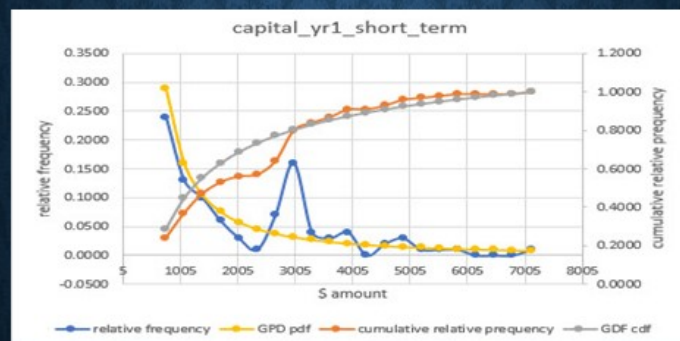


Texts in Decision Support and Management

# INFORMING DECISIONS

## Using Spreadsheets



William Foote



# Informing Decisions Using Spreadsheets

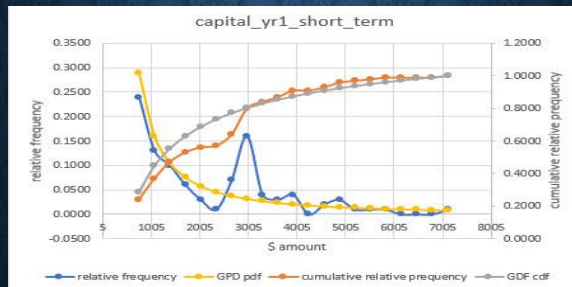
William G. Foote

2022

Texts in Decision Support and Management

# MAKING INFORMED DECISIONS

## Using Spreadsheets



William Foote



To the patience, kindness and wisdom of my family, friends, students, and  
colleagues.



---

# *Contents*

---

List of Figures	ix
List of Tables	xiii
Preface	xv
About the Author	xix
Part 1 – Spreadsheet Engineering	1
1 Tortuous Pie-making in the Sky	3
1.1 Spreadsheets? Really? . . . . .	3
1.2 Questions, questions . . . . .	4
1.3 Count the errors of our ways . . . . .	5
1.4 Prevailing recommended practices . . . . .	7
1.4.1 Do not ever do this . . . . .	7
1.4.2 Instead practice these . . . . .	7
1.5 Pie-in-the-Sky . . . . .	8
1.6 Where’s the paper and pencils? . . . . .	10
1.7 Cost and volume . . . . .	11
1.8 Demand analysis . . . . .	12
1.9 Weekly profit . . . . .	13
1.10 Profit sensitivity to price . . . . .	14
1.11 Lo and behold . . . . .	16
2 Chaotic Pie-making in the Sky”	17
2.1 How many? . . . . .	17
2.2 What’s new? . . . . .	18
2.2.1 The costs they are a-changing . . . . .	18
2.2.2 Demand takes a step back . . . . .	21
2.2.3 A new profit dawning . . . . .	22
2.3 An algebra of pie . . . . .	23
2.3.1 A little lite algebra . . . . .	23
2.3.2 PV’s last stand . . . . .	24
2.3.3 The model in the mist . . . . .	26
2.3.4 How sensitive? . . . . .	27

2.3.5	One way or the other . . . . .	30
2.4	Lo and behold yet again . . . . .	33
<b>3</b>	<b>Case: Salmerón Solar Systems LLC</b>	<b>35</b>
3.1	Photovoltaic systems . . . . .	35
3.2	The particulars . . . . .	36
3.3	Some definitions are in order . . . . .	36
3.4	Requirements . . . . .	36
<b>Part 2</b>	<b>– Optimization</b>	<b>39</b>
<b>4</b>	<b>Bringing Pie to Earth</b>	<b>41</b>
4.1	It all started here . . . . .	41
4.2	From humble beginnings . . . . .	43
4.2.1	1. Decisions, decisions . . . . .	45
4.2.2	2. Resources they are a'constraining . . . . .	46
4.2.3	3. Not without constraint . . . . .	51
4.2.4	Let us eat pie . . . . .	54
4.3	An easier way? . . . . .	55
<b>5</b>	<b>Expanding horizons</b>	<b>59</b>
5.1	Dido's bullhide . . . . .	59
5.2	Making dough . . . . .	62
5.2.1	Model me an EOQ . . . . .	64
5.2.2	Implementing the simple EOQ . . . . .	67
5.2.3	Life constrains our simple model . . . . .	68
5.3	Back to the future . . . . .	72
5.3.1	Bottom line up front . . . . .	72
5.3.2	Tighter, and tighter . . . . .	74
5.3.3	Implementing the model . . . . .	75
5.3.4	A revised rendering . . . . .	76
5.4	Next steps? . . . . .	78
<b>6</b>	<b>Case: Pricing Production at Make-A-Pie</b>	<b>79</b>
6.1	Yet another ask! . . . . .	79
6.2	A process unfolds . . . . .	79
<b>Part 3</b>	<b>– Simulation</b>	<b>81</b>
<b>7</b>	<b>Probability for Real People</b>	<b>83</b>
7.1	Can we rationally reason? . . . . .	83
7.1.1	Priors: what we think might happen . . . . .	84
7.1.2	Likelihoods: thinking about the data . . . . .	86
7.1.3	Altogether now . . . . .	87
7.1.4	Updating beliefs . . . . .	91
7.2	Enough with the shovels! . . . . .	92

<i>Contents</i>	vii
7.3 What did we accomplish? . . . . .	94
7.4 Try this out . . . . .	95
<b>8 Waiting on a Simulation</b>	<b>97</b>
8.1 A story of chance . . . . .	97
8.1.1 A short digression on volatility . . . . .	97
8.1.2 Back Solitaire . . . . .	98
8.2 A piece of pie, please . . . . .	99
8.3 A brief interlude with Cholesky . . . . .	101
8.4 Solution . . . . .	101
8.5 Now we can simulate . . . . .	102
8.5.1 Excel or bust! . . . . .	103
8.6 Back to the eatery robot . . . . .	104
8.6.1 The main course . . . . .	105
8.6.2 Exploring wait times . . . . .	106
8.7 Any next steps? . . . . .	108
<b>9 The Outer Limits</b>	<b>111</b>
9.1 Tales of tails . . . . .	111
9.2 It always starts with data . . . . .	112
9.3 A distribution to remember . . . . .	116
9.4 What does it all mean, so far? . . . . .	119
9.5 The joint's ajumpin' . . . . .	120
9.6 A sinking feeling? . . . . .	122
9.7 Simulate and stimulate . . . . .	124
9.8 Finally, exhaustively, where are the results? . . . . .	126
<b>10 Case: Forecasting Workers Compensation Claims</b>	<b>129</b>
10.1 Some background . . . . .	129
10.2 The ask . . . . .	129
10.3 Some requirements . . . . .	130
<b>Part 4 – Optimal Decision Simulation</b>	<b>133</b>
<b>11 Vegan Workers Insurance Fund: case and notes</b>	<b>135</b>
11.1 Tales of tails . . . . .	135
11.2 It always starts with data . . . . .	136
11.3 A distribution to remember . . . . .	140
11.4 What does it all mean, so far? . . . . .	143
11.5 The joint continues to jump . . . . .	145
11.6 A sinking feeling? . . . . .	146
11.7 Simulate and stimulate . . . . .	149
11.8 Finally, exhaustively, where are the results? . . . . .	150
11.9 Another scenario emerges . . . . .	151



---

## *List of Figures*

---

1	Sample 7 week course schedule. . . . .	xvi
1.1	Pie volume-cost relationship . . . . .	9
1.2	What influences profit? . . . . .	10
1.3	Pie cost driver plot . . . . .	11
1.4	Pie demand estimation . . . . .	13
1.5	Weekly profit model . . . . .	14
1.6	Pie price profit sensitivity: gateway to optimization. . . . .	15
1.7	Interactive profit dashboard . . . . .	16
2.1	Pie cost data - again. . . . .	19
2.2	Pie cost estimation with LINEST(). . . . .	19
2.3	Probability intervals illustrated. . . . .	20
2.4	Linear pie demand. . . . .	21
2.5	Weekly pie profit. . . . .	22
2.6	Pie profit sensitivity. . . . .	22
2.7	Drawing present value. . . . .	24
2.8	Just one calculation, many times over. . . . .	25
2.9	Yet another present value calculation. . . . .	25
2.10	Altogether now. . . . .	26
2.11	Net Present Value, in all its glory. . . . .	27
2.12	Another shot at pie profit sensitivity. . . . .	28
2.13	Profit-price plot settings. . . . .	28
2.14	The plot thickens. . . . .	29
2.15	NPV price sensitivity using Data Table . . . . .	29
2.16	NPV price x filling cost 2-way Data Table setup. . . . .	30
2.17	NPV price x filling cost 2-way Data Table results. . . . .	31
2.18	A faster way? VBA user-defined function (UDF). . . . .	31
2.19	Gaussian probabilities (aka Normal Distribution) . . . . .	32
2.20	The wonderful world of color. . . . .	32
3.1	The solar system - package version. . . . .	35
4.1	Origins of algebra. . . . .	42
4.2	Product margin calculations. . . . .	44
4.3	Product availability drivers. . . . .	45
4.4	Bounds on product decisions. . . . .	46

