac{R^j e^{-R}}{j!} \times \pi^j = \frac{e^{-R}}{e^{-R\pi}} \sum_{j=0}^{\infty} \frac{(R\pi)^j e^{-R\pi}}{j!} = e^{-R(1-\pi)}$$

where we recognize the second summand above as a Poisson distribution with mean $R\pi$, and thus the sum itself equals 1 leaving $g(\pi) = e^{-R}/e^{-R\pi} = e^{-R(1-\pi)}$. So now you don't need the sumproduct command, but you still need to either use trial-and-error or perhaps the solver to find the smallest value of π that solves $\pi = e^{-R(1-\pi)}$ (note that $\pi = 1$ is always a solution since the right hand side becomes $e^{-0} = 1$). With $R = 1.29$ you will once again discover that $\pi = 0.59$.

4.5.2 Condom Failures Revisited

Recall that if more than 4 condoms out of 1000 are found defective, the batch from which the 1000 condoms were drawn is rejected. Also, 12% of all batches tested in this manner were rejected. A different way to estimate the probability that a randomly selected condom would be found defective is to simply reason that since condom failures are rare events, the number of condom failures in a batch should approximately follow a Poisson distribution with some mean λ that in turn must be equal to $1000 \times p$ where p is the per condom failure probability. Again defining X is the number of condoms that fail in a batch of size 1000, let's presume that X has a Poisson distribution mean λ, and find the value of λ that sets the batch failure probability equal to 12%. That is, find that value of λ that solves

$$\Pr\{\text{Batch Fails}\} = \Pr\{X > 4\} = 1 - \Pr\{X \leq 4\} = 1 - \sum_{x=0}^{4} \frac{\lambda^x e^{-\lambda}}{x!} = 0.12.$$

This is equivalent to solving

$$\sum_{x=0}^{4} \frac{\lambda^x e^{-\lambda}}{x!} = 0.88$$

and doing so (for example, in Excel) yields $\lambda = 2.\,58$. Setting $\lambda = 1000 \times p = 2.58$ results in a per condom failure probability of $2.58/1000 = .00258$ in agreement with the Binomial model considered earlier.

4.5.3 Bloodbanking for a Rare Disease

A rare disease has appeared. New cases occur in your health district in accordance with a Poisson process with rate $\lambda = 50$ per month. It turns out that 80% of these cases can be treated inexpensively, but 20% of the cases require an immediate transfusion with a special blood product that itself is rare and costly to keep. You can only obtain monthly replenishments of this blood product from the one national collection center. Also, once delivered the "shelf life" of the blood product is only one month, so any product not used in the month of delivery must be discarded.

Denote the quantity of blood product used in any transfusion as a "transfusion unit" and assume that all transfusions require the same amount. The cost of the blood product per transfusion unit equals \$1,000, while in addition the cost of actually performing the transfusion (which is high because of the need for associated drugs in such cases) equals \$5,000 per procedure. If a case requiring a transfusion arrives at your hospital but you are out of blood product, the person is sent as an emergency to the national center at a cost of \$15,000.

The problem you face is to determine the amount of blood product to store in your local blood bank.

(a) Suppose that you store exactly 5 transfusion units of blood, and this month only 3 cases arrive that require transfusions. What is your cost of treating the rare disease this month?

The costs are as follows: blood product acquisition and storages costs=$5 \times \$1,000$; procedure costs=$3 \times \$5,000$; and emergency transfer costs=$0 \times \$15,000$; so the total is \$20,000.

(b) Suppose that you store exactly 5 transfusion units, but instead 7 cases arrive this month. What is your cost of treating the rare disease in this instance?

Well, we have: acquisition and storage costs=$5 \times \$1,000$; procedure costs=$5 \times \$5,000$ (for you can only perform 5 procedures as you only have 5 transfusion units!); and emergency transfer costs=$2 \times \$15,000$ (as two cases must be treated as emergencies); so the total is \$60,000.

(c) Back to the original problem. What is the probability distribution for the number of cases that require transfusion in any given month?

Let's see – new cases arrive in accordance with a Poisson process with $\lambda = 50$ per month, but only 20% of these cases require transfusion. Therefore, the rate with which cases that require transfusion arrive is given $50 \times .2 = 10$ cases per month. And, since we are splitting a Poisson process according to a Bernoulli trial (with probability .2, a case requires transfusion independent of all other cases), the resulting probability distribution for the number of cases that require transfusion in any month is itself Poisson with $\lambda = 10$ per month. So, if N is the number of cases that require transfusion in a month, the probability distribution is given by

$$\Pr\{N = n\} = \frac{10^n e^{-10}}{n!} \text{ for } n = 0, 1, 2, ...$$

(d) Let N denote the number of cases that require a transfusion in any given month, and now suppose you store exactly β (for *βlood*) transfusion units. What is the cost of treating the disease as a function of N and β?

Well, it depends of course on whether N is greater than β. We need to consider the sum of three things: acquisition and storage costs; procedure costs; and emergency costs. If we stock β transfusion units, the acquisition and storage costs are given by \$1,000β. Now, the procedure costs will equal \$5,000N if $N \leq \beta$; otherwise the procedure costs will equal \$5,000β (as you can't perform more procedures than transfusion units!). So, the procedure costs are given by $\$5,000\min(N, \beta)$. As for the emergency costs, you will only experience them if $N > \beta$, in which case you need to transfer $N - \beta$ persons to the national center at a cost of \$15,000 each. So, emergency costs are given by \$15,000 $\max(0, N - \beta)$. So the cost of treating the disease is given by

$$COST = \$1,000\beta + \$5,000\min(N, \beta) + \$15,000\max(0, N - \beta).$$

(e) What value of β leads to the *smallest expected cost* of treating the disease? What is this expected cost? Using this choice of β, on average how many patients will you refer as emergencies to the national center each month?

What you need to find is that value of β that minimizes the expected total costs, that is

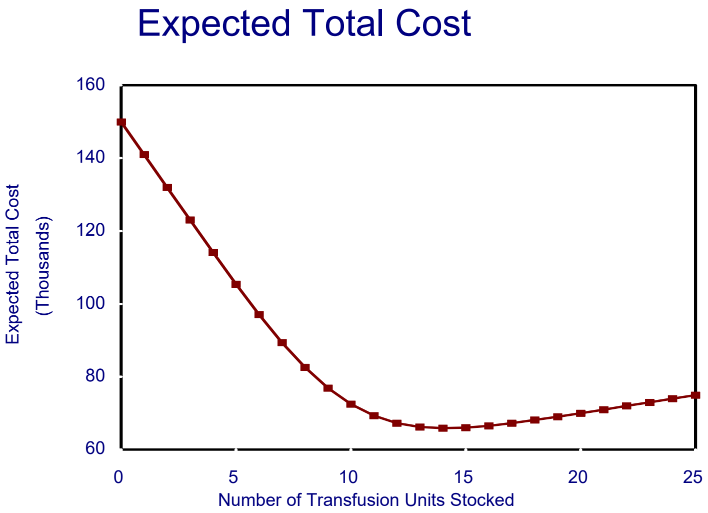
$$\min_{\beta} \sum_{n=0}^{\infty} \{1,000\beta + 5,000 \min(n, \beta) + 15,000 \max(0, n - \beta)\} \frac{10^n e^{-10}}{n!}.$$

To solve this problem, you need to evaluate the above summation for different values of β and find that value that costs the least! This is easily done in a spreadsheet. The optimal stocking level $\beta^* = 14$. When 14 transfusion units are stocked each month, the expected total cost each month is \$65,869. The expected number of emergencies is given by $E(\max(0, N - 14)) = 0.187$. In only 8.3% of all months will more than 14 cases requiring transfusion arrive.

The attached figure shows how expected total costs vary with stocking level. Note that when $\beta = 0$, all cases requiring transfusion are treated as emergencies. As on average 10 such cases arrive each month, the expected total cost when $\beta = 0$ is given by $10 \times \$15,000 = \$150,000$ (as seen in the figure). Now suppose β gets large. In this case, $E(\min(N, \beta)) \approx E(N) = 10$, while $E(\max(0, N - \beta)) \approx 0$. Thus, for large values of β the expected total costs grow linearly as

$$\$50,000 + \$1,000\beta.$$

Note, for example, that at $\beta = 25$ the expected total costs on the graph equal \$75,000, in accord with the simple formula above.



Chapter 5

Renewal Process Models

5.1 The Bus Stop

Consider a bus line with a stop near you place of work (or study or home or...). Bus lines might be scheduled so that they arrive once every five or ten minutes, but as anyone who rides a bus knows, the actual time between consecutive buses, which we will refer to as the *headway* and denote by H, is a random variable. Now, imagine a bus line where the headway probability distribution is given by

$$\Pr\{H = h\} = \begin{cases} 8/9 & h = 1 \\ \\ 1/9 & h = 10 \end{cases}$$

and $\Pr\{H = h\} = 0$ for values of h not equal to 1 or 10. We will call the 1 minute headways "short gaps" and the 10 minute headways "long gaps." You need to take the bus somewhere, and you have arrived at the bus stop just in time to see the door close as the bus drives off. What is your expected waiting time until the next bus? Clearly, you need to wait for an entire headway to expire, and thus your expected waiting time is the same as the expected headway $E(H)$ which is given by

$$E(H) = \frac{8}{9} \times 1 + \frac{1}{9} \times 10 = \frac{18}{9} = 2 \text{ minutes.}$$

Consider now a different situation – you need to catch the bus, you arrive at the bus stop, and...you don't see the bus! You look to the left, no bus. Look to the right, no bus. You realize that you've arrived somewhere inside a bus headway. At first you think "yay – I did *not* just miss the bus!" You surmise that your waiting time for the bus must be less than 2 minutes, because that is the average time to the next bus *when you just missed the last one*, so having *not* missed the last bus, your wait will be shorter.

But...you'd be wrong to think this way. Why? First, you surmise that you are randomly located within whatever headway you are in as the only information that you have is that you are not at the beginning (or end) of a headway (if you were you'd see a buss opening or closing the door!). Based on this reasoning, your expected waiting time just equals half the length of whatever headway you are in. So far so good.

But, there are two headway types, short and long. A short headway lasts one minute, so if you knew that you were in a short headway, your expected wait for the bus would equal 30 seconds. A long headway lasts 10 minutes, so if you knew that you were in a long headway, your expected wait would equal 5 minutes. *But which headway are you in*?

You might think that since only $1/9^{th}$ of all headways are long, you only have a $1/9$ chance of falling into a long headway (and an $8/9$ chance of falling into a short headway), and from this reasoning you would anticipate an expected wait until the next bus of $1/9 \times 5 + 8/9 \times 1/2 = 5/0 + 4/9 = 1$ minute, which of course is exactly half of the expected headway $E(H) = 2$ minutes. But you would be wrong! After all, *long headways are 10 times as long as short headways!* This means that if the *number* of long and short headways were equal, arriving at random would make landing in a long headway 10 times higher than the chance of landing in a short headway! Now the number of long and short headways are not equal; there are 8 times as many short headways as long headways. Still, since a single long headway is 10 times longer than a short headway, we see that the chance of landing in a headway of a given length must be proportional to both the *duration* of the headway and its relative frequency of occurrence! For the bus stop, this means that

$$\Pr\{\text{Enter long headway}\} \propto 10 \times \frac{1}{9}, \text{ and}$$

$$\Pr\{\text{Enter short headway}\} \propto 1 \times \frac{8}{9}$$

from which we conclude that

$$\Pr\{\text{Enter long headway}\} = \frac{10 \times \frac{1}{9}}{10 \times \frac{1}{9} + 1 \times \frac{8}{9}} = \frac{5}{9}$$

and

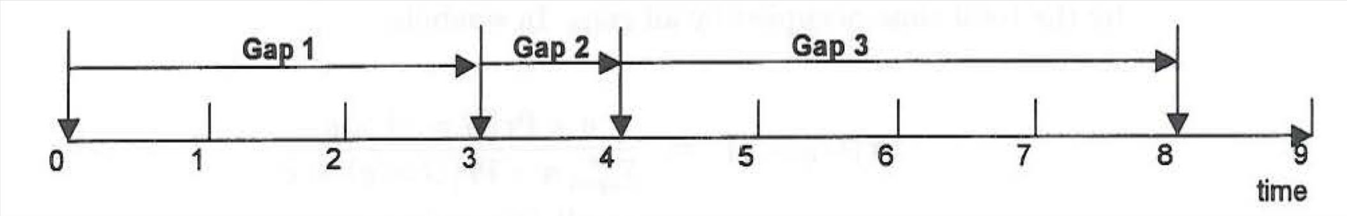
$$\Pr\{\text{Enter short headway}\} = \frac{1 \times \frac{8}{9}}{10 \times \frac{1}{9} + 1 \times \frac{8}{9}} = \frac{4}{9}$$

which makes the expected wait until the next bus given a random arrival equal to $5/9 \times 5 + 4/9 \times 1/2 = 3 > 2 = E(H)$ (!!!). To have a shorter expected waiting time for the bus, it turns out that you are better off just missing the bus than you are arriving at a random time!

The bus stop problem is an example of a general class of problems know as...

5.2 Random Incidence

Consider any random process that can be construed as a sequence of "arrivals and gaps" occurring over time; such a process is known as a "point process." In particular, suppose that the probability distribution for the lengths of the gaps is unchanging over time, and that the length of any particular gap is independent of the length of any other gap. These are known as "renewal processes," for the process "renews itself' whenever an arrival occurs. The situation is illustrated in the figure below.



The downward arrows denote "arrivals" while the times between successive arrivals comprise the "gaps." Let G be the random variable denoting the length of a gap; in the picture above, G assumes the values 3, 1, and 4 respectively for gaps 1, 2 and 3. The gap lengths follow some probability distribution denoted by $\Pr\{G = g\}$; for example, the gap lengths could follow a Geometric distribution (though they don't have to!!).

Suppose that we observe this process over a long period of time. Clearly, the mean and variance of the gap lengths, $E(G)$ and $Var(G)$, are given by the usual formulas:

$$E(G) = \sum_{g=1}^{\infty} g \times \Pr\{G = g\}$$

and

$$Var(G) = \frac{\sum_{g=1}^{\infty}(g - E(G))^2 \times \Pr\{G = g\}}{} = E(G^2) - E(G)^2.$$

So, for example, if gap lengths do follow the Geometric distribution, then $E(G) = 1/p$ and $Var(G) = (1 - p)/p^2$.

Now, suppose that the process in question has been operating for a long period of time. You show up at a random point in time, independent of the process in question. What is the probability that the length of the gap you will enter by virtue of random incidence, G_R, has length 1, 2, 3, or in general, g? To answer this question, we will use the "vanishing n" trick: consider n gaps. Of these, roughly $n \times \Pr\{G = g\}$ will be of length g, thus gaps of length g occupy roughly $n \times \Pr\{G = g\} \times g$ units of time. The total time occupied by all n gaps must then equal (roughly)

$$\text{Total Time} = \sum_{g=1}^{\infty} n \times g \times \Pr\{G = g\} = n \times E(G).$$

Thus, the likelihood of encountering a gap of length g by random incidence, $\Pr\{G_R = g\}$, must equal the total time occupied by gaps of length g, divided by the total time occupied by all n gaps. In symbols:

$$\begin{aligned}
\Pr\{G_R = g\} = & \quad \frac{n \times g \times \Pr\{G = g\}}{\sum_{g=1}^{\infty} n \times g \times \Pr\{G = g\}} \\
= & \quad \frac{n \times g \times \Pr\{G = g\}}{n \times E(G)} \\
= & \quad \frac{g \times \Pr\{G = g\}}{E(G)} \text{ for } g = 1, 2, 3...
\end{aligned}$$

So, if we know the probability distribution of gap lengths, we can find the corresponding probability distribution for gap lengths entered by random incidence.

Let's use this distribution to find $E(G_R)$, the expected length of a gap entered by random incidence. We obtain by direct summation:

$$\begin{aligned} E(G_R) &= \sum_{g=1}^{\infty} g \times \Pr\{G_R = g\} \\ &= \frac{\sum_{g=1}^{\infty} g^2 \times \Pr\{G = g\}}{E(G)} \\ &= \frac{E(G^2)}{E(G)}. \end{aligned}$$

Noting that $E(G^2) = E(G)^2 + Var(G)$, we can write $E(G_R)$ as

$$\begin{aligned} E(G_R) &= \frac{E(G)^2 + Var(G)}{E(G)} \\ &= E(G) + \frac{Var(G)}{E(G)}. \end{aligned}$$

As $Var(G)$ is always non-negative, we see that $E(G_R) \geq E(G)$, always!

Finally, suppose we seek the average time from random incidence to the process until the next arrival occurs. As random incidence places you, on average, in the middle of the gap in which you arrive, we see that the waiting time to the next arrival from random incidence, W_R, has a mean given by

$$E(W_R) = \frac{E(G_R)}{2}.$$

Returning to the bus stop problem, note that the headway between successive buses, H, correspond to gaps between arrivals in a renewal process, G. Substituting the results from the bus stop headway distribution into the general formulation above, we see that

$E(G) = 2$

$Var(G) = 8$

$\Pr\{G_R = 1\} = 4/9$; $\Pr\{G_R = 10\} = 5/9$

$E(G_R) = 6$

$E(W_R) = 3 > 2 = E(G)$ (and thus the conclusion that you are better off just missing the bus!)

Is it always true that you are better off just missing the bus? No - it is only true when $E(W_R) > E(G)$. When does this happen? You can verify that this condition occurs if (and only if) $E(G) < \sqrt{Var(G)}$, that is, if the average gap length is less than the standard deviation of the gap length.

5.3 Examples

5.3.1 A Museum Problem

In a small college town (and unknown to the new director of the town museum), 50% of the museum-going public visit the museum once per year, 30% visit twice per year, 15% visit three times per year, and 5% visit four times per year. Some of these museum-goers are willing to donate money to help support the museum. Again unknown to the new museum director, 20% of those who visit once per year are willing to donate, 40% of those visiting twice per year, 60% of those visiting three times per year, and 80% of those visiting four times per year.

(a) Among all museum-goers, what is the average number of visits to the museum per year?

If X is the number of visits per year among museum-goers, you are just being asked to find the expected value $E(X)$. That's easy:

$E(X) = 1 \times 0.5 + 2 \times 0.3 + 3 \times 0.15 + 4 \times 0.05 = 1.75$

(b) Among all museum-goers, what is the variance of the number of visits to the museum per year?

Now you are asked to find $Var(X)$. Since we already have $E(X)$ let's use the formula $Var(X) = E(X^2) - E(X)^2$ and first compute

$E(X^2) = 1^2 \times 0.5 + 2^2 \times 0.3 + 3^2 \times 0.15 + 4^2 \times 0.05 = 3.85$ and consequently $Var(X) = 3.85 - 1.75^2 = 0.7875$.

(c) What fraction of all museum-goers are willing to donate money to support the museum?

Let D be the event that a museum-goer is willing to donate; from the problem statement we are given the conditional probabilities $\Pr\{D|x\}$ for $x = 1, 2, 3$ and 4 visits per year. So, we just uncondition to discover that

$\Pr\{D\} = 0.2 \times 0.5 + 0.4 \times 0.3 + 0.6 \times 0.15 + 0.8 \times 0.05 = 0.35$. So 35% of museum-goers are willing to donate money to help support the museum.

One day, the new director decides to administer a survey to those who happen to be visiting the museum on that day. When people enter the museum, they are asked two questions:

(i) How many times per year do you come to the museum?

(ii) Would you be willing to donate money to help support the museum?

Given *your* knowledge of the visitation and donation probabilities from the problem statement:

(d) What is the average number of visits per year that would be reported among survey respondents?

This is random incidence! Someone who visits the museum x times per year is x times more likely to be in the survey compared to someone who only visits once per year. So let X_R denote the number of visits per year we would observe among persons sampled on a random day *at the museum.* The probability distribution of X_R is given by

$$\Pr\{X_R = x\} = \frac{x \Pr\{X = x\}}{E(X)} \text{ for } x = 1, 2, 3, 4.$$

Rolling this through with the underlying museum-going public visitation probabilities (and recalling from part (a) that $E(X) = 1.75$) we get:

$\Pr\{X_R = 1\} = 1 \times 0.5/1.75 = 50/175$; $\Pr\{X_R = 2\} = 2 \times 0.3/1.75 = 60/175$; $\Pr\{X_R = 3\} = 3 \times 0.15/1.75 = 45/175$; and $\Pr\{X_R = 4\} = 4 \times 0.05/1.75 = 20/175$. Getting the average visits per year reported in the survey can now just be found from

$E(X_R) = 1 \times 50/175 + 2 \times 60/175 + 3 \times 45/175 + 4 \times 20/175 = 2.2$ visits per year. You could also have just answered this directly from the formula

$$E(X_R) = \frac{E(X^2)}{E(X)} = \frac{3.85}{1.75} = 2.2$$

(e) What fraction of survey respondents would report that they are willing to donate money to help support the museum?

We already know the fraction of people willing to donate as a function of the number of visits they make to the museum each year. However, we now need to adjust for the visitation probabilities we would see in the survey. Happily, we computed those probabilities in part (d) above. Thus, letting D_R be the event that a museum-goer *sampled at the museum* is willing to donate, we obtain:

$$\Pr\{D_R\} = 0.2\times 50/175+0.4\times 60/175+0.6\times 45/175+0.8\times 20/175 = 0.44.$$

Note that this is higher than the true fraction of the museum-going population that is willing to donate, which from part (c) is given by 0.35. So we see how random incidence could lead our new museum director astray: on average people visit less often (1.75 visits per year) than the survey would suggest (2.2 visits per year), while the fraction willing to donate is lower (0.35) than the survey would suggest (0.44). Does this mean that taking surveys in the museum is a bad idea? Well, no actually. Suppose the museum director understood how random incidence works. With that knowledge, one could work backwards from the (observable) survey visitation frequencies to estimate the (unknown) population visitation frequencies, and use those to estimate the population donation base (as the survey would still reveal the relationship between willingness to donate and visitation frequency).

5.3.2 Checks and Balances

Many commercial banks in the United States offer overdraft protection to their customers. Overdraft protection allows customers to write checks, make ATM withdrawals, or make debit card purchases even if the customer does not have sufficient funds in their account to cover the expense (the alternative is for the check, debit card purchase, etc. to simply be declined).

The practice of enrolling customers automatically into overdraft protection programs is widespread: the Federal Deposit Insurance Corporation (FDIC) estimates that nearly 80% of large U.S. banks engage in this practice. In 2008, the FDIC conducted a study of roughly 40 large banks to learn more about the prevalence of overdrafts within the banking system. They divided the population of banking customers into five distinct groups based on the number of overdrafts each customer made within a given year. The FDIC reported the following results: roughly 74% of the population incurred no overdrafts during a given year; 12% of the population incurred 1-4 overdrafts; 5% incurred 5-9 overdrafts; 4% incurred 10-19 overdrafts; and 5% incurred 20 or more overdrafts.

(a) Based only on the data above, estimate the average annual number of overdrafts per customer per year. To simplify matters, assume that each person within a given group incurred the *lowest* number of overdrafts listed within that group. For example, assume that a person in the “1-4 overdrafts” group incurred exactly one overdraft during the year, a person in the "5-9 overdrafts" group incurred exactly five overdrafts, and so on.

Well, this is pretty easy – given the simplification, we have that a randomly selected *customer* incurred 0, 1, 5, 10 or 20 overdrafts with probabilities 0.74, 0.12, 0.05, 0.04 and 0.05 respectively. Letting X denote the annual number of overdrafts per customer, we have

$$E(X) = 0 \times .74 + 1 \times .12 + 5 \times .05 + 10 \times .04 + 20 \times .05 = 1.77 \text{ overdrafts.}$$

(b) Based only on the data above and using the same simplification as in part (a), estimate the variance of the annual number of overdrafts per customer per year (treating the percentages given as true probabilities as opposed to statistical estimates).

The easiest way to do this is to use the formula $Var(X) = E(X^2) - E(X)^2$ (this is the soundbite "variance = mean square minus squared mean"). So, let's find the mean square $E(X^2)$:

$$E(X^2) = 0^2 \times .74 + 1^2 \times .12 + 5^2 \times .05 + 10^2 \times .04 + 20^2 \times .05 = 25.37$$

and thus the variance of the annual number of overdrafts per customer equals

$$Var(X) = 25.37 - 1.77^2 = 22.237.$$

(c) What is the probability that a randomly selected overdraft transaction originated from a customer with five *or more* overdrafts per year?

Ah – here we must be careful, as the question has switched from customers to overdrafts. The likelihood that a randomly selected overdraft stems from a customer with x overdrafts, call this $g(x)$, is related the fraction of all customers with x overdrafts (call this $f(x)$) by the random incidence formula

$$g(x) = \frac{xf(x)}{E(X)}.$$

Now, the question asks for the likelihood that a randomly selected overdraft originated from a customer with five or more overdrafts. The easy way to do this is to look at the probability that a sampled overdraft originated from a customer with *fewer* than five overdrafts. And the only way a sampled overdraft could originate from a customer with fewer than five overdrafts is if the customer in question had 1 overdraft (if the customer had zero overdrafts, they could not have an overdraft sampled, could they?). So, the likelihood that a sampled overdraft corresponds to a customer with fewer than five overdrafts just equals $g(1)$ which is given by

$$g(1) = \frac{1 \times f(1)}{E(X)} = \frac{1 \times 0.12}{1.77} = 0.0678.$$

From this, the probability that a randomly selected overdraft originated from a customer with five or more overdrafts is given by $1 - g(1) = 1 - 0.0678 = 0.932\,2$ or about 93%.

(d) What is the expected number of overdraft transactions per year for a customer who originated a *randomly selected overdraft transaction*?

Well, we can do this the Fresca way or the Champagne way. Let X_R denote the number of overdrafts corresponding to a customer who originated a randomly selected overdraft. The Fresca approach says use the probability distribution $g(x)$ to find the mean, that is, compute

$$E(X_R) = \sum_x xg(x).$$

You'd first have to compute all the probabilities $g(x)$ and then toss them into the expected value formula above. You would discover that the probabilities corresponding to the number of overdrafts from the customer who originated a randomly selected overdraft equal 0.0678, 0.1412, 0.226, and 0.565 for 1, 5, 10 or 20 overdrafts respectively, and thus $E(X_R) = 14.33$, which is a *lot* bigger than the average number of overdrafts per customer.

Or, we could take the Champagne approach and recall from our discussion of random incidence in class (and in the notes) that

$$\begin{aligned} E(X_R) &= E(X) + \frac{Var(X)}{E(X)} \\ &= 1.77 + \frac{22.237}{1.77} = 14.33 \end{aligned}$$

as before (recall we previously computed $Var(X) = 22.237$ in part (b)!). Clearly sampling from overdrafts gives a rather different picture than sampling from customers!

5.3.3 SIDS

Of the roughly 3.9 million babies born in the United States each year, about 3,000 die of sudden infant death syndrome (SIDS), for a death-by-SIDS probability of 0.000769 per birth. A recent study notes that the fraction of SIDS casualties who had siblings also victimized by SIDS is higher than the figure computed above. Some feel that the parents of multiple SIDS victims should be investigated for murder as a result.

Now, suppose that among parents who have children, the mean number of births per family equals 1.86 with a standard deviation of 2.13. Suppose that in fact, all babies really do have the same probability of being victimized by SIDS independently of other events (and that there are no parental murderers using SIDS as a masquerade). Given this assumption, and using no other data, what is the probability that a randomly chosen baby would have at least one older sibling who also died of SIDS?

Some hints:

- If we are told that a baby is born into a family that ultimately experiences k births (including the baby!), then the probability that the baby is the $1st$, $2nd$, $3rd$, ..., k^{th} kid born equals $1/k$ for each possibility. Worded differently, any baby born is equally likely to be the $1st$, $2nd$, ..., k^{th} baby born in the family, *conditional* on the information that ultimately the family in question has k births.

- If p is the probability of SIDS (which we have estimated as 0.000769), and a newborn baby follows s older sibling births in the same family, then the probability that at least one of these siblings succumbed to SIDS equals (via our assumptions) $1 - (1 - p)^s$ (THINK THIS THROUGH!). In addition, since p is so small, it turns out that for any reasonable value of s, $1 - (1 - p)^s \approx sp$, an approximation you can use here without apology!

- Let B_f equal the ultimate number of births experienced by a randomly chosen *family*, and B_b equal the ultimate number of births experienced by the family belonging to a randomly chosen newborn *baby* (note: you know the mean (1.86) and standard deviation (2.13) of B_f). How are these random variables related?

First, some intuition. Suppose that a baby is the first birth in a family. That baby might die of SIDS, but that baby could not possibly have an older sibling who died of SIDS (for the baby in question is the first born!). Now suppose the baby was the second born. Then that baby has exactly one older sibling. *Second born children have exactly the same probability of having an older sibling who died of SIDS as a randomly chosen baby has of dying of SIDS!!* As stated in the second hint above, since $p = 0.000769$ is such a small number, the probability of having an older sibling who dies of SIDS for babies with s prior births (i.e. s older siblings) is just equal to sp (and this approximation is excellent). So in general, let S equal the number of older siblings for a randomly chosen new born. Then the probability q that this baby has at least one older sibling who died of SIDS is just given by

$$q = E(S)p$$

(and no, you did not need ESP to figure this out!).

OK – so now the whole problem reduces to figuring out what $E(S)$ is equal to. Suppose we consider a baby from a family with (ultimately) k births. If the baby was born first, then there are no older siblings. If the baby was born second, then there is one older sibling. If the baby is the j^{th} birth, there are $j-1$ older siblings. Now, given a baby from a family with (ultimately) k births, the baby is equally likely to be the first, second, ..., k^{th} birth (conditional on the ultimate number of births; this was your first hint). So, the expected number of older siblings for babies born to families with (ultimately) k births, $E(S|k)$, is given by

$$E(S|k) = \sum_{j=1}^{k} \frac{1}{k} \times (j-1) = \frac{1}{k} \times \frac{k(k-1)}{2} = \frac{k-1}{2}.$$

So we see that the average number of older siblings from a family with (ultimately) k births just equals $(k-1)/2$.

Here is another way to understand this result. In a family with (ultimately) k births, there are $\binom{k}{2} = k(k-1)/2$ sibling pairs (this is the number of different ways to chose two objects from k objects!). Now for each sibling pair, there is of course exactly one older sibling! There is also one younger sibling, but we don't care about that. So, the *total* number of older siblings (counted as older siblings in any sibling pair) also equals $k(k-1)/2$, and hence the *average* number of older siblings *per birth* in a family of (ultimately) k births is given by

$$E(S|k) = \frac{\text{Total older siblings}}{\text{Total births}} = \frac{k(k-1)/2}{k} = \frac{k-1}{2}\cdot$$

Pretty cool.

Now, we need to figure out the unconditional expectation $E(S)$. Clearly, we want to do this with reference to the probability distribution of the (ultimate) number of births in the family *corresponding to a randomly chosen newborn baby.* This, of course, is where the third hint comes in. What we want to compute is

$$E(S) = \sum_{k=1}^{\infty} E(S|k) \times \Pr\{B_b = k\}$$

where B_b is the (ultimate) number of births experienced by the family corresponding to a randomly chosen newborn. We don't have any data regarding B_b, but we do know the mean (1.86) and the standard deviation (2.13) of B_f, the (ultimate) number of births in a family as sampled from *families* (i.e. the census data). How are these related? Suppose we have a family that (ultimately) has 10 births, versus a family with only one. If there were equal numbers of these two types of families, and we pick a newborn *baby* at random, then the chance that newborn comes from the family of 10 *must be ten times higher* than the chance the baby came from the family of one! Why? The former families contribute 10 times as many babies to the baby pool! *Random incidence*, plain and simple. The random variable B_b is thus obtained by a *length biased sample* of the random variable B_f (where the "length" in this case corresponds to the number of babies per family). This being the case, what do we know about the relationship between these variables? Well, all we need to know is this:

$$E(B_b) = \frac{E(B_f^2)}{E(B_f)} = \frac{Var(B_f) + E(B_f)^2}{E(B_f)} = E(B_f) + \frac{Var(B_f)}{E(B_f)}\cdot$$

The formulas above all appear in your handout on random incidence.

How does this relate to the above? Well,

$$E(S) = \sum_{k=1}^{\infty} E(S|k) \times \Pr\{B_b = k\} = \sum_{k=1}^{\infty} \frac{k-1}{2} \times \Pr\{B_b = k\} = \frac{E(B_b) - 1}{2}\cdot$$

Combining the expressions we have

$$E(S) = \frac{1}{2} \times (E(B_b) - 1) = \frac{1}{2} \times \left(E(B_f) + \frac{Var(B_f)}{E(B_f)} - 1 \right)$$

and thus the probability that a randomly chosen newborn has at least one older sibling who died of SIDS equals

$$q = E(S)p = \frac{p}{2} \times \left(E(B_f) + \frac{Var(B_f)}{E(B_f)} - 1 \right).$$

Plugging in the data ($p = 0.000769, E(B_f) = 1.86, Var(B_f) = 2.13^2$) we obtain

$$q = \frac{0.000769}{2} \times \left(1.86 + \frac{2.13^2}{1.86} - 1 \right) = 0.000769 \times 1.65 = 0.001269.$$

Now suppose we sampled SIDS babies (instead of any baby), and asked what fraction of SIDS babies should have older siblings with SIDS. Well, if it is really the case that all babies have the same probability of death from SIDS, independent of other events, then the likelihood a SIDS baby has older siblings who died of SIDS also equals q as computed above. This number is 65% larger than the fraction of all babies who die of SIDS.

So, should parents of multiple SIDS victims be investigated for murder? *It depends upon whether the observed fraction of SIDS babies with siblings who have died of SIDS is comparable to 0.001269!!* According to an article titled “Killers may fake infant death syndrome” on p. D1 of the October 28, 1997 issue of *USA Today, about 3% had siblings who also died of SIDS!* Hmmmmmm.......