4.5	Iso-profit analysis. . . . .	48
4.6	Bounds meet iso-profits. . . . .	50
4.7	A solution? . . . . .	50
4.8	Product and process together at last. . . . .	52
4.9	The principle of parsimony at work. . . . .	53
4.10	Keeping it simple. . . . .	53
4.11	Process yielding profit solution. . . . .	54
4.12	Solving with Solver. . . . .	55
4.13	More from Solver. . . . .	56
4.14	Solver constraints. . . . .	57
4.15	Solver gets a feasible and optimal solutions! . . . . .	57
5.1	Dido's bullhide; getting the most out of a surface. . . . .	59
5.2	EOQ on the board. . . . .	64
5.3	EOQ neo-classical analysis. . . . .	66
5.4	EOQ a la Solver . . . . .	67
5.5	EQO Solver setup. . . . .	68
5.6	EOQ cost map. . . . .	68
5.7	Bounding the EOQ solution. . . . .	69
5.8	Marginal values of each constraint. . . . .	70
5.9	Multiperiod production template. . . . .	73
5.10	Where the template came from. . . . .	73
5.11	Some data is in order. . . . .	76
5.12	A network view. . . . .	76
5.13	Setting up Solver version. . . . .	77
5.14	A bang-bang solution presents itself. . . . .	77
7.1	Daily Pie demand and the weather. . . . .	93
7.2	What are the odds? . . . . .	93
8.1	Cholesky does a demo. . . . .	104
8.2	Simulating wait time parameters. . . . .	105
8.3	Simulating wait time setups. . . . .	106
8.4	Building EDA . . . . .	107
8.5	Finally, a picture. . . . .	109
9.1	Some workers compensation claims data. . . . .	112
9.2	Mean excess plot calculations. . . . .	115
9.3	MEP claims. . . . .	116
9.4	A grid to approximate claims distribution plausibility. . . . .	118
9.5	In the cross-hairs: variability and uncertainty. . . . .	120
9.6	Simulating Xi and sigma jointly. . . . .	121
9.7	Further setting up the xi and sigma simulation. . . . .	122
9.8	A model to design a monoline insurance fund. . . . .	123
9.9	On run of the monoline fund design. . . . .	124
9.10	A modified simulation automated flow. . . . .	125



9.11 EDA for the claims simulation. . . . .	126
11.1 The new claims data. . . . .	136
11.2 EDA for claims - again. . . . .	137
11.3 MEP for claims - again. . . . .	139
11.4 MEP for claims- the eplot this time - again. . . . .	140
11.5 Xi and sigma grid approximation. . . . .	142
11.6 Xi x sigma joint distribution and optimal values. . . . .	144
11.7 Begins the xi-sigma simulation. . . . .	145
11.8 sampling xi and sigma to generate a single claims forecast. . .	146
11.9 Claims forecast as input to the monoline design. . . . .	147
11.10A dry run of fund design. . . . .	148
11.11Automating the simulation work flow for fund design. . . . .	149
11.12Alas! The simulation of so many fund designs. . . . .	149
11.13Making some initial sense of the simulations. . . . .	151
11.14Setting up a more useful plot. . . . .	152
11.15Trying another fund design scenario for comparison and sensi- tivity. . . . .	152
11.16Short-run funding vehicle scenario sample run. . . . .	153
11.17Simulating the new scenario. . . . .	153
11.18An EDA comparison of the two scenarios. . . . .	154
11.19Even better: a plot emerges. . . . .	154



---

## *List of Tables*

---

1.1	Decision analytical process. . . . .	5
7.1	Priors by hypotheses . . . . .	86
7.2	data meets hypotheses . . . . .	86
7.3	Both data and hypotheses . . . . .	89
7.4	both data and hypotheses in days in January . . . . .	89
7.5	Unobserved belief tempered by observed data = posteriors. .	92



---

## *Preface*

---

---

### **Prolegomena**

The material in this book arises from several years of quantitative consulting for management and executive decision makers across several industries including government entities. Chapters and the flow of models derives from teaching MBA courses at Manhattan College which form the content for an eponymous focused one term, 3-credit graduate core course. As the title indicates the primary goal is to model decisions organizations and their decision makers face. The modeling approach is decidedly quantitative. The decisions include hiring and laying off, making and buying, marketing, financing, procuring, distributing, developing, and strategizing what and when to produce and serve to customers, employees, investors, vendors, and, yes, regulators.

---

### **Spreadsheets?**

The primary computer software platform is the Microsoft Excel 2019 spreadsheet environment with the Solver add-in. The course platform consists of companion workbooks. I do not recommend the use of Google Spreadsheets, yet. They often do not support some of the more basic structures of Excel, including the charting object (plots), the pivot table object, and macro recording. There is the smallest amount of programming in Visual Basic for Applications to automate the heavier computational loads of simulation. The goal, computationally, has been to use wherever possible what is native to the spreadsheet environment without use the of external calls to vast computational environments where algorithms alas are behind the proverbial curtain.

---

### **Using this book**

Here is a sample weekly schedule for a 7 week course.

Week	Topic
Week 1	Spreadsheet <a href="#">engineering</a> I: Building simple decomposition models
Week 2	Spreadsheet engineering II: Multi-variable sensitivity analysis
Week 3	Optimization I: archetypal linear programming decision models
Week 4	Optimization II: Multi-period planning and portfolio selection
Week 5	Simulation I: Uni- and bi-variate simulation of a customer model
Week 6	Simulation II: A multi-variate risk model
Week 7	Compendium: Final case preparation and presentation

**FIGURE 1:** Sample 7 week course schedule.

The book may be used to supplement much larger tomes devoted to the entire span of management science and computing techniques. The approach taken in this book is to titrate, curate, and focus students and practitioners on the essentials of using spreadsheets carefully in developing rapid prototype models of decisions in the field. The panoply of models that could have made the cut, but did not, underscores the need for a principled focus on essential model building and implementation in a spreadsheet.

---

### A bottom line, or so

With the essentials in hand, and in practice, analysts will have a solid basis to build prototypical models that can be expanded horizontally across other model configurations and vertically through various decision making requirements. As such, for example, network models are not discussed but are used to present the basics of a planning model. Also the range of Monte Carlo simulation is barely touched, but its utility is exploited with an uncommonly found model of a thick tail event distribution.

---

## Acknowledgments

Decades of courses, workshops, simply lots of colleagues and students on not a few continents in the private and public sectors, helped me with all of the material in this book. Thanks to all of you for your patience. If you have any comments, or would like to start a discussion, you may initiate a thread at [{}\(https://github.com/wgfoote/book-decision/issues\)](https://github.com/wgfoote/book-decision/issues). Again, thanks for your support.





---

## *About the Author*

---

My work meanders the seams between economics, finance, and underlying value chains, commodities, operations, and strategies.

I have had the privilege of being on the faculties of Pace, Clarkson, Syracuse Universities, Le Moyne and Manhattan College where I taught analytics, finance, and economics in evolving online and executive learning programs. My goal has always been to prepare students to take leadership positions through the rigorous application of data driven decision making.

During my career, I have helped build an energy and commodities practice in Europe for Ernst & Young, led the quantitative capital markets team at Deloitte, and helped to establish new practices in enterprise risk management at Deloitte, Charles River Associates, and Alvarez & Marsal, as well as in my own firm.

Thanks,

**Bill Foote**



---

## ***Part 1 – Spreadsheet Engineering***

---

Getting our arms around

- Design models with influence diagrams
- Deploy principles of good spreadsheet practice
- Use named ranges with INDIRECT, search with INDEX, MATCH, aggregate with SUMPRODUCT
- Basic, non-probabilistic, sensitivity analysis with Data Table
- Basic, non-probabilistic, sensitivity analysis with VBA
- One variable optimization by grid approximation with Data Table
- Plotting to communicate with a simple dashboard



# 1

---

## *Tortuous Pie-making in the Sky*

---

### 1.1 Spreadsheets? Really?

Yes, emphatically! George Gilder says we should waste transistors (that is chips).<sup>1</sup> Gilder makes the fairly obvious point that we must use transistors (lots of them in an integrated circuit) or go out of business. They are ubiquitous. And arrived everywhere in a very short amount of time to boot. If you do not use them you lose control of your cost structure. Anything you build will be too expensive, too heavy, too big, too slow, too lacking in quality.

The same idea goes with Michael Schragge builds on Gilder's ironic hyperbole about transistors and analogizes that we should "waste simulations."<sup>2</sup> If we do not so-called waste prototyping, rapid development, simulating potential problems, solutions, we will also go out of business. We must simulate until we drop! The alternative is that we will miss the one opportunity to improve or the one error that eliminates us. Of course the point he makes is that it is not a waste, rather we should never shy away from working the problem, simulating the art of the possible.

So what is the value added of a prototype, which is simply a working model? It is about information, and information is a surprise, a deviation from a trend. Schragge believes that testing a hypothesis just gets us to the point of saying we seem, in probability that is, to have a trend going on here. In the world of growth, opportunity, error and ignorance, having a trend is barely the beginning of our journey. It is the deviation from the trend that matters.

Are we still talking about spreadsheets? Schragge quotes Louis Pasteur: "Chance favors the prepared mind." Here the prepared mind is a product of simulations, the rapidly developed prototypes, Fleming used agar and discovered penicillin – completely unexpected! Dan Bricklin developed the spreadsheet

---

<sup>1</sup>See Gilder's comments here. He goes on to a further idea: build billions of 1-chip interconnected systems (our mobile phones that are really computers) and waste chips that way instead of manufacturing billion chip data centers. According to Moore's law we will eventually get to a near zero-cost chip.

<sup>2</sup>Here is a taste of Schragge's points of view. He compiled the "wasting prototyping" paradigm into this book a couple of decades ago.

IBMDOS/Apple IIe program Visicalc.<sup>3</sup> As a complete surprise this product was able to be used by millions of people to rapidly simulate other products and services. Steve Jobs credited Visicalc with the success of the Apple IIe and the Macintosh in 1985. IBM credited it with the success of the PC. Now people had a reason to buy the PC.