Chapter 6

Markov Models

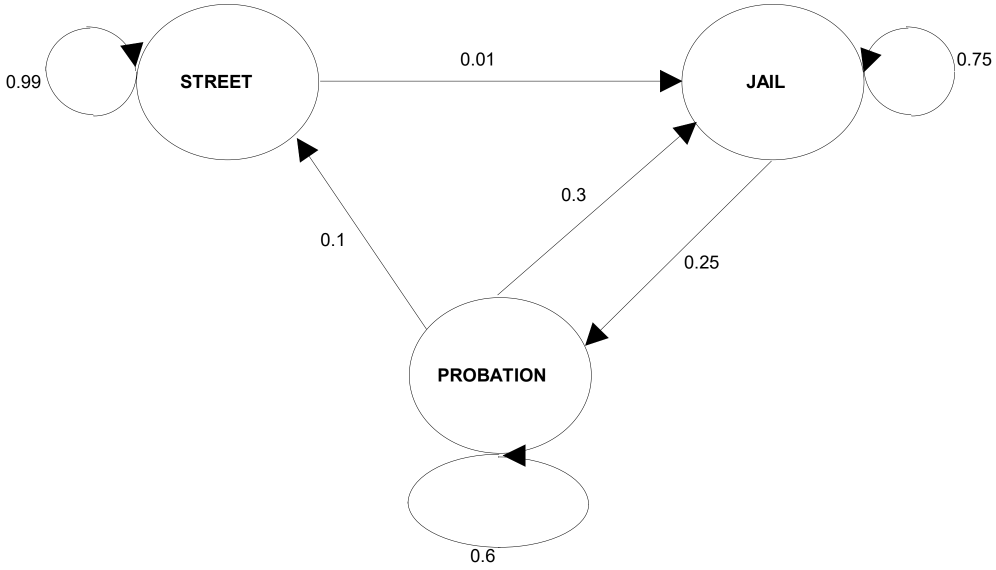
6.1 Examples

6.1.1 Prison Planning

As the aide to the newly elected governor of your state, you have assumed responsibility for implementing the campaign promise to define a set of changes in the operation of the penal system. You have given some thought to the matter and have decided that you have essentially one instrument that you can control and one resource that you must provide. In the first instance, you could persuade the courts and probation boards to raise or lower the average length of incarceration. The average time in prison after each conviction is now four years, but you believe that you could move that average up or down by one year. The resource you must provide is the staff of probation officers. Each probation officer can handle an average of 25 cases at a time.

Conditions in your state result in the annual conviction of about 1% of the population not in jail or on probation. All prisoners who are released from jail must enter probationary status. In any given year, 30% of the ex-convicts on probation are convicted of a new crime and returned to prison, but 10% are fully rehabilitated and return to the “general population” (that is, those not in jail or on probation); the rest remain on probation. Assume that the current policies have been in operation for a long time, and thus the distribution of the population has reached an equilibrium. It costs \$50,000 per year to maintain a prisoner in jail, but only \$10,000 per year (including the salary costs of probation officers) to keep a prisoner on probation. There are one million people in your state, including those in prison or on probation.

(a) Draw a diagram to indicate the probabilistic flows through this system. Make sure to indicate clearly the states and annual transition probabilities.



(b) What is the expected number of people who are neither in jail nor on probation?

Well, first we need to find the steady state of the system. Denote p_S, p_P and p_J as the steady state probabilities of finding a citizen on the Street (i.e. not in jail or on probation), on Probation, or in Jail. We have from (a):

$$p_S = .99p_S + .1p_P + 0p_J$$

$$p_P = 0p_S + .6p_P + .25p_J$$

$$p_S + p_P + p_J = 1$$

These solve to yield $p_S = 0.794$, $p_P = 0.079$, and $p_J = 0.127$. Now, the question asks for the expected number of people on the street which is simply $1,000,000 \times p_S = 794,000$.

(c) What is the expected number of persons in jail?

$1,000,000 \times p_J = 127,000$.

(d) How many probation officers are needed to maintain the average caseload?

Well, there are on average $1,000,000 \times p_P = 79,000$ persons on probation, and if each probation officer can handle 25 cases, then we need $79,000/25 = 3160$ officers. Actually, roundoff error shows up here – $p_P = 0.079365$ to more decimal places, and hence we need 3175 officers (but what are 15 parole officers between friends?).

(e) What is the annual cost of operating the state penal system?

Well, we pay $50,000 for each prisoner and $10,000 for each person on probation annually. So, in total we pay $\$50,000 \times 127,000 + \$10,000 \times 79,000 = \$7.14$ *billion* (!)

(f) What is the annual conviction rate in the state?

Convictions occur with probability .3 each year for parolees and with probability .01 for those on the street. So, the conviction rate is $.3p_P + .01p_S = 0.032$ (i.e. 3.2% of the population gets convicted each year on average).

(g) Given what you know about the system, what changes might you recommend to the governor? Remember, you can control (within limits) the average prison sentence, and you also have flexibility with respect to the number of probation officers. In discussing your recommended changes, make sure to provide estimates of the annual conviction rate and the cost of operating the penal system.

Basically what you control is the average sentence length; call this t. Now if you resolve the model for the steady states as a function of t you will discover that:

$p_S = 10/(11 + .4t)$
$p_P = 1/(11 + .4t)$
$p_J = .4t/(11 + .4t)$.
(Try substituting $t = 4$ in the above expressions to see how you reproduce your results in (a)).

The total operating cost of the prison system per capita with an average sentence length of t is just $Prison\ Cost = \$50,000p_J + \$10,000p_P$ while the conviction rate is given by $.01p_S + .3p_P$. Now, it is clear that as a function of t, prison costs are going up with t while convictions are going down with t. How should the tradeoff be managed? Well, suppose that the average cost society suffers from crime *per conviction* is just denoted by c. Then the sum of crime plus prison costs per capita would be given by

$$Total\ Cost = 50,000p_J + 10,000p_P + c(.01p_S + .3p_P)$$

which is equal to

$$Total\ Cost = \frac{10,000 + 50,000 \times .4 \times t + c(.01 \times 10 + .3)}{11 + .4t} = \frac{10,000 + .4c + 20,000t}{11 + .4t}.$$

Now this expression is of the form $(\alpha + \beta t)/(\gamma + \delta t)$ where $\alpha = 10,000 + .4c$, $\beta = 20,000$, $\gamma = 11$ and $\delta = .4$. Convince yourself that the following is true: $Total\ Cost$ is *increasing* in t if $\beta/\delta > \alpha/\gamma$, otherwise $Total\ Cost$ is decreasing in t. If the total costs are increasing in t then we want to *shorten* prison sentences, while the reverse is true if total costs are *decreasing* in t (in which case we wish to *lengthen* prison sentences).

The condition for $Total\ Cost$ to be increasing in t is thus equivalent to $c < \$1,350,000$ while for $c > \$1,350,000$ $Total\ Cost$ is decreasing in t. What this says is that if the cost per conviction is relatively cheap ($c <$ \$1.35 million) then we should allow more convictions and save on prison time, and thus set prison sentences as short as we can. On the other hand, if the cost per conviction is relatively expensive ($c >$ \$1.35 million) then we should cut down on the number of convictions by increasing prison time. Note that what is really going on is that if you increase prison time by increasing t, you reduce the number of *crimes* that are occurring (as there are fewer potential criminals to commit such crimes). In other words, it is not the convictions per se that cost money - we have simply looked at the cost of crime per conviction.

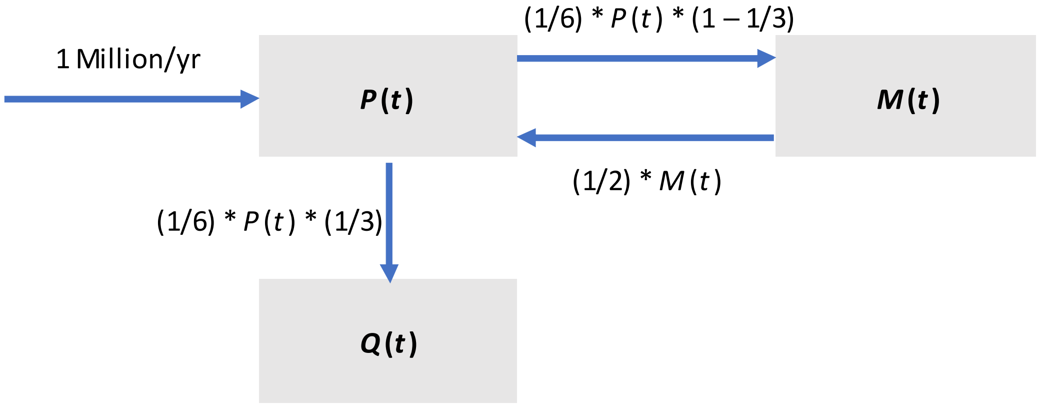
So, in a nutshell, under the assumptions of this model, I would advise the governor to decrease the length of prison sentences if crimes are costing society less than \$1.35 million on average per conviction, and to increase the length of prison sentences if crimes cost more than \$1.35 million per conviction on average. Of course, there are several features of this model which are not realistic. The most unrealistic feature is that the Markov transition probabilities are independent of t, the length of a prison sentence. If we have longer sentences, for example, then the reality that the population really consists of would be criminals and those who would never commit crimes would cause the transition rate from the street to prison to decline as t increases (as those not in jail or on probation would be less likely to commit crimes). Also, we are using a single average sentence length over all crimes, something which convicted murderers might enjoy but tax evaders might protest. Nonetheless, the model does capture some realistic features, including the tradeoff between paying for the penal system and paying for crime. Indeed, it seems reasonable that if crimes are not that expensive to society, then society should be willing to suffer more of them (and hence we have shorter prison sentences) relative to the case where crimes cost more to society so that lowering the crime rate (via increasing the length of stay in prison and associated prison costs) is an appropriate policy.

Markov Migration Model

Imagine a scenario in which in expectation, one million new (i.e. first time or rookie) undocumented immigrants cross the border into the US for the first time each year (all immigrants in this problem are undocumented). In addition to these first time border crossers, repeat (or circular) migrants re-cross the border having previously visited the US and departed in a manner to be described shortly. Assume that irrespective of whether an arriving migrant is a rookie or a repeat visitor, the average duration of stay in the United States equals six years (equivalently, within the Markov model you are to construct, in expectation one sixth of all undocumented immigrants in the US leave each year). After any visit to the US (whether rookie or repeat), migrants retire from all further migration (that is, quit the migration process) with probability 1/3 and never return to the US. With probability 2/3, migrants who just left the US remain "active" in their home country. In any year, active migrants return to the US from wherever they are with probability 1/2. Equivalently within the Markov model you are to construct, migrants who leave the US but have not yet retired return to their home country, and in any year, in expectation 50% of such active migrants return to the US. Note that all migrants retire outside of the US, so none remain in the US indefinitely.

(a) Draw a state transition diagram for this discrete time, three state Markov process letting the states $P(t)$, $M(t)$, and $Q(t)$ denote the (expected) number of migrants who are in the US, outside the US but active, and outside the US and retired respectively as of the end of year t. Note that your diagram must account for the expected arrival of a million rookie migrants to the US each year, in addition to repeat (circular) migrants.

Here's my picture:



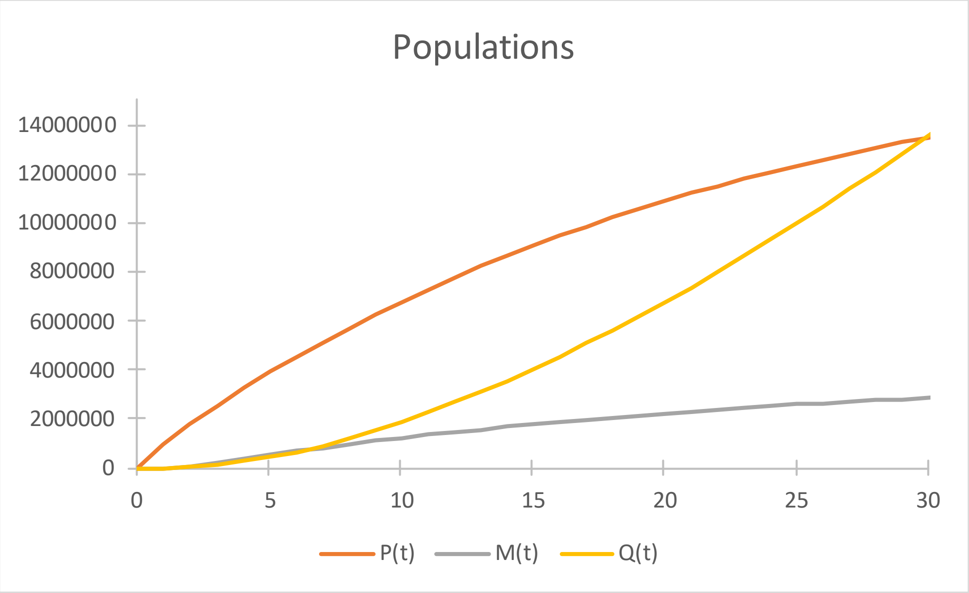
There is an external arrival rate of 1 million new migrants per year directly to the undocumented population in the US, in addition to repeat migrants who arrive at rate $M(t)/2$ (as the problem states that active migrants return in any year with probability $1/2$, and there are $M(t)$ such active migrants in expectation).

(b) Suppose that at time $t = 0$, the process has not yet begun and thus $P(0) = M(0) = Q(0) = 0$. Starting at time $t = 1$ (time is measured in years), the first wave of rookie migrants arrives, and the process evolves as described from that point. Produce a graph showing $P(t)$, $M(t)$, and $Q(t)$ over the first 30 years of this process. As of time $t = 30$ years, how many undocumented immigrants are there in the US? How many active migrants are there outside of the US? How many migrants have quit the migration process?

Producing the graph requires executing the model from (a), which means the state transition diagram must be translated into equations. You need three equations, one for each of $P(t)$, $M(t)$ and $Q(t)$. Here they are:

$$\begin{aligned} P(t+1) &= P(t) + 1,000,000 + M(t)/2 - P(t)/6 \\ M(t+1) &= M(t) + (1/6) \times (2/3) \times P(t) - M(t)/2 \\ Q(t+1) &= Q(t) + (1/6) \times (1/3) \times P(t). \end{aligned}$$

Iterating these equations from $t = 0$ for 30 years yields the graph shown below:



As of year 30, the (expected) numbers of undocumented immigrants in the US, active migrants outside the US, and retired migrants are given by $P(30) = 13,515,021.4$, $M(30) = 2,905,589.3$, and $Q(30) = 13,579,389.3$. Note that summing these equations reveals that

$$P(t+1) + M(t+1) + Q(t+1) = P(t) + M(t) + Q(t) + 1,000,000,$$

which says that the total population of people who have ever visited the US increases by a million persons a year on average. And, $P(30) + M(30) + Q(30) = 30$ million.

(c) Now imagine that this process continues unabated for a very long time. What are the limiting values P and M of $P(t)$ and $M(t)$ as $t \to \infty$? Provide an answer using each of the following two approaches, and show that your answers are the same: (i) apply the Markov model to achieve steady state solutions for P and M, and (ii) deduce the values of P and M directly via Little's Theorem from queueing theory.

(i) Using the Markov model, the steady state equations for P and M can be found by substituting P and M into the state transition equations and solving. So, you need to solve:

$$P = P + 1,000,000 + M/2 - P/6$$

and

$$M = M + (1/6) \times (2/3) \times P - M/2.$$

The second equation implies that $M = 2\times(1/6)\times(2/3)\times P = 2P/9$. Plugging this result into the first equation yields

$$1,000,000 + (2P/9)/2 - P/6 = 0,$$

which implies that $P = 1,000,000/(1/6-1/9) = 1,000,000/(3/18-2/18) =$ 18 million. Substituting back into $M = 2P/9$ implies that $M = 4$ million. Thus, if this process continued indefinitely, there would be on average 18 million undocumented immigrants in the US, plus another 4 million active migrants outside the country.

(ii) We can also get these results directly using queueing theory. The arrival rate of first time undocumented immigrants to the US is one million per year; think of this as λ. Over time, each of these will average 3 visits to the US (since the probability of quitting is 1/3 after each visit), and on each visit, the average duration of stay is 6 years, for a total of 18 years spent in the US; think of this as W. So, using Little's Theorem, $L = \lambda W$ implies that there are 1 million/year × 18 years = 18 million undocumented immigrants in the US, in agreement with the Markov result from (i). Similarly, over time, each of the 1,000,000 first time migrants each year will spend an average of two episodes as an active migrant outside the US (because if there are three visits until quitting on average, then there must be an average of two active migrant spells), and each such active migrant episode averages 2 years (since for active migrants there is a 1/2 probability of returning to the US each year). Thus, the total time active migrants spend active outside of the US averages 2 episodes × 2 years = 4 years in total. Thus, Little's Formula now yields 1,000,000 × 4 = 4 million active migrants in steady state, again in agreement with the Markov result from (i).

(d) After t years have passed, what is the expected total number of migrants who visited the United States *at least once* (i.e. the expected number of distinct migrants who have ever crossed the border, including retired migrants; your answer must be a very simple formula involving t)?

A gift, I tell you, a gift! The number who visit the United States at least once is exactly the same as the number of first time migrants, and with a million rookies each year, the expected total number of migrants who visit the US at least once just equals $1,000,000 \times t$.

(e) Let $m(t)$ denote the mean number of times a migrant has crossed the border to the United States up to the end of year t, where this average is taken over *all* migrants who have *ever* crossed the border (that is, all migrants considered in the solution to part (d) above, which again includes retired migrants). After t years have passed, produce a simple formula for the expected *total* number of times migrants have crossed the border *into* the United States in terms of $m(t)$. Based on this result, produce a simple formula for the expected total number of migrants who have retired by the end of year t (that is, a simple formula for $Q(t)$) in terms of $t, m(t)$ and $P(t)$.

We just learned from (d) that the expected total number of migrants who visited the US at least once after t years is equal to 1,000,000 × t. Now we are told that over all migrants who have ever crossed the border, which is the same thing as over all migrants who visited the US at least once, the average number of trips per migrant equals $m(t)$. This means that the expected *total* number of border crossings (whether first time or repeat) must equal $1,000,000 \times t \times m(t)$. Now, all of these border crossings into the US will eventually have an associated exit from the US. But, as of time t, of the $1,000,000 \times t \times m(t)$ border crossings into the US, $P(t)$ have yet to exit – these are the undocumented immigrants still in the US at the end of year t! We thus have a simple formula for the total number of *exits* from the US at the end of year t, namely $1,000,000 \times t \times m(t) - P(t)$. Since we know that one third of all migrants retire after any trip to the US, we conclude that

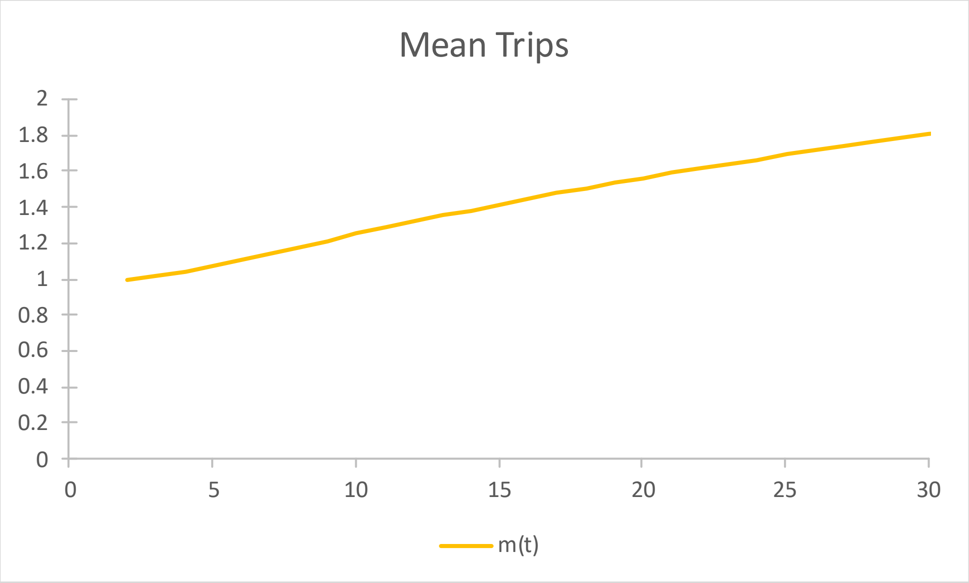
$$Q(t) = \frac{1,000,000 \times t \times m(t) - P(t)}{3}.$$

(f) Invert your result from (e) to produce a simple formula for $m(t)$ in terms of t, $Q(t)$ and $P(t)$, and plot $m(t)$ for the first 30 years of this process. What is $m(30)$?

Simply solving the result from part (e) for $m(t)$ yields

$$m(t) = \frac{3Q(t) + P(t)}{1,000,000 \times t}.$$

Plotting over 30 years yields the graph below, while $m(30) = 1.81$.



(g) As $t \to \infty$, what happens to $m(t)$? Answer this based on the problem formulation, not by continuing your numerical calculation from part (f) ad infinitum.

As $t \to \infty$, we see that the retired population will keep growing by just under a million persons per year while the number of migrants who are in the US or active outside the country approach the constant values deduced in part (c) above. As virtually all those who have ever visited the US will have retired as $t \to \infty$, we expect that the average number of trips to the US over all those who ever visited will converge to 3 (as that is the average number of visits until quitting). This is indeed the case, but it takes a very long time until it happens, as illustrated in the optional problem below.

(h) Define $m_P(t)$, $m_M(t)$, and $m_Q(t)$ as the average number of trips to the US by the end of year t among migrants in the US ($P(t)$), active migrants not in the US ($M(t)$), and retired migrants ($Q(t)$) respectively at the end of year t. In part (f) you deduced $m(t)$, the overall average number of trips to the US among all those who ever visited the US by the end of year t. Clearly the following "conservation of trips" equation must hold:

$$m(t) = \frac{m_P(t)P(t) + m_M(t)M(t) + m_Q(t)Q(t)}{P(t) + M(t) + Q(t)}.$$

Using your Markov migration model, deduce a recursive scheme to compute $m_P(t)$, $m_M(t)$, and $m_Q(t)$, apply your scheme to plot $m_P(t)$, $m_M(t)$, and $m_Q(t)$ for the first 30 years of the process, and show that the conservation equation above holds numerically. (HINT: think first about computing the expected *total* number of trips to the US to the end of year t for migrants in $P(t)$, $M(t)$, and $Q(t)$ respectively.)

We like hints! Let's define $\tau_P(t)$, $\tau_M(t)$, and $\tau_Q(t)$ as the expected total number of trips (i.e. departures) to the US to the end of year t for migrants in $P(t)$, $M(t)$, and $Q(t)$ respectively, just as the hint says. If we knew the values of these three quantities, then we could easily compute $m_P(t)$, $m_M(t)$, and $m_Q(t)$ via the ratios $\tau_P(t)/P(t)$, $\tau_M(t)/M(t)$, and $\tau_Q(t)/Q(t)$ respectively. Note that with these ratios, the conservation of trips law expressed above clearly holds, as we would have

$$\begin{aligned} m(t) &= \frac{\tau_P(t)/P(t) \times P(t) + \tau_M(t)/M(t) \times M(t) + \tau_Q(t)/Q(t) \times Q(t)}{P(t) + M(t) + Q(t)} \\ &= \frac{\tau_P(t) + \tau_M(t) + \tau_Q(t)}{P(t) + M(t) + Q(t)} \\ &= \frac{\text{Total Trips to the US by the end of year } t}{\text{Total Migrants as of the end of year } t}. \end{aligned}$$

So, we need to figure out $\tau_P(t)$, $\tau_M(t)$, and $\tau_Q(t)$.

These actually follow quite directly from the Markov process! Let's start with $\tau_Q(t)$, the total trips to the US up to the end of year t taken by retired migrants as of the end of year t. I claim that

$$\tau_Q(t+1) = \tau_Q(t) + \frac{1}{6} \times \frac{1}{3} \times \tau_P(t).$$

Why? Because if $\tau_P(t)$ is the expected total number of trips taken to the US by those in the US as of the end of year t, in the next year, one-sixth of migrants in the US are expected to leave, and one third of those leaving are expected to retire, and all of these new retirees will bring their total trips to date with them! This means that the expected total trips to date to the US among those retired at the end of year $t+1$ gets increased over the total number of such trips among those retired at the end of year t by $\frac{1}{6}\times\frac{1}{3}\times\tau_P(t)$. This is great, if we can figure out $\tau_P(t)$.

Similarly, considering total expected trips to the US among active migrants outside the US, I claim that

$$\tau_M(t+1) = \tau_M(t) + \frac{1}{6}\times\frac{2}{3}\times\tau_P(t) - \frac{1}{2}\tau_M(t).$$

Here we track the expected number of people leaving the US who do not retire by the end of year t; these people bring their trips-to-date with them, which by definition equals $\tau_P(t)$ in total. But, half of the active migrants return to the US in the following year, so we need to subtract out half of the expected trips-to-date in that population. The result is the equation shown above.

Almost done! What about migrants in the US? I claim that (and this is a bit trickier)

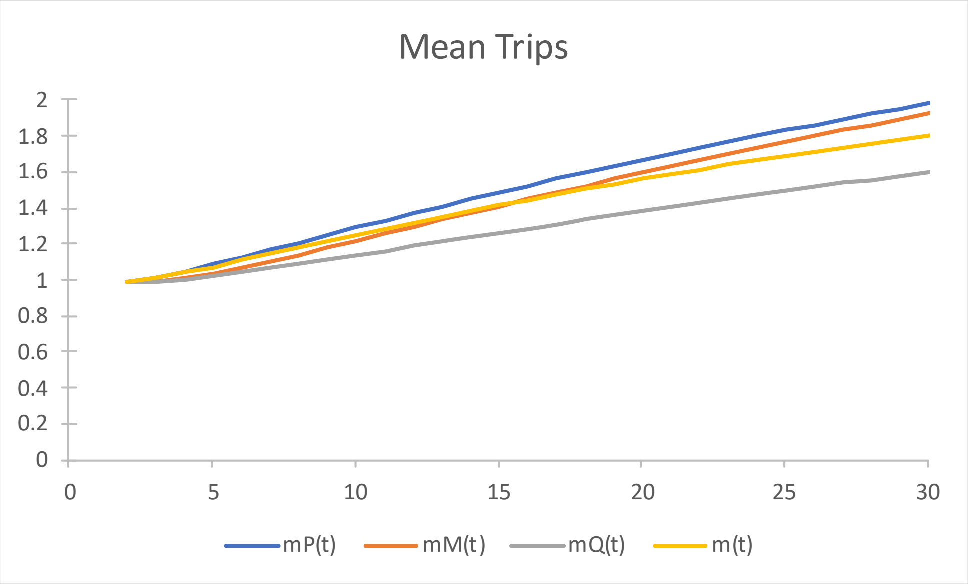
$$\tau_P(t+1) = \tau_P(t) + 1,000,000 + \frac{1}{2}(\tau_M(t) + M(t)\times 1) - \frac{1}{6}\tau_P(t).$$

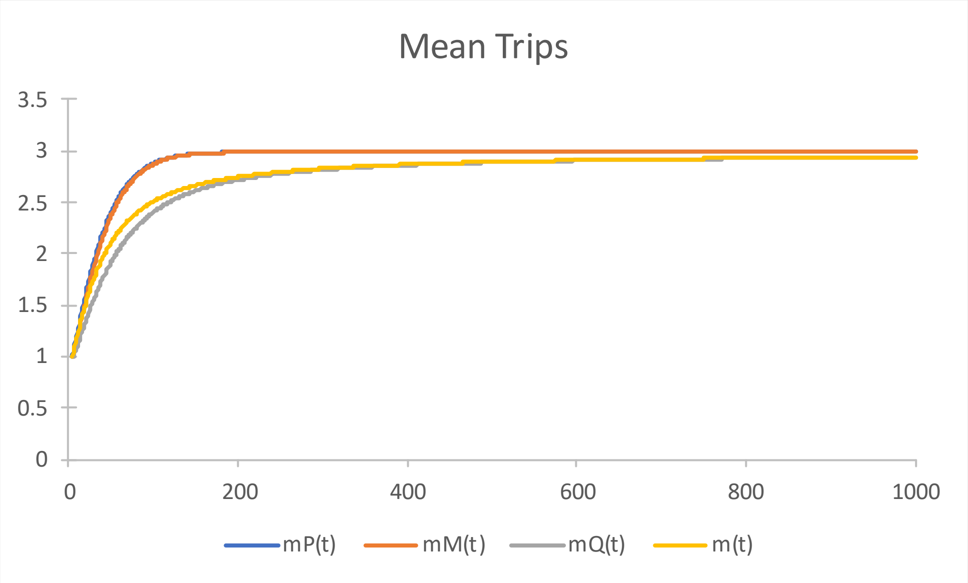
What is different in this equation is that whenever a migrant enters the US, they increase the number of trips to the US by one per migrant. SO, since on average there are one million rookie (i.e. first-time) visits to the US per year, we need to add in one million new trips! Plus, since on average $(1/2)\times M(t)$ active migrants return to the US over year t, they bring an expected $\tau_M(t)/2$ total trips to date with them, plus an additional $M(t)/2$ trips, one for each active migrant crossing back into the US. Finally, since on average one sixth of the migrants in the US leaves in any year, the total trips to date among migrants in the US must be reduced by $\tau_P(t)/6$. The result is the equation shown above.

Iterating these three equations for 30 years starting with $\tau_P(0)$, $\tau_M(0)$, and $\tau_Q(0)$ all equal to zero yields the expected total trips in each population subgroup over time, and dividing by the subpopulation sizes to obtain our ratios yields the sought after estimates of mean trips to date in each subpopulation. Plotting over time gives the graph below, while numerically it is clear that applying the conservation of trips equation

$$m(t) = \frac{m_P(t)P(t) + m_M(t)M(t) + m_Q(t)Q(t)}{P(t)+M(t)+Q(t)}$$

yields exactly the same results for mean trips to date overall as found previously in part (f). Eventually, all of these average trips per migrant in different subpopulations converge to 3, but it takes a very, very long time, as shown in the second figure below.





6.1.2 HIV Infection and Paid Plasma Donation

Unlike voluntary blood donation, the bulk of plasma donations stem from paid plasma donors. It is perhaps not surprising, then, that the rate of new HIV infections among paid plasma donors is about 20 times greater than the corresponding rate for voluntary blood donors. In this problem you will develop a simple Markov model that describes the HIV risk from paid plasma donors; models to evaluate policies for mitigating such risks (e.g. holding policies) follow from the basics you will develop below, but are a bit beyond the methods of this class.

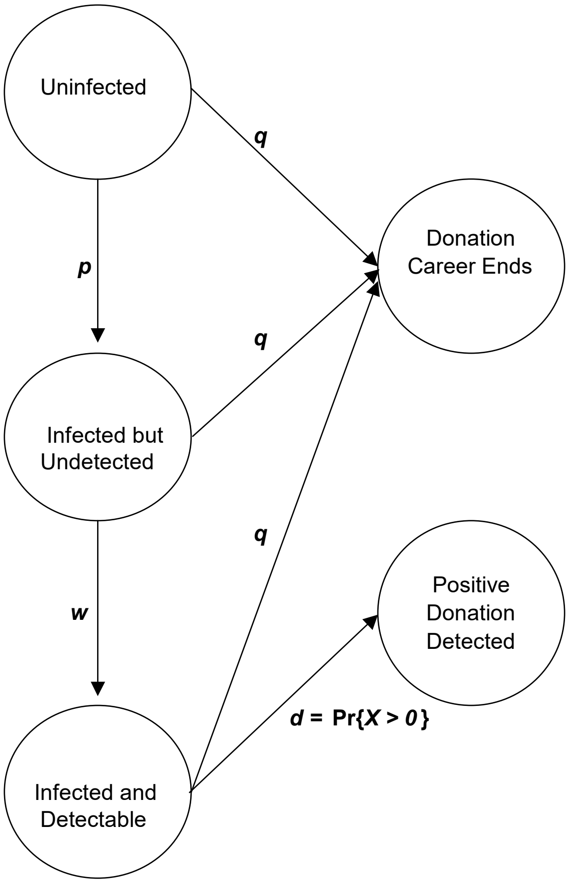
Let us discretize time to a weekly scale, and focus on a new paid plasma donor just beginning her plasma donation "career." We will assume that such new donors are HIV$^-$ as determined by "qualifying" HIV antibody tests. Each week, there is a probability p that this new donor will become infected with HIV (via whatever risk), a probability q that this donor will quit donating plasma, and a probability $1 - p - q$ that the donor will not have become infected and continue donating plasma. The actual number of plasma donations in a given week, X, follows a Poisson distribution with mean λ donations per week.

If a donor becomes infected, the infection is not detectable for a certain period of time (the "window period") until the HIV antibody level rises above the detection level of the antibody test employed. While a donor is infected but not detectable, we assume that paid plasma donations continue at following the same Poisson distribution with a mean of λ donations per week. We continue to assume that there also remains the same probability q of quitting each week, but now assume that the window period expires with probability w each week (and thus with probability $1 - w - q$ the donor continues in the next week to donate while infected but not detectable as infected).

If the donor does leave the window period (a probability w event) while still an active donor, we assume that donations continue in accord with a Poisson process, but that the infection will be detected if any infectious donations are made in a week. We will assume that all test results are determined at the end of the week, so for an infected donor who is out of the window period and thus able to be detected as infected, in a given week, either the donor quits donating (with probability q again), donates and is detected as infected (with a probability d that you will determine below), or continues to the next week without having quit (or having been detected) with probability $1 - d - q$.

(a) Carefully formulate this problem as a five-state Markov chain, where the states correspond to donors being uninfected and donating, infected but not detectable and donating, infected and detectable and donating, having voluntarily quit donating plasma, or detected as having made an infected plasma donation. Draw the state-transition diagram, and label the transition probabilities between states using the notation described above.

The state-transition diagram (ignoring self-loops) is shown below:



(b) An infected donor who is out of the window and hence able to be detected will get detected if any donations are made in a given week. Given that the weekly number of donations X is a Poisson random variable with mean λ donations, what is d, the probability that an detectable infected donor is actually detected?

This is simply given by $d = \Pr\{X > 0\} = 1 - \Pr\{X = 0\} = 1 - e^{-\lambda}$.

(c) What is the probability that a new paid plasma donor will quit donating plasma before becoming infected with HIV?

Directly from the state-transition diagram, this is just given by $q/(q+p)$.

(d) What is the expected number of undetectable, infected plasma donations made over the course of a paid plasma donation career?

First, the probability a new donor becomes infected before quitting is (by subtracting the answer to (c) from 1) $p/(q+p)$. Second, the expected time a donor spends in the infected but not detectable state is just equal to $1/(w+q)$ – why? Because the probability of leaving the infected but not detected state each week equals $w + q$, so we have a geometric distribution for the time during which undetectable infected donations can be made, with mean $1/(w+q)$ as claimed. Third, while donating, the expected number of donations per week is just λ. So, the expected number of undetectable, infected plasma donations over the course of a career equals

$$\frac{p}{p+q} \times \frac{1}{w+q} \times \lambda \cdot$$

(e) What is the probability that a new plasma donor will eventually end up infected and be detected via at least one HIV-infected plasma donation?

For this to happen, the donor must be infected, pass through the window, and then be detected. Directly from the transition diagram, this occurs with probability

$$\frac{p}{p+q} \times \frac{w}{w+q} \times \frac{d}{d+q}$$

where $d = 1 - e^{-\lambda}$ from part (b).

(f) Consider the following parameter values that are representative of paid plasma donors in the US (as well as the HIV antibody test): $p = 1.2 \times 10^{-5}$; $q = 0.01$; $w = 0.3$; and $\lambda = 1.3$ donations per week. There are about 13 million paid plasma donations each year. If there were no additional safety measures in place beyond basic antibody screening, how many infected plasma donations would enter the plasma supply? (HINT: what is the probability that a randomly selected plasma donation is infected and not detected? HINT for the HINT: on average, how many donations are there in a donation career, and of these, on average how many are infected but unable to be detected?)

First, what is the average number of donations per career? The answer is just given by

$$\lambda \times (\frac{1}{p+q} + \frac{p}{p+q} \times \frac{1}{w+q} + \frac{p}{p+q} \times \frac{w}{w+q} \times \frac{1}{d+q}).$$

But for the data in this problem, note that $p = 1.2 \times 10^{-5}$ is so small that the expression above essentially collapses to $\lambda/q = 1.3/.01 = 130$. Now, how about the expected number of infected but undetected donations per career? This is just given by

$$\lambda \times (\frac{p}{p+q} \times \frac{1}{w+q}) = 0.005.$$

Make sure you understand why these two expressions are correct!

So if there are on average 0.005 infected but undetected donations out of every 130 in a career, then a randomly selected donation will be infected but undetected with probability 0.005/130. So why worry? Because there are about 13 million paid plasma donations each year, so on average, the number of undetected but infected donations that would enter the plasma supply absent further safety measures would equal

$$13,000,000 \times \frac{0.005}{130} = 500.$$

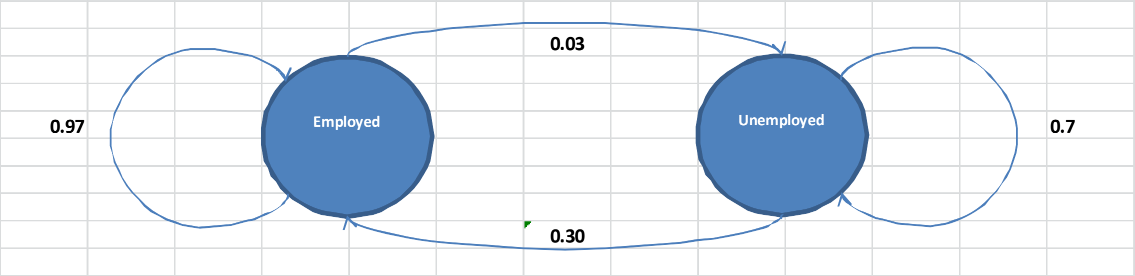
And infected plasma is bad news, since plasma products are pooled (e.g. Factor VIII for treating hemophilia, which is why there was a devastating HIV epidemic among hemophiliacs before blood and plasma donor screening was initiated).

6.1.3 Working For The Government

Suppose you are a well-educated modeler who earns \$20,000 per month working for the Bureau of Labor Statistics. Given the recent government shutdown you want to plan how much to set aside for a rainy day. Luckily you have accurate data for this. You know that your chance of being furloughed (that is, losing employment) after any month in which you were working is 3% (and your chance of staying employed in the next month is 97%). When unemployed, the probability that you remain unemployed in the next month is 70% (and with probability 30% the furlough ends and you get your job back). Thus, in any month you are either employed or unemployed, and this month you are employed. Your minimum necessary expenditures when unemployed (rent + chicken noodle soup, no movies, cappuccinos or fancy cocktails) are $12,500 per month.

(a) Clearly if you are employed this month, the chance you are unemployed next month is 3%, so you have a low expected cost from unemployment next month. But you are planning for the long-term! What is your equilibrium unemployment rate (probability of being unemployed in any single month over the long-run)?

First things first – this a Markov chain where the states correspond to being employed or unemployed, and the big picture is:



Now that we understand what is going on, we can answer the question. Let $p = \Pr\{\text{Unemployed}\}$ in the long run. That means that the probability of being employed must equal $1 - p$. To find p, we note that to be unemployed in some month far in the future (which happens with probability p by definition), it could be that you were employed in the prior month (that happens with probability $1-p$) and given that you lose your job in the next month with probability 0.03, or you could have been unemployed in the prior month (with probability p) and given that you don't regain your job in the next month with probability 0.7. Thus,

$$p = (1 - p) \times 0.03 + p \times 0.7$$

Solving for p yields

$$p \times (1 + .03 - 0.7) = 0.03, \text{or}$$

$$p = \frac{0.03}{0.33} = \frac{1}{11}.$$

(b) If you become unemployed next month, how much money should you have saved to cover your expected costs until you get your job back?

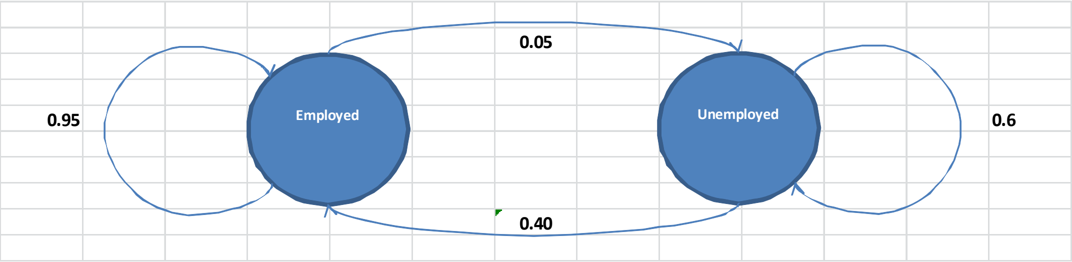
Well, once unemployed, you stay unemployed in the next month with probability 0.7 and get your job back with probability 0.3. This is like tossing a coin where the probability of heads equals 0.3, and you want to know how many times you have to toss the coin until you get heads for the first time. And we know the answer to that, since this is a garden variety geometric distribution (just like the Nation of Shoplifters). So, the expected duration of unemployment just equals the expected time until regaining employment, and that is given by $1/0.3 \approx 3.33$ months. Now, you have already determined that your minimum monthly expenditure is \$12,500, so your expected total expenditures over your duration of unemployment equals \$12,500 $\times E(\text{duration of unemployment}) = \$12,500/0.3 = \$41,666.67$.

(c) Good news: you have saved the amount of money indicated by your answer to (b)! Bad news: you have just been furloughed. What is the probability you will run out of money before you get back to work at the Bureau of Labor Statistics?

From part (b), you saved enough money to get by for about 3.33 months. If you are unemployed for 1, 2 or 3 months, you will have saved enough to get by. However, if your unemployment spell lasts four or more months, you will run out of money. So, the probability that you will run out of money is just the probability that you will be unemployed for more than three months, and that is given by $0.7^3 = 0.343$. Note that since hiring/firing occurs only at the end of each month, it is not possible to be unemployed for exactly 3.33 months in this model. This little example shows why it is not always a great idea to make decisions in accord with averages. Maybe instead of saving enough to meet expected expenses over an unemployment spell, you should instead save enough so that there is only a 5% (or maybe only a 1%) chance of running out of money.

(d) Now you are offered the opportunity to take a different position doing web-design. The chance you get re-employed in the next month following one you start as unemployed would go up to 40%, while your chance of being laid off in a month following one where you are working would equal 5%. What is the minimum monthly salary from the web-design job that would justify switching jobs and taking the offer? Please answer purely on the basis of whichever career would offer the highest expected monthly salary over the long run.

Let's start by first computing your expected monthly salary over the long run for your job at the Bureau of Labor Statistics. From (a) we learned that the chance you are unemployed in any month equals 1/11, which means that you are earning money in any month with probability 10/11, and consequently your expected monthly salary from the BLS equals $(10/11) \times \$20K = \$18,182$. Now let's consider the web-design job. This can also be represented as a Markov process, just like the BLS job, but now your probabilities of getting re-employed or laid off differ. Just changing the figure from part (a) to the new option yields:



So with the web-design job, your chance of getting rehired after being laid off are higher than with the BLS, but on the other hand there is also a higher chance of getting laid off in a month you are working. Following exactly the same logic as in part (a), the steady state probability of being employed is equal to $0.40/(0.40 + 0.05) = 8/9$. Suppose your monthly salary from web-design while working is equal to s. Then your expected monthly salary from web-design would just equal $(8/9) \times s$. In order for you to prefer the web-design job on the basis of having the highest expected monthly salary, it must be that

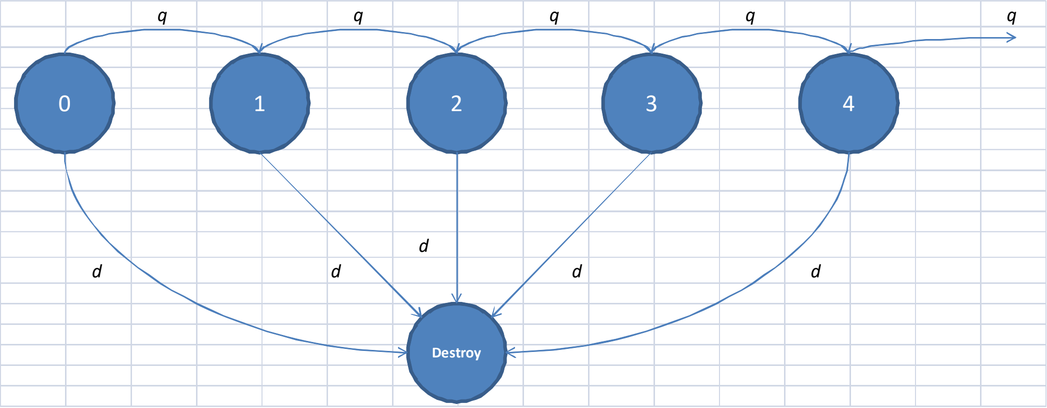
$$\frac{8}{9} \times s > \frac{10}{11} \times \$20,000 \approx \$18,182.$$

This means that you would prefer the web-design job if it offers a monthly salary $s > \frac{9}{8} \times \frac{10}{11} \times \$20,000 \approx \$20,455$, or about a 2.3% raise over your BLS salary.

6.1.4 More Cyberscadaddle

(d) Recall the HOL model, and suppose that upon infection of the HOL, the government discovers the cyber attack but cannot destroy it immediately. Specifically, suppose that starting from time 0 when the HOL is infected, the government is able to destroy the virus and halt all further SCADA infections with probability d per period. However, there remains a probability q per period that the virus spreads to the next vulnerable SCADA, and consequently a $1 - d - q$ probability that the situation remains unchanged to the next time period (so in any period, either the virus is destroyed, a new SCADA is infected, or nothing changes). Under these assumptions, determine the expected total number of additional SCADAs that are infected beyond the HOL until the government destroys the computer virus. Do this using a Markov chain model with states corresponding to the incremental number of infected SCADAs beyond the HOL plus a trapping state depicting destruction of the virus.

OK – my Markov chain is diagrammed below, minus the self-loops on each of the numbered states with probabilities $1-d-q$. The numbers correspond to how many SCADAs have been infected beyond the HOL. Note that there is a probability q that the chain advances from i to $i+1$ infected SCADAs per period (for $i = 0, 1, 2, \ldots$), and a probability d that the government destroys the cybervirus and all subsequent infection ends.



Now define τ as the expected number of SCADAs infected beyond the HOL. Starting with only an infected HOL, there are three things that can happen: the government destroys the virus with probability d in which case *no* SCADAs beyond the HOL are infected, the virus spreads to the next SCADA with probability q in which case one more SCADA gets infected, or nothing happens with probability $1 - d - q$ and the number of infected SCADAs remains the same. Using the repetition method, this leads to

$$\tau = d \times 0 + q \times (\tau + 1) + (1 - d - q) \times \tau$$

which simplifies to

$$\tau = q + \tau - d\tau$$

or

$$\tau = \frac{q}{d}.$$

(e) Look carefully at your answer to (d) above – it should be a very simple formula. Explain how you could have deduced this little result without resorting to a Markov model!

There is indeed a simple way to see this result. Ask yourself how much time passes before the government detects and destroys the cybervirus. Since the government will detect the virus in each period with probability d, the expected time until the virus is detected just equals $1/d$ time periods. Of course, the time until the government destroys the virus is the same as the time during which the virus can spread, and since the chance that the virus spreads equals q per time period, we have

$$\begin{aligned} E(\text{SCADAs infected beyond HOL}) &= \Pr\{\text{SCADA infected per period}\} \\ &\quad \times E(\#\text{ periods virus circulates}) \\ &= q \times \frac{1}{d} = \frac{q}{d}. \end{aligned}$$

Chapter 7

Queueing Models

7.1 Motivation

Queueing theory is used to model any situation that can broadly be construed as an interaction between "customers" and "servers." Canonical examples include waiting in line at a bank, post office, or a grocery store. More modern applications include call centers, hospital emergency departments, airport security, or the department of motor vehicles. But there are numerous other situations which, at first blush, look nothing like service systems but nonetheless have all the characteristics of customer/server interactions, which means that skill in *identifying* or *recognizing* different situations as queueing processes can lead to rapid understanding/prediction via application of queueing theory. Some examples of this include estimating the annual number of new HIV infections, or estimating the number of undetected terror plots-in-progress.

7.2 Questions and Answers in Queueing Processes

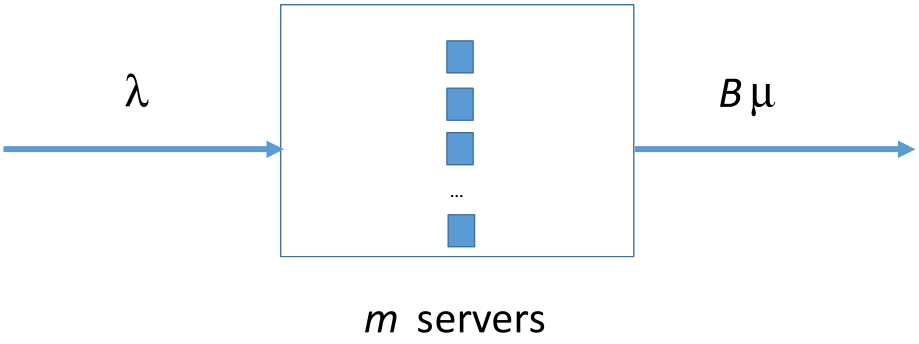
Typical questions queueing models try to answer include: on average how many servers are busy? how many customers are waiting in queue? on average, how long must a newly arriving customer wait to receive service? how many servers are necessary to ensure that the probability a customer waits longer than some pre-specified delay (e.g. 5 minutes) is less than some threshold (e.g. 5%)? if customers drop out before receiving service due to impatience (e.g. waiting for bagels) or worse (e.g. waiting for a public housing unit, waiting for a kidney), what fraction of customers actually do receive service versus the fraction that drop out? how many servers are needed to ensure that the fraction receiving service exceeds some threshold?

Queueing theory answers these questions via the construction of appropriate probability models. Naturally one makes certain assumptions, in which case the answers provided are specific to those assumptions. However some questions can be answered in great generality as we will see, while others do indeed need more assumptions. Typical assumptions made govern the arrival process (e.g. customers arrive in accord with a Poisson process, meaning that the probability a customer arrives in the next Δt time units is just proportional to Δt), the service process (e.g. servers are identical with service times that are independently and identically distributed according to some probability law, often taken as exponential), the queueing discipline (when a server becomes free, which customer gets served next – is service provided First-Come First-Served, Last-Come First-Served, in Random Order, by some Priority, etc.), and if reneging (dropout) is allowed, the reneging process (e.g. aggregate dropout occurs in proportion to the number of persons waiting in queue). Such assumptions enable direct mathematically-derived answers to questions like those posed above.

Finally, the most common queueing models focus on equilibrium (or steady-state) behavior, where steady-state here has the same meaning that it has in Markov Processes – enough time has passed for the process to settle into a form of probabilistic stability where the different system state probabilities (e.g. distribution of number of busy servers, or number of waiting customers) no longer change over time. There are also important time-dependent problems when arrival and/or service and/or dropout rates change over time, but we won't get into such problems in this class.

7.3 The Basics: Canonical System Description

Consider the figure below:



In this figure, λ represents the customer arrival rate (in customers per unit time) to the service facility (the big box). No customers renege (drop out) in this process. Inside the service facility, there are m identical servers (the blue boxes) that work independently, and each server has service capacity μ, meaning that a *busy* server is capable of processing μ customers per unit time (*idle* servers of course process nobody). Another way to interpret μ is as the reciprocal of the mean service time: on average, once a server begins working on a customer, on average that server requires $1/\mu$ time units to process the server.

Again, a single *busy* server processes μ customers per unit time. Suppose two servers are busy; together they process 2μ customers per unit time. There are a total of m servers, thus the maximum processing rate for the facility equals $m\mu$ customers per unit time. Clearly, this facility can only process all arriving customers per unit time if $\lambda < m\mu$ (why??).

Now, define B as the expected number of busy servers. Can we figure out what B equals? Sure we can. Recall that each busy server processes customers at rate μ per unit time. This being the case, if on average there are B busy servers, then the average processing rate of the facility must equal $B\mu$. However, this same processing rate must match the arrival rate λ. If $\lambda > B\mu$, the queue would explode, while if $\lambda < B\mu$, well, how could the rate customers leave the facility exceed the rate with which customers arrive on average? We thus conclude that:

$$\lambda = B\mu \tag{7.1}$$

which of course tells us what the expected number of busy servers, B, must equal:

$$B = \frac{\lambda}{\mu}. \tag{7.2}$$

A technical aside: this result is correct providing $\lambda < m\mu$. If instead $\lambda \geq m\mu$, then $B = m$, that is, all servers are always busy (why??).

7.4 Little's Theorem

In our canonical queueing model, we are given three pieces of information: the customer arrival rate (λ), the service rate (μ) (or equivalently the mean service time $1/\mu$), and the number of servers (m). We have already shown you how to find the average number of busy servers (B). Usually one is next interested in the *four fundamental quantities*: the average number of customers in queue (denoted by L_q), the average number of customers in the service system (denoted by L), the average time spent per customer waiting in queue (denoted by W_q), and the average time spent per customer in the service system (denoted by W). Here is an amazing result: if in addition to λ, μ, and m, you are given any single one of these four fundamental quantities, you can instantly deduce the values of the other three. This is due a bit of magic known as Little's Theorem, which states that:

$$L_q = \lambda W_q. \tag{7.3}$$

First, let's see why this equation is true. Then, let's use it to find the other three fundamental quantities. Finally, let's discuss why this is so intriguing.

Imagine for now that in our canonical system, customers are processed First-Come First-Served (they need not be for Little's Theorem to be true, but this is the easiest way to explain why it works). Suppose a new customer arrives to the system, and goes to the back of the queue. Just before joining, she looks ahead and counts the number of customers waiting. On average, how many are there? Answer: L_q, since by definition that's the average number of customers waiting in queue. Now, on average, how long must this new customer wait to begin service? Answer: W_q, since by definition that's the average waiting time per customer in queue. Now, our hero has finally arrived at the front of the queue. Just before entering service, she casts a furtive glance back towards the line she has just exited. How many customers does she see on average? Answer: L_q, again by definition. Now for the fun part: all of these L_q customers arrived after she did, and we know she waited W_q, so given that the customer arrival rate equals λ customers per hour, how many customers arrived while she was waiting? Answer: λW_q (!!). So, we have just seen that $L_q = \lambda W_q$.

Now, what about the other fundamental quantities? Let's start with L, the average number of customers in the system, which simply equals the average number of customers in queue (L_q) plus the average number of customers in service. But, for each customer in service, there is a busy server! This means that the average number of customers in service just equals B, from which we deduce:

$$L = L_q + B. \tag{7.4}$$

How about W, the average time a customer spends in the system? Well, W must be the sum of average time in queue (W_q) and average time in service. But, the average time in service is the same as the average service time, which equals $1/\mu$. We thus see that

$$W = W_q + \frac{1}{\mu}\cdot \tag{7.5}$$

Finally, let's see what happens if we multiply equation (7.5) above by λ. We get:

$$\begin{aligned} \lambda W &= \lambda(W_q + \frac{1}{\mu}) \\ &= \lambda W_q + \frac{\lambda}{\mu} \\ &= L_q + B = L \text{ (!!)} \end{aligned} \tag{7.6}$$

The result is that $L = \lambda W$, which is actually the formula most often used when referring to "Little's Theorem."

Now, note that if you are given any one of L, L_q, W or W_q, along with λ, μ and m, you can immediately deduce the other fundamental quantities. For example, suppose you know W_q. That means that you immediately know W (via equation 7.5), L_q (via equation 7.3), and L (via equation 7.6).

The power of Little's Theorem stems from the endless number of situations that can be construed as customers waiting for service. Suppose λ is the rate a new product enters a certain production stage, and W is the mean time spent per unit product in that production stage; then $L = \lambda W$ is the average work-in-process inventory for that production stage. If λ is the average rate with which airplanes take off, and W is the average time spent airborne per flight, then $L = \lambda W$ is the average number of airplanes in the sky. If λ is the annual mean number of illicit guns that enter circulation, and W is the mean time an illicit gun remains in use, then $L = \lambda W$ is the average number of illicit guns in circulation that are available for use. And so on.

7.5 A Little Pollution

Water flows through Lake Mephitic at the rate of 80,000 cubic meters of water per day. The volume of water in Lake Mephitic remains stable at approximately 40 million cubic meters. Also along the shore of Lake Mephitic is an industrial processing plant that dumps 0.16 tonnes per day of the completely soluble and foul-smelling pollutant from which Lake Mephitic derives its name; the pollutant is thus evenly mixed throughout the water in the lake (that is, you can assume that the concentration of the pollutant is the same throughout the lake). Ignoring evaporation of both water and pollutant, what is the average amount (in tonnes) of pollutant in the lake? (HINT: the name of this problem!)

Answering this question largely involves writing down the information in the problem. Focus first on the water. We are told that water flows through the lake at a rate of 80,000 cubic meters per day, and that the volume of water remains stable at 40 million cubic meters. Ignoring evaporation, what goes in equals what goes out, so equate the water inflow = outflow rate λ to 80,000 cubic meters/day. Also, identify the volume of water in the lake $L = 40$ million cubic meters. Relying on the hint, we recognize the water flow and lake volume relation as Little's Theorem and write

$$L = \lambda W$$

or

$$40 \times 10^6 = 80 \times 10^3 \times W$$

which means that W, the average residence time for water in the lake, is equal to

$$W = \frac{40 \times 10^6}{80 \times 10^3} = 500 \text{ days.}$$

Now, we are told that the industrial plant dumps 0.16 tonnes per day of pollutant into the lake, and that the pollutant is evenly mixed with the water in the lake. That means that on average, a unit of pollutant will spend the same amount of time in the lake as a unit of water (since the pollutant concentration is constant throughout the lake, and the pollutant is completely soluble – that is, it dissolves in the water, as opposed to forming something like an oil slick on top of the water and traveling at perhaps a different rate). So, we can now turn our focus from the water to the pollutant, recognize that the appropriate arrival rate for the pollutant is $\lambda = 0.16$ tonnes per day, and recognize $W = 500$ days as the mean residence time for the pollutant (from our earlier calculation plus the assumption of perfect solubility and even mixing). The average amount of pollutant in the lake, therefore, is equal to

$$L = \lambda W = 0.16 \times 500 = 80 \text{ tonnes.}$$

And apparently 80 tonnes of that stuff really stinks! (NOTE: problem borrowed from Harte, J. *Consider a Spherical Cow*, Sausalito, CA: University Science Books, 1988)

7.6 $M/M/m/\infty$ Queueing Models

Most of the time, sadly, you won't know one of the four fundamental quantities even if those are really what you are interested in. A reason for that is you are often interested in changing the number of servers in search of the optimal service level. So, it would be nice to be able to produce a *probability model* that yields the four quantities, and only depends upon the arrival rate λ, service rate μ, and number of servers m.

To do this, we need to make a few more assumptions. Here's what we'll assume for now: the customer arrival process will be assumed to be *Poisson*, as we discussed a few classes ago. So, in a time period of duration τ, the number of arriving customers will have a Poisson distribution with mean (and variance) equal to $\lambda\tau$. We will also assume that all service times follow the *exponential distribution*. This means that for any server, the probability that a service time T exceeds some duration t is given by $\Pr\{T > t\} = e^{-\mu t}$. It is still the case that the mean service time $E(T) = 1/\mu$. Finally, we will continue to assume that there are m servers, each working independently and with identical exponential service time distributions.

These assumptions give rise to the $M/M/m/\infty$ family of queueing models. The first "M" refers to "memoryless" (or "Markov"), and means that the arrival process is Poisson. The second "M" refers to the exponential service time distribution assumption (see "Is it live or is it memoryless?" in the Poisson process handout). The "m" refers to the number of servers, while the "∞" refers to the waiting capacity in the system (so there is no limit on the number of people in queue).

The simplest instance of such a model is the single server queue ($M/M/1/\infty$). For this model there are well-known results: assuming Poisson arrivals, and a single server who processes customers with exponential service times averaging $1/\mu$, define $\rho = B = \lambda/\mu$ as the server's *utilization*, that is, the fraction of time the server is busy (which is the same as the expected number of busy servers when there is only one server to begin with – why???). One can prove that:

$$W_q = \frac{\rho/\mu}{1-\rho} \text{ providing } \rho < 1 \tag{7.7}$$

So, given λ and μ, you can find ρ. From μ and ρ, equation (7.7) tells you W_q. And, once you have W_q you can find the other fundamental quantities from equations (7.4)–(7.6).

7.6.1 Example: Small Town in Iowa with One Ambulance

Suppose in a small town in Iowa with exactly one ambulance, on average there is one call-for-service per hour (so $\lambda = 1$), while the mean service time equals a half hour (so $1/\mu = 1/2$, which means $\mu = 2$). The utilization $\rho = \lambda/\mu = 1/2$, thus by equation (7.7) we see that the average time spent waiting for an ambulance equals

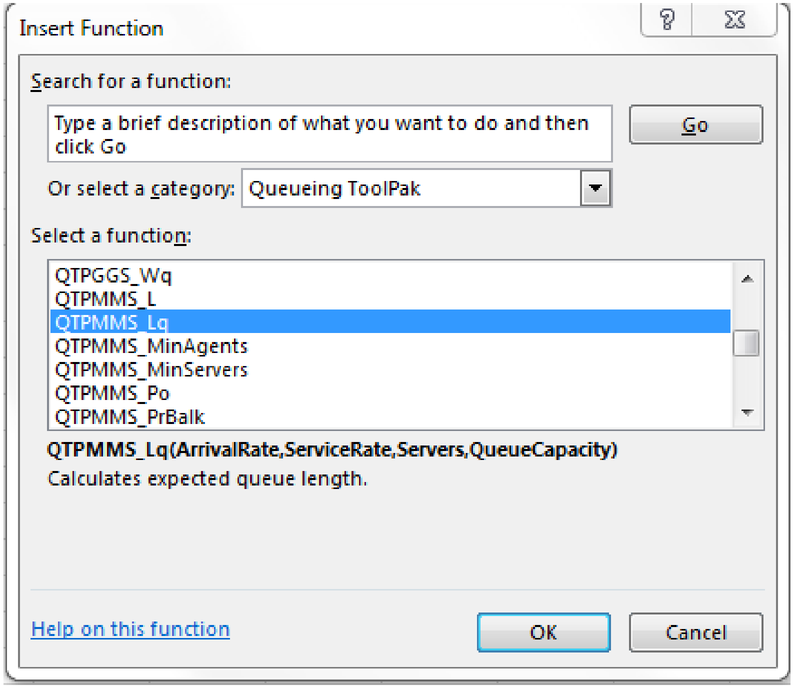
$$\frac{(1/2)/2}{1-1/2} = \frac{1}{2} \tag{7.8}$$

or 30 minutes. So even though the ambulance is only busy half of the time, on average a new "customer" (accident victim? heart attack victim?) must wait 30 minutes until the ambulance arrives!

7.6.2 $M/M/m/\infty$ and Queueing Toolpak

Now we will keep the same assumptions as the ambulance example, except we will allow an arbitrary number of servers, denoted by m. There are also formulas for this model, and they are displayed in "Queueing Theory's Greatest

Hits" in the coursepack, but they are admittedly ugly formulas. However, it is easy to employ the $M/M/m/\infty$ model via the Queueing Toolpak (which can be downloaded from http://queueingtoolpak.org/download.shtml). The toolpak functions all begin with QTP and are very easy to use. For example, to find L_q for the $M/M/m/\infty$ model, one uses the Queueing Toolpak function QTPMMS_Lq. You can access it in Excel as:



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Queueing ToolPak

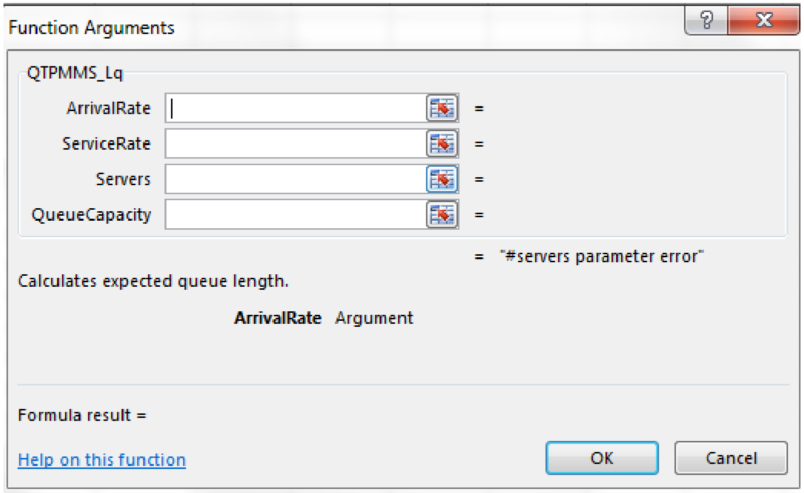
Select a function:

QTPGGS_Wq
QTPMMS_L
QTPMMS_Lq
QTPMMS_MinAgents
QTPMMS_MinServers
QTPMMS_Po
QTPMMS_PrBalk

QTPMMS_Lq(ArrivalRate,ServiceRate,Servers,QueueCapacity)

Calculates expected queue length.

Once you click "OK" you'll see:



Function Arguments

QTPMMS_Lq

ArrivalRate =
ServiceRate =
Servers =
QueueCapacity =

= "#servers parameter error"

Calculates expected queue length.

ArrivalRate Argument

Formula result =

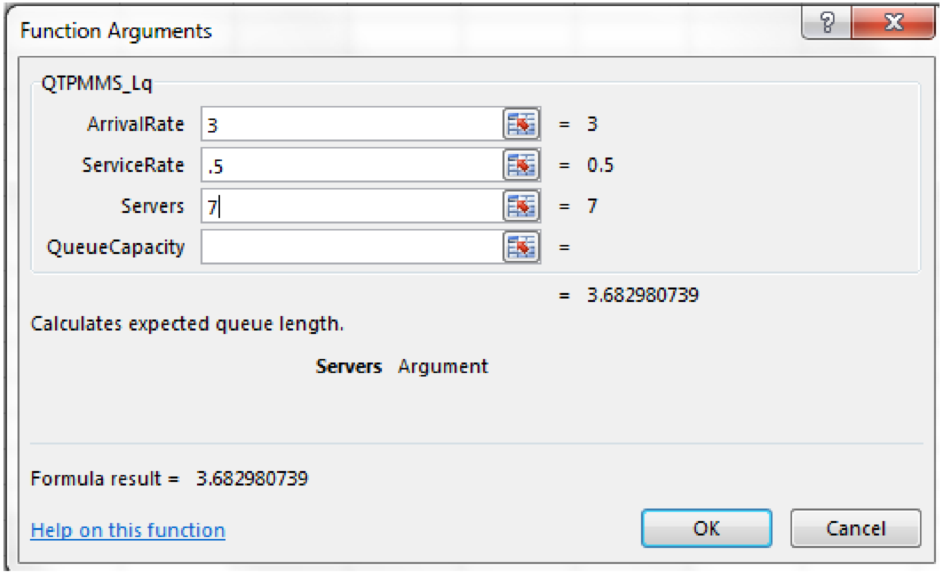
Just enter the required inputs and click “OK.” Note: just leave the QueueCapacity blank, and it will understand that you have unlimited queue capacity.

7.7 Examples

7.7.1 Winnie the Queue

At the local library, requests arrive for a copy of Winnie the Pooh at an average rate of 3 requests per week. The average circulation time equals 2 weeks per check-out, circulation times are exponentially distributed, and the library presently has 7 copies of the book. What is the average number of requests waiting for a copy of Winnie the Pooh?