Using Visicalc we were able 40 years ago to build practical, plottable, usable option pricing models which transparently allowed us to visualize the calculations directly. Financial analysts built interactive pro forma balance sheet, income statements, and cash flow statements fed from managers' expectations, scenarios, and expert knowledge of markets. These models literally paid for themselves in days, not years. The main criterion for innovation success has always been the customer's payback, not the investors. How long did it take for the customer to recoup her investment? That's the innovation criterion. The spreadsheet is a sophisticated scratchpad some have used to be a production ready system.

But what is the most important message? A working prototype should be a sandbox where everyone is willing to get in and play. It has at least to be durable enough to get to the first slate of useful comments and suggestions for further improvement. Development continues! Rick Lamers recently open sourced his Grid Studio spreadsheet product with deep integration with Python.

Yes, let's play.



## 1.2 Questions, questions

Some questions come to mind as we begin.

- **What will we use the spreadsheet for?** We will rapidly prototype decision support models for business.
- **What's a business?** We will use a working definition: an organization of resources to reach a common goal. Yes, there is more to it than that, but it's good enough for now.
- **What's a decision, you might ask?** Again we will use a working definition for the moment: a commitment by people to deploy resources over time and space in support of a common goal. Decisions work inside of interconnected processes. They are the levers that allow, or refuse, inputs

---

<sup>3</sup>Here is a summary of his work. His innovation with Visicalc was to transform 20 hours of work into 15 minutes, almost of play at the time. Visicalc first ran on the Apple IIe. Dan is working on a web-based WikiCalc these days.

- to become outputs. All very abstract, but we are on course to build out our use of a spreadsheet as a prototyping tool.
- **What’s a model?** Any model is a human being’s abstraction from reality. We will not get into the cognitive, epistemological, ontological, or even methodological questions that might arise from this definition. Even saying a word like **sweet** is an abstraction of a perception, given experience with tasting anything, and tantalizingly informed by furtive imaginations. But one point must be made. **A model is not reality. Knowing is not just taking a look.** After all that’s why we build models. But again a decision is the result of an affirmation, a judgment. Gosh, it gets epistemological really fast. We need to visit the Philosophy Department next and soon. Back to our geocentric way of thinking.

TABLE 1.1: Decision analytical process.

activity	Inputs	Outputs	Decision
Pose	Requirements	Defined terms	Objective
Hypothesize	Terms	Conjectures	Choices
Sample	Data	Database	Drivers
Model	Relationships	Distributions	Choices   Drivers
Infer	Plausibile hypotheses	Most plausible hypotheses	Range of optimal choices

\end{document}

We will use this paradigm extensively, and creatively, usually from the outside in to ensure we achieve mission critical objectives given the contributions of the many drivers of those objectives.

### 1.3 Count the errors of our ways

- Spreadsheets are dangerous when in the wrong hands, with stubby fingers, bad memories, lack of structure, ramshackle governance, in short, poorly engineered. Here are some indicative **horror stories** from the European Spreadsheets Risk Interest Group.
- Covid: Test error ‘should never have happened’: 2020-10-05. The health secretary has said a technical glitch that saw nearly 16,000 Covid-19 cases go unreported in England “should never have happened”. The error meant that although those who tested positive were told about their results, their close contacts were not traced. By Monday afternoon, around half of those who tested positive had yet to be asked about their close contacts. The error was the use of an old Excel format. Counts exceeded the number

of rows in the spreadsheet. Labour said the missing results were “putting lives at risk”.

- AG: State overpaid Stroudsburg PA nearly \$500K: 29/07/2019. The district used cumulative mileage totals rather than running calculations on a sample average for vehicles, which resulted in the district significantly over reporting total mileage data, causing the subsidy overpayments. In some cases the spreadsheet was double counting total days for some of the activity runs. -Emailed spreadsheet contained private data in ‘hidden’ columns: 22/2/2017. A company employee mistakenly emailed a spreadsheet full of 36,000 coworkers’ personal details to his spouse in November, 2016, including Social Security numbers and dates of birth, all in hidden columns.
- SolarCity adviser Lazard made mistake in Tesla deal analysis: 2016-09-01. Lazard Ltd (LAZ.N), the investment bank that advised SolarCity Corp (SCTY.O) on its \$2.6 billion sale to Tesla Motors Inc (TSLA.O), made an error in its analysis that discounted the value of the U.S. solar energy company by \$400 million. This was the result of a computational error in certain SolarCity spreadsheets setting forth SolarCity’s financial information that Lazard used in its discounted cash flow valuation analyses.
- Report identifies lack of spreadsheet controls, pressure to approve, at JP Morgan: 18 January 2013-01-13. See pages 131-132 of the JP Morgan Task Force Report “...further errors were discovered in the Basel II.5 model, including, most significantly, an operational error in the calculation of the relative changes in hazard rates and correlation estimates. Specifically, after subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the modeler had intended. This error likely had the effect of muting volatility by a factor of two and of lowering the VaR.” As reported in [“A tempest in a spreadsheet”] (<http://ftalphaville.ft.com/2013/01/17/1342082/a-tempest-in-a-spreadsheet/>) Lisa Pollack comments that “On a number of occasions, he asked the trader to whom he reported for additional resources to support his work on the VaR model, but he did not receive any. Also it appears that he (had to?) cut a number of corners, which resulted increased operational risk and artificially low volatility numbers ... pressure was put on the reviewers to get on with approving the model.”

Could many of these errors have been performed in a programming language like Visual Basic for Applications (underlies Excel spreadsheets), or R, or Python? Sure and they have. What they share in common are violations of basic software engineering design principles and practices, let alone well known and implemented risk management and governance.

Of course Dilbert might calm us all down a bit. Then again, maybe it isn’t



the spreadsheet after all, but the strange idea about ant farms that someone might come up with in the first place.

---

## 1.4 Prevailing recommended practices

We might refrain from using the term *best practices* as that means, literally, there are no possibilities of improvement. At best, then, we might see some aspirational **leading practices**. here are some at least recommended practices for our consideration. They are based on this selection.

- Susan Allen's work at Lloyds Bank
- Payette's documenting spreadsheets,
- Various guidelines compiled by Raffensperger
- An all time favorite still Read and Batson's IBM spreadsheet modeling design practices from way back in 1999.

### 1.4.1 Do not ever do this

1. Hard code data in a formula
2. Take the word *spread* in spreadsheet literally
3. Put more than one major task or component on a worksheet
4. Guess the length or dimensions of any array, and everything in a spreadsheet is an array;
5. Use more than 3 IF-THEN-ELSE's nested in a single formula
6. Calculate parameters inside a chart or presentation table
7. Use Excel as the standard system of record database
8. Pretend you don't know what a named range is.

### 1.4.2 Instead practice these

1. Work flow. Paper and pencil the work flow for a model first. Document the scope, timing, user and system requirements, testing criteria first. Put data in one worksheet, task 1 in another, task 2, in another, and so on, plot set-up, table set-up in other separate worksheets, end-user presentation in another worksheet. Treat the spreadsheet model like a 3rd normal form data base. But for good-

ness sake avoid using Excel as a system of record data base if you can!

2. Definitions. Use named ranges to refer to any cell or array by name. Refer to cells and arrays where appropriate with INDIRECT(). This is what programming languages do. Each object has a unique name and scope of operations (e.g., integer, floating point, character data, class, slot) with descriptions in comments. Named ranges has some of this capability to document what the cells are and a bit of what they are meant to do. Named ranges are effectively row and column absolute addresses where cell data resides in worksheets. Attributes of cells include value, format, data type.
3. Testing. Be ready to test the model. Paper and pencil calculations should yield the same calculations for bits of the model, often called a *unit test*. But does the whole model stand up to scrutiny? A testing plan would stretch every assumption even to the breaking point of the model. Such *stress testing* and *system testing* are critical to the credibility both of the model and the model-builder in the eyes of the consumer of the model's results. The resiliency of models to changes in assumptions can be tested by installing form controls to sensitize results to those changes.
4. There are many more good practices, and even more bad ones. We should keep working, testing, improving, and communicating with one another about our various attempts at implementation. Perhaps a step in the right direction is the extensive use of FORMULATEXT() to display formulas in cells. Documentation can lead to immediate improvements.

---

## 1.5 Pie-in-the-Sky

Simone Tortiere makes pies: gluten-free, vegan, full of micro-nutrients. She targets a highly under-served niche market: nutritarians. She has a dilemma. She would like to expand her business. The problem is what price should she charge? Simple economics usually indicates the price that yields the highest profit.

Here is some data at hand Tortiere finds credible enough to use along with some questions to which she wants answers.