Answer: well, this is an $M/M/m/\infty$ model with $\lambda = 3/\text{wk}$, $\mu = 0.5/\text{wk}$ (since the mean circulation time is 2 weeks), and $m = 7$ books. So, entering these quantities into the Queueing Toolpak function QTPMMS_Lq yields:



Function Arguments

QTPMMS_Lq

ArrivalRate	3	= 3
ServiceRate	.5	= 0.5
Servers	7	= 7
QueueCapacity		=

= 3.682980739

Calculates expected queue length.

Servers Argument

Formula result = 3.682980739

So on average there are 3.68 people waiting for a copy of Winnie the Pooh. How long does someone have to wait on average to get a copy? The answer is given by

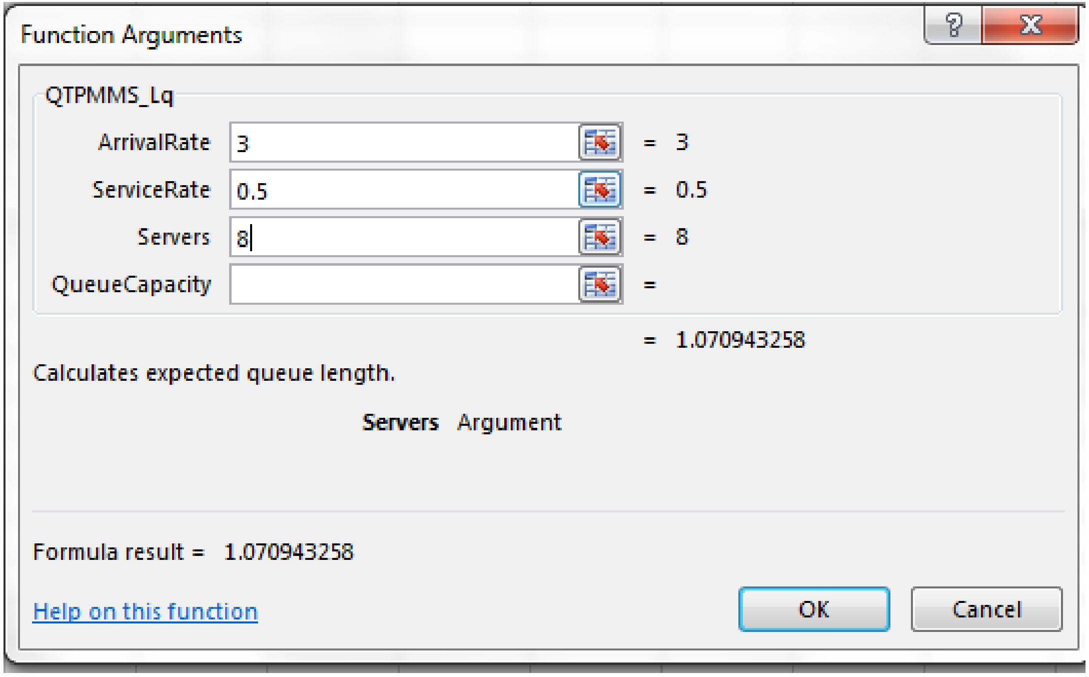
$$W_q = \frac{L_q}{\lambda} = \frac{3.68}{3} = 1.23 \text{ weeks.}$$

Bummer.

Now suppose that you are told that the library imagines an average cost of waiting equal to \$5/per waiting customer (most likely a kid)/month, while it costs \$20 per year to maintain a single copy of the book. How many copies should the library stock? To solve this, note that if you have m books, then it costs \$20 × m to maintain them annually. Also, the average number of waiting "customers" would now equal $L_q(m)$, that is, the average queue length depends on the number of servers! And, if you have an average of $L_q(m)$ waiting customers, the waiting cost imagined by the library equals \$5 × $L_q(m)$ per month, or \$60 × $L_q(m)$ per year. So, the goal is to choose m, the number of copies, to minimize

$$\$20 \times m + \$60 \times L_q(m)$$

You can do this by just using the Queueing Toolpak and plugging in different numbers of m. At the present number of copies (7), total costs are given by $20 \times 7 + 60 \times 3.68 = \$360.8$ per year. Increasing the number of books to 8 you discover that



Function Arguments

QTPMMS_Lq

ArrivalRate 3 = 3

ServiceRate 0.5 = 0.5

Servers 8 = 8

QueueCapacity =

= 1.070943258

Calculates expected queue length.

Servers Argument

Formula result = 1.070943258

and total costs drop to $20 \times 8 + 60 \times 1.07 = \$224.2$. Increasing to 9 copies, proceeding in the same fashion you will discover that total costs drop further to \$203.52 (the average number waiting only equals 0.392). Ten copies, however would raise costs to \$209.12 (though only 0.15 would be waiting on average). So, the optimal number of copies is equal to 9.

7.7.2 Gun Control

Violence due to guns is prevalent in our society. One policy proposed to curb such violence is gun control. One method of gun control would be to severely limit the sales of new guns, decreasing the rate with which new guns are introduced to the population of guns in circulation. Suppose that at present, the rate with which new guns enter circulation is given by λ guns per year. Also, assume that once in circulation, a gun remains in use for an average τ years.

(a) On average, how many guns are circulating in society?

This is simple: if the arrival rate is λ guns per year, and the mean circulation time is τ years, then the average number of guns in circulation simply equals $\lambda\tau$. Pretty easy!

(b) Suppose that the passage of a gun control act reduces the rate with which new guns enter circulation by 80%. However, in response to gun control, those using guns find ways to increase the length of time guns remain functional, effectively increasing the average gun circulation time. By how much would the average gun circulation time have to increase in order to exactly offset the benefits of gun control (in terms of the number of guns in circulation)?

Let's let g equal the current average number of guns in circulation. If λ is reduced by 80% due to gun control, how much does τ have to increase to offset the benefits? Clearly we wish to keep g constant, so the average circulation time has to increase by a factor of k where

$$g = \lambda\tau = (1 - .8)\lambda \times k\tau.$$

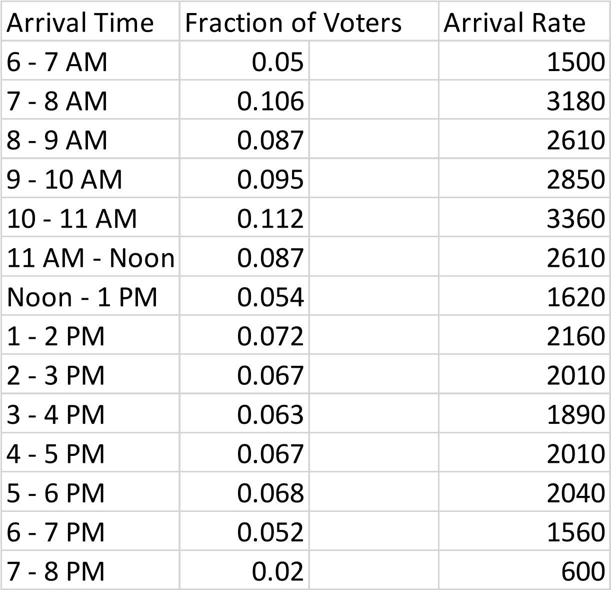
So we see that $k = 1/(1 - .8) = 1/.2 = 5$. Guns would have to remain in circulation 5 times as long to offset the benefits.

7.7.3 Election Queues

The Town of Hamden is expecting 30,000 in-person voters on election day. Polls will be open from 6 AM through 8 PM, and based on past voter behavior the town expects that the fraction of all voters who will arrive in each hour to vote is as shown in the table below. The time necessary to vote averages 5 minutes. The town wants to staff a sufficient number of booths each hour such that on average, voters do not have to wait more than 1 minute in queue before voting. Presuming that steady state conditions govern system performance during each hour, determine the smallest number of voting booths that must be operational in each hour. You may assume that voters arrive in accord with a Poisson process, and that the actual time required to vote follows an exponential distribution.

Hour	Fraction of Voters
6 - 7 AM	0.050
7 - 8 AM	0.106
8 - 9 AM	0.087
9 - 10 AM	0.095
10 - 11 AM	0.112
11 AM - Noon	0.087
Noon - 1 PM	0.054
1 - 2 PM	0.072
2 - 3 PM	0.067
3 - 4 PM	0.063
4 - 5 PM	0.067
5 - 6 PM	0.068
6 - 7 PM	0.052
7 - 8 PM	0.020

The first thing you need to do is compute the hourly voter arrival rates. This simply means multiplying the fraction of voters in the table above by 30,000, which is the expected total number of voters on election day. You should discover the following:



Arrival Time	Fraction of Voters		Arrival Rate
6 - 7 AM	0.05		1500
7 - 8 AM	0.106		3180
8 - 9 AM	0.087		2610
9 - 10 AM	0.095		2850
10 - 11 AM	0.112		3360
11 AM - Noon	0.087		2610
Noon - 1 PM	0.054		1620
1 - 2 PM	0.072		2160
2 - 3 PM	0.067		2010
3 - 4 PM	0.063		1890
4 - 5 PM	0.067		2010
5 - 6 PM	0.068		2040
6 - 7 PM	0.052		1560
7 - 8 PM	0.02		600

Now you need to determine the smallest number of voting booths in each hour to ensure that the average waiting time in queue, W_q, is less than one minute.

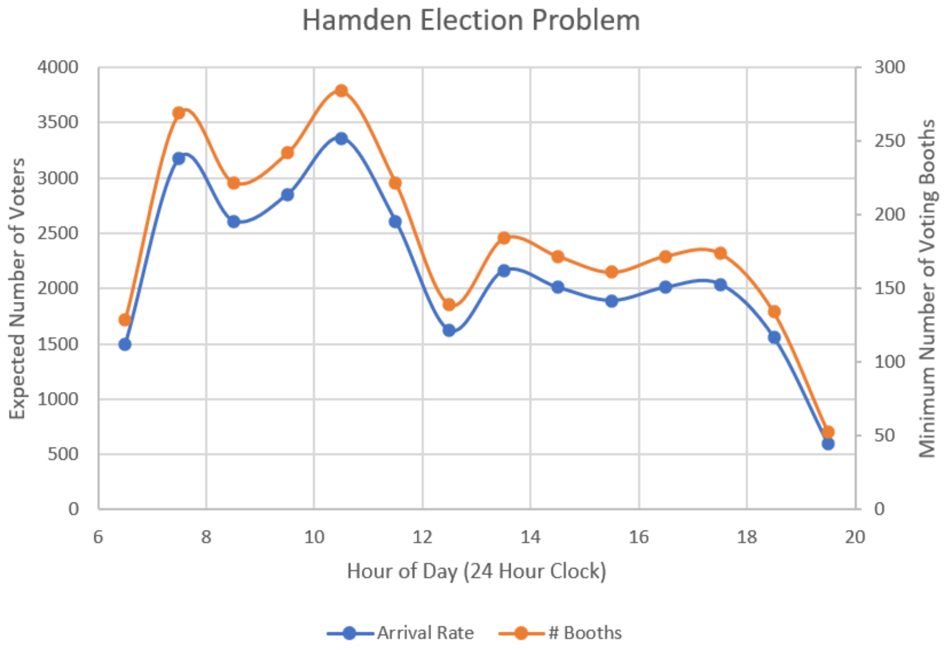
Here's a three-step approach to solving this problem:

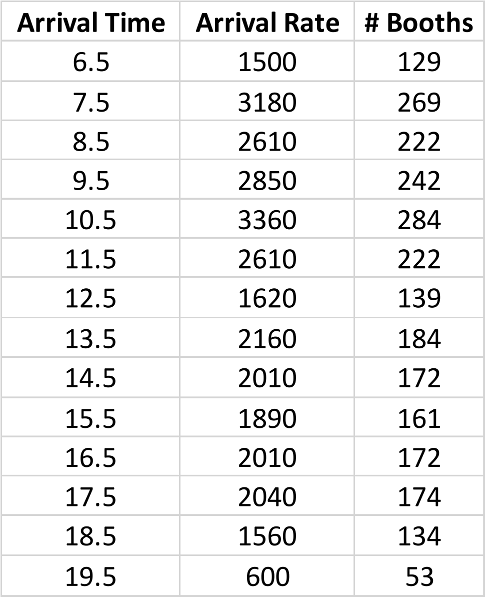
(a) For each hour let λ be the expected number of voters arriving that hour, and set $1/\mu = 5$ minutes which means that the service (voting) rate $\mu = 12/\text{hr}$.

(b) Set m, the number of servers (voting booths), to the smallest integer greater than $\lambda/\mu = \lambda/12$. This is the smallest number of voting booths you can consider and still be in steady state.

(c) Using either the Queueing Tool Pak function QTPMMS_Wq or the QUEUEBOOK equivalent, evaluate the value of W_q. Is it ≤ 1 minute (or 0.016666 hours)? Since the arrival and service rate information are in hours while your waiting time target is in minutes, you can make this easier on yourself if you multiply the resulting value of W_q from whatever queueing package you are using by 60; then you can just ask if you got a number less than one. If the answer is yes, STOP. If not, try another value of m.

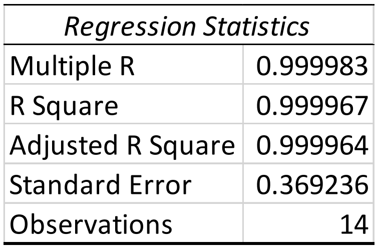
Now, there are boring and cute ways to search for the smallest value of m that works. The boring way is to just keep incrementing m by one over the value you start with in (b), and keep going until $W_q \leq 1$ minute. One cute way is to use a "binary chop" search (i.e. bisection). Assuming the value of m in (b) (call this $m^{(1)}$) is too small (that is W_q is too large), try a larger value of m; call this $m^{(2)}$. Now suppose this results in $W_q < 1$ minute. Would a smaller value of m work? Set $m^{(3)}$ = smallest integer $\geq (m^{(1)} + m^{(2)})/2$ and try that. If it is still true that $W_q < 1$ minute, you know that all values of $m > m^{(3)}$ are too big as well. Now set $m^{(4)}$ = smallest integer $\geq (m^{(1)}+m^{(3)})/2$. Suppose it turns out that for this number of servers $W_q > 1$ minute. Then $m^{(4)}$ is too small, and so is any value of $m < m^{(4)}$. So what next? You set $m^{(5)}$ = smallest integer $\geq (m^{(4)} + m^{(3)})/2$. And you keep going until you hit the smallest integer m such that $W_q \leq 1$ minute. *This approach cuts the number of eligible solutions in half at each step*! As it turns out, for this problem, the boring method is just as fast, so there are no real efficiency gains (though there could be for other problems...). In particular, for those of you using the QTPMMS_Wq function in Excel, simply incrementing m until you get the waiting time $W_q < 1$ minute is easy, as you can set up a spreadsheet to do the calculations in parallel (that is, simultaneously vary W_q in all time periods as a function of m). My solutions appear below:



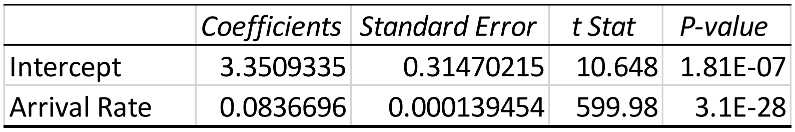


Arrival Time	Arrival Rate	# Booths
6.5	1500	129
7.5	3180	269
8.5	2610	222
9.5	2850	242
10.5	3360	284
11.5	2610	222
12.5	1620	139
13.5	2160	184
14.5	2010	172
15.5	1890	161
16.5	2010	172
17.5	2040	174
18.5	1560	134
19.5	600	53

Note that the relationship between the required number of voting booths and the arrival rate is practically linear in this application. In fact, a regression of the number of voting booths required versus the arrival rate yields the results:



Regression Statistics	
Multiple R	0.999983
R Square	0.999967
Adjusted R Square	0.999964
Standard Error	0.369236
Observations	14



	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.3509335	0.31470215	10.648	1.81E-07
Arrival Rate	0.0836696	0.000139454	599.98	3.1E-28

This is interesting – it says that in essence, $m \approx 3.35 + \lambda/\mu$ since the mean service time $1/\mu$ of 5 minutes equals 0.833 hours (the regression coefficient on the arrival rate is 0.0837, really close!). So basically, all you need to do to satisfy the requirements is figure out the expected number of occupied voting booths λ/μ, and add 4 as an insurance measure (since 4 is the smallest integer greater than the intercept of 3.35). Note that in the actual solution above, the average difference between the minimum number of booths required to keep the waiting time in queue under a minute and the expected number of occupied booths λ/μ equals 4.07.

7.7.4 Big Data at the NSA

According to *Washington Post* investigative reporters, "Every day, collection systems at the National Security Agency intercept and store 1.7 billion e-mails, phone calls and other types of communications." (D. Priest and W. M. Arkin, 2010, "National Security Inc.," *Washington Post*, (July 20) A1.) Suppose that the NSA intercepts communication messages (SIGINT) in accord with a Poisson process where the mean message arrival rate is as stated above. Assume that the NSA initially processes such intercepts for further study (or discard) using 341 automated algorithmic data analyzers (i.e. dedicated computers), each of which is capable of analyzing and classifying at most 5 million intercepts per day. Assume further that the duration of time required for a single data analyzer to process a single intercept follows an exponential distribution (equivalently, when busy, a data analyzer processes intercepts in accord with a Poisson process).

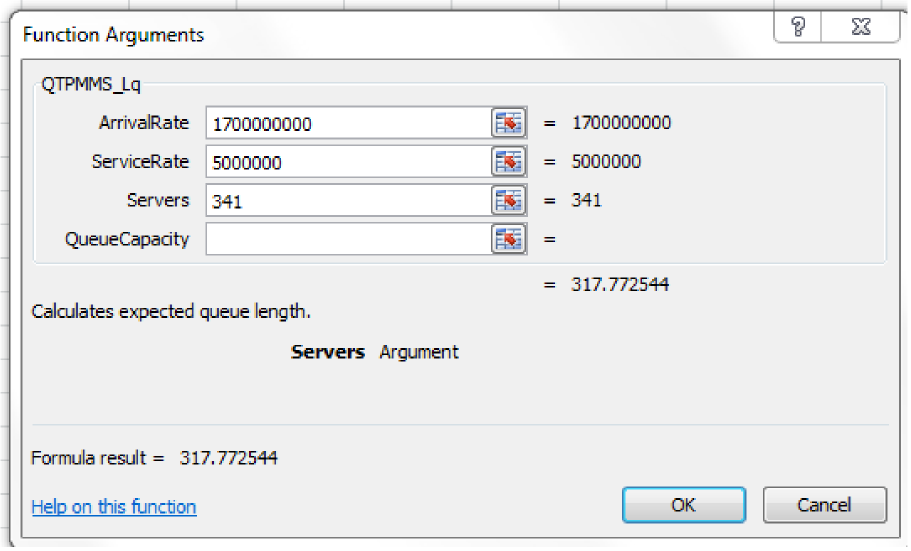
(a) What is the standard deviation of the daily number of messages intercepted by the NSA?

Since new messages are intercepted in Poisson fashion with mean 1.7 billion per day, and since for any Poisson variable, the variance equals the mean, we have that the variance of the daily number of messages intercepted also equals 1.7 billion. The standard deviation is just the square root of the variance, which is $\sqrt{1.7 \text{ billion}} = \sqrt{1,700,000,000} \approx 41,231$. The only thing tricky here is that you have to remember to take the square root of the "billion" and not just the 1.7!

(b) What is the expected number of intercepts in queue waiting for an available data analyzer?

The problem statement clarifies that in addition to the Poisson arrival rate of 1.7 billion messages per day, there are 341 "servers" each of which is capable of processing up to 5 million intercepts per day. That the service time required to process a single intercept is exponentially distributed (equivalently, that the analyzers process intercepts in Poisson fashion when busy) coupled with the fact that the customers are intercepted e-mails who are not going to drop out on account of impatience implies that the entire system can be modeled as an $M/M/m/\infty$ queue with $m = 341$, $\lambda = 1.7$

billion per day, and $\mu = 5$ million per day. This is great, because you can use the Queueing ToolPak QTPMMS_Lq function to figure out the expected number of intercepts waiting in queue! I get:



Function Arguments

QTPMMS_Lq

ArrivalRate 1700000000 = 1700000000

ServiceRate 5000000 = 5000000

Servers 341 = 341

QueueCapacity =

= 317.772544

Calculates expected queue length.

Servers Argument

Formula result = 317.772544

So there are, on average, 317.77 or about 318 intercepts waiting in the NSA queue for initial processing. The probability that all 341 servers are busy turns out to equal 0.935. There's a lot going on at the NSA!

Note that if you did not have access to the Queueing ToolPak, you could also have determined the answer directly from the formula for L_q that applies for the $M/M/m/\infty$ queue as it appears in Queueing Theory's Greatest Hits in your course pack:

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

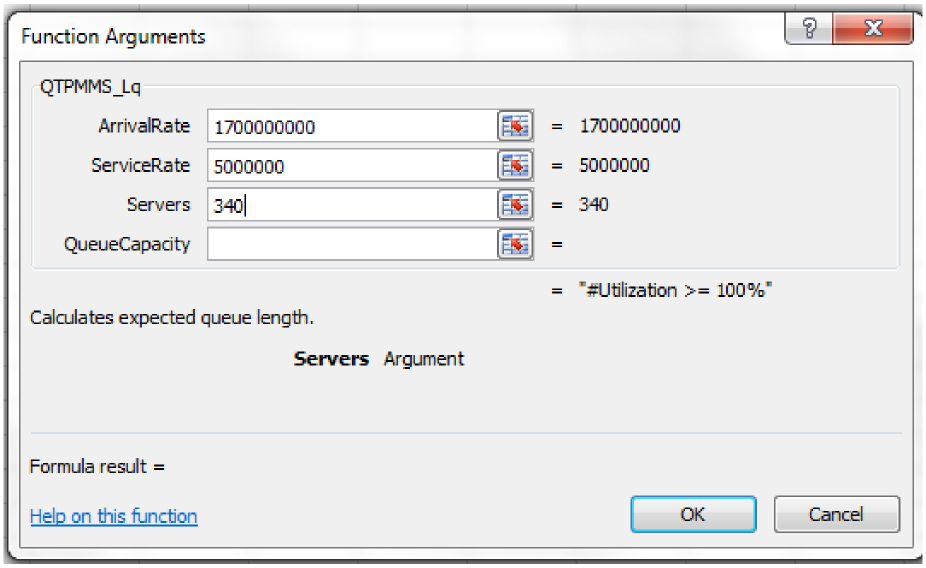
where

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

It took me about 5 minutes to implement the formulas above directly in Excel, and guess what? I get $L_q = 317.77$. So there.

(c) A newly transferred human intelligence (HUMINT) specialist has, thanks to the recent government shutdown, been placed in charge of the NSA's SIGINT collection efforts. Bored, the new boss hijacks one of the 341 automated data analyzers from SIGINT intercept processing to play Candy Crush, leaving 340 servers to handle incoming NSA intercepts from around the globe. What happens to the queue of NSA intercepts awaiting initial processing? Be precise.

So what happens to national security when the head of NSA SIGINT collection decides to play Candy Crush? Well, since each server can process at most 5 million intercepts per day, reducing the number of servers from 341 to 340 means that the total daily NSA intercepted message processing capacity is reduced to 5 million × 340 $= 1,700,000,000 = 1.7$ billion. Uh oh...for queues of this type, if the maximum total processing rate exactly equals the arrival rate, the queue explodes! To illustrate, watch what happens inside the Queueing ToolPak if you try to calculate the expected number of messages in queue with only 340 servers:



QTPMMS_Lq

ArrivalRate	1700000000	= 1700000000
ServiceRate	5000000	= 5000000
Servers	340	= 340
QueueCapacity		=

= "#Utilization >= 100%"

Calculates expected queue length.

Servers Argument

Formula result =

Help on this function

Not good. You can also see this directly from the equation for L_q shown above: if $\lambda = m\mu$, your formula for L_q involves division by zero, and you know what that does (to really see what happens, you need to plug the formula for P_0 into the formula for L_q and see what happens; you still end up dividing by zero). Moral: if you are in charge of the NSA, don't play Candy Crush.

7.7.5 A Backup Service Model

The SCII (Small City in Iowa) hospital system currently fields three ambulances. Calls arrive in Poisson fashion at an average rate of 10 per hour. Each ambulance requires an average of 20 minutes to service a call. If all three ambulances are busy, the SCII staff alerts the Big Bucks Backup Corporation (BBBC), which immediately dispatches one of its many, many, many private ambulances. BBBC ambulances also require an average of 20 minutes to service a call.

(a) What percentage of the time are all three SCII ambulances busy?

This is an instance of the Erlang Loss Model – i.e. the $M/G/3/3$ queue. The analysis is similar to our homeless shelter and (the first part of the) ICU examples from class. Arrivals occur at rate $\lambda = 10$/hour, while the average service time equals 20 minutes, which means that the service rate μ equals 3 per hour. Using the Erlang loss formula, the percentage of time that all three SCII ambulances are busy is given by:

$$P_3 = \frac{(\lambda/\mu)^3/3!}{\sum_{k=0}^{3}(\lambda/\mu)^k/k!} = \frac{6.17}{1+3.33+5.56+6.17} = 0.38.$$

So we see that all three ambulances are busy 38% of the time. You could also have obtained this result using the function QTPMMS_PrFull(10,3,3,0).

(b) What is the expected number of busy BBBC ambulances servicing SCII calls?

Let's look at a "block diagram" of this system:

From the picture above, it is clear that the expected number of busy BBBC ambulances, call this B(2), is given by

$$B(2) = \frac{3.8}{3} = 1.27$$

(c) BBBC makes money doing this sort of thing. If the cost per busy BBBC ambulance per minute is *twice* the overall cost per minute for SCII ambulances (i.e. SCII pays for their own ambulances on a 24 hour basis, but only pays BBBC ambulances on an as needed basis), and neglecting the fixed

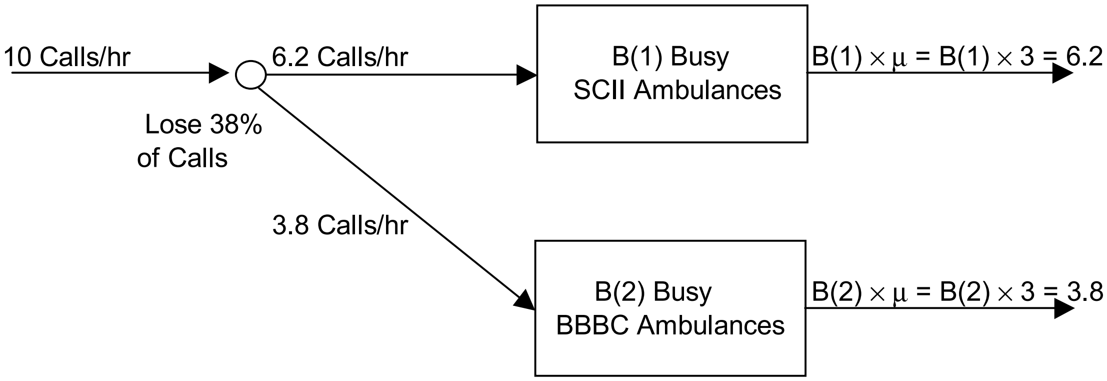

10 Calls/hr
6.2 Calls/hr
Lose 38% of Calls
3.8 Calls/hr
B(1) Busy SCII Ambulances
B(1) × μ = B(1) × 3 = 6.2
B(2) Busy BBBC Ambulances
B(2) × μ = B(2) × 3 = 3.8

Figure 7.1:

costs of purchasing new ambulances, how many ambulances should SCII field to minimize expected costs per unit time?

Well let's see. The total cost per hour will equal the sum of the costs for SCII ambulances (primary) plus the costs for BBBC ambulances (backups). Let c be the cost per hour for SCII ambulances, and suppose we staff m of them. Then the costs per hour for the SCII fleet will just equal cm on average. Now, if we have m ambulances, then the probability that all of them will be busy is given by the Erlang loss formula

$$P_m = \frac{(10/3)^m/m!}{\sum_{k=0}^m (10/3)^k/k!}.$$

This means that the arrival rate on calls to the BBBC ambulances will equal $10 \times P_m$, and consequently the expected number of busy BBBC ambulances serving SCII calls will equal $10 \times P_m/3$ (see the block diagram above). Now, the costs of a busy BBBC ambulance are twice the average cost per SCII ambulance, so the total costs of using BBBC ambulances when we staff m SCII vehicles is given by $2 \times c \times 10 \times P_m/3$. So, the problem is:

$$\min_m cm + 2 \times c \times 10 \times P_m/3.$$

This is easily accomplished in a spreadsheet, either using the formula for P_m above, or more conveniently using QTPMMS_PrFull(10, 3, m, 0). Note that the specific value of c is immaterial, so we can set it to one. I obtain:

m	P_m	Total Cost
0	1	6.67
1	0.77	6.13
2	0.56	5.73
3	0.38	5.53
4	0.24	5.6

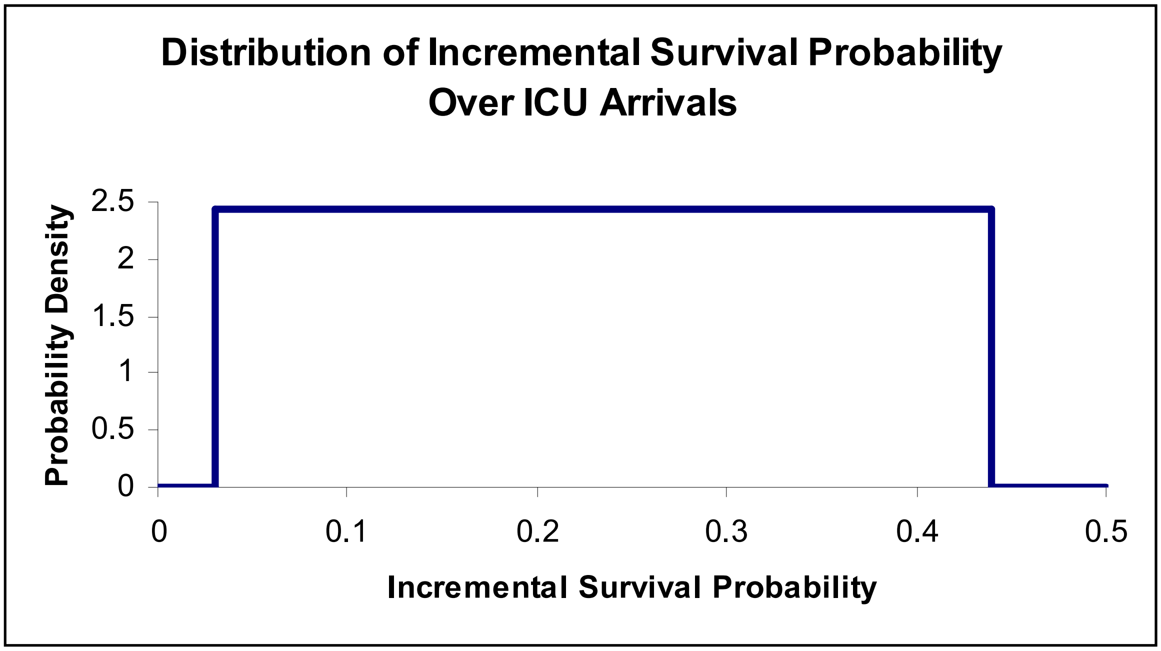
So, set $m = 3$ – hey! They are already operating optimally!

7.7.6 Optimizing ICU Admissions

Each arrival to the ICU receives an APACHE score (APACHE stands for Acute Physiological and Chronic Health Evaluation); the APACHE scores can in turn be converted to survival probabilities conditional upon whether the patient is, or is not, admitted to the ICU. Figure 1 on p. 134 of the paper shows the survival probabilities as a function of the APACHE scores with (top curve) or without (bottom curve) the ICU derived from data collected at the Hebrew University-Hadassah medical center in Jerusalem. Note that for any APACHE score, patients admitted to the ICU always have a higher survival probability than patients not admitted (the top curve is always higher than the bottom curve). Also note that whether admitted to the ICU or not, the survival probability declines as the APACHE score increases – those with the lowest APACHE scores are in good shape, while those with the highest scores are in serious trouble.

To determine the incremental survival probability gained from admission to the ICU as a function of the APACHE score, simply subtract the bottom curve from the top curve in Figure 1. The result is shown in Figure 2 on p. 134 of the article. Note that the largest incremental gains in survival accrue to those with intermediate APACHE scores, rather than those at either end of the severity spectrum.

Now, there is a distribution of APACHE scores across all those who arrive at the ICU (see Figure 6 on p. 135). This distribution can be converted to a distribution of incremental survival probabilities by filtering the probability distribution of APACHE scores in Figure 6 through the transformation from APACHE scores to incremental survival probability offered in Figure 2. It turns out that after doing this, the distribution of incremental survival probabilities across ICU arrivals is *uniformly distributed* between 0.03 and 0.44. That is to say, a randomly sampled ICU arrival would, if admitted to the ICU, gain an incremental survival benefit (probability) that is equally likely to fall between 0.03 and 0.44. The fraction of all arrivals with survival benefit of at least b is shown as a function of b in Figure 3 on p. 134. Presented in probability density form, the uniform distribution of survival benefits across ICU arrivals appears as shown below. Note that the average survival benefit from this distribution is simply given by $(0.03 + 0.44)/2 = 0.235$, that is, a randomly selected ICU arrival would gain an additional 23.5% chance of surviving – that's pretty good (which is why we have ICUs!).



Deterministic Reasoning: Flow Analysis

NOTE: DETERMINISTIC SECTION SHOULD BE IN FLUID MODELS SECTION

Let λ denote the arrival rate to the ICU; μ represent the service rate (departure rate) per ICU bed; and m denote the number of beds in the ICU. For the example in the paper, $\lambda = 2.297$ arrivals/day; $\mu = 0.164$ departures/bed/day; and $m = 8$ beds. The capacity of the ICU, that is, the maximum rate with which patients can be processed (and hence with which patients can be admitted) is equal to $m\mu = 8 \times 0.164 = 1.\,31$ persons per day.