- **Value capture.** Make-A-Pie Co. generates profit by combining two purchased ingredients (fruit and dough) into pies, processing the pies (cooking,

packaging, & delivering), and selling them to local grocery stores. **What are the determinants of profit?**

- **Processing.** Make-A-Pie keeps track of weekly processing costs. The table below shows expense for various output levels. **What is the relationship between output volume and expense?**

Number of Pies (x1000)	Actual Cost (\$/1000 pies)
X	Y
8	12.8
10	17.5
12	21.6
14	25.9
16	32.8
18	39.6
20	46.0
22	56.1
24	63.6

**FIGURE 1.1:** Pie volume-cost relationship

- **Demand.** Make-A-Pie has also experimented with demand elasticity. At a price above \$12, there will be no demand for her pies, but demand increases by about 4,000 pies per week for each dollar price decrease below \$12. For example, at a price of \$10 Make-A-Pie could expect demand of 8,000 pies. **What is the average relationship between price and revenue?**
- **Price.** Savory vegetable filling costs \$3.48/pie, and sprouted flour dough costs \$0.30/pie. Overhead expenses are \$12,000 per week. **At what pie price will Make-A-Pie maximize profit? At what price will Make-A-Pie break even?**
- **Sensitivity.** Tortiere believes in her estimates, but she well knows that markets, property tax assessments, customer sentiment, and supplier costs can change. **How sensitive is weekly profit when any of the drivers of cost and revenue change?**

Tortiere muses further. She needs to understand if her business is profitable

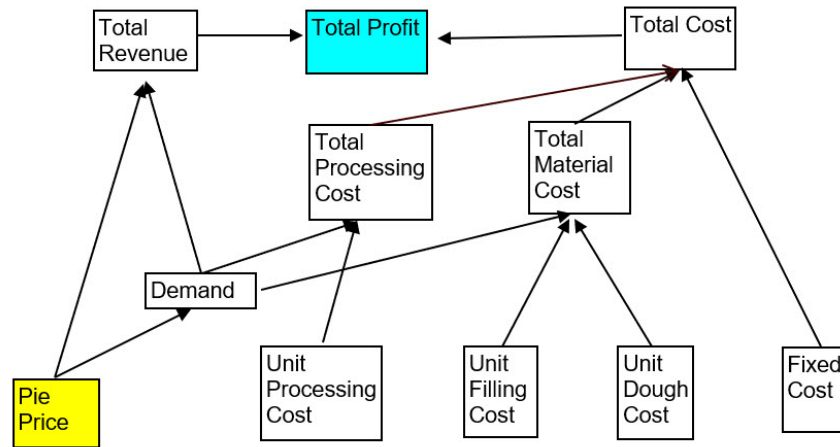
over the next few years and at what very sticky price should she charge her customers. She has already invested \$2.5 million.

- **Multi-period value.** Suppose that savory vegetable filling cost rises by 7% per year, and sprouted dough rises (pun included) by 5%. Overhead expenses rise by 2% per year. Prices will also rise by 2% per year. **At what year 1 pie price will Make-A-Pie maximize profit across a 3 year horizon if the business is worth \$2,500,000 today?**

Tortiere hires us to help her with her analysis. She asks us what is the first thing we should do?

## 1.6 Where's the paper and pencils?

First and foremost we map the requested analysis. We use a decompositional technique called an influence diagram, as developed by Howard and Matheson. The diagram extends the ideas behind a similar structure called a decision tree, both of which are examples of directed acyclic graphs and causal inference, e.g., Judea Pearl's work Software such as dagitty can greatly aid visualization of causal relationships among decisions (e.g., price), criteria (e.g., profit), and drivers (e.g., unit costs). After all of that consideration here we will simply use boxes and arrows.



**FIGURE 1.2:** What influences profit?

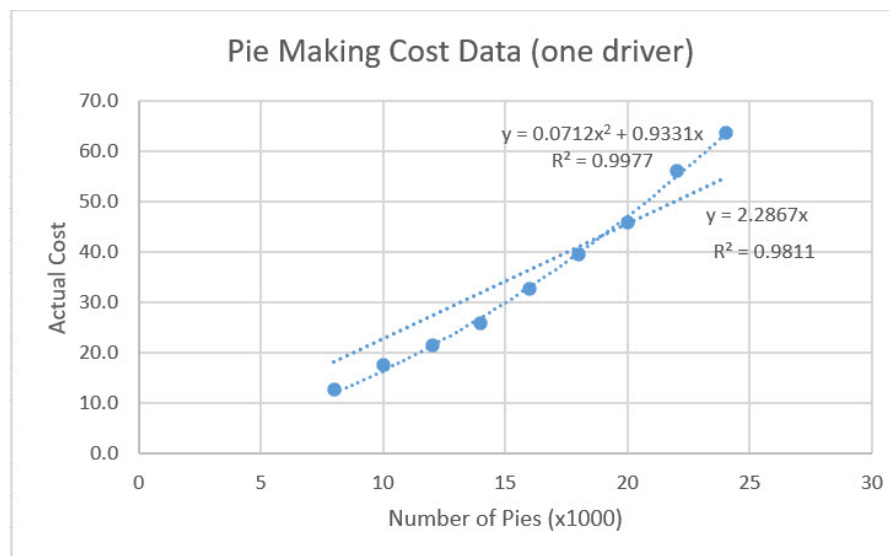
We identify the decision as price in the yellow box and the decision criterion profit in the blue box. Price, unit and fixed costs are somehow exogenous since

they do not depend on any variable, but they do determine other variables. Unit and fixed costs will be assumed and thus condition the rest of the variables. Price on the other hand is the one variable for which we solve. Literally we will pick a price and see what happens to profit.

Tortiere agrees with our approach and analysis. She sees that her efforts to understand how the demanded volume of pies sold depends on her pricing decision. But she wants to get at how volume influences, conditions, cost. We move to the next task, the cost structure.

## 1.7 Cost and volume

Make-a-Pie's experience with expense and volume provides some inside into the way volume, as determined by number of pies sold, will influence, condition, cost. Here is a scatter plot with linear and quadratic lines fit to the data.<sup>4</sup>



**FIGURE 1.3:** Pie cost driver plot

Right-clicking one of the blue observations will reveal a dialogue box with a Trendline feature. This produced the lines and the equations. We can check

<sup>4</sup>See Winston (2019), Chapter 54 Charting Tricks is a one-stop destination for a variety of charts, including the scatter plot here.

these results by hand, but let's use the LINEST array function instead in the cost worksheet.<sup>5</sup>

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q			
1	Pie Making Cost Data (one driver)																			
2							Ordinary Least Squares Regression													
3		Number of Pies (x1000)	Number of Pies (x1000) <sup>2</sup>	Actual Cost	Predicted Cost			$Y = b_1 \cdot X + b_2 \cdot X^2 + e$												
4		X	X <sup>2</sup>	Y	Y_hat			X <sup>2</sup>	X	constant			X	X <sup>2</sup>						
5		8	64	12.8	\$12.02			b	0.07117	0.933138	0		0.933138	0.07117						
6		10	100	17.5	\$16.45			se	0.004071	0.079489										
7		12	144	21.6	\$21.45			R <sup>2</sup> / se_e	0.999576	0.906556	#N/A		M4:	=INDEX(\$B\$4:\$J\$4, 1, MATCH(M3, \$B\$3:\$J\$3, 0))						
8		14	196	25.9	\$27.01			F / df	8246.01	7	#N/A		N4:	=INDEX(\$B\$4:\$J\$4, 1, MATCH(N3, \$B\$3:\$J\$3, 0))						
9		16	256	32.8	\$33.15			SSR / SSE	13553.88	5.752912	#N/A			Note: Reversed for order of regressors X, X <sup>2</sup> for use in SUMPRODUCT()						
10		18	324	39.6	\$39.86			H4,J8												
11		20	400	46.0	\$47.13			(=LINEST(D5:D13, B5:C13, FALSE, TRUE))												
12		22	484	56.1	\$54.98			LINEST() documentation												
13		24	576	63.6	\$63.39			<a href="http://www.mit.edu/~mbarker/formula1/f1help/04-g-m60.htm">http://www.mit.edu/~mbarker/formula1/f1help/04-g-m60.htm</a>												
14		C5:		E5:																
15		=B5^2		=SUMPRODUCT(\$M\$4:\$N\$4, B5:C5)																
16																				
17																				
18																				
19																				
20																				
21																				

The FALSE setting in LINEST sets the intercept to zero in the Ordinary Least Squares (OLS) estimation. Predicted costs will be used in the profit calculation using volume demanded as an input to the quadratic cost structure. We also use the INDEX-MATCH combination to wrangle coefficients for use in the SUMPRODUCT calculation.<sup>6</sup>

## 1.8 Demand analysis

Tortiere's experiments with pricing provide valuable windows into customer preferences. Even a simple change of price can reveal a conjecture about the range of preferences. In this situation we need to anchor the demand analysis around two important points:

- the give-away price and volume, and
- the no-show price where volume is just zero.

The give-away price is zero at least. The volume associated this lowest price

<sup>5</sup>Winston (2019) discusses several array functions and formulas in Chapter 91 and LINEST at page 578. LINEST is discussed here as well along with references to array functions. To enter an array function: 1. Select a results range. Some functions require the results range to contain a certain number of columns or rows. A an abridged results range will not show all of the function results; a results range that is longer than the function's capability will display error messages in some cells. Check the documentation on the specific function for information about the required size and shape of the results range. 2. In the results range you selected, type in the equals sign, the array function keyword, and the argument(s) or argument range(s), in parentheses. 3. To enter the function DO NOT PRESS ENTER! Instead press SHIFT and CONTROL together and while they are pressed down then press ENTER. (If you simply press ENTER, the array function will return the #VALUE! error.)

<sup>6</sup>See Winston (2019) Chapters 4 and 5 for INDEX-MATCH and Chapter 59 for multiple regression.

becomes the intercept of a down-ward sloping straight-line demand curve in price. The increment downward is estimated using a simple rise-over-run rate of change of volume with respect to price.