First Come First Served

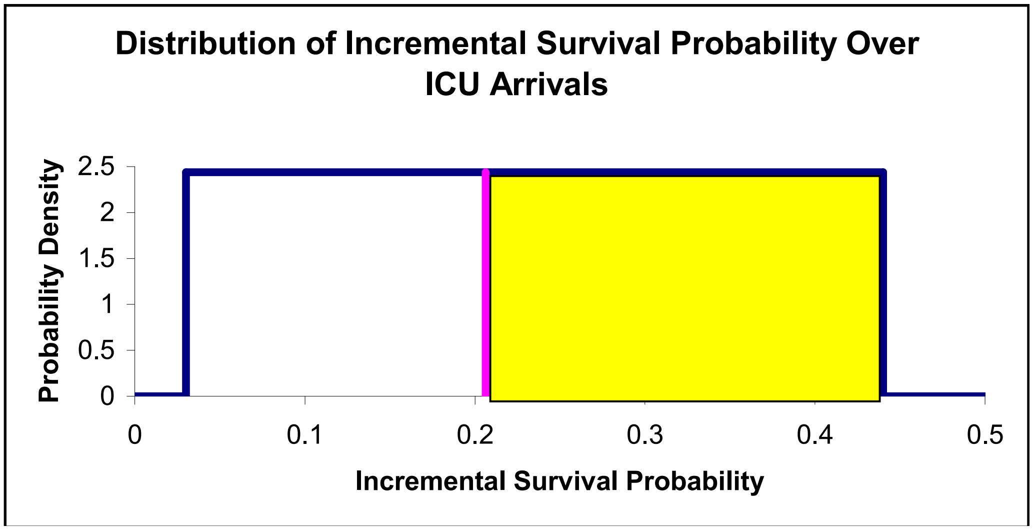
Suppose that any patient who arrives is allowed entry if there is an empty bed; otherwise the patient is denied admission. Since there is no discrimination among patients, any patient admitted would be expected to gain the average survival benefit of 0.235. Thus, the annual expected number of incremental lives saved on account of operating the ICU would equal

$$\text{Annual Lives Saved } = 1.31 \times 0.235 \times 365 = 112.4.$$

So the deterministic benchmark suggests that the ICU would save 112.4 lives per year on average if ICU admission was strictly first come first served.

Cherry Picking

However, if the idea is to save as many lives as possible, why not admit those with the highest incremental survival benefits within the bed-capacity constraint? There are 2.297 arrivals per day, but the ICU can only process 1.31 per day. Note that $1.31/2.297 = 0.57$, so deterministic reasoning would suggest "cherry picking" the 57% of the arrivals with the highest survival benefits. That amounts to selecting only those arrivals with survival benefits in the yellow shaded region in the figure below (the yellow region is 57% of the total). This is equivalent to selecting only those patients with survival benefits at least as large as b^* where b^* is the solution to



$$\frac{0.44 - b^*}{0.44 - 0.03} = 0.57.$$

This solves to yield $b^* = 0.44 - 0.57 \times (0.44 - 0.03) = 0.206$. Now, if only those with incremental survival benefits of at least 0.206 are admitted, the average benefits for such admits is simply $(0.206 + 0.44)/2 = 0.323$. Thus, the annual expected incremental lives saved under cherry picking is given by

$$\text{Annual Lives Saved } = 1.31 \times 0.323 \times 365 = 154.4.$$

So, deterministic reasoning suggests that if only those with incremental survival benefits of at least 0.323 are admitted (that is, if the 57% of all arrivals with the highest benefits are admitted), the expected annual number of lives saved can be increased from 112.4 per year under first come first served to 154.4 per year under cherry picking, a substantial increase.

(i) What is the conditional expected survival benefit for admitted patients?

The key is to note that the survival benefit distribution will now be uniform between b^* and 0.44 as opposed to between 0.03 and 0.44. Therefore, the conditional expected survival benefit for those admitted will be given by

$$E(B|B > b^*) = \frac{b^* + 0.44}{2}.$$

(ii) Produce a formula for the total ICU admission rate in patients per day that applies for all potential values of b^* between 0.03 and 0.44. (HINT: be careful!)

Clearly what will happen is that as b^* increases from 0.03, the number of patients eligible for admission will fall. However, with $b^* = 0.03$, the arrival rate is $\lambda \approx 2.3 > 1.31 = m\mu$. So we have to be careful in distinguishing between the arrival rate and the admission rate. Let's start with the arrival rate: if we apply a lower cutoff of b^*, then the arrival rate $\lambda(b^*)$ will be given by

$$\lambda(b^*) = 2.3 \times \Pr\{B > b^*\} = 2.3 \times \frac{0.44 - b^*}{0.44 - 0.03} \text{ for } b^* \text{ between 0.03 and 0.44 inclusive.}$$

Now, what about the admission rate? Clearly the maximum admission rate is 1.31 patients per day as that is the capacity of the ICU. However, if b^* is large enough, then the implied arrival rate $\lambda(b^*)$ will be *less* than 1.31 per day, and the facility cannot admit more patients than those that arrive! So, your formula needs to be of the form:

$$\text{Admission Rate } (b^*) = \begin{cases} 1.31 & \text{if } \lambda(b^*) \geq 1.31 \\ \\ \lambda(b^*) & \text{if } \lambda(b^*) < 1.31 \end{cases}$$

which simplifies to

$$\text{Admission Rate}(b^*) = \min(1.31, \lambda(b^*)) = \min(1.31, 2.3 \times \frac{0.44 - b^*}{0.44 - 0.03}).$$

Note that the admission rate switches from 1.31 patients per day to $\lambda(b^*)$ patients per day once $\lambda(b^*) < 1.31$, or equivalently when

$$\begin{aligned} 2.3 \times \frac{0.44 - b^*}{0.44 - 0.03} &< 1.31, \text{ or when} \\ b^* &> 0.206. \end{aligned}$$

So another way to express the admission rate is

$$\text{Admission Rate } (b^*) = \begin{cases} 1.31 & \text{if } b^* \leq 0.206 \\ \\ 2.3 \times \frac{0.44 - b^*}{0.44 - 0.03} & b^* > 0.206 \end{cases}.$$

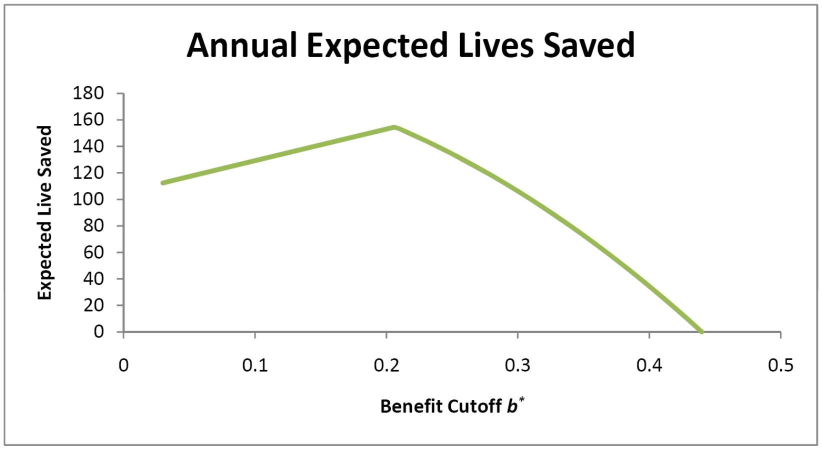
(c) You hope to maximize the expected annual number of lives saved from operating the ICU, so you use this objective to guide your choice of the cutoff value b^*.

(i) What value of b^* maximizes the expected annual number of lives saved? Explain your reasoning!

Using the results from part (b), the expected annual number of lives saved from operation the ICU as a function of the cutoff b^* is given by

$$E(\text{Annual Lives Saved}) = \text{Admission Rate } (b^*) \times \; E(B|B > b^*) \times 365.25.$$

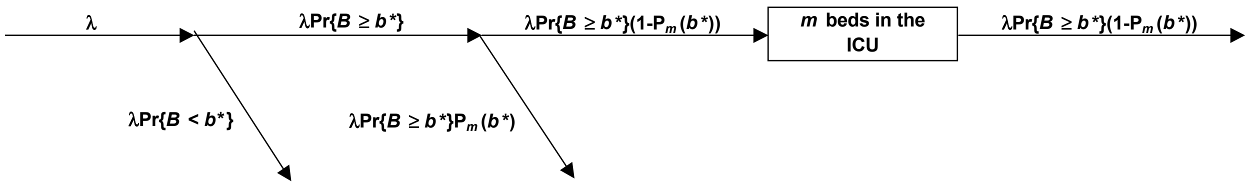
A graph of the expected annual lives saved as a function of the cutoff b^* appears below:



Looking at this figure, it is clear that the annual expected lives saved increases linearly as a function of b^* before it declines. And look what the magic maximizing value of b^* turns out to be: 0.206, that value of $b^*$where the admission rate exactly equals the ICU capacity of 1.31 patients per day! Why is this happening? Simple: under the fluid assumption, the ICU can simply take the best patients, that is, those with maximal incremental survival benefits. So, starting with $b^* = 0.44$, the ICU starts *lowering* the value of b^* until they can no longer accept patients, which happens when the admission rate exactly equals 1.31 patients per day (which happens when $b^* = 0.206$). It makes no sense to use a higher cutoff than 0.206, as then the ICU has excess capacity while admitting *all* patients with survival benefits in excess of the cutoff. It makes no sense for the ICU to choose a cutoff lower than 0.206, as then not all patients can be admitted, and the ICU might admit patients with benefits lower than 0.206 while losing some with with higher benefits. So, the best thing for the ICU to do is to “cherry pick” the best patients (from the standpoint of incremental survival), which leads to only accepting patients with $b^* > 0.206$.

Probabilistic Reasoning: Erlang Loss Model

The real problem is probabilistic: arrivals do not “flow” into the ICU, but rather arrive discretely according to some random process which we will assume to be Poisson. This leads to the $M/G/m/m$ Erlang loss model as the natural representation for the problem. Potential ICU admits arrive in accord with a Poisson process with rate λ ($= 2.297$ in our example). Now, we can still implement a cherry picking policy by specifying a minimum incremental survival benefit b^* that is required for admission. For any particular value of b^*, this yields a new arrival rate of potential admits given by $\lambda \Pr\{B \geq b^*\}$ where B is the random variable denoting incremental survival benefits (the distribution of which is uniform, that is equally likely to fall, between 0.03 and 0.44 in our example). However, even after applying the benefit threshold, there is still a probability given by $P_m(b^*)$ that all m beds will be busy; this follows from an Erlang loss model with arrival rate $\lambda \Pr\{B \geq b^*\}$, and m servers each working with rate μ. Thus, the actual rate with which patients admitted to the ICU is given by $\lambda \Pr\{B \geq b^*\}(1 - P_m(b^*))$. To get in, an arrival must both have a benefit of at least b^* and find at least one free bed. The flow pattern is illustrated in the figure below.



Now, recall from Queueing Theory’s Greatest Hits (or equivalently from equation (3) on p. 132 of the paper) that the probability that i servers out of m in an Erlang loss model are busy is given by

$$P_i = \frac{(\lambda/\mu)^i/i!}{\sum_{j=0}^{m}(\lambda/\mu)^j/j!} \text{ for } i = 0, 1, 2, ..., m.$$

The probability that all m beds are filled when a benefit threshold (or *hurdle*)

b^* is in force is found from this formula by substituting $\lambda \Pr\{B \geq b^*\}$ for the arrival rate and evaluating the formula above for $i = m$. Specifically, we have

$$P_m(b^*) = \frac{(\lambda \Pr\{B \geq b^*\}/\mu)^m/m!}{\sum_{j=0}^{m}(\lambda \Pr\{B \geq b^*\}/\mu)^j/j!}.$$

For the example at hand where the incremental benefits B are equally likely to fall between 0.03 and 0.44, note that

$$\Pr\{B \geq b\} = \begin{cases} 1 & b < 0.03 \\ \frac{.44-b}{.44-.03} & .03 \leq b \leq .44 \\ 0 & b > 0.44 \end{cases} \cdot$$

Finally, if a threshold b^* is applied, note that the expected survival benefit for those admitted to the ICU simply equals $(b^* + .44)/2$ (providing b^* falls between 0.03 and 0.44).

Example: First Come First Served ($b^* = 0.03$)

To evaluate the First Come First Served policy, simply set the benefit threshold $b^* = 0.03$, which means that anyone arriving to the ICU will be admitted providing that there is a bed available (since everyone has an incremental survival benefit of at least 0.03 in this example). Using the Erlang loss model spreadsheet available on the course web site, and plugging in the appropriate parameter values ($\lambda = 2.297$, $\mu = 0.164$, and $m = 8$) you will discover that the probability that all 8 beds are filled is given by

$$P_8 = \frac{(2.297/0.164)^8/8!}{\sum_{j=0}^{8}(2.297/0.164)^j/j!} = 0.491.$$

There is almost a 50% chance that all beds are filled! Note that on average, the number of beds filled equals (from Little's Theorem after tossing out those ICU arrivals who do not get admitted due to finding all beds filled)

$$2.297 \times (1 - 0.491)/0.164 = 7.13.$$

Also, as before the average benefit is just given by $(0.03 + 0.44)/2 = 0.235$. Thus, under the First Come First Served policy, the expected annual number of lives saved by the ICU is given by

$$\text{Annual Lives Saved } = 2.297 \times (1 - 0.491) \times 0.235 \times 365 = 100.3.$$

Note that this is smaller than the equivalent figure computed under deterministic reasoning, which was equal to 112.4.

Finding the Optimal Benefit Threshold

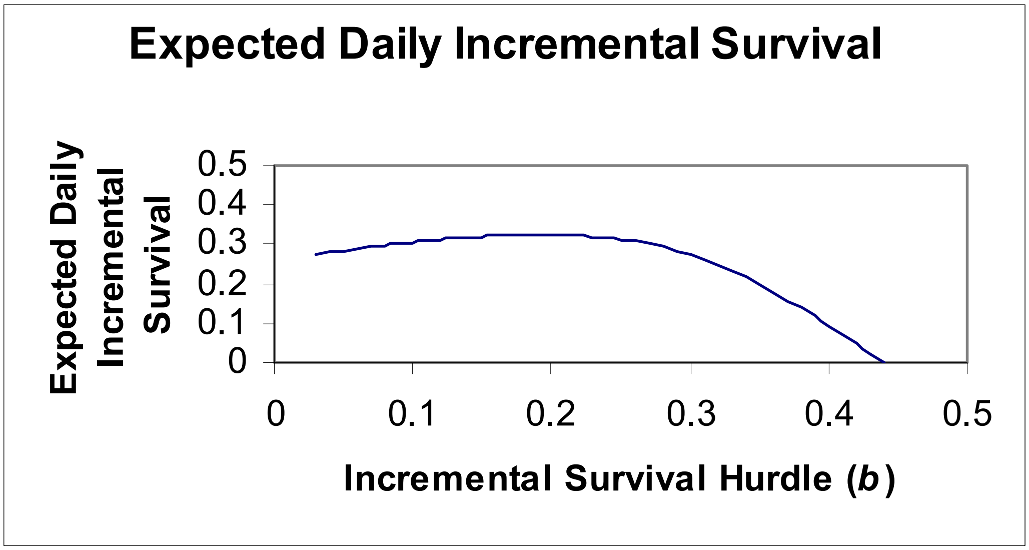
The general formula for the expected annual number of incremental lives saved as a function of the minimum benefit threshold b^* is given by

$$\text{Annual Lives Saved } = \lambda \times \Pr\{B \geq b^*\} \times (1 - P_m(b^*)) \times E(B|B \geq b^*) \times 365.$$

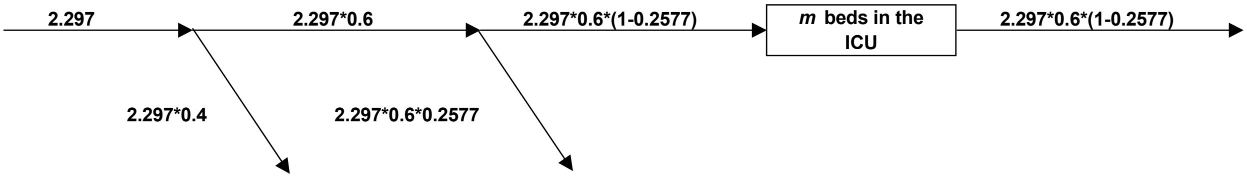
Specialized to the parameters and benefit distribution of our example, we have

$$\text{Annual Lives Saved } = 2.297 \times \frac{.44 - b^*}{.44 - .03} \times (1 - P_8(b^*)) \times \frac{b^* + .44}{2} \times 365$$

where the formula for the loss probability $P_8(b^*)$ appears earlier in these notes. The optimal benefit threshold is the one that maximizes the annual expected number of lives saved. The optimal benefit threshold is found by graphing the expected lives saved as a function of the benefit threshold and choosing that benefit that maximizes expected lives saved (the graph shows daily as opposed to annual lives saved, but this will yield the same optimal benefit threshold):



The optimal threshold is given by $b^* = 0.194$. This leads to excluding 40% of arrivals on the basis of insufficient benefit (that is, $\Pr\{B < 0.194\} = 0.4$), while admitting 60% of arrivals still yields a loss probability $P_8(b^* = 0.194) = 0.2577$. The resulting probabilistic flows are shown below:



The expected annual lives saved with the optimal threshold $b^* = 0.194$ equals

$$\text{Annual Lives Saved } = 2.297 \times 0.6 \times (1 - .2577) \times \frac{.194 + .44}{2} \times 365 = 118.4.$$

This is an improvement over First Come First Served by 18 lives per year, which seems worthwhile.

Note that in comparison to the deterministic model, the stochastic model is less optimistic (albeit more realistic): the queueing model estimates expected annual lives saved of 118.4 versus the 154.4 from the deterministic model. However, it is interesting to note that both the deterministic cherry picking model and the Erlang loss model produce nearly identical optimal benefit thresholds: 0.194 in the Erlang loss model versus 0.206 in the deterministic model.

Bed-Specific Hurdles

The paper considers an even more general admission control, namely, one where the optimal benefit threshold depends on how many beds are filled. This model is developed on p. 133 of the paper. The idea is simple: as beds fill up, you should raise the benefit threshold since capacity has become scarcer. One can formulate a more general Erlang model (see equations (8)-(10) on p. 133) to determine the optimal bed-specific thresholds and the expected lives saved. Of interest in this specific application was that the optimal bed-specific hurdles only saved an additional expected 1.4 lives beyond the policy with a single benefit threshold. Now, if you think there are ethical issues applying a single benefit cutoff, imagine what happens with bed-specific thresholds – the same patients could be admitted or rejected based on the simple order in which they arrive! In the Jerusalem example, the extra benefits from the more complicated policy are not sufficient to justify arguing for bed-specific hurdles, but in other situations, it could be that substantial additional lives could be saved this way. Something to think about!

7.7.7 Repeat COVID-19 Testing

In the Fall of 2020, Yale determined that repeat SARS-CoV-2 asymptomatic testing of all students and high-contact faculty and staff would be necessary to enable the safe opening of the university. The Yale Health planning team determined that ten testing sites would be necessary to handle the 2,320 daily tests it would take to cover the population screened. Assume that these tests were uniformly spread over the ten testing sites (so each site conducted one-tenth of the daily tests), and suppose that each testing site operated for 8 hours per day. In planning how to deliver this service, it was assumed that within each site there would be some number m testing stations, each of which would be occupied by at most a single person at a time. Further, after several trial runs, it was assumed that the average time required for an individual to enter a testing station, self-swab, and depart was equal to ten minutes. With these assumptions:

(a) What is the smallest number of testing stations required at each testing site to process the demand for testing? Remember that the number of stations must be an integer!

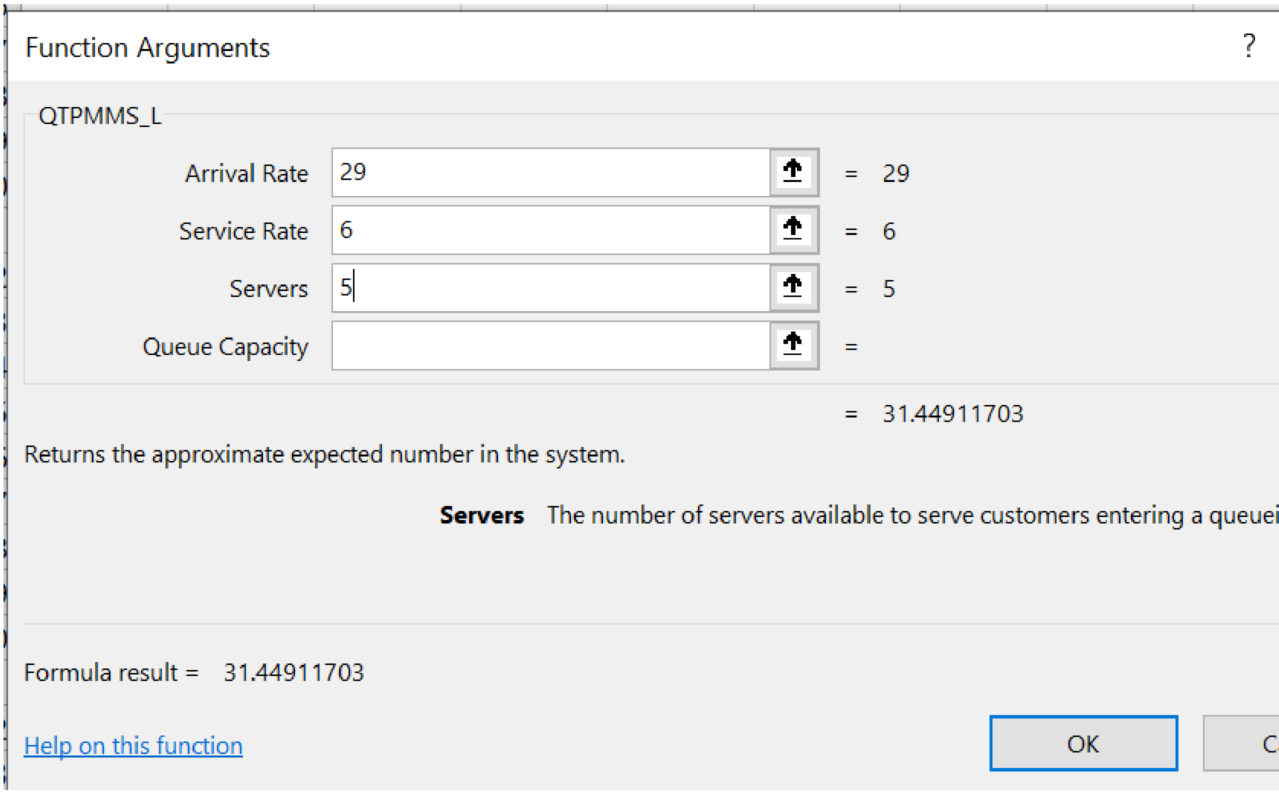
Given that a testing site must process 2,320/10 = 232 tests per day, and given that each site will operate for 8 hours, the hourly demand for testing is given by $\lambda = 232/8 = 29$ tests per hour. Furthermore, we are told that the service time $1/\mu = 10$ minutes, which means that the hourly processing capacity per testing station is given by $\mu = 6$ tests per hour. Consequently, the expected number of busy servers (or occupied testing stations) is given by

$$B = \frac{\lambda}{\mu} = \frac{29}{6} = 4.83$$

which means that each site needs *at least 5* testing stations.

(b) Initially planners imagined persons simply showing up at random (that is, in accord with a Poisson process) at the daily demand rate implied by the assumptions above. Further, the duration of time spent "in service" (that is, the time an individual occupies a testing station) was presumed to be exponentially distributed with mean equal to ten minutes as described above. On average, how much time would each person screened spend at a testing site (that is, the sum of the time spent "in service" plus waiting time to enter a testing station) if the number of testing stations was equal to your answer from part (a) above?

The assumptions above correspond to those of an $M/M/m/\infty$ queue, and you are being asked to find the average total time spent per visit to the testing site, or W, when the number of servers is $m = 5$ (since that's the answer to part (a)). There are several ways you could do this: you can use the formula for L_q from Section 3 of Queueing Theory's Greatest Hits in the coursepack, then compute $L = L_q + B$ (where $B = 29/6$ from part (a)), and then compute $W = L/\lambda$ (Little's Theorem). Or you could use the Queueing Toolpak or Queuebook to just compute W directly. For example, using the Queueing Tookpak to compute L yields



Function Arguments

QTPMMS_L

Arrival Rate 29 = 29

Service Rate 6 = 6

Servers 5 = 5

Queue Capacity =

= 31.44911703

Returns the approximate expected number in the system.

Servers The number of servers available to serve customers entering a queuei

Formula result = 31.44911703

or 31.449 persons in the system, which implies that $W = 31.449/29 = 1.08$ hours. Whichever way you do it, you'll discover that $W = 1.08$ hours (65.07 minutes)!! That sucks!!

(c) Planners felt that the population would rebel against getting tested if the average time required per visit to a testing site exceeded 15 minutes. This being the case, what is the smallest number of testing stations possible that would satisfy this waiting time constraint (consistent with all of the assumptions and data presented above)?

You are still in $M/M/m/\infty$ land, but you've discovered from (c) that $m = 5$ won't work because the total time in the system is 65 minutes, far in excess of the 15 minute target. So, you need to increase the number of servers. Let's try adding one more so $m = 6$. Again you are free to use any of your myriad available approaches to get the answer, but done correctly you'll discover that just adding one server reduces the average time spent in the system from 65 minutes to 14.5 minutes! That makes 6 the smallest number of testing stations per site for which the average total time spent per visit is less than 15 minutes.

(d) One of the Yale Health staffers suggested that rather than assume those getting screened would arrive in accord with a Poisson process, it would be possible to *schedule* arrivals giving everyone a specific arrival time. Furthermore, this staffer thought that with practice, the time spent in testing stations, rather than being exponentially distributed, could be brought to essentially equal ten minutes with probability one! Under these assumptions, what would be the smallest number of testing stations to keep the total time spent at a testing site below 15 minutes?

Well, with an arrival rate of 29 per hour, in principle one could schedule an arrival once every 1/29 of an hour or once every 2.069 minutes. We know that at least 5 testing stations are needed from (a) above. At a capacity of 6 tests per hour for each station, having five perfectly scheduled stations could process 30 tests per hour. This means that if everyone was perfectly scheduled and it took exactly ten minutes to process each test, you could get by with only five stations. In fact, if appointments and processing times were really adhered to flawlessly, you could handle 30 tests per hour with five stations – suppose five people show up on the dot once every ten minutes – those five people would enter the five testing stations just as they empty out their previous occupants, and lo and behold you'd be processing 30 tests per hour. BUT – note what is required for this to work – appointments plus processing times have to be perfect! This can be achieved in mechanized production systems (e.g. robots and conveyer belts), but once you add variability into the process (in arrivals or service times or both), this all falls apart. Given that the expected number of occupied testing stations equals 4.83 without regard to the probability distributions of the interarrival and service times (keeping their means fixed), trying to get by with only five stations per testing site would be a recipe for disaster. Six stations on the other hand? Works just fine.

7.7.8 The Oldest Person in the World

A *supercentenarian* is someone who lives to age 110 or beyond. The demographer Jutta Gampe has estimated that the mortality rate for supercentenarians is constant at approximately 0.7 per supercentenarian per year for both men and women (though more women reach age 110 than men; see http://www.demogr.mpg.de/books/drm/007/3-1.pdf if interested). The

Gerentology Research Group has estimated that on average, there are 375 living supercentenarians on the planet.

(a) Suppose that on average, the number of persons who reach age 110 each year is equal to λ. Based on the information provided thus far, provide a numerical estimate for λ.

The key here is to recognize this as a queueing system! λ is the arrival rate of new supercentenarians (the rate at which people reach their 110^{th} birthday). The service time corresponds to the remaining life expectancy of new supercentenarians. We are told that "...the mortality rate for supercentenarians is constant at approximately 0.7 per supercentenarian per year..." which means that the mean "service time," which is equivalent to the expected remaining duration of life, is equal to $1/0.7 = 1.43$ years. So, recognizing that the number of living supercentenarians is just equal to the number of "customers in the system," we have via Little's Theorem that

$$L = \lambda W$$

which, for this problem, works out to

$$375 = \lambda \times (1/0.7)$$

and consequently

$$\lambda = 375 \times 0.7 = 262.5$$

new supercentenarians per year.

(b) Suppose a person has just reached her 110^{th} birthday. On average, how much longer will this person live?

Well, if you answered part (a), then you have already answered part (b)! Since the mortality rate is constant at 0.7 for anyone aged at least 110 years, the expected remaining life is just given by $1/0.7 = 1.43$ years.

(c) Let X denote the actual number of supercentenarians alive; X is a random variable. What is the probability distribution of X, that is, produce a mathematical expression for

$$\Pr\{X = x\} \text{ for } x = 0, 1, 2, 3, ...$$

Let's see – we have an infinite server queueing system here owing to "self-service" – each supercentenarian serves him/herself. Checking out Queueing Theory's Greatest Hits on the course website reveals that for an infinite server queue, the probability distribution for the number of customers in service is just Poisson with mean λ/μ, which in this case is given by 375 (as we are told that on average there are 375 living supercentenarians; or note that $\lambda/\mu = (375/1.43)/0.7 = 375$ (since $1/0.7 = 1.43!$). So, we conclude that

$$\Pr\{X = x\} = \frac{375^x e^{-375}}{x!} \text{ for } x = 0, 1, 2, 3, ...$$

This will look just like a normal distribution with a mean of 375 and a standard deviation of $\sqrt{375} = 19.4$.

(d) Suppose that the oldest person in the world, who of course is a supercentenarian, has just died (e.g. https://www.nytimes.com/2022/04/27/world/asia/kane-tanaka-japan-worlds-oldest-person.html). On average, how much time will pass until the next oldest person in the world passes away?

Well, since the mortality rate for all supercentenarians is constant at about 0.7 per person per year, from the time the oldest person in the world dies, on average it will take $1/0.7 = 1.43$ years until the next oldest person passes away.

Chapter 8

Deterministic Fluid Models

8.1 Overloaded Queues with Reneging: Queues in Public Housing

The salient feature of public housing systems is that the rate at which households apply for public housing greatly exceeds the rate at which housing units become available. This is the basic cause of the long waits for housing assignments cited in the introduction. Denoting the average applicant arrival rate by λ and the average unit turnover rate (which is equivalent to the assignment rate) by μ, it is true that $\lambda >> \mu$ for most large public housing authorities. For example, from April 1984 through March 1985, 4,422 eligible applicants arrived at the Boston Housing Authority, while only 771 households moved out of public housing units [I].

While waiting lists are large, they are not infinitely so. The reason for this is that many applicants drop out (or renege) from public housing waiting lists before receiving a unit assignment. Unlike customer behavior in some queuing systems where reneging occurs due to impatience, housing applicants renege only when they have the opportunity. In other words, housing applicants will leave a waiting list if alternative housing has been secured. A reasonable assumption which reflects this states that the rate at which applicants renege is proportional to the number of applicants waiting in queue. More formally, we assume that if n applicants are waiting in queue, the dropout rate of applicants from the queue equals $n\delta$; the parameter δ is called the "household-specific dropout rate." Of course, this assumption is exceedingly difficult to verify empirically as dropout is only detected when applicants fail to respond to unit offers.

In saturated queuing systems, one expects the percentage deviation of queue length from average queue length at any given time to become quite small (9). This is true without regard to the underlying arrival and service processes. In such circumstances, a deterministic model of the queuing system (commonly referred to as a "fluid approximation") leads to the same conclusions as the "correct" probabilistic model. This is the case in our public housing system.

Consider the system described above operating in steady state. Let L_q denote the expected length of the public housing waiting list. As the dropout rate is proportional to queue length, the dropout rate in steady state equals $L_q\delta$. Also, since all arriving applicants are housed or drop out, conservation of flow requires that

$$\lambda = L_q\delta + \mu.$$

We may think of equation (1) as the law of conservation of housing applicants. This law implies the expected number of households on the waiting list is given by

$$L_q = \frac{\lambda - \mu}{\delta}\cdot$$

Equation 2 provides a simple explanation for the large waiting lists cited in the introduction. For example, if 1,000 households apply yearly for housing but only 100 units vacate annually, and if the average time to find alternative housing equals 4 years (implying $\delta = 0.25$), then an average of 3,600 households would be waiting for public housing.

Now consider a transient analysis of this system. To make matters more specific, we will assume a first-come, first-housed tenant assignment policy (such policies are used; see (7) for examples). Let $n(t)$ denote the size of the waiting list at time t. The transient behavior (according to the fluid approximation) is described by

$$\frac{dn(t)}{dt} = \lambda - \mu - \delta n(t).$$

This simple differential equation has the solution

$$n(t) = n(0)e^{-\delta t} + \frac{\lambda - \mu}{\delta}(1 - e^{-\delta t})$$

and it is clear that $n(t) \to L_q$ as $t \to \infty$.

As mentioned in the introduction, one of the reasons for modeling housing systems is to predict waiting times. Assuming first-come first-housed, the time a newly arriving applicant who finds n households in queue must wait to receive a housing assignment is equal to the time required to deplete this queue from n to 0. Letting $q(t)$ denote the number of the original n households remaining on the waiting list t units after the arrival of a new application, the queue depletion equation is given by

$$\frac{dq(t)}{dt} = -\mu - \delta q(t)$$

with solution

$$q(t) = q(0)e^{-\delta t} - \frac{\mu}{\delta}(1 - e^{-\delta t}).$$

The waiting time to deplete an initial queue of size n, W_n, is then found by setting $q(0) = n$ and solving for $q(W_n) = 0$ which yields

$$W_n = \frac{1}{\delta}\log(\frac{n\delta + \mu}{\mu}).$$

In steady state, a new arrival would expect to find L_q applicants waiting, so the equilibrium waiting time W is, after substituting L_q for n, given by

$$W = \frac{1}{\delta}\log(\frac{\lambda}{\mu}).$$

An alternative derivation of this result can be found as follows: the fraction of all applicants who eventually receive a housing assignment is, via the law of conservation of housing, given by

$$\Pr\{\text{Receive housing}\} = \frac{\mu}{\lambda}\cdot$$

A different expression for the fraction of applicants who receive housing follows from the proportional reneging assumption, which is equivalent to assuming that the likelihood an applicant will wait longer than some time t is given by the exponential distribution $\exp(-\delta t)$. In equilibrium, if W is the waiting time for people receiving housing, then the fraction of all applicants who receive housing must equal $\exp(-\delta W)$. Equating to the earlier expression for the likelihood of receiving a housing assignment yields the identity

$$\Pr\{\text{Receive housing}\} = \frac{\mu}{\lambda} = e^{-\delta W}$$

and solving for W yields $(1/\delta \log(\lambda/\mu))$ as before.

8.2 Examples

8.2.1 Treatment on Demand

Once upon a time in San Francisco, about 1,400 persons were waiting to enter drug treatment (RE: Kaplan and Johri, Treatment on Demand). There were roughly 6,300 treatment slots available citywide. Approximately 50% of those seeking drug treatment dropped off of the waiting list before (indeed without!) receiving treatment, while for those seeking treatment who did gain entry to the system, the waiting time from request until the initiation of was one month on average. Drug treatment advocates wanted the system to provide *treatment on demand*, meaning that anyone requesting treatment should be able to enter immediately. Since there were clearly an insufficient number of treatment slots, as evidenced by the facts that 1,400 people were waiting to enter treatment while half of all applicants dropped out of the treatment queue, the advocates thought that providing an additional 1,400 treatment slots on top of the 6,300 that already existed would take care of the backlog. Indeed, after expanding the number of treatment slots the queue quickly diminished, but over time (and to the dismay of treatment advocates), the waiting list rebounded to average about 1,100 persons waiting, the fraction of applicants dropping off the waiting list decreased from 50% to 40%, and the waiting time for those who received treatment dropped from one month to three weeks. What happened?

Thinking of the treatment system as an overloaded queue with reneging, let λ denote the annual arrival rate of new drug treatment requests, and let μ denote the number of treatment episodes the 6,300 slot system could process per year, that is, the treatment capacity. Note that if the average duration of a single drug treatment episode is equal to τ years, then the annual turnover per treatment slot would equal $1/\tau$, and with 6,300 slots that are always presumed full (due the fact that the demand for treatment clearly outstrips the available supply), the treatment capacity is given simply by

$$\mu = \frac{6,300}{\tau}.$$

We don't know the numerical value of τ, but as we'll see we won't need to in order to understand what transpired.

Now recall that 50% of treatment applicants drop out of queue without receiving treatment. This means that the fraction of applicants who do receive treatment is also equal to 50%, which is to say that the demand for treatment is twice the available capacity. Worded differently,

$$\Pr\{\text{Dropout}\} = \Pr\{\text{Receive Treatment}\} = \frac{\mu}{\lambda} = 0.5,$$

from which we again see that $\lambda = 2\mu$ – the demand for treatment is twice the available capacity.

What happened when 1,400 treatment slots were added? Clearly the treatment capacity increased to some new level, say μ'. Presuming that the average duration of a treatment episode τ remained unchanged (as there was no change to the nature of drug treatment offered), the new treatment capacity can be determined by

$$\begin{aligned}\mu' &= \frac{6,300 + 1,400}{\tau} \\ &= \frac{7,700}{6,300} \times \frac{6,300}{\tau} \\ &\approx 1.22 \times \mu.\end{aligned}$$

As a consequence, the fraction of applicants who were able to receive treatment, assuming that the demand for treatment λ remained unchanged, was given by

$$\begin{aligned}\Pr\{\text{Receive Treatment}\} &= \frac{\mu'}{\lambda} \\ &= 1.22 \times \frac{\mu}{\lambda} \\ &= 1.22 \times 0.5 = 0.61\end{aligned}$$

or about 60%. Right away we see why the dropout rate only declined from 50% to 40% of the annual applicant rate!

What about the number of people waiting to enter drug treatment? Assuming that the aggregate dropout rate from the treatment queue is proportional to the number waiting by analogy to the overloaded public housing model described in the previous section, we see that before the addition of the 1,400 new treatment slots,

$$\begin{aligned} L_q &= \frac{\lambda - \mu}{\delta} \\ &= \frac{\lambda - 0.5\lambda}{\delta} \\ &= 0.5 \times \frac{\lambda}{\delta} = 1,400 \end{aligned}$$

where $1/\delta$ is the average time a new treatment applicant is willing (or able) to wait in queue before reneging. After adding the 1,400 new slots, the new queue length, say L'_q, is given by

$$\begin{aligned} L'_q &= \frac{\lambda - \mu'}{\delta} \\ &\approx \frac{\lambda - 0.6\lambda}{\delta} \\ &= 0.4 \times \frac{\lambda}{\delta} \\ &= \frac{0.4}{0.5} \times 0.5 \times \frac{\lambda}{\delta} \\ &= 0.8 \times 1,400 = 1,120 \end{aligned}$$

which is very close to the reported 1,100 persons waiting on average after 1,400 treatment slots were added.

What about the waiting time for those who received treatment? Before the new slots were added, we were told that this average waiting time, call it W was equal to 1 month, so again by analogy to the public housing model we have

$$\begin{aligned} W &= \frac{1}{\delta} \ln(\frac{\lambda}{\mu}) \\ &= \frac{1}{\delta} \ln(\frac{\lambda}{0.5\lambda}) \\ &= \frac{1}{\delta} \ln(2) = 1 \text{ month.} \end{aligned}$$

After adding the 1,400 slots however, the new waiting time for those who receive treatment, say W', is given by

$$\begin{aligned} W' &= \frac{1}{\delta}\ln(\frac{\lambda}{\mu'}) \\ &\approx \frac{1}{\delta}\ln(\frac{\lambda}{0.6\lambda}) \\ &= \frac{1}{\delta}\ln(1.67). \end{aligned}$$

From this we deduce that

$$\begin{aligned} \frac{W'}{W} &= \frac{\ln(1.67)}{\ln(2)} \\ &\approx 0.74 \end{aligned}$$

and thus

$$\begin{aligned} W' &= 0.74 \times W \\ &= 0.74 \times 1 \text{ month} \\ &\approx 3 \text{ weeks} \end{aligned}$$

as observed. Oh – note that from the equation $W = \frac{1}{\delta}\ln(2) = 1$ month, we can also deduce that $\frac{1}{\delta} = 1/\ln(2) \approx 1.44$ months. On average, those seeking drug treatment are not able to survive long in queue before dropping out!

8.2.2 A Public Housing Problem

At present, there are 100 households one the waiting list to receive a three bedroom public housing apartment. Three bedroom units only become available at a rate of 20 per year. However, the quality of these units is such that applicants are, on average, willing to wait ten years to receive an assignment. Assuming that the public housing application and assignment process has reached equilibrium:

(a) What is the annual number of new applicants for three bedroom units?

We are told that the system is in equilibrium. Letting L_q, μ, and $1/\delta$ denote the mean number of households waiting (the mean queue length), the annual unit turnover, and a household's average willingness to wait, the problem identifies $L_q = 100$ households, $\mu = 20$ three bedroom apartments per year, and $1/\delta = 10$ years. The "law of conservation of housing applicants" then states that the annual application rate λ is given by

$$\lambda = L_q\delta + \mu = 100 \times \frac{1}{10} + 20 = 30 \text{ new applicants per year.}$$

(b) What fraction of new applicants eventually drop out of the waiting list without receiving a public housing assignment?

If 30 applicants apply each year but only 20 households receive an apartment (as that is the annual unit turnover), it must be that $30 - 20 = 10$ applicants drop out of the waiting list without receiving an assignment each year (note that $L_q\delta = 100/10 = 10$). Consequently, the fraction of new applicants that eventually drop out of the waiting list without receiving a public housing assignment is given by $10/30 = 1/3$.

(c) For those applicants lucky enough to receive a three bedroom unit, on average how years must they wait from application until assignment, assuming that households are assigned in first-come first-housed fashion?

Given that households are assigned in first-come first-housed fashion, the waiting time for those lucky enough to receive housing is just given by

$$W = \frac{1}{\delta}\log\frac{\lambda}{\mu} = 10\log\frac{30}{20} = 4.05 \text{ years}$$

(note that log refers to the natural logarithm, also denoted by ln).

8.2.3 Housing Is Their Middle Name

The Boston Housing Authority (BHA) was in the news during 1988 in a big way. Stories gracing both the *Boston Globe* and the *New York Times* surfaced relating tales of discriminatory action in assigning public housing applicants to available units. The federal Department of Housing and Urban Development (HUD) filed charges against the BHA following an investigation of certain assignment practices. HUD cited the following specific statistics in accusing the BHA of discriminating against minority applicants:

- On average, non-whites who receive public housing must wait nine months longer for public housing than white applicants.

- The above is the case, even though non-whites account for 83.5% of the waiting list for public housing.

- In addition, only 48% of all assignments are made to minorities (in spite of the fact that 83.5% of those waiting for public housing are minorities).

That these facts are consistent with a discriminating agency appears self-evident. However, do these facts *imply* discrimination on behalf of the BHA? Consider the following proposition: suppose that the fraction of all BHA assignments that are received by minorities is exactly equal to the fraction of all BHA applications that are submitted by minorities. In this instance, whites and non-whites would have exactly the same probability of receiving public housing assistance. Surely such a balance of assignment probabilities is one reasonable definition of a "fair" assignment process. Assuming such a balance actually exists, can the three HUD charges stated above be reconciled? If so, what are the implied waiting times (in months) for whites and non-whites who receive housing? You may assume that the housing system is in steady state, enabling the application of two fluid models for white and non-white applicants respectively.

First some notation. Let:
λ_1, λ_2 = applicant rate for non-whites, whites respectively
W_1, W_2 = waiting time to enter housing for non-whites, whites respectively
μ_1, μ_2 = assignment (move in) rates for non-whites, whites respectively
δ_1, δ_2 = dropout (reneging) rates for non-whites, whites respectively
L_{q1}, L_{q2} = queue lengths for non-whites, whites respectively

The three HUD claims can be summarized as follows:
(1) $W_1 = W_2 + 9$ (with time measured in months)
(2) $L_{q1}/(L_{q1} + L_{q2}) = 83.5\%$ (non-whites are 83.5% of the waiting list)
(3) $\mu_1/(\mu_1 + \mu_2) = 48\%$ (48% of assignments go to non-whites)

In addition, the proposition states that

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\mu_1}{\mu_1 + \mu_2} = 48\%$$

or that the fraction of all applicants who are non-white equals the fraction of all assignments who are non-white. Note that this implies that the probability of receiving a housing assignment is the same for whites and for non-whites, an arguably fair process.

From the fluid model discussed in class, note that

$$W_1 = \frac{1}{\delta_1}\log\frac{\lambda_1}{\mu_1} \Longrightarrow \mu_1 = \lambda_1 e^{-\delta_1 W_1}$$

and

$$W_2 = \frac{1}{\delta_2}\log\frac{\lambda_2}{\mu_2} \Longrightarrow \mu_2 = \lambda_2 e^{-\delta_2 W_2}.$$

These relationships imply that

$$\begin{aligned}\frac{\mu_1}{\mu_1+\mu_2} &= \frac{\lambda_1 e^{-\delta_1 W_1}}{\lambda_1 e^{-\delta_1 W_1} + \lambda_2 e^{-\delta_2 W_2}}\\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2 e^{(\delta_1 W_1 - \delta_2 W_2)}}\\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}\end{aligned}$$

with the last equality following from the proposition that the fraction of assignments that are non-white equals the fraction of applicants who are non-white. This in turn implies that

$$\delta_1 W_1 = \delta_2 W_2.$$

Now consider charge (2):

$$\frac{L_{q1}}{L_{q1}+L_{q2}} = \frac{(\lambda_1-\mu_1)/\delta_1}{(\lambda_1-\mu_1)/\delta_1 + (\lambda_2-\mu_2)/\delta_2} = 0.835.$$

However, substituting in our prior expressions for μ_1 and μ_2 leads to the result that

$$\frac{L_{q1}}{L_{q1}+L_{q2}} = \frac{\lambda_1/\delta_1}{\lambda_1/\delta_1 + \lambda_2/\delta_2}$$

(recall that $\delta_1 W_1 = \delta_2 W_2$ which will help with the last result).

Now, let $\lambda = \lambda_1 + \lambda_2$ be the total application rate. By the proposition,

$$\lambda_1 = .48\lambda \text{ and } \lambda_2 = .52\lambda$$

which means that

$$\frac{L_{q1}}{L_{q1} + L_{q2}} = \frac{.48\lambda/\delta_1}{.48\lambda/\delta_1 + .52\lambda/\delta_2} = 0.835.$$

Solving we see that this expression is consistent as long as

$$\frac{\delta_1}{\delta_2} = 0.1824$$

which says that the ratio of the average time until an outside housing opportunity becomes available to non-whites $(1/\delta_1)$ relative to whites $(1/\delta_2)$ is given by $1/0.1824 = 5.5$. It takes 5.5 times as long for non-whites as whites to find alternatives to public housing!

Now consider charge (1), which says that $W_1 = W_2 + 9$. We have already showed that, under our proposition of assignments proportional to application rates, $\delta_1 W_1 = \delta_2 W_2$. So, we must have

$$\delta_1 W_1 = \delta_1(W_2 + 9) = \delta_2 W_2$$

or

$$\frac{\delta_1}{\delta_2} = \frac{W_2}{W_2 + 9} = .1824.$$

We find that

$$W_2 = 2 \text{ months, and } W_1 = W_2 + 9 = 11 \text{ months.}$$

So, the charges could be consistent with a fair assignment policy! The differences in waiting time could be due to differences in reneging rates. That $\delta_1/\delta_2 = .1824$ suggests that it takes non-whites $5\frac{1}{2}$ times longer than whites to find housing on the open market which in turn reflects the greater number of alternatives to public housing whites have relative to non-whites in Boston!

This does not prove that the BHA did *not* discriminate against non-whites. What it does show, however, is that the HUD claims alone cannot establish guilt (as long as you believe that assignment in proportion to application rates constitutes a fair policy).

8.2.4 Waiting for a Kidney

Recent data from Israel suggest that on average there are about 800 persons waiting to receive a kidney transplant. However, the actual number of Israelis who receive kidney transplants roughly equals 140 per year. Now, not all those who enter the waiting list for kidney transplants receive them (you can guess why), thus there is substantial reneging from the transplant waiting list. For those lucky enough to receive kidney transplants, the average waiting time is 3.25 years.

(a) Based solely on the figures above, estimate the annual demand for kidney transplants in Israel (HINT: public housing models).

The hint says "public housing models" and that's a good hint! Let λ be the annual demand for kidney transplants (i.e. the arrival rate), μ be the actual kidney transplant rate per year, and δ equal the reneging rate from the transplant queue (which by part (c) is the same as the mortality rate from kidney failure). Then we have two equations, one for the queue length and the other for the waiting time to receive a transplant for those lucky enough to get a new kidney. The equations are:

$$L_q = \frac{\lambda - \mu}{\delta}$$

and

$$W = \frac{1}{\delta} \log \frac{\lambda}{\mu}.$$

Now, the problem statement identifies $L_q = 800$, $\mu = 140$ per year, and $W = 3.25$ years. Substituting into the equations above yields

$$800 = \frac{\lambda - 140}{\delta}$$

and

$$3.25 = \frac{1}{\delta} \log \frac{\lambda}{140}.$$

So divide the first of these equations by the second to obtain

$$\frac{800}{3.25} = 246.15 = \frac{\lambda - 140}{\log(\lambda/140)}$$

and solve for λ. You need to do this numerically (e.g. using Excel, or via trial-and-error). The result is given by $\lambda = 395.8$ or about 400, which provides an estimate for the annual demand for kidney transplants. Now, this estimate does assume that the reneging rate is proportional to the queue length, but that's not an unreasonable assumption for this problem!

(b) What fraction of those in need of kidney transplants actually receive them?

Well, once we have an estimate of λ, the fraction of those in need of kidney transplants who actually receive them is given simply by

$$\frac{\mu}{\lambda} \approx \frac{140}{400} = 0.35.$$

That's not so good...

(c) Suppose that all reneging from the transplant waiting list is due to death from end-stage renal disease (i.e. kidney failure). If this were the case, what would the mean survival time for those in need of kidney transplants be (assuming it is no longer possible to obtain a kidney transplant)? Note that your answer must be consistent with the rest of the data in this problem.

This question is a wordy way of asking for the value of $1/\delta$. You can get this from either of the equations for the queue length or the waiting time once you know the arrival rate $\lambda \approx 400$. From the queue length equation you obtain

$$\frac{1}{\delta} = \frac{L_q}{\lambda - \mu} = \frac{800}{400 - 140} \approx 3.1 \text{ years.}$$

From the waiting time equation you obtain

$$\frac{1}{\delta} = \frac{W}{\log(\lambda/\mu)} = \frac{3.25}{\log(400/140)} \approx 3.1 \text{ years.}$$

8.2.5 Time Dependent Tenant Assignment Policy

This question pertains to the article *Tenant Assignment Policies with Time Dependent Priorities* that is contained in the Course Packet on our policy modeling website. Suppose that a small housing authority has a turnover of roughly 100 units per year. There are three main applicant groups. For each of these three groups, the authority has computed the difference between the average rent group members would pay to the authority, and the (higher) average rent group members must pay in their current living arrangements. The authority decides to use these differences as the cost of waiting for public housing, under the argument that, for example, a low income household paying \$500/month on the private market that would only have to pay \$200/month to the housing authority is suffering a cost of \$300/month of waiting in queue. The authority thus seeks a tenant assignment policy that will equalize the waiting costs for those who do gain a public housing assignment.

The arrival rates (λ), dropout rates (δ), and rent differentials (r) are shown in the table below:

	λ (applicants/yr)	δ (per waiting household/yr)	r (in \$/waiting household/mon
Group 1	150	1	100
Group 2	100	0.67	200
Group 3	50	0.5	300

(a) Of the 100 expected moveouts, how many units should the housing authority assign to each of the three groups annually?(HINT: use a spreadsheet!)

Let W_i equal the waiting time for applicants from group i that receive housing. From the fluid model discussed in class (and in the article), we know that

$$W_i = \frac{1}{\delta_i} \log \frac{\lambda_i}{\mu_i}$$

where μ_i is the (to be determined) assignment rate for group i households. Using the equation above to solve for μ_i we get

$$\mu_i = \lambda_i e^{-\delta_i W_i}.$$

Now, the time-dependent priority model discussed in the article works as follows: waiting applicants gain points according to group specific rates, corresponding to the rent differentials in this example. The "score" for someone in group i who has waited W_i time units is just $r_i W_i$ (though you have to convert the waiting costs to annual units, or equivalently work with waiting times in months, for the actual scores to work out properly). In equilibrium, all applicants who receive housing will have the same score (for recall that the authority seeks a tenant assignment policy that will equalize the waiting costs for those who do gain a public housing assignment). Call this score

s. Then $r_1W_1 = r_2W_2 = r_3W_3 = s$. This means that for the i^{th} group, applicants who get housed have waiting times given by

$$W_i = s/r_i$$

(where again you need to keep your time units straight)!

Substituting this into the equation for the assignment rates, we have

$$\mu_i = \lambda_i e^{-\delta_i s/r_i}.$$

Since the total number of assignments per year equals 100, it must be true that

$$\mu_1 + \mu_2 + \mu_3 = 100$$

which is equivalent to stating that

$$\lambda_1 e^{-\delta_1 s/r_1} + \lambda_2 e^{-\delta_2 s/r_2} + \lambda_3 e^{-\delta_3 s/r_3} = 100.$$

There is only one unknown in this equation: s, the equilibrium score. And there is one value of s that will solve this equation, as is clear from the figure below (which plots the total assignment rates implied by different values of s).

The solution is given by $s = \$2612.932$. Plugging this value of s into the formula for the assignment rates yields:

$\mu_1 = 17/yr$, $\mu_2 = 48.2/yr$, and $\mu_3 = 34.8/yr$. Note that the assignment rates sum to 100!

(b) What will the waiting times be for each applicant group?

Now just plug the assignment rates from (a) into the waiting time equations $W_i = 1/\delta_i \log(\lambda_i/\mu_i)$. The results are:

$W_1 = 2.2$ yrs, $W_2 = 1.1$ yrs, and $W_3 = 0.73$ yrs (about 9 months). Even more simply, use the equation $W_i = s/r_i$ after recalling that $s = \$2612.932$ (and after converting the given r's from monthly to annual costs). Again you will find that $W_1 = 2612.932/1200 = 2.2, W_2 = 2612.932/2400 = 1.2$,and $W_3 = 2612.932/3600 = 0.73$.

(c) What fraction of the applicants in each group will receive public housing?

Easy! For each group, this is given by μ_i/λ_i, resulting in assignment probabilities for each group of:

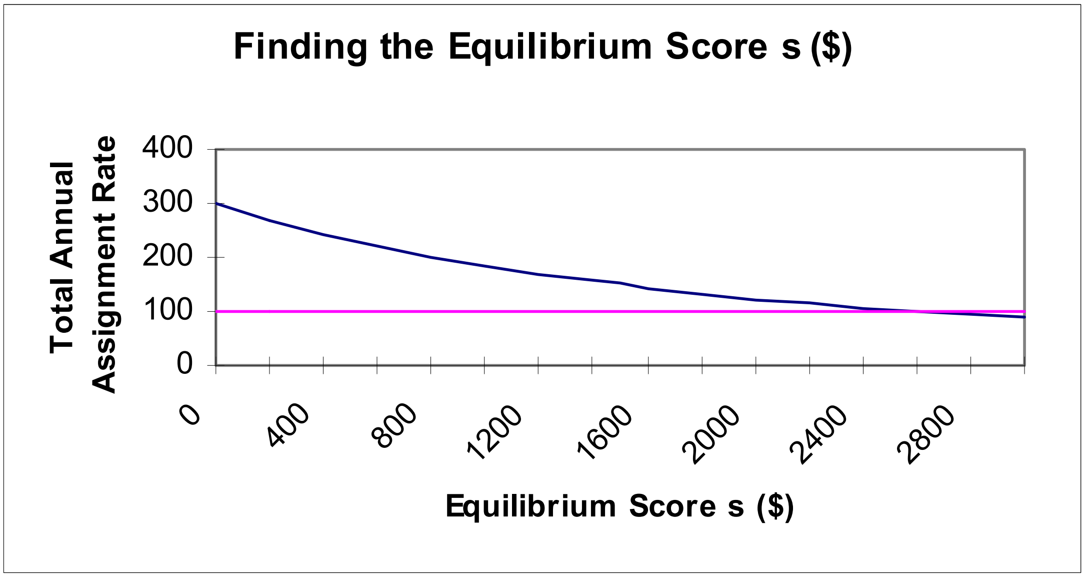


Figure 8.1:

Group 1: 11.3%, Group 2: 48.2%, Group 3: 69.6%.

(d) What is the expected cost per accepted household under this policy (measured in terms of the rent differential)?

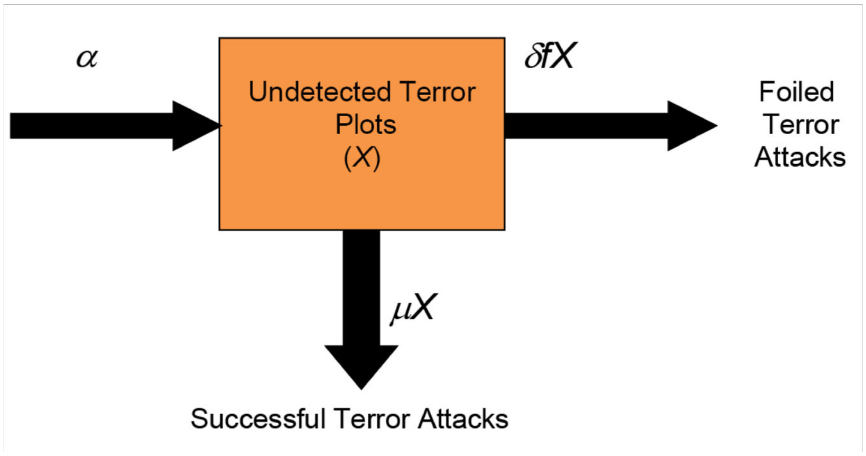
This is a giveaway, as you already solved this in part (a). The expected cost per accepted household is just the equilibrium score, which equals $2612.932.

8.2.6 The US Terror Queue?

Consider the terror queue approximation that applies when all agents are almost always available, that is, when $f - Y \approx f$ (i.e. $Y \approx 0$) as described in more detail in the paper Terror Queues in the course readings online (Section 3.1, p. 778 – see the discussion preceding and including equation (27)).

(a) Let X denote the number of undetected terror plots, W denote the time from when a new terror plot is hatched until it is either executed (i.e. carried out) or foiled (i.e. detected and interdicted), α equal the arrival rate (i.e. initiation rate) of new terror plots, μ be the rate at which terror plots are executed per plot per unit time, f denote the number of agents/informants available for detection, and δ denote the terror plot detection rate per plot per agent/informant. The approximation provides a simple formula for $E(X)$, the expected number of undetected plots, in terms of α, μ, f, and δ. Your job is to produce a simple formula for $E(W)$, the mean time from terror plot initiation until it is either foiled or executed.

First let's get a flow diagram of what's going on. Working from the info in the terror queue slides or paper or both, we can visually depict the model this way:



This leads to the following "terror flow balance" equation (where the flow is in units of terror plots per unit time):

$$\alpha = \delta f E(X) + \mu E(X) = (\delta f + \mu)E(X).$$

Now, this looks *exactly* like the basic flow balance for how many busy servers there are in a standard queueing model, and recalling that in such model, each busy server is matched with a customer in service, we immediately see that the expected number of customers (undetected terror plots)

$$E(X) = \frac{\alpha}{\delta f + \mu}.$$

Interpreting this equation as "Little's Theorem" where α is the customer arrival rate (think "λ" for standard queueing models) and $E(X)$ is the mean number of customers in the system (think "L" for standard queueing models), it must be that the mean time from the initiation of a terror plot (customer arrival) until the plot is foiled or executed (time customer leaves the system) is given by

$$E(W) = \frac{1}{\delta f + \mu}$$

(or just plain "W" in $L = \lambda W$). That is, we must have

$$E(X) = \alpha E(W) = \frac{\alpha}{\delta f + \mu}\cdot$$

This has an additional interesting interpretation: the mean time required from the initiation of a terror plot to its execution *in the absence of any intelligence effort* (that is, if $f = 0$, or equivalently completely ineffective intel effort, in which case $\delta = 0$) just equals $1/\mu$. Conversely, if all plots took *forever* to plan ($\mu = 0$), then the average time it would take to foil a plot would just equal $1/(\delta f)$. The time required to either foil or execute a plot is then the *minimum* of the time to execute and the time to foil, and in expectation this is given by $1/(\delta f + \mu)$.

(b) The figure below reports publicly disclosed foiled and executed terror attacks in the *United States* from 1999 to 2009 inclusive. Executed attacks include terror plots that failed at the last minute due to malfunctions of some sort but otherwise escaped detection (examples include Faisal Shahzad's attempted car bomb in Times Square, and the Christmas attempt of "Captain Underpants" to bring down Northwest Airlines Flight 253). Suppose that the data in this figure do in fact represent all foiled and executed terror attacks in the United States over the time period shown.

(i) Estimate the annual terror plot arrival rate from these data.

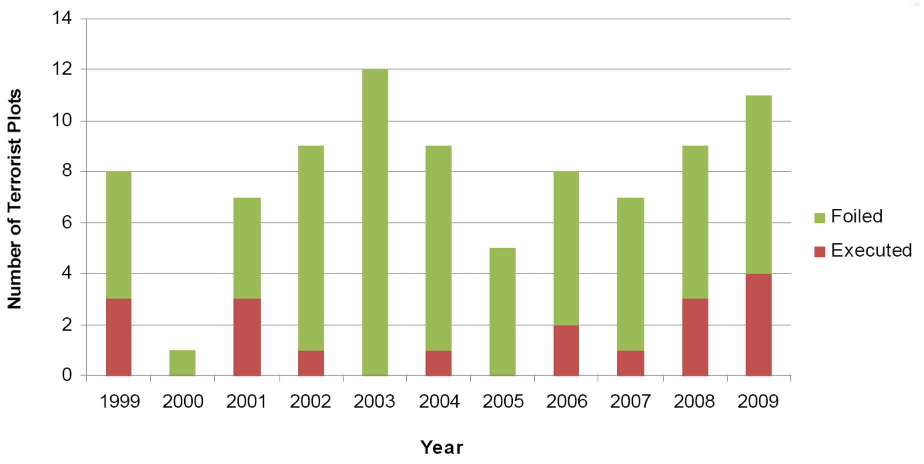
The model employed presumes that the terror queue is in steady state, in which case the terror plot arrival rate equals the terror plot departure rate. A terror plot departs the system whenever it is executed or foiled, so the problem just reduces to finding the average of the observed number of *departures* from the terror queue. Since the data in the figure below correspond to departures from the terror queue, all you have to do is find the average number of plots that were either foiled or executed over the eleven years shown. That reduces to just adding up the heights of the bars and dividing by eleven! There were a total of 86 plots that were foiled or executed, so the annual terror plot arrival rate is estimated as $86/11 = 7.8$ plots per year.

(ii) Estimate the fraction of terror plots that are foiled.

This is also pretty easy: we know that the total number of plots foiled or executed over the eleven years shown equals 86; of these we can easily see (from tallying the red bars) that a total of 18 were executed, which means that $86 - 18 = 68$ were foiled. Thus, our estimate of the fraction of terror plots that are foiled is given by $68/86 = 0.79$ or 79%. In rough terms, about 1 in 5 plots were executed with the remaining 4 in 5 foiled.

(iii) What mathematical expressions (in terms of α, μ, f, and δ) do the quantities in (i) and (ii) above correspond to?

Figure 1. Foiled and Executed Terrorist Plots by Year, 1999–2009



The annual terror plot arrival rate from (i) is just $\alpha = 7.8$ plots per year. To understand how the model corresponds to the fraction of terror plots that are foiled in (ii), look back at the flow diagram presented in discussing the solution to part (a) of this problem. Note that the rate with which plots are foiled is given by δf *per plot* per unit time, while the rate with which plots are executed equals μ *per plot* per unit time. Thus, the fraction of plots that are foiled will equal the rate with which plots are foiled, divided by the rate with which plots are foiled or executed, that is

$$\text{Fraction of terror plots foiled } = \frac{\delta f}{\delta f + \mu}.$$

We can also write this ratio by splitting the annual terror plot arrival rate into those that are foiled and those that are executed: on average α new plots arrive per year, and of these $\delta f E(X)$ are foiled per year on average (look at the flow diagram!), which yields

$$\text{Fraction of terror plots foiled } = \frac{\delta f E(X)}{\alpha}.$$

I'll leave it to you to substitute the formula for $E(X)$ into the expression above to show that you arrive at the same result for the fraction of terror plots foiled derived previously.

(c) Suppose that based on interrogating captured terrorists and investigations conducted after successful attacks, intelligence officials estimate that the mean time from terror plot initiation until interdiction or execution $E(W) = 1.3$ years. Given this figure and your answer to part (a), estimate the mean and variance of the number of undetected active terror plots targeting the United States.

Well, we previously saw that $E(X) = \alpha E(W) = \alpha/(\delta f + \mu)$, and we've just been given the estimate that $E(W) = 1.3$ years. So, to get the mean number of undetected terror plots, we just compute

$$E(X) = \alpha E(W) = 7.8 \times 1.3 = 10.1.$$

The intelligence estimate $E(W) = 1.3$ years combined with our model and the estimated value of the terror plot arrival rate α from Figure 1 suggest that there are about 10 expected undetected plots targeting the United States!

Now, the question also asked for an estimate of the *variance* of the number of undetected plots. This required some careful reading of the class materials on your part. On slide 21 of the terror queue presentation, you'll see the direct statement

$$E(X) = Var(X) = \frac{\alpha}{\delta f + \mu}$$

which pretty much gives it away, no? Or, on p. 778 of the paper, just before equation (27) where it was suggested that you look, you'll see the statement that X has a Poisson distribution with mean and variance equal to $\alpha/(\delta f + \mu)$. So, the estimated variance of the number of undetected terror plots equals the estimated mean of 10.1.

(d) Using the same model, what is the probability that there are more than 5 undetected terror plots targeting the United States?

As just stated in answering part (c), the key is that the number of undetected plots X has a Poisson distribution with mean (and variance) roughly equal to 10 (OK, 10.1 to one decimal place). So, to find the probability that there are more than 5 undetected plots, the easiest thing to do this is find the probability that there are *at most* 5 undetected plots and then subtract from 1, that is

$$\Pr\{X > 5\} = 1 - \Pr\{X \leq 5\}.$$

You can do this directly using the Poisson formula – using $E(X) = 10.1$ you would compute

$$\Pr\{X \leq 5\} = \sum_{x=0}^{5} \frac{(10.1)^x e^{-10.1}}{x!} = 0.0634$$

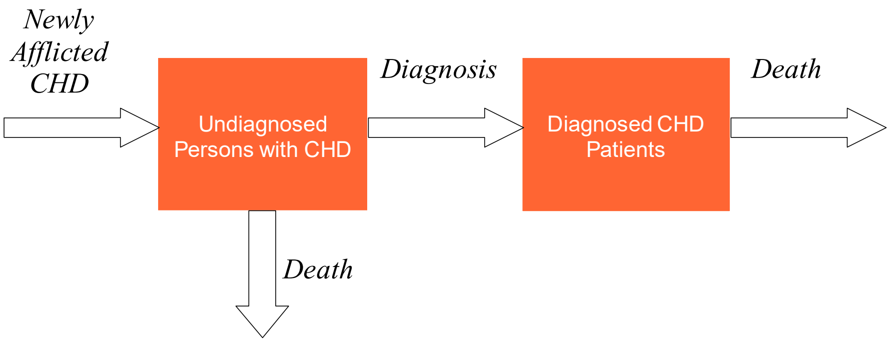
and hence $\Pr\{X > 5\} = 1 - 0.0634 = 0.936\,6$ or about 94%. Or, you could do this in one step using the Excel command = 1− poisson(5,10.1,1) = 0.9366. So, according to the model (via the intel estimate and data from Figure 1), there is about a 94% chance that there are more than 5 undetected terror plots targeting the United States! We better catch them!