	A	B	C	D	E	F	G	H	I
1									
2									
3			intercept: a_0 means zero price sensitive demand					Demand	Price
4			Price_0	12	Demand decrease	4000		48000	0
5			Demand_0	4000				44000	1
6			a_0	48000	=D4*D5			40000	2
7								36000	3
8			slope: "rise over run" demand sensitivity to price					32000	4
9			a_1	-4000	=(H14-H13)/(I14-I13)			28000	5
10								24000	6
11			Demand = a_0 + a_1*Price					20000	7
12								16000	8
13			Demand = 48000 - 4000*Price					12000	9
14								8000	10
15			C13:					4000	11
16			="Demand = "&D6&" - "&ABS(D9)&"*Price"					0	12
17								-4000	13
18								-8000	14
19								H4:	
20								=D6	
21								H5:	
22								=H4-\$F\$4	
23									

**FIGURE 1.4:** Pie demand estimation

At a price above \$12, there will be no demand for her pies, but demand increases by about 4,000 pies per week for each dollar price decrease below \$12. For example, at a price of \$10 Make-A-Pie could expect a demand of 8,000 pies in a week. The intercept is then  $12 \times 4000 = 48000$ . The formula in D9 confirms the negative slope  $-4000$ . Thus we have a demand equation which serves as a schedule that will feed both revenue and processing cost.<sup>7</sup>

## 1.9 Weekly profit

Weekly profit is only a snapshot of performance. It is defined as total revenues minus total costs with various drivers defined mapped in the influence diagram. The calculations depend on assumptions provided by another worksheet, the finale of the model, the dashboard. Cost and demand volumes derive from estimated algebraic relationships we already peaked at.

Price comes from a sensitivity analysis we have yet to review and is an input from that analysis of the profit maximizing price. Three unit and fixed cost

<sup>7</sup>Winston (2019) Chapters 87 and 89 provide further approaches to demand analysis.

	A	B	C	D
1	<b>Make-A-Pie Co. -- Weekly Profit Model</b>			
2	Decision Variable			
3	Pie Price	\$9.38		=sensitivity!C3
4				
5	Model Parameters			
6	Filling cost/pie	\$ 3.48		=dashboard!L7
7	Dough costs/pie	\$ 0.30		=dashboard!L8
8	Fixed Cost	\$ 12,000.00		=dashboard!L9
9				
10	Operations			
11	Consumer: Pie Demand (X) = a_0 + a_1*Price	10,500.00 + 1000		=B12+B13*B3
12	Intercept: a_0	48000		=demand!D6
13	Slope: a_1	-4000		=demand!D9
14	Production: Processing cost = b_1*X + b_2*X^2	\$17.64		=B15*(B11/1000)^2+B16*(B11/1000)
15	Quadratic term: b_2	0.07117018	x (Pie Demand / 1000)^2 +	=cost!I4
16	Linear term: b_1	0.932137568	x (Pie Demand / 1000)	=cost!J4
17	Total Revenue	\$98,437.50		=B11*B3
18	Total Material Cost	\$ 39,690.00		=B11*(B6+B7)
19	Processing Cost	\$ 17,644.46		=B14*1000
20	Fixed Cost	\$ 12,000.00		=B8
21	Total Cost	\$69,334.46		=SUM(B18:B20)
22	Profit	\$29,103.04		=B17-B21
23				

FIGURE 1.5: Weekly profit model

assumptions also wander in from the dashboard. The user moves slide bars to choose those levels.

## 1.10 Profit sensitivity to price

Perhaps the most important activity in the building of a model, and the reason for a model in the first place, is to understand how model elements are sensitive to one another. Not emotionally of course as our models are robots and golems. BUT really there is no one so-called right answer! It is a range of answers dependent on the instructions we gave to our spreadsheet robots. Yes, the decision maker has to sign the contract, write a cheque, hire (or fire) a designated vendor, employee, partner. But before these decisions are cast in stone, we owe it to ourselves to waste simulations, again to recall George Gilder and Michael Schrage's ironic hyperbole.

In this graph we use data tables (Data > What if > Data Table) to take the column of possible prices and run them against the profit model. For each price (yes this is a for loop) the data table calculates a new profit.<sup>8</sup>

<sup>8</sup>Winston (2019) Chapter 17 deftly covers this vast topic. I do not like to use data tables too extensively and without further some thought. They are dynamic. Whenever you change a cell, the table recalculates. That's fine for as simple data table as we have here. The problem is when we have a 10000 by 100 table. It simply take a while to update. The easiest way to change the calculation mode is on Excel's Formula ribbon. In the Calculation grouping, on the right side of the ribbon, is a drop-down button for **Calculation Options**. We choose an option that works best for the current file. There are two other important buttons in the Calculation grouping. **Calculate Now** (equivalent to pressing function key



	A	B	C	D	E	F
1	<b>Make-A-Pie Co. -- Weekly Profit Model: price sensitivity</b>					
2						
3		Pie Price	9.00			
4						
5		Pie Price	Profit		max	12
6			25713.84		min	3
7		3.00	-176349.51		nodes	9
8		4.13	-110279.94		width	1.125
9		5.25	-57217.78			
10		6.38	-17163.00			
11		7.50	9884.39			
12		8.63	23924.38			
13		9.75	24956.98			
14		10.88	12982.18			
15		12.00	-12000.00			
16						
17		B7:	C6:		F6:	
18		=F6	=profit!B22		= ( F5 - F6 ) / ( F7-1)	
19		B8:				
20		=B7+\$F\$8				
21						

**FIGURE 1.6:** Pie price profit sensitivity: gateway to optimization.

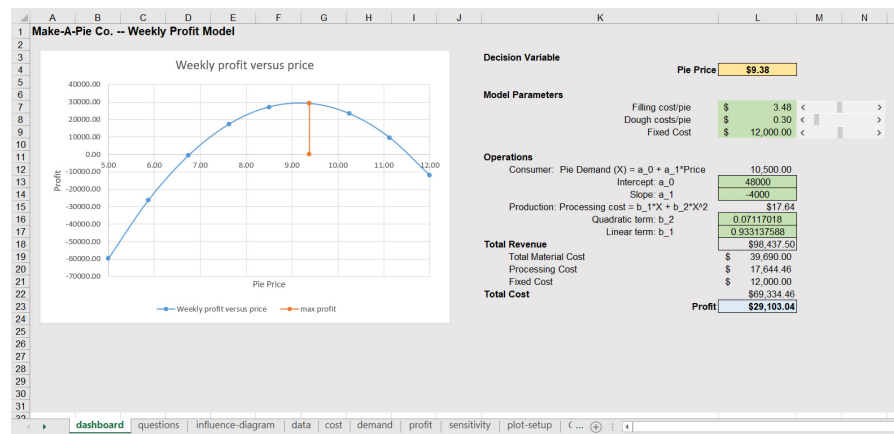
A plot of price and profit is more than a little instructive. Here is a setup worksheet with titles, additional calculations and lookups. Shown only is the sensitivity table setup. Any other plots in this workbook have a setup grid devoted to their charts.

	A	B	C	D	E	F
1						
2		title	Weekly profit versus price			
3		y-axis	Profit			
4		x-axis	Pie Price			
5		max profit	24956.98			
6		price	9.75			
7		abline	x	y		
8		a	9.75	0		
9		b	9.75	24956.98		
10						

F9) updates the entire workbook. **Calculate Worksheet** updates only the current sheet, which is much faster in a large file.

### 1.11 Lo and behold

This is what we have been waiting for. We must first notice the modular nature of this spreadsheet application. The dashboard worksheet interacts with all but the questions worksheet, at least in this iteration of the application. All Simone Tortiere needs to do, after paying our invoice, is to put the cursor on the slide bars that help her understand how the profit maximizing price changes with changes in the unit and fixed cost assumptions.<sup>9</sup>



**FIGURE 1.7:** Interactive profit dashboard

Is there too much in this dashboard? Perhaps the cost and demand equations could be left to their respective worksheets. We might replace those with the underlying assumptions of both structures. We leave that for another day and week.

What price is best? All we have to do is live with whatever assumptions we make and read the dashboard.

<sup>9</sup>Winston (2019) Chapter 27 has a discussion on the implementation of user form controls. These are located on the Developer ribbon at the Insert button. Here are directions to show this ribbon in your workbook.

## 2

---

### *Chaotic Pie-making in the Sky”*

---

#### 2.1 How many?

According to the Irish Technology Review and quoting Microsoft over 750 million users enjoy (?) Excel. I am enamored of points 3 and 4. I first ran into Excel with Martha Grabowski at LeMoyne College in 1989. I *enjoyed* Lotus 1-2-3, used Visicalc, and Supercalc since 1981. I was not yet convinced about Excel. I learned C by building a very stripped down version of Borland’s Quattro Pro with the idea of building option pricing models from finance directly into the DNA of the spreadsheet. We already knew that spreadsheets would inhabit if not infect the earth. My first balk was the use of the Excel = instead of the @ of Lotus formulae. Easy to overcome, but still the @ (shift-2) is in my fingers’ muscle memory.

I say all of this because we will see many changes in the expression of technology over the life cycle called our careers. Prepare Ye For The Changes To Come! I have experienced a lot of software products (not an expert at all and mostly a dilittante!) from very simple machine language to assembler to FORTRAN (with IMSL, LINPACK, EISPACK) to APL, APL2 (data shaping on steroids!), PLI and SAS, very early stage Matlab on DOS, S, S-Plus, LISP, PROLOG, BASIC, VBA (skipped Visual Basic interface with Oracle!), JavaScript (also skipped Java), C, C++, now R, Python, and, my almost favorite, Julia. Anything that works! Among all of these, spreadsheet products seem to have passed muster and kept going strong for the past 4 decades.

I must mention separately, Arthur Whitney’s kx, A+ and Shakti as the RISC (reduce instruction set computing) version, and maximizing time series vectors with relational databases and SQL (think vectoring FORTRAN). This is in the avant-garde of analytics. What a list of developments over the past several decades that lead us to the mixmaster and symbiosis of technology, technique, and skills we call analytics.