8.2.7 Terror Queue of the Heart (with apologies to Bonnie Tyler)

In the United States, the annual estimated all-cause number of deaths for those with Coronary Heart Disease (CHD), including both known cases and those undiagnosed, approximately equals 670,000. There are an estimated 17.6 million known CHD patients. It is also known that those newly afflicted with CHD average 46.3 years until either diagnosis or death, whichever comes first.

(a) Given these data, what is the mean number of undiagnosed persons with CHD?

With only the data in the problem to go by, let's see where a steady-state analysis gets us, taking advantage of the hint provided by the title of the problem. Just like in the Terror Queue model, here we have an arrival rate of persons undiagnosed with CHD. Such an undiagnosed person will either die or be diagnosed, whichever comes first. Once diagnosed, death still awaits, but hopefully at a lower rate given treatment for CHD. In steady state, then, the arrival rate of new undiagnosed persons with CHD will equal the all-cause number of deaths for all with CHD, whether diagnosed or not. See figure below:



This means that the arrival rate is approximately 670,000 per year. And, we are told that those newly afflicted with CHD average 46.3 years until diagnosis or death. So, the mean number of undiagnosed persons with CHD just follows from Little's Theorem and is given by

$$E(\#\text{ persons with CHD and undiagnosed}) = 670,000 \times 46.3 = 31,021,000.$$

That's a lot of undiagnosed CHD!

(b) Suppose that the all-cause per capita mortality rate is 3.5 times higher for undiagnosed persons with CHD than the per capital mortality rate for known CHD patients. What are the per capita all-cause mortality rates for undiagnosed versus diagnosed persons with CHD?

Let μ_U and μ_D be the all-cause mortality rates for persons with CHD who are undiagnosed and diagnosed respectively. From the problem statement and our answer to (a) above, we see that the total all-cause deaths per year is given by

$$31,021,000\mu_U + 17,600,000\mu_D = 670,000.$$

However, we are now told that $\mu_U = 3.5\mu_D$ (the all-cause per capita mortality rate is 3.5 times higher for undiagnosed persons with CHD than the per capital mortality rate for known CHD patients), which leads to the equation

$$31,021,000 \times 3.5\mu_D + 17,600,000\mu_D = 670,000$$

which implies that the all-cause mortality rate for those known to have CHD is given by

$$\mu_D = \frac{670,000}{31,021,000 \times 3.5 + 17,600,000} = 0.0053$$

or about half a percent per year. This implies that the all-cause mortality rate for undiagnosed persons with CHD equals $3.5 \times 0.0053 = 0.018\,6$ or almost 2%. To check, note that

$$31,021,000 \times 0.0186 + 17,600,000 \times 0.0053 \approx 670,000$$

as required.

(c) What fraction of those newly afflicted with CHD will be diagnosed?

To figure out the fraction of all those newly afflicted with CHD who will be diagnosed, just divide the annual number of all-cause deaths among persons diagnosed with CHD by the annual number of all-cause deaths among all persons with CHD. This yields

$$\Pr\{\text{Person Newly Afflicted with CHD is Diagnosed}\} = \frac{17,600,000 \times 0.0053}{670,000} = 0.139$$

or about 14%.

(d) What fraction of all persons living with CHD have been diagnosed?

By contrast, the fraction of all those living with CHD who have been diagnosed is given by

$$\frac{17,600,000}{31,021,000 + 17,600,000} = 0.362$$

or just over 36%. Clearly, if someone has CHD, it is really important to get diagnosed, as diagnosed CHD patients have a much lower mortality rate than undiagnosed persons with CHD, meaning that CHD treatment keeps one alive longer.

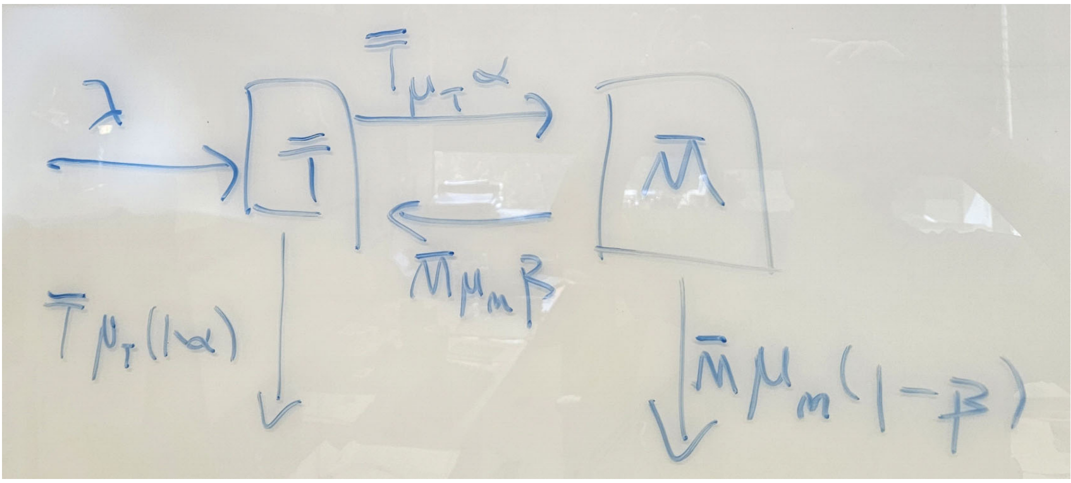
8.2.8 Treatment and Monitoring

Patients arrive to a specialty hospital that provides patient treatment and monitoring in accord with a Poisson process with arrival rate λ. The hospital has a large number of beds relative to the number of admitted patients, so there is no bed capacity constraint. Newly admitted patients proceed directly to the treatment unit. Whenever a patient arrives to the treatment unit, the length of time a patient spends in treatment is exponentially distributed with mean $1/\mu_T$ (so treatment episodes are completed at rate μ_T per patient per unit time). When a patient completes a treatment episode, they are discharged from the hospital with probability $1-\alpha$, but with probability α the patient is transferred to the observation unit for monitoring. Whenever a patient is sent to the observation unit, the length of time spent being monitored is exponentially distributed with mean $1/\mu_M$ (so monitoring episodes are completed with rate μ_M per patient per unit time). When a patient completes a monitoring episode, they are discharged from the hospital with probability $1-\beta$, but with probability β the patient is transferred back to the treatment unit to begin a new treatment episode.

Suppose that the patient arrival, treatment and monitoring processes have been operating for a long time and that the hospital has reached steady state. Let the random variables T and M refer to the number of patients in the treatment and monitoring units respectively, and let $\overline{T} = E(T)$ and $\overline{M} = E(M)$ denote the expected number of patients in the treatment and monitoring units respectively.

(a) Draw a diagram that indicates all patient flows in the hospital. Your diagram must show new arrivals, discharges from both the treatment and monitoring units, and transfers from treatment to monitoring and vice-versa. Identify $\overline{T}$ and $\overline{M}$ on the diagram, and show formulas for all patient flows along each flow branch using the notation described in the problem statement ($\lambda, \mu_T, \mu_M, \alpha, \beta, \overline{T}$, and $\overline{M}$).

Here's a picture of the flow diagram I drew on my office whiteboard:


λ
$\overline{T}\mu_T\alpha$
$\overline{T}$
$\overline{M}$
$\overline{M}\mu_m\beta$
$\overline{T}\mu_T(1-\alpha)$
$\overline{M}\mu_m(1-\beta)$

Note that patients arrive directly to the treatment unit (which is why the λ arrow connects to the box containing the average number of patients in treatment, $\overline{T}$). Also note that patients leave the treatment unit at aggregate rate $\overline{T}\mu_T$ per unit time, but are then (Bernoulli) split to go to monitoring (with probability α) or discharge (with probability $1-\alpha$). Similarly, patients leave the monitoring unit at aggregate rate $\overline{M}\mu_M$ per unit time, but are then (Bernoulli) split to go back to treatment (with probability β), or discharge (with probability $1-\beta$).

(b) Produce flow equations and solve them to determine formulas for $\overline{T}$ and $\overline{M}$ respectively in terms of the other parameters in the problem (meaning that if you knew the numerical values of $\lambda, \mu_T, \mu_M, \alpha$, and β, you would be able to directly compute the numerical values of $\overline{T}$ and $\overline{M}$).

The equations needed follow directly from the flow diagram above. For each of the two hospital units, you must have the entering flow set equal to the exit flow. So, the Inflow = Outflow equations for the treatment and monitoring units respectively are given by:

$$\lambda + \overline{M}\mu_M\beta = \overline{T}\mu_T$$

and

$$\overline{T}\mu_T\alpha = \overline{M}\mu_M.$$

Solving the second equation yields

$$\overline{M} = \frac{\mu_T\alpha}{\mu_M}\overline{T},$$

and substituting this into the first equation gives

$$\lambda + \frac{\mu_T \alpha}{\mu_M}\overline{T} \times \mu_M \beta = \overline{T}\mu_T.$$

Solving for $\overline{T}$ we obtain

$$\overline{T} = \frac{\lambda}{\mu_T}\frac{1}{1-\alpha\beta}$$

and consequently from the third equation we obtain

$$\overline{M} = \frac{\lambda\alpha}{\mu_M}\frac{1}{1-\alpha\beta}.$$

(c) Produce a formula for the fraction of all patients in the hospital who are undergoing treatment, and by implication the fraction of all patients in the hospital who are being monitored, in terms of the other parameters of the problem (meaning $\lambda, \mu_T, \mu_M, \alpha$, and β).

This is easy – the fraction of all patients undergoing treatment is just given by

$$\begin{aligned}\frac{\overline{T}}{\overline{T}+\overline{M}} &= \frac{\frac{\lambda}{\mu_T}\frac{1}{1-\alpha\beta}}{\frac{\lambda}{\mu_T}\frac{1}{1-\alpha\beta} + \frac{\lambda\alpha}{\mu_M}\frac{1}{1-\alpha\beta}} \\ &= \frac{\mu_M}{\mu_M + \mu_T\alpha}.\end{aligned}$$

The fraction of all patients in the hospital undergoing monitoring is found by subtracting the result above from unity and thus equals $\mu_T\alpha/(\mu_M+\mu_T\alpha)$. Note that if $\alpha = 0$, all patients in the hospital would be in treatment as patients would never enter monitoring, while if $\alpha = 1$, patients would always go through both treatment and monitoring before either being discharged (with probability $1-\beta$) or being sent back for more treatment (with probability β). In this case, the fraction of patients in treatment would be the same as the probability that a patient selected at random was in the treatment portion of a treatment-monitoring pair. Since the per episode average time in treatment (monitoring) equals $1/\mu_T$ $(1/\mu_M)$, the chance of finding a patient in the treatment part of a paired treatment/monitoring episode just equals the ratio of the (per episode) average time in treatment to the sum of the (per episode) average treatment and monitoring times. This ratio is given by

$$\frac{1/\mu_T}{1/\mu_T + 1/\mu_M} = \frac{\mu_M}{\mu_M + \mu_T}.$$

(d) Again in terms of the other parameters of the problem (meaning $\lambda, \mu_T, \mu_M, \alpha$, and β), produce a formula for the fraction of all discharged hospital patients who are released from the treatment unit (equivalently, the probability that when a patient is discharged, it is from the treatment unit as opposed to the monitoring unit) following two different approaches:

(i) By working with the flows of patients exiting the hospital, and

Here we just need to look at the ratio of the discharged-from-treatment flow to the sum of the discharged-from-treatment and discharged-from-monitoring flows. This ratio is given by

$$\begin{aligned}\frac{\overline{T}\mu_T(1-\alpha)}{\overline{T}\mu_T(1-\alpha) + \overline{M}\mu_M(1-\beta)} &= \frac{\overline{T}\mu_T(1-\alpha)}{\overline{T}\mu_T(1-\alpha) + \frac{\mu_T\alpha}{\mu_M}\overline{T}\mu_M(1-\beta)} \\ &= \frac{1-\alpha}{1-\alpha+\alpha-\alpha\beta} \\ &= \frac{1-\alpha}{1-\alpha\beta}.\end{aligned}$$

This is interesting – the result only depends upon α and β. It does not depend in any way on the amount of time spent in treatment or monitoring. So, there should be a more direct way to get this result, which leads us to...

(ii) Using the repetition method (HINT: East Rock).

The hint says East Rock and it's a pretty good hint. Think of being discharged from treatment as going off the "left cliff" in the East Rock problem, and think of being discharged from monitoring as going off the "right cliff." Also, think of being in treatment as being in "position 1" in the East Rock problem, while monitoring corresponds to position 2. Starting in treatment (position 1), what is the probability that the process ends by going off the left cliff? Let's call that q_1 and write

$$q_1 = (1-\alpha) \times 1 + \alpha \times q_2$$

where we recognize that starting in position 1, there is a $(1-\alpha)$ chance of going off the left cliff (in which case we stay there), or an α chance of going to position 2 (treatment) from which there must be some probability q_2 of going off the left cliff given you start in position 2! It's our usual policy modeling trick – when you don't know something, assume you do! Now, from position 2, we can write

$$q_2 = (1-\beta) \times 0 + \beta q_1$$

since with probability $(1-\beta)$ you'll go off the right cliff (be discharged from monitoring) and thus there is no chance of going off the left cliff, while with probability β you return to position 1 (treatment). So, substituting the expression for q_2 into the equation for q_1 we obtain

$$q_1 = (1-\alpha) \times 1 + \alpha \times \beta q_1$$

which solves to yield

$$q_1 = \frac{1-\alpha}{1-\alpha\beta}.$$

How cool is that?

(e) A new patient has just been admitted to the hospital. What is the expected total number of treatment episodes this patient will spend in the hospital before discharge? What is the expected total number of monitoring episodes this patient will spend in the hospital before discharge? Produce formulas for each in terms of the other parameters of the problem (meaning $\lambda, \mu_T, \mu_M, \alpha$, and β) (HINT: repetition method).

The hint says repetition method, but it could just as easily have said East Rock again. Let's stick with the analogy made in part (d)-(ii) above. Let τ_1^T equal the expected number treatment episodes until discharge for a patient who just entered treatment. Similarly, let τ_2^T be the expected number of treatment episodes until discharge for a patient who just entered monitoring. Using our East Rock logic we have

$$\tau_1^T = 1 + (1-\alpha) \times 0 + \alpha \times \tau_2^T$$

and

$$\tau_2^T = (1-\beta) \times 0 + \beta \times \tau_1^T.$$

In the first of these equations, since a patient has just entered treatment, we augment the expected number of treatment episodes until discharge by one. If the patient is discharged immediately following treatment, there are no more treatment episodes (this happens with probability $(1-\alpha)$ but with probability α the patient goes to monitoring from which there will be an average of τ_2^T additional *treatment* episodes. Now, note that in the equation for τ_2^T we do *not* start by counting 1 as it could be that there will not be *any* additional treatment episodes (indeed that is the case with probability $1-\beta$), but with probability β there is a return to treatment from which an *additional* expected τ_1^T treatment episodes will occur. Substituting the second equation into the first we obtain

$$\tau_1^T = 1 + \alpha\beta \times \tau_1^T$$

which solves to

$$\tau_1^T = \frac{1}{1-\alpha\beta}.$$

We have just learned that on average a newly admitted patient will experience $1/(1-\alpha\beta)$ treatment episodes.

What about monitoring? Here we need to be careful. Using similar notation to denote the expected number of additional monitoring episodes for patients newly admitted to treatment or to monitoring we obtain

$$\tau_1^M = (1-\alpha) \times 0 + \alpha \times \tau_2^M$$

and

$$\tau_2^M = 1 + (1-\beta) \times 0 + \beta \times \tau_1^M.$$

Notice the slight difference in these equations from those used to find the average number of treatment episodes. Here, we note that a person who newly enters treatment might not experience any additional monitoring episodes, while a patient who newly enters monitoring does undergo an additional monitoring episode. The question again asks for the expected number of monitoring episodes for a patient who was just admitted to the hospital, and all newly admitted patients first go to treatment, so the quantity we seek is τ_1^M. We find that

$$\tau_1^M = \alpha \times \tau_2^M = \alpha \times (1 + \beta\tau_1^M)$$

which solves to

$$\tau_1^M = \frac{\alpha}{1-\alpha\beta}.$$

Note that if $\alpha = 0$, patients are discharged after treatment with certainty, and thus the expected number of treatment episodes until discharge is equal to one while the expected number of monitoring episodes is equal to zero – the patient is never monitored. On the other hand, if $\alpha = 1$, then patients are always monitored following treatment, and thus the expected number of treatment and monitoring episodes both equal $1/(1-\beta)$.

(f) In terms of the other parameters of the problem (meaning $\lambda, \mu_T, \mu_M, \alpha$, and β), produce a formula for the expected total time that a newly admitted patient from the outside will spend in the hospital that accounts for all possible treatment and monitoring episodes until discharge following two different approaches:

(i) Use your results from part (e) combined with the average duration of treatment and monitoring episodes.

Wow, this is really easy after answering part (e)! Since we have already figured out the expected number of treatment and monitoring episodes, and we know from the problem statement the expected time spent per treatment and monitoring episode, we see immediately that the expected total time a newly admitted patient spends in the hospital, let's call this W (you'll see why in part (ii)), is given by

$$\begin{aligned} W &= \frac{\tau_1^T}{\mu_T} + \frac{\tau_1^M}{\mu_M} \\ &= \frac{1}{1-\alpha\beta} \times \left(\frac{1}{\mu_T} + \frac{\alpha}{\mu_M} \right). \end{aligned}$$

That was easy!

(ii) Use Little's Theorem and your results from part (b).

Little's Theorem says $L = \lambda W$. In this problem, the expected number of "customers" in the "system" (L) is just the expected number of patients in the hospital, which in turn is the expected total number of patients in treatment plus the expected number in monitoring. We figured out both of these quantities in part (b), so plugging in we have

$$\begin{aligned} L &= \overline{T} + \overline{M} \\ &= \frac{\lambda}{\mu_T} \frac{1}{1-\alpha\beta} + \frac{\lambda\alpha}{\mu_M} \frac{1}{1-\alpha\beta} \\ &= \frac{\lambda}{1-\alpha\beta} \times \left(\frac{1}{\mu_T} + \frac{\alpha}{\mu_M} \right). \end{aligned}$$

So, now using Little's Theorem we discover (!) that

$$\begin{aligned}
W &= \frac{L}{\lambda} \\
&= \frac{1}{\lambda} \times \frac{\lambda}{1-\alpha\beta} \times \left(\frac{1}{\mu_T} + \frac{\alpha}{\mu_M}\right) \\
&= \frac{1}{1-\alpha\beta} \times \left(\frac{1}{\mu_T} + \frac{\alpha}{\mu_M}\right).
\end{aligned}$$

Wow!! Exactly the same answer as in (f)-(i) above. See how all this stuff fits together?

8.2.9 Terror Queue Staffing Models

The terror queue model described in the article of the same name in your course pack defines α as the terror plot initiation rate (new plots per unit time), μ as the terror plot completion rate (so in the absence of any counterterror detection effort, terror plots would average $1/\mu$ time units from inception to execution), δ as the detection rate per plot per available agent per unit time, ρ as the interdiction rate per detected plot per unit time (so a newly detected plot requires on average $1/\rho$ time units to be interdicted), and f as the total number of undercover agents deployed (that is, counting both agents actively involved in interdicting known plots, and agents search for undetected plots).

(a) Suppose that the counterterror agency (e.g. the FBI) wants to deploy enough agents to detect (and interdict) a fraction θ of all new terror plots. That is, the counterterror agency seeks to determine the value of f such that a newly arriving plot will be detected (and interdicted) with probability θ. Determine f using the following three-step argument:

(i) The number of agents $f = f_a + f_b$ where f_a = # of agents available for detection (on average), and f_b = average # of agents who are busy interdicting plots (= # of detected plots being interdicted).

(ii) The mean number of busy agents f_b must, via Little's Theorem, satisfy

$$f_b = \alpha\theta \times \frac{1}{\rho}\cdot$$

(iii) The mean number of agents available for detection, f_a, must satisfy

$$\frac{\delta f_a}{\delta f_a + \mu} = \theta\cdot$$

You are to justify each of the three steps presented above, and then determine the function $f(\theta)$ that reports the number of agents required to detect (and interdict) a fraction θ of all new terror plots.

Step (i) is true by definition, since each agent is either available for detecting existing undetected plots, or is "busy" and in the act of interdicting a detected plot. For step (ii), note that if a fraction θ of new plots are detected, then the rate with which plots are interdicted must equal $\alpha\theta$ plots per unit time (that is, of all new plots, a fraction θ are interdicted). So, think of $\alpha\theta$ as the arrival rate of customers (terror plots) to "receive service" (get interdicted), and $1/\rho$ as the mean service time which is the same as the mean time that a detected plot spends getting interdicted. Note that detected plots never have to wait in an "interdiction queue" as for every detected plot, there is an accompanying server (agent) who immediately begins the interdiction process. For step (iii), suppose that there are x undetected plots. Then the total rate at which plots are detected when there are f_a servers available equals $\delta f_a x$ while the total rate with which plots execute equals μx. The fraction of plots that are detected is then given by the total plot detection rate, divided by the sum of the total plot detection and execution rates; this ratio is given by

$$\frac{\delta f_a x}{\delta f_a x + \mu x} = \frac{\delta f_a}{\delta f_a + \mu}\cdot$$

But, the goal is to detect (and interdict) a fraction θ of all new plots, so equating our two expressions for the fraction of plots that are detected yields

$$\frac{\delta f_a}{\delta f_a + \mu} = \theta\cdot$$

Solving the equation above for f_a in terms of θ yields

$$f_a = \frac{\mu}{\delta}\frac{\theta}{1-\theta}$$

while the Little's Theorem expression for f_b from step (ii) gives

$$f_b = \frac{\alpha}{\rho}\theta \cdot$$

Finally, step (i) says that $f = f_a + f_b$ so our final result for $f(\theta)$ is

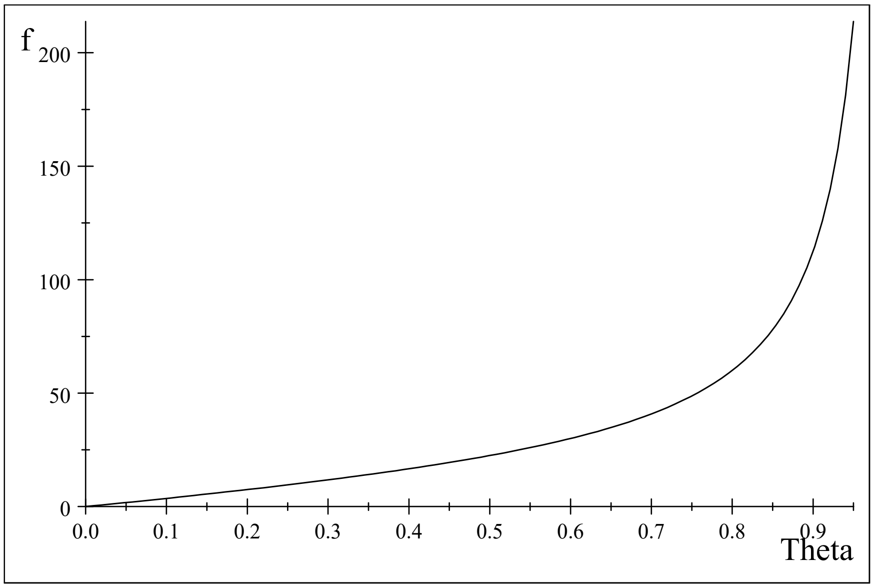
$$f(\theta) = \frac{\mu}{\delta}\frac{\theta}{1-\theta} + \frac{\alpha}{\rho}\theta \cdot$$

(b) Suppose that the agency in question is operating in an environment where $\alpha = 100/yr$, $\mu = 1/yr$ (so terror plots on average require one year to plan and execute), $\rho = 4/yr$ (so the average time to interdict a plot post-detection equals 3 months), and $\delta = 0.1$ per plot per agent per year (so if, for example, there were 10 available agents and 5 plots in progress, the total plot detection rate would equal $5 \times 10 \times .1 = 5$ detections per year, or 2.4 months until detection on average). For these parameters, produce a plot of f as a function of θ. How many agents are required to interdict 50% of all plots ($\theta = 0.5$)? How many are required to prevent 90% of all agents ($\theta = 0.9$)? What can you say in general about the shape of $f(\theta)$?

Substituting in the values for α, μ, ρ and δ into the result from (a) yields

$$\begin{aligned} f &= \frac{1}{.1}\frac{\theta}{1-\theta} + \frac{100}{4}\theta \\ &= 10\frac{\theta}{1-\theta} + 25\theta \cdot \end{aligned}$$

Plotting as a function of θ yields



$$f = 10\frac{\theta}{1-\theta} + 25\theta$$

To interdict 50% of all plots requires $10\frac{.5}{1-.5} + 25 \times .5 = 22.\,5$ agents, while to interdict 90% of all plots requires $10\frac{.9}{1-.9} + 25 \times .9 = 112.\,5$ agents. The shape of $f(\theta)$ is convex, with the required number of agents exploding in unbounded fashion as $\theta \longrightarrow 1$. Pretty sobering reminder of why it is not realistic to expect that all terror attacks can be prevented!

(c) In the United States, it has been estimated that of 35 attempted Jihadi terror attacks in the United States between 9/11/2001 and 6/30/2011, seven successfully evaded detection while the remaining 28 were interdicted. Considering the 9.8 years over which these 35 attempted attacks took place, let $\alpha = 3.57$ new Jihadi plots per year. Also, given that 28 of 35 plots were successfully interdicted, take $\theta = 0.80$. Finally, suppose that $\mu = 4/15$, $\delta = 1/1500$, and $\rho = 16$. With these parameter assignments, how many undercover agents would you estimate have been devoted to finding and interdicting Jihadi terror plots in the United States?

This is just plug and play: $f = \frac{\mu}{\delta}\frac{\theta}{1-\theta} + \frac{\alpha}{\rho}\theta = \frac{4/15}{1/1500} \times \frac{0.8}{1-0.8} + \frac{3.57}{16} \times 0.8 =$ 1600.2. And the number becomes interesting given the following: according to an FBI report (http://www.fbi.gov/stats-services/publications/fbi_ct_911com_0404.pdf), since 9/11 the FBI "...increased the number of Special Agents working terrorism matters from 1351 to 2398." Hmmmmmmm.....

(d) Suppose that the government estimates a benefit b for each terror plot that is interdicted, but also incurs a cost c for each agent deployed. Argue that the net benefits of deploying f agents are then given by

$$NB(f) = b\alpha\theta - cf$$

where as before, θ is the fraction of plots that are interdicted. Using your model from part (a) above that expresses the number of agents that must be deployed to interdict a fraction θ of all attacks, and using the parameters for α, μ, δ and ρ from part (c) above, what fraction of all terror plots should be prevented if the government seeks to maximize net benefits when the benefit to cost ratio $r = b/c = 2,100$? What is the implied optimal number of counterterror agents that the government should deploy? You can do this numerically, or you can produce an exact formula.

First, if there is a benefit b for each attack prevented, and a fraction θ of α attacks per year are prevented, then the annual benefit of preventing attacks at this level clearly equals $b\alpha\theta$, while the costs of doing so equal cf, so the formula is correct. What is important to note is that f can be expressed as a function of θ via our formula from part (a), so plugging in and rewriting the net benefits as a function of θ yields

$$\begin{aligned} NB(\theta) &= b\alpha\theta - cf(\theta) \\ &= b\alpha\theta - c\left(\frac{\mu}{\delta}\frac{\theta}{1-\theta} + \frac{\alpha}{\rho}\theta\right). \end{aligned}$$

We want to maximize the net benefits as a function of θ, and once we figure out what the right value of θ is, we can just plug it into the result from part (a) to figure out the optimal staffing level f. We can do this analytically – differentiating the net benefits with respect to θ yields

$$\begin{aligned} \frac{d}{d\theta}NB(\theta) &= b\alpha - c\left(\frac{\mu}{\delta}\left(\frac{1}{1-\theta} + \frac{\theta}{(1-\theta)^2}\right) + \frac{\alpha}{\rho}\right) \\ &= b\alpha - c\left(\frac{\mu}{\delta\,(1-\theta)^2} + \frac{\alpha}{\rho}\right). \end{aligned}$$

Equating to zero and solving for θ yields

$$b\alpha = c\left(\frac{\mu}{\delta\,(1-\theta)^2} + \frac{\alpha}{\rho}\right)$$

which is the same as

$$b\alpha - \frac{c\alpha}{\rho} = \frac{c\mu}{\delta\left(1-\theta\right)^2},$$

which is the same as

$$(1-\theta)^2 = \frac{c\mu}{\delta}\frac{\rho}{b\alpha\rho - c\alpha}$$

which is the same as

$$1-\theta = \sqrt{\frac{\mu\rho}{\alpha\delta}\frac{1}{\frac{b}{c}\rho - 1}}.$$

If we denote the benefit-to-cost ratio b/c by r, we arrive at

$$\theta = 1 - \sqrt{\frac{\mu\rho}{\alpha\delta}\frac{1}{r\rho - 1}}$$

as the general result. Note that the larger this benefit-to-cost ratio, the larger the value of θ (the greater the fraction of plots we prevent by deploying more agents in accord with the result of part (a)). With the parameter values given, the optimal fraction of terror plots to prevent is given by

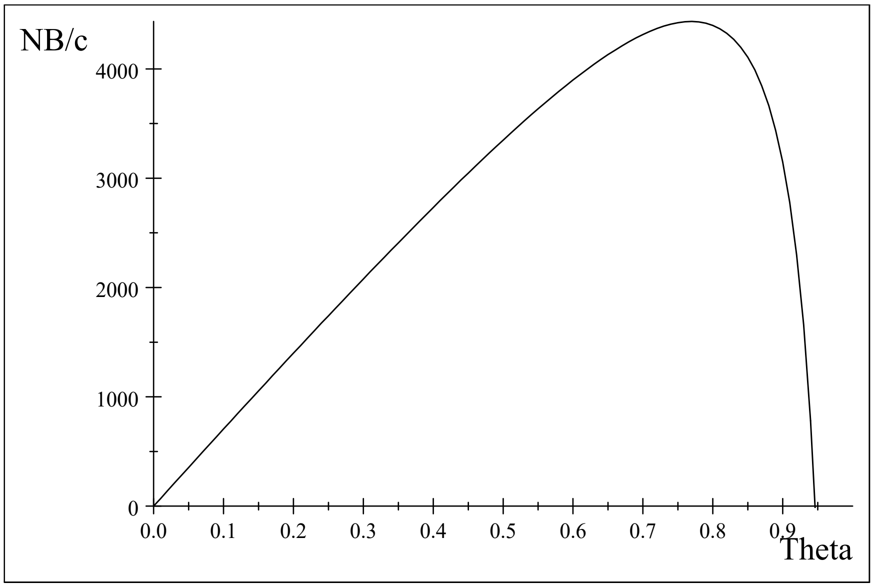
$$\theta = 1 - \sqrt{\frac{4/15 \times 16}{3.57 \times 1/1500}\frac{1}{2100 \times 16 - 1}} = 0.77.$$

From this, the result of part (a) tells us that we should assign $f = \frac{\mu}{\delta}\frac{\theta}{1-\theta} + \frac{\alpha}{\rho}\theta = \frac{4/15}{1/1500} \times \frac{0.77}{1-0.77} + \frac{3.57}{16} \times 0.77 = 1339.\,3$ or about 1,340 agents.

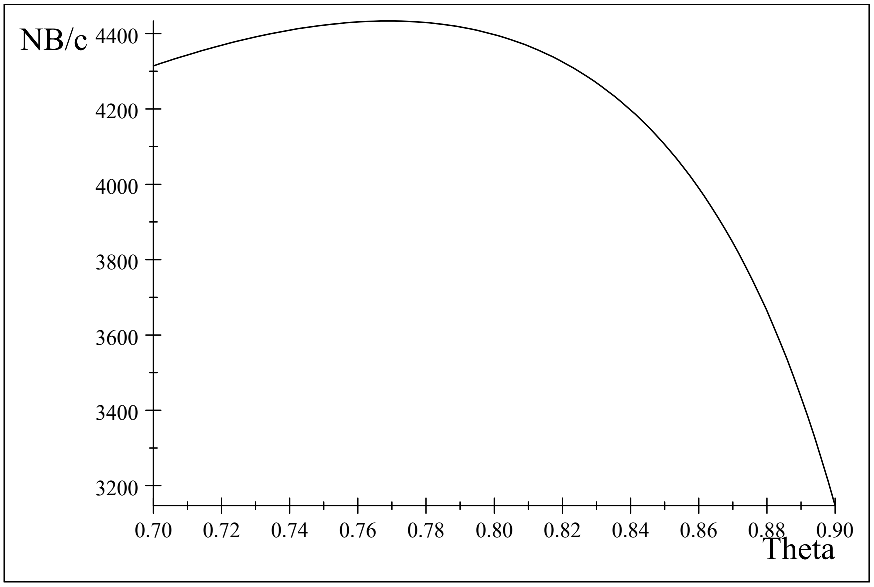
And what if you don't know calculus? Just plot $NB(\theta)$ as a function of θ and find the maximum. But wait – we don't know b and c, we only know that $r = b/c = 2,100$. No problem – just divide both sides of the net benefit equation by the (unknown) constant cost c to produce

$$\begin{aligned}\frac{NB(\theta)}{c} &= r\alpha\theta - \left(\frac{\mu}{\delta}\frac{\theta}{1-\theta} + \frac{\alpha}{\rho}\theta\right) \\ &= 2,100 \times 3.57 \times \theta - \left(\frac{4/15}{1/1500} \times \frac{\theta}{1-\theta} + \frac{3.57}{16} \times \theta\right).\end{aligned}$$

Plotting this against θ yields



Zooming in between $\theta = .7$ and $\theta = .9$ yields



which makes it pretty clear that the optimal value of θ equals 0.77 (to two decimals). Plugging into our formula from part (a) again yields the optimal number of agents $f = 1,340$.

8.2.10 An Adoption Agency

Prospective adoptive parents (PAPs) apply to an adoption agency at an average rate of 30 per month. Immediately upon application, newly arriving PAPs are placed on the agency's waiting list. Of course, many of these PAPs are looking at other adoption options, and some are continuing their attempts at achieving pregnancy on their own. As a consequence, we assume that PAPs drop off of the waiting list at a rate of 0.2 per month (or worded differently, on average PAPs are willing to wait 5 months before dropping out).

New babies available for adoption are discovered by (or referred to) the agency at an average rate of 10 babies per month.

Whenever a new baby arrives, the agency notifies the PAP on the waiting list who has been waiting the longest, and admits that PAP to the selection stage of the adoption process (one PAP at a time, so PAPs could be single adults or couples). The selection stage works as follows: all PAPs in the selection stage continuously review all available babies. Following a review of medical records, birth family history and the like, the probability that any PAP agrees to adopt a randomly selected baby is only equal to 1% (one reason for the low probability is the opportunity to interact with lots of babies, in addition to the specific desires/requirements of individual PAPs). Note that the number of PAPs in the selection stage is always equal to the number of babies being reviewed for adoption. Eventually, all PAPs who survive the waiting list and enter the selection stage will adopt a baby (and equivalently, all babies are eventually adopted by some PAP in the selection stage).

(a) Let Q_t denote the number of PAPs on the waiting list in month t. Given the information in the problem, complete the following equation:

$$Q_{t+1} = Q_t + ...$$

What we have is:

$$\begin{aligned} Q_{t+1} &= Q_t + \text{New PAPs - Dropouts - PAPs entering selection stage} \\ &= Q_t + 30 - 0.2Q_t - 10 \\ &= 0.8Q_t + 20. \end{aligned}$$

(b) Let B_t denote the number of babies under review for adoption in month t. Note that by the design of the program, B_t also denotes the number of PAPs in the selection stage of the adoption process. Given that *all* PAPs in the selection stage review *all* available babies, in month t what is the aggregate rate with which babies are adopted? (That some PAPs could select more than one baby, or that some babies could be selected by more than one PAP is not your problem so don't worry about it.)

There are B_t PAPs. Each PAP looks at each of the B_t babies, and adopts with probability 0.01. Thus the expected number of adoptions per PAP equals $0.01B_t$. Repeating over all PAPs, the aggregate rate of adoption is thus given by $B_t \times 0.01B_t = 0.01B_t^2$.

(c) Having answered part (b), complete the following equation for the number of babies available for adoption (equivalently, PAPs in the selection stage) over time:

$$B_{t+1} = B_t + ...$$

We have:

$$\begin{aligned} B_{t+1} &= B_t + \text{New Babies - Adopted Babies} \\ &= B_t + 10 - 0.01B_t^2. \end{aligned}$$

(d) Suppose that the program begins with 10 available babies and 10 PAPs in the selection process, and no other PAPs on the waiting list. Using your results from (a) and (c), produce a graph showing the number of PAPs on the waiting list and the number of babies available for adoption for the first 24 months (i.e. two years) of the program.

(e) Suppose that the program has reached a steady state, that is, the number of PAPs on the waiting list and the number of babies available for adoption no longer change with time. What is the steady state number of PAPs on the waiting list? What is the steady state number of babies available for adoption (which equals the steady state number of PAPs in the selection stage)?

Let Q denote the steady state length of the waiting list. Substituting into part (a) we get

$$Q = 0.8Q + 20$$

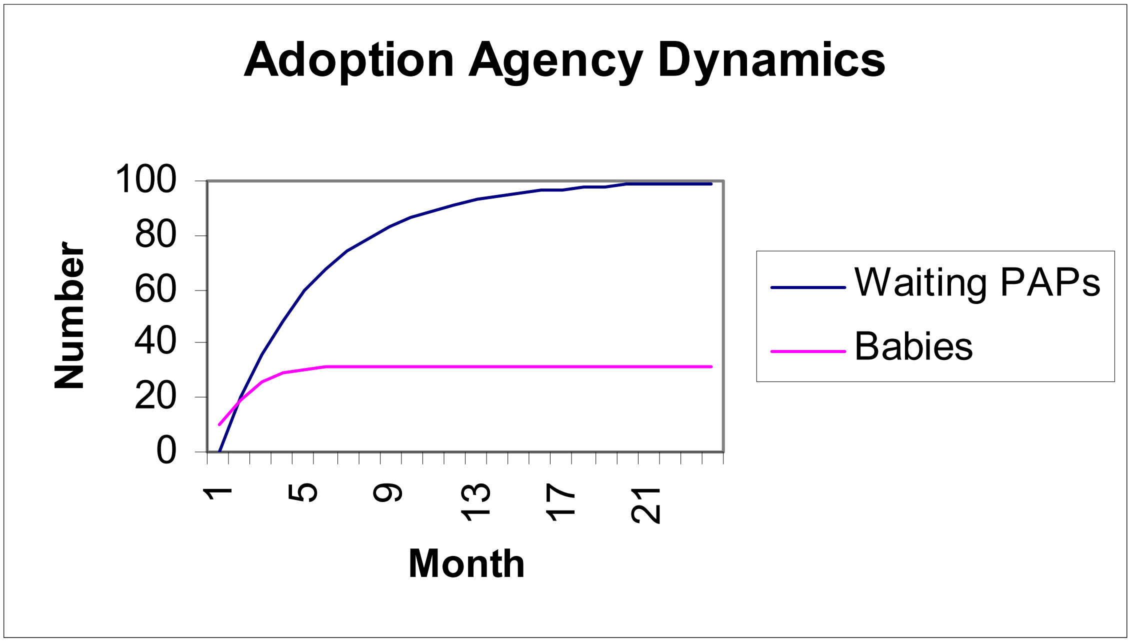

Adoption Agency Dynamics
Number
100
80
60
40
20
0
1
5
9
13
17
21
Month
Waiting PAPs
Babies

Figure 8.2:

which solves to yield $Q = 100$. Note from the figure in part (d) that this steady state is reached after 24 months. As for the number of babies up for adoption, denote this by B. From part (c) we get

$$B = B + 10 - 0.01B^2$$

which solves to yield $B = \sqrt{1000} = 31.623$.

(f) How long did those PAPs who survived the waiting list wait to enter the selection stage?

Note that the waiting list of PAPs evolves exactly like the waiting list in the public housing models we discussed in class. Thus, for those PAPs who survive the waiting list (similar to those applicants who receive public housing), the waiting time in equilibrium is given by

$$\begin{aligned} W &= \frac{1}{\delta}\log\frac{\lambda}{\mu} \\ &= \frac{1}{0.2}\log\frac{30}{10} \\ &= 5.49 \text{ months.} \end{aligned}$$

(g) On average, how long does it take for a newly arriving baby to get adopted?

Let μ equal the per-baby departure rate from the adoption agency. Whatever μ is, in equilibrium the average time babies spend in the selection stage is given by $1/\mu$. Now, how do we find μ? We know that the arrival rate of babies is equal to 10 babies per month. We also know (from part (d)) that the average number of babies up for adoption, B, is equal to $\sqrt{1000}$. Balance of flow (the rate with which babies come in equals the rate with which babies are adopted and leave) requires that

$$\begin{aligned} Inflow &= Outflow \\ 10 &= B\mu \\ 10 &= \sqrt{1000}\mu \\ \frac{1}{\mu} &= \frac{\sqrt{1000}}{10} = \sqrt{10} = 3.16 \text{ months.} \end{aligned}$$

8.3 Epidemic Models

8.3.1 Modeling Disease Progression

Let T_L, T_A and T_I denote the durations of the latent, asymptomatic infectious, and symptomatic infectious periods for some infectious disease. Assume that these three random variables are mutually independent, that a person infected at time 0 immediately enters the latent stage, then progresses sequentially through the latent, asymptomatic infectious, and symptomatic infectious stages, and recovers once all three stages of infection are complete. The probability distributions for the durations of each stage of infection are given by:

$$T_L = \begin{cases} 1 \text{ with probability } \frac{1}{4} \\ 2 \text{ with probability } \frac{1}{2} \\ 3 \text{ with probability } \frac{1}{4} \end{cases}$$

$$T_A = \begin{cases} 1 \text{ with probability } \frac{1}{3} \\ 2 \text{ with probability } \frac{2}{3} \end{cases}$$

and

$$T_I = \begin{cases} 3 \text{ with probability } \frac{1}{2} \\ 4 \text{ with probability } \frac{1}{3} \\ 5 \text{ with probability } \frac{1}{6} \end{cases} .$$

(a) What is the shortest possible course of infection (that is, time from infection through recovery)? What is the probability that a newly infected person experiences the shortest possible course of infection?

Well, the shortest course of infection occurs when each of the stages of infection takes on the shortest possible time. These are 1 day, 1 day, and 3 days respectively (for T_L, T_A and T_I), thus the shortest possible course of infection equals $1 + 1 + 3 = 5$ days. And, since the stage durations are independent by assumption, the probability that someone has a duration of infection equal to 5 days equals $\frac{1}{4} \times \frac{1}{3} \times \frac{1}{2} = \frac{1}{24}$.

(b) What is the longest possible course of infection? What is the probability that a newly infected person experiences the longest possible course of infection?

Same idea. The longest stage durations are 3, 2 and 5 days, thus the longest course of infection possible equals 3+2+5 = 10 days. The probability of this equals $\frac{1}{4} \times \frac{2}{3} \times \frac{1}{6} = \frac{1}{36}\cdot$ So the shortest course possible is more likely than the longest!

(c) What is the expected duration of the latent, asymptomatic infectious, and symptomatic infectious periods? What is the expected duration of the entire course of infection?

From the definition of the expected value of a random variable we have:

$$E(T_L) = \frac{1}{4} \times 1 + \frac{1}{2} \times 2 + \frac{1}{4} \times 3 = 2.$$

$$E(T_A) = \frac{1}{3} \times 1 + \frac{2}{3} \times 2 = \frac{5}{3}\cdot$$

$$E(T_I) = \frac{1}{2} \times 3 + \frac{1}{3} \times 4 + \frac{1}{6} \times 5 = \frac{11}{3}\cdot$$

And, since the expected value of a sum equals the sum of the expected values, the expected overall duration of infection equals $2 + \frac{5}{3} + \frac{11}{3} = \frac{22}{3}$ or $7\frac{1}{3}$ days.

(d) The *incubation time* is defined as the time from infection until symptoms develop, which in this case is the sum $T_L + T_A$. What is the probability distribution of the incubation time (that is, what is the probability that the incubation time equals 1 day, 2 days, 3 days, ... for all possible days)? What is the mean incubation time? Can you verify this from results you already obtained earlier?

We need to find the probability distribution of $T_L + T_A$. The table below reports all possible combinations of latent and asymptomatic infectious periods, and records their joint probability of occurrence along with the implied duration of the incubation time. The final distribution is just found by adding up the probabilities for all incubation times with the same duration. It's that simple! (In the table, "wp" means "with probability")

	$T_A = 1$ wp $\frac{1}{3}$	$T_A = 2$ wp $\frac{2}{3}$
$T_L = 1$ wp $\frac{1}{4}$	$T_{Inc} = 1 + 1 = 2$ $p = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$	$T_{Inc} = 1 + 2 = 3$ $p = \frac{1}{4} \times \frac{2}{3} = \frac{2}{12}$
$T_L = 2$ wp $\frac{1}{2}$	$T_{Inc} = 2 + 1 = 3$ $p = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} = \frac{2}{12}$	$T_{Inc} = 2 + 2 = 4$ $p = \frac{1}{2} \times \frac{2}{3} = \frac{2}{6} = \frac{4}{12}$
$T_L = 3$ wp $\frac{1}{4}$	$T_{Inc} = 3 + 1 = 4$ $p = \frac{1}{4} \times \frac{1}{3} = \frac{1}{12}$	$T_{Inc} = 3 + 2 = 5$ $p = \frac{1}{4} \times \frac{2}{3} = \frac{2}{12}$

Adding up over the table we find that the incubation time $T_{Inc} = T_L + T_A$ takes on the values 2, 3, 4 or 5 with probabilities $1/12$, $2/12 + 2/12 = 4/12$, $4/12 + 1/12 = 5/12$, and $2/12$. These probabilities sum to one, as indeed they must. The mean incubation time, evaluated from this distribution, equals

$$E(T_{Inc}) = 2 \times \frac{1}{12} + 3 \times \frac{4}{12} + 4 \times \frac{5}{12} + 5 \times \frac{2}{12} = \frac{11}{3}.$$

Note that this is equal to $E(T_L) + E(T_A) = 2 + \frac{5}{3}$ based on our earlier results from part (c).

(e) Similarly, find the probability distribution of the total time spent infectious (that is, the sum $T_A + T_I$), along with the mean duration of time spent infectious.

Using exactly the same approach, we find:

	$T_I = 3$ wp $\frac{1}{2}$	$T_I = 4$ wp $\frac{1}{3}$	$T_I = 5$ wp $\frac{1}{6}$
$T_A = 1$ wp $\frac{1}{3}$	$T_{Inf} = 1 + 3 = 4$ $p = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6} = \frac{3}{18}$	$T_{Inf} = 1 + 4 = 5$ $p = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9} = \frac{2}{18}$	$T_{Inf} = 1 + 5 = 6$ $p = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}$
$T_A = 2$ wp $\frac{2}{3}$	$T_{Inf} = 2 + 3 = 5$ $p = \frac{2}{3} \times \frac{1}{2} = \frac{2}{6} = \frac{6}{18}$	$T_{Inf} = 2 + 4 = 6$ $p = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9} = \frac{4}{18}$	$T_{Inf} = 2 + 5 = 7$ $p = \frac{2}{3} \times \frac{1}{6} = \frac{2}{18}$

Again the sum of the probabilities equals one as it must. Adding over the table, we find that the infectious period equals 4, 5, 6, or 7 days with probabilities 3/18, 8/18, 5/18, and 2/18. The mean infectious period equals

$$E(T_{Inf}) = 4 \times \frac{3}{18} + 5 \times \frac{8}{18} + 6 \times \frac{5}{18} + 7 \times \frac{2}{18} = \frac{16}{3}.$$

Note that this is equal to $E(T_A)+E(T_I) = \frac{5}{3} + \frac{11}{3}$ based on our earlier results from part (c).

(f) Considering the incubation time and the total time spent infectious, are these two quantities independent of each other, positively correlated (so a longer incubation period is associated with a longer total time spent infectious), or negatively correlated (so a longer incubation period is associated with a shorter total time spent infectious)? Answer by finding the covariance of the incubation time and total time spent infectious.

Clearly these quantities cannot be independent, since they both contain the asymptomatic infectious period T_A. So, the larger T_A is, the larger both the incubation and infectious periods are, while the smaller T_A is, the smaller both the incubation and infectious periods are. Thus, the incubation and infectious periods are positively correlated. Further, the asymptomatic infectious period T_A is the only source of correlation, since T_L and T_I are independent of each other (and of T_A).

To find the covariance between the incubation and infectious periods, first recall that the covariance for any two random variables X and Y is given by

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Applying this formula to $X = T_L + T_A$ and $Y = T_A + T_I$ we find that

$$\begin{aligned} Cov(T_L + T_A, T_A + T_I) &= E[(T_L + T_A)(T_A + T_I)] - E(T_L + T_A)E(T_A + T_I) \\ &= E[T_L T_A + T_L T_I + T_A T_I + T_A^2] \\ &\quad -[E(T_L)E(T_A) + E(T_L)E(T_I) + E(T_A)E(T_I) + E(T_A)^2] \\ &= E(T_A^2) - E(T_A)^2 = Var(T_A) \end{aligned}$$

because the crossproduct terms all cancel out due to independence – that is, $E(T_L T_A) = E(T_L)E(T_A)$, and similarly for the other crossproducts. Since the variance of T_A is positive, the covariance of the incubation time and the total time spent infectious is positive, and thus these two random durations are positively correlated as claimed above.

8.3.2 Minor Outbreaks Again

(a) A person infected with Nonovirus has just entered the community. The probability distribution for X, the number of infections transmitted by this initial case, is given by

$$\Pr\{X = j\} = \begin{cases} 1/4 & j = 0 \\ & \text{for} \\ 3/4 & j = 2 \end{cases}$$

and zero probability for any other value of j.

(i) What is the mean number of secondary infections $E(X)$?

Applying the definition of expected value, we have $E(X) = 0 \times 1/4 + 2 \times 3/4 = 6/4 = 1.5$.

(ii) What is the probability π that whatever outbreak ensues will be self-extinguishing (i.e. a minor outbreak)?

We need to solve the equation

$$\pi = \frac{1}{4} + \frac{3}{4}\pi^2$$

for π. Note that $\pi = 1$ is one solution, but it is not the solution we seek. Multiply both sides of the equation above by 4 and rearrange the terms to get the quadratic equation

$$3\pi^2 - 4\pi + 1 = 0.$$

The solution is given by the smallest root of the quadratic above, which we find from

$$\pi = \frac{4 \pm \sqrt{16 - 12}}{6} = \frac{2}{3} \pm \frac{1}{3} = (\frac{1}{3}, 1)$$

so we conclude that $\pi = 1/3$. There is a 1/3 chance that Nonovirus will self-extinguish without any intervention.

(b) Oh great, now a person infected with Ohnovirus has just entered the community. The probability distribution for X, the number of infections transmitted by this initial case, is given by

$$\Pr\{X = j\} = \begin{cases} 1/2 & j = 1 \\ & \text{for} \\ 1/2 & j = 2 \end{cases}$$

and zero probability for any other value of j.

(i) What is the mean number of secondary infections $E(X)$?

Applying the definition of expected value, we have $E(X) = 1 \times 1/2 + 2 \times 1/2 = 3/2 = 1.5$. Ah, so the expected number of Ohnovirus infections transmitted from the index case equals 1.5, just like Nonovirus.

(ii) What is the probability π that whatever outbreak ensues will be self-extinguishing (i.e. a minor outbreak)?

We need to find the smaller root of the equation

$$\pi = \frac{1}{2} \times \pi + \frac{1}{2}\pi^2.$$

You can reduce this equation to $\pi = \pi^2$ from which it is obvious that the two solutions are $\pi = 1$ and $\pi = 0$. We want the smaller root, which means that $\pi = 0$. There is *zero chance* of a minor Nonovirus outbreak!

(c) Rats! Whoknows virus just found its way into the community. Less is known about this virus except that $\Pr\{X = 0\} = 0$ and $\sum_{j=1}^{\infty} \Pr\{X = j\} = 1$.

(i) What is the probability π that whatever outbreak ensues will be self-extinguishing (i.e. a minor outbreak)?

We seek the smaller root of the equation

$$\pi = \sum_{j=1}^{\infty} \Pr\{X = j\}\pi^j$$

and by inspection, we see that $\pi = 1$ is a solution (since $\sum_{j=1}^{\infty} \Pr\{X = j\} = 1$), as is $\pi = 0$ (since $0^j = 0$ for $j = 1, 2, 3, ...$, and thus we get $0 = 0$). Bad news: Whoknows virus cannot lead to a minor outbreak, just like Nonovirus.

(ii) Using this result for Whoknows virus, explain the different public health threats posed by Nonovirus and Ohnovirus.

Well, we just learned from Whoknows virus that if $\Pr\{X = 0\} = 0$, there is zero chance of a minor outbreak. Ohnovirus is like Whoknows virus this way, since for both there is zero chance that an infected person fails to infect at least one other person. On the other hand, the fact that there is a 1/4 chance that a person with Nonovirus fails to infect anyone enables the possibility of self-extinction (the probability of which equals 1/3 as you found earlier). Nonovirus is thus less of a threat than Ohnovirus from the standpoint of triggering a major outbreak. Of course if Nonovirus is fatal while Ohnovirus is inconsequential, you'd rather have a major Ohnovirus outbreak for sure than a 2/3 chance of a Norovirus epidemic!

8.3.3 Estimating the Length of a Chain of Infection: *Vaccinia*

From class and your coursepack readings on HIV/AIDS, you know that for an infectious disease (such as HIV), the *reproductive rate of infection*, denoted by R_0, is the expected number of secondary infections transmitted directly by a single infected person early in the epidemic (when the fraction of the population infected is negligible). Suppose a single infectious person is introduced to a huge population (so we will take it as infinite in size) of otherwise uninfected (i.e. susceptible) persons. The infection in question can be characterized by R_0 as discussed above. Suppose that $R_0 < 1$ (so the expected number of infections generated by this new infected person is less than one).

(a) Let τ be the expected *total* number of infections generated over *all* time starting with this single infected person (that is, τ is 1 (for the initially infected person) plus the sum of the number of infections directly transmitted by the initially infected person, plus the number of infections generated by each of these secondary infections, plus the sum of all tertiary infections, etc. etc. etc.). Use the repetition method to determine a simple expression for τ in terms of R_0.

This is easy – a newly infected person will generate R_0 direct infections, and for each newly infected person, the future looks exactly the same as it did starting with the first infected person, and thus each new infection will him/herself generate a total of τ infections over all time, including him/herself! This implies that the expected total number of infections generated over all time from an initially infected person is given by

$$\tau = 1 + R_0\tau$$

from which we obtain

$$\tau = \frac{1}{1 - R_0}.$$

Note how much easier this is than taking the “Fresca” approach of summing the geometric series, that is

$$\begin{aligned}\tau &= 1 + R_0(1 + R_0(1 + R_0) + ... \\ &= 1 + R_0 + R_0^2 + R_0^3 + R_0^4 + ... \\ &= \frac{1}{1 - R_0} \text{ providing } R_0 < 1.\end{aligned}$$

(b) A concern raised during the Great Smallpox Debates of 2002 was that the *vaccinia* virus used in smallpox vaccination could itself be transmitted from a recently vaccinated person to others, with possibly problematic complications resulting. Indeed, an alarming presentation was made at an important national meeting suggesting that 20% of all smallpox vaccine complications stemmed from transmitted *vaccinia* infections. However, a still balding but formerly fat Yale professor suggested that these same data implied that 80% of all complications were *not* due to *vaccinia* transmission (and hence due to direct vaccination), implying that the ratio of complications due to *vaccinia* transmission to complications unrelated to *vaccinia* transmission equals 20%/80% or 0.25. Using the result from (a) above, what is the implied expected *total* number of vaccine complications over *all* time per direct vaccine complication (that is, the sum of the initial direct complication plus all future *transmitted* vaccine complications resulting from the original direct complication)? And, given a death rate of 1 per million from vaccination (due to direct complications), how should this death rate be adjusted to account for deaths that could result from complications due to transmitted *vaccinia*?

First, let’s just model vaccinia transmission, and then ask how one would see the consequences of vaccinia transmission in data reporting vaccine complications. Suppose that vaccinia transmits with $R_0 < 1$. Then from part (a) we know that the expected total number of vaccinia infections resulting from a person initially infected with vaccinia (i.e. a person vaccinated against smallpox) would equal $\tau = 1/(1 - R_0)$.

Now, suppose that anyone with vaccinia develops a vaccinia complication with probability p. Then the expected total number of vaccinia complications that would result from a single person initially infected with vaccinia would just equal τp (which is just the expected total number of vaccinia infections that lead to a complication). If n persons were directly vaccinated, then you would expect a total of $n\tau p$ vaccine complications, with np of these among those directly vaccinated. This means that the number of vaccinia complications attributable to transmitted vaccinia just equals $n\tau p - np = np\left(\frac{1}{1-R_0} - 1\right) = np\frac{R_0}{1-R_0}$. Consequently, the ratio of indirectly transmitted vaccinia complications to complications among those vaccinated directly would be given by

$$\frac{\#\text{ indirect complications}}{\#\text{ direct complications}} = \frac{np\frac{R_0}{1-R_0}}{np} = \frac{R_0}{1-R_0}.$$

Now, the data reported at the conference implied that 20% of all complications resulted from vaccinia transmission, which in turn implies that 80% of all complications resulted from direct vaccination. This means that the ratio of indirect to direct complications would solve

$$\frac{R_0}{1-R_0} = \frac{0.2}{0.8} = 0.25$$

which implies that $R_0 = 0.2$. Returning to part (a), the total number of vaccinia infections that would result from a single directly vaccinated person equals $\tau = 1/(1 - R_0) = 1/.8 = 1.25$. The total number of complications that would result then equals $1.25p$ *per person directly vaccinated.* Since the death rate among those directly vaccinated was previously estimated as 1 per million, the argument above shows that the effect of indirectly transmitted vaccinia on deaths amounts to increasing the death rate per person vaccinated from 1 per million to 1.25 per million (i.e. set $p = 1/million$).

While it is clear that the reported ratio of transmitted to direct vaccine complications equals 0.25, the timing over which this ratio was obtained is not clear. The argument above assumed that the indirect complications were observed over all time. Let's instead consider the other extreme, and assume that all of the data resulted from a single "generation" of transmission, that is, that n persons were vaccinated, that each of these n persons generated on average an additional R_0 vaccinia infections, that vaccinia infections led to complications with probability p, but that only complications among either those directly vaccinated or those who became infected with vaccinia from someone directly vaccinated were observed (because the observation period only amounted to one generation, namely, those initially vaccinated in the military along with a small number of first responders – this was before the vaccination of 500,000+ military personnel plus 40,000 or so civilians). This argument would imply that the total number of observed complications equals $np + nR_0p$, and that the observed ratio of transmitted to direct complications would equal $nR_0p/np = R_0$. In this case, you would model vaccinia transmission as an infection process with an R_0 of 0.25. From part (a), this results in a total over all time of

$$\tau = \frac{1}{1 - R_0} = \frac{1}{1 - 0.25} = \frac{4}{3} = 1.33$$

vaccinia infections *per person directly vaccinated* (and thus $1.33p$ complications).

To summarize, taking the death rate due to vaccinia as 1 per million infected with vaccinia ($p = 1/million$), then accounting for all vaccinia transmission beyond direct vaccination, we can safely say that the death rate should increase from 1 per million vaccinated to a number between 1.25 and 1.33 per million vaccinated. If you vaccinate a population of 10 million, instead of expecting 10 deaths from vaccine complications, now you should expect 12.5 to 13.3. Not a huge excess to worry about in the event of a smallpox attack (which would really be the only event in which such large-scale vaccination would take place)!

8.3.4 Repeat Screening for HIV Infection

Throughout the history of the HIV/AIDS epidemic, certain populations have been subjected to repeat HIV testing. These include, but are not limited to, commercial sex workers, prisoners, and persons in military service. The reasons for such testing are varied: commercial sex workers are typically screened because they are potentially at high risk for the transmission and acquisition of HIV, while persons in military service are considered part of a “walking blood bank.” In some sense, persons in the population at large who engage in high risk behavior and regularly have themselves tested also constitute a repeatedly screened population.

This problem considers one aspect of this issue: how often should persons be tested? Let us agree that undetected infections are costly, and identify a “cost of infection” c per person-year of undetected infection in the population.

Let us also agree that testing is costly, at a charge of k per test. For simplicity, we will assume that HIV testing is perfectly sensitive and specific: if the test says HIV+, then the person tested is truly infected, while an HIV- result means that the person is free of infection. Let the incidence rate of new infections (per uninfected person per unit time) remain constant and equal to r, and let the time in between successive HIV tests (the "screening interval") be denoted by τ.

(a) Recall that each person in this population is screened once every τ time units. Suppose an individual has just become infected. On average, how much time will pass from the moment of infection until the next screening test (when the infection is detected)? (HINT: the timing of infection is completely independent of the timing of HIV tests, the latter happening once every τ time units without fail!)

Since persons in the population are getting tested every τ time units, the time until the next screening test is clearly something between 0 and τ inclusive. Furthermore, since the timing of infections is completely random with respect to testing times, a newly infected person is equally likely to be anywhere within the interval bounded by the last and next testing times. This means that the average location of the time of infection within a screening interval is smack in the middle! Consequently, the expected time from infection until the next screening test equals $\tau/2$.

(b) Suppose that there are n persons in the population, and that the number of infected persons is sufficiently small relative to the population size that the aggregate rate of new infections in the population can be written as nr per unit time. What is the expected number of undetected infected persons in the population? (HINT: it obviously has something to do with the length of time that a newly infected person remains undetected!!)

Well let's see. Suppose that there are u persons who are infected but undetected in the population on average. Then the rate with which persons are detected must equal $u/(\tau/2) = 2u/\tau$, since on average each undetected infected person remains as such for $\tau/2$ time units. However, it must also be the case that the rate with which persons are detected equals nr since all infections are eventually detected, and new infections are occurring in the aggregate at rate nr. We thus have the balance of infection flow result:

$$nr = \frac{2u}{\tau}$$

from which we obtain

$$u = \frac{nr\tau}{2}\cdot$$

This result assumes that the persons stay in the population until they are detected with infection. Of course, in reality people are arriving and departing the population, but this model still works as long as the arrivals and departures cancel to leave the number of persons in the population constant and equal to n.

(c) Now, if it costs k per test, and there are n persons in the population each getting tested once every τ time units, what is the aggregate cost of testing per unit time? (HINT: recall the EOQ model.)

Clearly, if each person is tested once every τ time units, then the testing rate equals $1/\tau$ per person per unit time. This means that the total number of tests conducted in the population equals n/τ per unit time, and consequently the total aggregate cost of testing is given by kn/τ per unit time.

(d) In the State of Nevada, commercial sex work is legal and regulated in certain counties. The HIV incidence rate among these sex workers has been estimated as 4 new HIV infections per thousand sex workers per year (that is, $r = 0.004/yr$). The state uses standard enzyme immunoassay tests (EIAs) to detect HIV antibody; these cost \$5 per test. Finally, suppose that the public health department has, using a variety of arguments, decided that the cost of undetected infection c equals \$360,000 per year of undetected HIV infection in a commercial sex worker. Total public health costs can be thought of as the sum of the total cost of undetected HIV infection and the total cost of screening. The former cost is simply equal to c times the number of undetected infected persons in the population, while the latter cost you found in part (c) above. Given the data values provided for r, c and k, what is the optimal screening interval? That is, what numerical value of τ minimizes the total public health costs? Feel free to answer using Excel, or try to do this analytically. (HINT: recall the EOQ model.)

Let's see - the aggregate cost of infection is just equal to $cu = cnr\tau/2$ per unit time, while the aggregate cost of testing is given by kn/τ per unit time. Thus, total public health costs are given by:

$$\text{Total Public Health Costs } = \frac{cnr\tau}{2} + \frac{kn}{\tau}\cdot$$

We want to minimize this as a function of the screening interval τ. Note that n doesn't matter in this problem so we can just set $n = 1$.

We'll do this analytically first. To minimize the total public health costs, differentiate the expression above with respect to τ and set the result equal to zero. We get:

$$\frac{d}{d\tau}\left(\frac{cr\tau}{2}+\frac{k}{\tau}\right)=\frac{cr}{2}-\frac{k}{\tau^2}=0$$

which leads to the solution

$$\tau^*=\sqrt{\frac{2k}{cr}}.$$

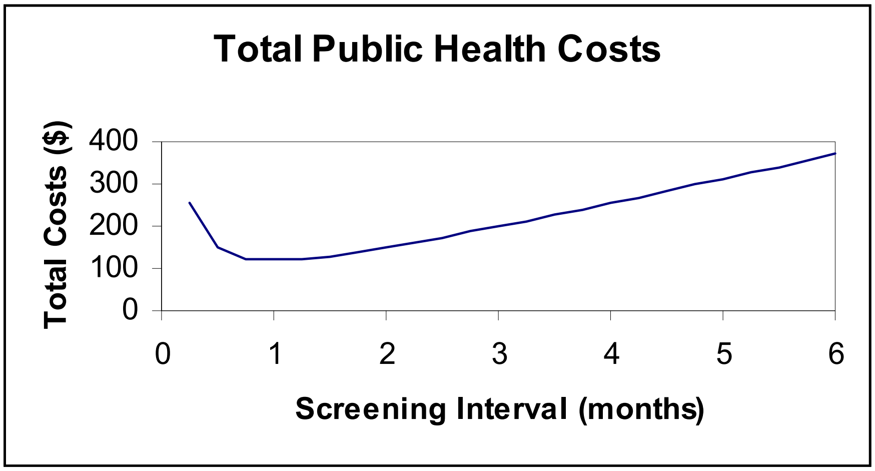
This makes sense: if the incidence rate r or the cost of infection c increase, the optimal screening interval τ^* decreases (that is, you screen more frequently). Conversely, if the test cost k increases, the optimal screening interval increases (that is, you screen less frequently). Note the analogy to the EOQ model: in the EOQ, the optimal order quantity is given by $q^*=\sqrt{2AK/H}$ where A is the item demand, K is the setup cost, and H is the holding cost. The time between orders (the cycle time) is given by $q^*/A=\sqrt{2K/HA}$. Comparing this to τ^* we see that the testing cost k plays the role of the setup cost in the EOQ, the cost of infection c is equivalent to the holding cost, and the infection rate r is equal to the demand rate A. Pretty cool, no? Repeat screening for HIV is an EOQ problem in disguise.

Continuing with the problem at hand, we plug in the parameter values ($r=0.004/yr, c=\$360,000,\ k=\$5$) and obtain

$$\tau^*=\sqrt{\frac{2\times 5}{360,000\times 0.004}}=0.083333\text{ years }=\frac{1}{12}\text{ years }=1\text{ month.}$$

Guess what? In Nevada, regulated commercial sex workers are screened once per month for HIV!

Now, suppose you don't know calculus and did not recognize the EOQ analogy. You could still have found this answer by plugging in the values of r, c, and k into the formula for Total Public Health Costs above, set $n=1$, and then calculated the resulting costs for different values of τ in a spreadsheet. Doing so, you would discover the following relationship between annual total costs and τ:



Note that Total Public Health Costs are minimized when the screening interval is set to 1 month, a result you could also have obtained via the Solver by minimizing Total Public Health Costs with τ as the decision variable using the GRG Nonlinear option.