APL2 taught me data shaping. Whatever heuristics and algorithms within algorithms we devise in Excel are completely influenced by the shaping capabilities of array-driven technology like APL2, whether overtly or covertly. Unfortunately for a mathematically and numerically challenged population of analysts, APL2 is about to be deprecated. Yes a historically interesting, but

practical issue exists? And yes, again, it does. we will develop analytically-based and at least influenced career paths that are really modeling with data paths. Musa Jafar keeps saying that there is a 10-15 year cycle of software hegemony. I agree. The issue for us analysts is to develop capacities, not just skills in specific environments, which are agnostic to software platforms. In computing jargon, our minds must be interoperable across the platforms that enable us to inject data and interpretable analysis into our analytical products for consumption by decision makers.

Enough for this bully pulpit! The key take-away, so to say, is that we must begin with mind, go to data, and end up with mind. The mind has the thinking faculties that dictate what data is or is not important, credible, useful to the mind's view of the next horizon, whatever the technological platform.

---

## 2.2 What's new?

Simone Tortiere liked the model but furrowed her brow. Supply costs are sky-rocketing and customer demand is flagging. She is thinking about refinancing, perhaps even selling the business to a company with more capital resources. In fact she is in contact with a ;Special Purpose Acquisition Company (SPAC).](<https://www.investopedia.com/terms/s/spac.asp>) Top of mind for her is three year plan. But first she wants to traverse the demand and cost terrain she might face. At risk is her \$2.5 million investment. She asks us to help her to revisit the pricing decision in this context.

### 2.2.1 The costs they are a-changing

Make-A-Pie is experiencing much more volatile changes in its processing costs. These represent everything from capital (think ovens) to labor (think bakers) and supporting infrastructure (think fixed costs here though). Here is a revised schedule of processing cost against volumes in thousands of pies produced in an average (actually a median) week.

The previous schedule appears for comparison purposes. All of this calls for a new estimation of the way cost varies with pies demanded, and sold. So far storage is not an issue, yet.

There is a new kid in this block of numbers. Columns F and G calculate lower  $L$  and upper  $U$  95% probability interval bounds. Make-a-Pie's policy is not to tolerate any more than 5% errors in any given week. This rendering helps to explain the policy.

We might observe costs distributed symmetrically about a proposed future, or even actual, value. If the distribution is Gaussian, also known as normal

	A	B	C	D	E	F
1						
2		Number of Pies (x1000)	Actual Cost (\$/1000 pies)		old processin g cost data	
3		X	Y		Y	
4		8	13.8		12.8	
5		10	16.0		17.5	
6		12	23.6		21.6	
7		14	28.9		25.9	
8		16	29.8		32.8	
9		18	44.3		39.6	
10		20	41.1		46.0	
11		22	50.3		56.1	
12		24	57.5		63.6	
13						
14						

FIGURE 2.1: Pie cost data - again.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S			
1	Make-A-Pie Weekly Profit: cost																					
2											Ordinary Least Squares Regression											
3											$Y = b_1 X + b_2 X^2 + e$										standard error of residuals $se_e$	
4	Number of Pies (x1000)	Number of Pies (x1000) <sup>2</sup>	Actual Cost (\$/1000 pies)	Y-hat	predicted processing cost (USD / 1000)	lower bound	upper bound															
5	X	X <sup>2</sup>	Y	Y-hat	U	L		b	X <sup>2</sup>	X	constant		X	X <sup>2</sup>								
6	8	64	13.8	\$13.80	\$7.97	\$19.59		0.041176	1.395234				1.395234	0.041176								
7	10	100	16.0	\$16.07	\$10.17	\$21.79		se	0.013314	0.259951	#N/A		M4									
8	12	144	23.6	\$22.67	\$17.80	\$29.42		R <sup>2</sup> / se_e	0.994964	2.964673	#N/A				=INDEX(\$B\$4:\$L\$4, 1, MATCH(O3, \$K\$3:\$L\$3, 0))							
9	14	196	28.9	\$27.80	\$23.14	\$34.76		F / df	691.5177		#N/A											
10	16	256	29.8	\$32.86	\$24.02	\$35.65		SSR / SSE	12155.89	61.52501	#N/A				=INDEX(\$B\$4:\$L\$4, 1, MATCH(P3, \$K\$3:\$L\$3, 0))							
11	18	324	44.3	\$38.46	\$38.52	\$50.14		H4_J8							Note: Reversed for order of regressors X, X <sup>2</sup> for use in SUMPRODUCT()							
12	20	400	41.1	\$44.38	\$35.32	\$46.94		(LINEST(D5:D13, B5:C13, FALSE, TRUE))														
13	22	484	50.3	\$50.62	\$44.45	\$56.07		LINEST() documentation <a href="http://www.mit.edu/~mbarker/formula1/1help/04-g-m00.htm">http://www.mit.edu/~mbarker/formula1/1help/04-g-m00.htm</a>										1.395234				
14	C5	24	576	57.5	\$57.20	\$51.67	\$63.29															
15	=data!B4	=B9*2	=data!C4	=SUMPRODUCT(\$D\$4:\$P\$4, B5:C5)	=B20 + NORM.INV(tolerance / 2, 0, 1)*se_e	=B20 + NORM.INV(1 - tolerance / 2, 0, 1)*se_e																
16																						
17																						
18																						
19																						

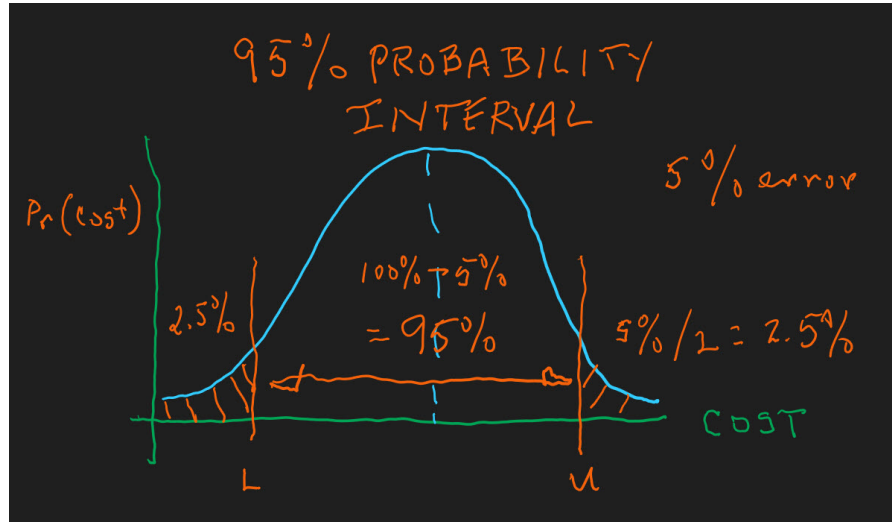
FIGURE 2.2: Pie cost estimation with LINEST().

(though there is nothing that normal about it!), then we can use Excel to calculate the lower and upper bounds of cost using the number of standard deviations  $z$  from the model mean of costs we estimated and the standard error of the model  $s_e$ .

$$L = \mu - |z_{0.025}|s_e \quad (2.1)$$

$$U = \mu + |z_{0.975}|s_e \quad (2.2)$$

For some mean  $\mu$  and standard deviation  $\sigma$  of observed outcomes  $X$  we can



**FIGURE 2.3:** Probability intervals illustrated.

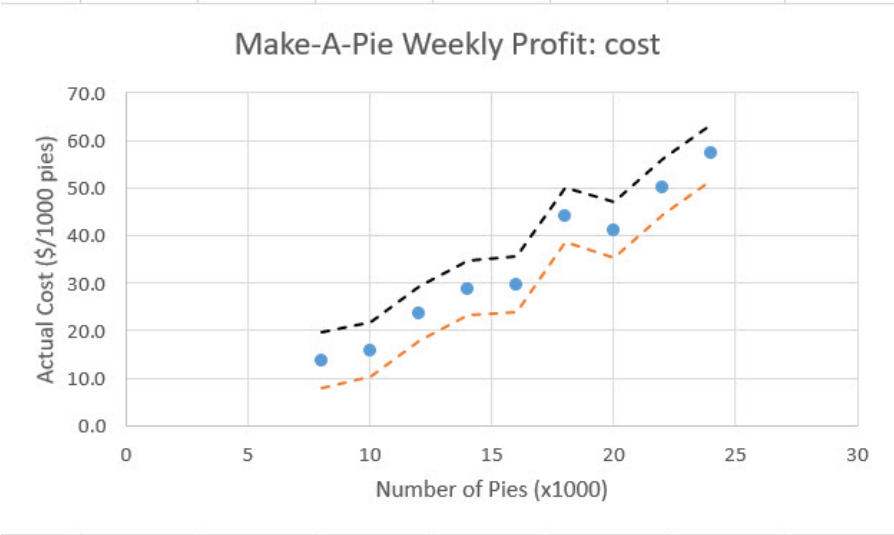
first first find the deviations of the outcome from the mean,  $X - \mu$ . Then we can calculate the number of deviations per standard deviation  $\sigma$ .

$$z = \frac{X - \mu}{\sigma}$$

We use Excel’s NORM.INV() to calculate the  $z$  values with a mean of 0 and standard deviation of 1 at lower bound 2,5% cumulative probability and upper bound 97,5% cumulative probability. This produces the  $97.5 - 2.5 = 95\%$  probability interval. All  $z$  distributions have this standard mean and standard deviation. That is very convenient.<sup>1</sup>

Of course some errors, we might call them, are intolerable, such as breaches of legal requirements and non-compliance with regulations such as food handling. The bounds are calculated around each actual, as opposed to predicted, processing costs. A picture is much in order.

<sup>1</sup>Winston (2019) Chapter 69 is a good and very basic introduction to random variables and Chapter 72 is a decent brief on the normal (Gaussian) distribution and  $z$  scores. We might skip ahead to Chapter 74 where he discusses the use of probability in making forecasting statements.



There is a wide enough variation to cause further identification and assessment of root causes and potential failure modes in the operation.

2.2.2 Demand takes a step back

Make-a\_Pie customers buy the pies from grocery stores. Tortiere plans to set up a program with restaurants and schools as well. But we will wait for that later. Suffice it to say, such a plan will garner increased regulatory scrutiny and requirements, all the more so-called overhead.

Recent store surveys reveal that there will be no demand for prices above \$10, but demand increases by about 4,500 pies per week for each dollar price decrease below \$10. For example, at a price of \$8 Make-A-Pie could expect demand of 9,000 pies. Here are the calculations.

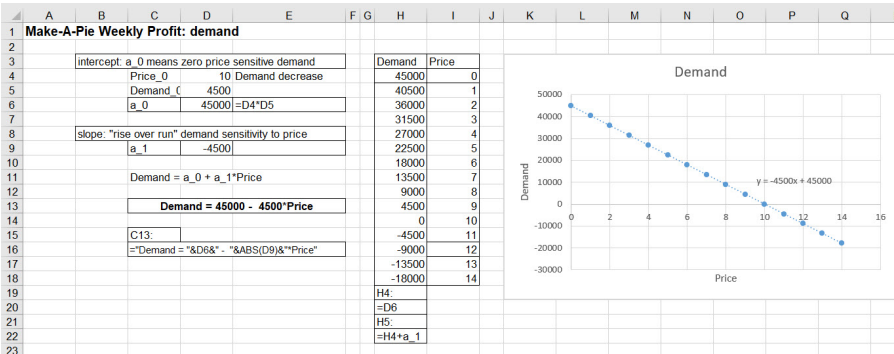


FIGURE 2.4: Linear pie demand.

The new conditions yield lower than previously thought volume intercept and steeper slope to meet the zero demand price of \$10. Consumers are definitely feeling the pinch in their household budgets. Grocery stores, responsible for handling, storage and insurance have increased their markups as well.

### 2.2.3 A new profit dawning

All of this yields a new weekly profit calculation and sensitivity analysis.

	A	B	C	D	E	F
1	<b>Make-A-Pie Co. -- Weekly Profit Model</b>					
2	<b>Decision Variable</b>					
3		Pie Price	\$7.63		=sensitivity!C3	
4						
5	<b>Model Parameters</b>					
6		Filling cost/pie	3.51		=settings!K3	
7		Dough costs/pie	0.26		=settings!K4	
8		Fixed Cost	13020.00		=settings!K5	
9						
10	<b>Operations</b>					
11		Consumer: Pie Demand (X) = a_0 + a_1*Price	10687.50 ÷ 1000		=C12+C13*C3	
12		Intercept: a_0	45000		=a_0	
13		Slope: a_1	-4500		=a_1	
14		Production: Processing cost = b_1*X + b_2*X^2	19.61		=C15*(C11/1000)^2+C16*(C11/1000)	
15		Quadratic term: b_2	0.041175804	x (Pie Demand / 1000)^2 +	=cost!K4	
16		Linear term: b_1	1.395234432	x (Pie Demand / 1000)	=cost!L4	
17		<b>Total Revenue</b>	81492.19		=C11*C3	
18		Total Material Cost	40291.88		=C11*(C6+C7)	
19		Processing Cost	19614.78		=C14*1000	
20		Fixed Cost	13020.00		=C8	
21		<b>Total Cost</b>	72926.65		=SUM(C18:C20)	
22		<b>Profit</b>	8565.53		=C17-C21	
23						
24						

**FIGURE 2.5:** Weekly pie profit.

Cells drive in from source worksheets. Notably the C3 price comes from the price sensitivity analysis worksheet.

	A	B	C	D	E	F	G	H
1								
2								
3		table reference: Pie Price	7.63					
4		price	profit		max	12		
5			8565.53		min	5		
6		5.00	-37583.03		nodes	9		
7		5.88	-14032.78		width	0.875		= (F4 - F5) / (F6-1)
8		6.75	1350.08					
9		7.63	8565.53		max profit	8565.53		=MAX(C6:C14)
10		8.50	7813.60		optimal price	7.625		=INDEX( B6:B14, MATCH( F9, C6:C14, 0 ) )
11		9.38	-1505.74					
12		10.25	-18792.47					
13		11.13	-44246.60					
14		12.00	-77868.13					
15								
16		B7:	C6:		F6:			
17		=F5	=profit!C22		=(F4 - F5) / (F6-1)			
18		B8:						
19		=B6+\$F\$7						

**FIGURE 2.6:** Pie profit sensitivity.

With these settings, Tortiere smiles at the weekly profit. But she knows full well this is not the whole story.



## 2.3 An algebra of pie

The rest of the story boils down to the relationships among variables of interest to Tortiere. She faces increased supplier costs and volatility of costs and consumer preferences. She believes her business is going under. Decisions, pricing, resources, timing, investment, must agily respond. She hunkers down with us, her analytical trusted advisors, to understand the breadth and depth of the situation.

### 2.3.1 A little lite algebra

First, she likes the influence diagram. Second, she wants to extend the ideas in the diagram across three years. She thinks this is a reasonable planning period. It can roll forward to future three year intervals. In her experience, three years is a cycle in her industry: feeding the hungry. The influence diagram generates this algebraic model of key relationships.

$$price_i = price_{set}(1 + price_g)^{i-1} \quad (2.3)$$

$$demand_i = (a_0 + a_1 price_i) * 52 \quad (2.4)$$

$$revenue_i = price_i demand_i \quad (2.5)$$

$$process_i = (b_1(demand_i/52000) + b_2(demand_i/52000)^2)(52)(1000) \quad (2.6)$$

$$doughcost_i = dough_0 demand_i (1 + dough_g)^i \quad (2.7)$$

$$fillingcost_i = filling_0 demand_i (1 + filling_g)^i \quad (2.8)$$

$$fixedcost_i = fixed_0 (1 + fixed_g)^i \quad (2.9)$$

$$cost_i = process_i + doughcost_i + fillingcost_i + fixedcost_i \quad (2.10)$$

$$profit_i = revenue_i - cost_i \quad (2.11)$$

$$pv_i = profit_i + pv_{i+1}/(1 + rate)^1 \quad (2.12)$$

The insight here is that the set price,  $price_{set}$ , gets the ball rolling. When chosen, price feeds demand, which, in turn, drops into the determination of processing cost, filling and dough costs as well. Total revenue is price times demand. Total cost is the sum of all component costs. Profit is revenue minus cost. All of these elements are tracked from year  $i = 1 \dots T$ . In our case Make-a-Pie has a  $T = 3$  year horizon.

Several variables grow across this horizon. If we set price in year  $i = 1$  to  $price_{set} = price_1 = 9.00$  and grow the price at  $price_g = 0.02$  per year, then by end of year  $i = 2$   $price_2$  is

$$price_2 = price_1(1 + price_g)^{2-1} \quad (2.13)$$

$$= price_1(1 + price_g)^1 \quad (2.14)$$

$$= 9.00(1 + 0.02)^1 \quad (2.15)$$

$$= (9.00)(1.02) \quad (2.16)$$

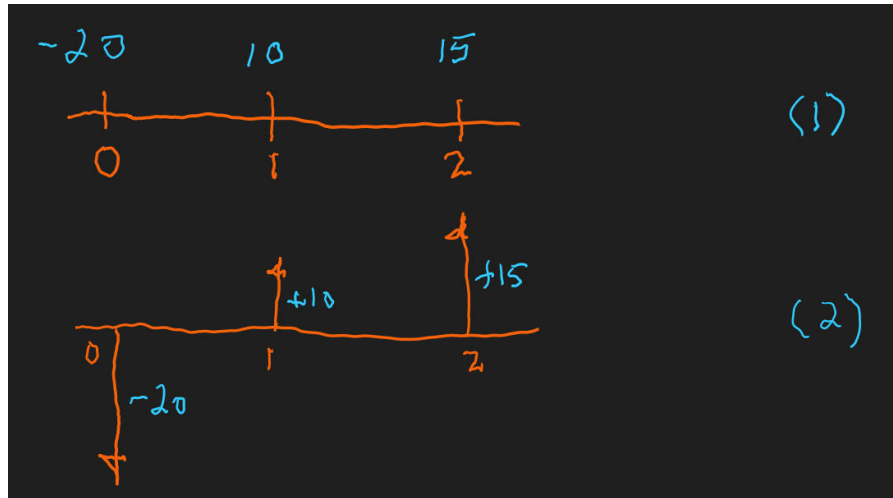
$$= 9.18 \quad (2.17)$$

In contrast we set  $filling_0 = 3.50$ , at  $i = 0$ , with growth rate  $filling_g = 0.07$  the initial 3.50 would grow to  $3.50(1 + 0.07)^1 = 3.57$  by the end of year  $i = 1$ . This amount will compound to  $3.50(1 + 0.07)(1 + 0.07) = 3.50(1 + 0.07)^2 = 4.0072$  by the end of year  $i = 2$ . And as they often say in Leipzig *und so weiter*, and so on with the other calculations subject to growth rates.

### 2.3.2 PV’s last stand

The last equation builds, even better, accumulates value from one yeat to another. Here is a simple example for a two period cash flow process. We can extend it to any number of forward periods. We can also think of periods as seconds, minutes, days, weeks, months, quarters, years, decades, centuries, and we get the idea! We conceive of getting or giving out cash flows (profits? losses?) right now, date 0, next period, as of date 1, and the following period, as of date 2.

This drawing depicts the process.



**FIGURE 2.7:** Drawing present value.

Panel (1) depicts three cash flows, -20, 10, and 15 occurring at the end of each periods with dates 0, 1, adn 2. Panel (2) uses directional arrows to animate the

conversation a bit. These directional arrows are frequently used in financial engineering applications and are part of the influence diagram tool-kit.

The next panel displays the calculation of the more or less obvious result that the present value of -20 paid out today is just that, -20.

$$\begin{array}{l} \text{PV}_0 \\ \hline -20 \end{array} \quad \begin{array}{l} \text{PV}(0, -20) = -20 / (1.10)^0 \\ = -20 / 1 \\ = -20 \end{array} \quad (3)$$

**FIGURE 2.8:** Just one calculation, many times over.

We notice the discounting mechanism at work here. We calculate  $-20/(1 + 0.10)^0$  and realize that anything raised to the zero power is just 1 and a thing times 1 is just the thing itself. Very philosophical.

This next panel shows a calculation of the present value of a cash flow in one period. In fact this is the kernel of all discrete cash flow present value calculations. It is all we need to know for any number of cash flows. Here it is.

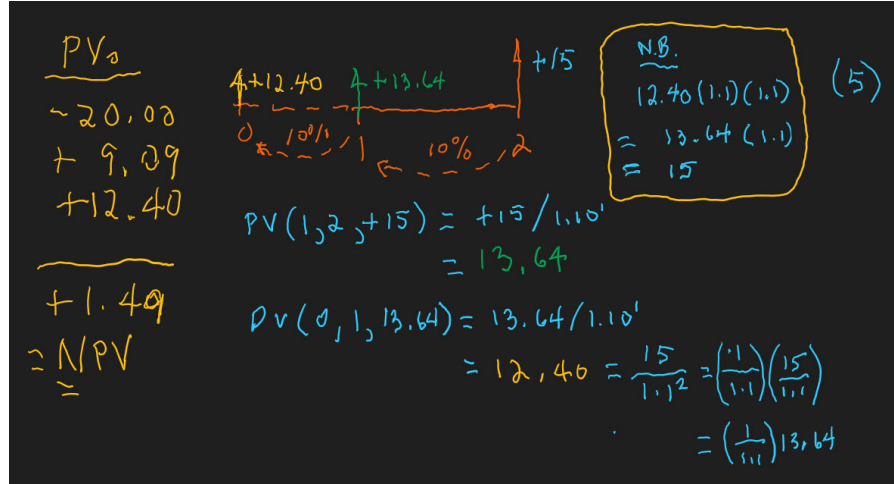
$$\begin{array}{l} + \text{PV}_0 \\ \hline + 9.09 \end{array} \quad \begin{array}{l} \uparrow 9.09 \quad \uparrow +10 \\ 0 \quad 10\% \quad 1 \end{array} \quad \begin{array}{l} \text{N.B.} \\ 9.09 + 0.1(9.09) \\ = 9.09(1 + 0.1) \\ = 9.09(1.10) \\ = 10 \end{array} \quad (4)$$

$$\begin{array}{l} \text{PV}(0, 10) = 10 / (1.10)^1 = \frac{F_1}{(1+r)} \\ = 10 / 1.1 \\ = 9.09 \end{array}$$

**FIGURE 2.9:** Yet another present value calculation.

We discount 10 by 10% as  $10/(1 + 0.10)^1$  to get 9.09. The *N.B* (that is, note well) comment works the process in the other direction. Post 9.09 bail at date 0 and receive 10% return of 0.91 and the return of principal of 9.09 to get 10.00 at the end of the next period at date 1. Discounting and growth are inverse processes.

The next, and thank goodness, last panel shows that all we need is the one period present value. The tally so far includes present values of -20 and 9.09. One more plank to walk down and we are done.



**FIGURE 2.10:** Altogether now.

A cashflow occurring at the end of two periods at date 2 is first discounted at 10% to date 1 for a date 1 value of 13.64. Then, and if all goes well (and the creek doesn't rise!), 13.64 discounts to 12.40, again at 10%. The PV tallies to 1.49. This is, finally, and forever more in our minds, **net present value** (NPV). NPV is just the present value of all cashflows, including today's cash flow.

Now we can add one more equation to the mix.

$$pv_i = cf_i + \frac{pv_{i+1}}{(1 + r_{i,1+1})^1}$$

Any date  $i$  present value equals the cash flow occurring in date  $i$  plus the discounted present value at the next date  $i + 1$  at a rate of return that is expected to occur from date  $i$  to date  $i + 1$ . This is powerful medicine for multi-period planning models. We will come back to this when we attempt to optimize a multi-period pricing structure for Simone Tortiere in the not so distant future.

### 2.3.3 The model in the mist

We will let this picture of the multi-period model begin to speak for itself (so to speak).

Two things we should notice. First, the model simply is a the profit model

Make-A-Pie Co. -- 3 year net present value model										
1	Year	0	1	2	3	Growth rate	=C2 + 1			
2	Decision Variable									
3	Pie Price	\$8.00	\$8.16	\$8.32	2%	=sensitivity!C10	=settings!S3			
4										
5	Model Parameters									
6	Fruit filling cost/pie	3.50	\$ 3.75	\$ 4.01	\$ 4.29	7%	=C7*(1+\$H7)	=settings!S4		
7	Dough cost/pie	0.26	\$ 0.27	\$ 0.29	\$ 0.30	5%	=C9*(1+\$H9)	=settings!S5		
8	Fixed Cost	677040	\$ 690,580.80	\$ 704,392.42	\$ 718,480.26	2%	=C9*(1+\$H9)	=settings!S6		
9										
10	Operations									
11	Consumer: Pie Demand (X) = a_0 + a_1*Price	458,000.00	430,560.00	392,371.20			=(D13+D14*D4)*\$2			
12	Intercept: a_0	45000	45000	45000			=a_0			
13	Slope: a_1	-4500	-4500	-4500			=a_1			
14	Production: Processing cost = b_1*X + b_2*X^2	15.89	14.38	12.87			=D16*(D12/52000)^2+D17*(D12/52000)			
15	Quadratic term: b_2	0.04	0.04	0.04			=cost!\$B4			
16	Linear term: b_1	1.40	1.40	1.40			=cost!\$L4			
17	Total Revenue	3744000.00	3513369.60	3265783.97			=D12*D4			
18	Total Material Cost	1890424.00	1848738.53	1800447.44			=(D7+D8)*D12			
19	Processing Cost	826402.20	747525.39	669358.03			=D15*D2000			
20	Fixed Cost	690580.80	704392.42	718480.26			=D9			
21	Total Cost	3397407.00	3300656.34	3189285.73			=SUM(D19:D21)			
22	Profit	346593.00	212713.26	77498.24			=D18-D22			
23										
24	Appraised Value	-2500000.00	346593.00	212713.26	77498.24		D25 =D23			
25	Discounted Value	-2500000.00	315084.54	175796.08	58225.58	10%	D26 =C25*(1+\$H26)^C2			
26	Net Present Value	-1950893.80					=SUM(D26:F26)+C25			
27							=NPV_pe(npv!\$C\$25, npv!\$H\$26, D4, npv!\$H\$4, npv!\$D\$13, npv!\$C\$14, npv!\$D\$17, npv!\$D\$16, \$C\$8, npv!\$H\$8, npv!\$C\$7, npv!\$H\$7, npv!\$C\$9, npv!\$H\$9, npv!\$F\$2)			
28										
29										
30										

FIGURE 2.11: Net Present Value, in all its glory.

with growth in filling, dough, and fixed costs, and let's not forget, a first year price that grows (or we can let decline). The algebraic representation maps directly into the formulae on the right. Also the price for the profit model can be set manually or comes from the sensitivity analysis worksheet.

Second, the calculation of discounted profits (really cash flow, we will have to consult with the accountants here), happens in two steps. We first calculate the discounted profit for each year. Then, second, we sum up all of the present values of cash flows (oops, profit) to arrive at the multi-period profit (cash-flow) criterion we called net present value in cell C27. We do not use Excel's NPV() function for expositional reasons (showing off the full calculation) and to eliminate any confusion about the calculation.<sup>2</sup>

### 2.3.4 How sensitive?

The model computes profit sensitivity again. Here is the setup to refresh our memories of days past.

The Data > What if > Data Table feature comes to our aid again. We use INDEX-MATCH to locate the profit maximizing price. We plot the results for human consumers of the analysis. Here is the setup for the plot.

<sup>2</sup>Winston (2019) Chapter 8 has many interesting details about the NPV() (and also the IRR()) functions in Excel. But if we calculate the =NPV(H26, C26:F26) we get the wrong answer, about -1,840,000. This is because Excel considers the first cash flow to occur in the first period and thus discounts the first flow by the discount rate, here 10%. Many have done this, and some have paid the price of operational error. So why not call the NPV function the PV function? Someone at Microsoft did this a long time ago but for cash flows of even amounts, sometimes called an annuity.

	A	B	C	D	E	F	G	H
1								
2								
3		table reference: Pie Price	7.63					
4		price	profit		max	12		
5			8565.53		min	5		
6		5.00	-37583.03		nodes	9		
7		5.88	-14032.78		width	0.875		$=(F4 - F5) / (F6 - 1)$
8		6.75	1350.08					
9		7.63	8565.53		max profit	8565.53		$=MAX(C6:C14)$
10		8.50	7613.80		optimal price	7.625		$=INDEX(B6:B14, MATCH(F9, C6:C14, 0))$
11		9.38	-1505.74					
12		10.25	-18792.47					
13		11.13	-44246.80					
14		12.00	-77868.13					
15								
16		B7:	C6:		F6:			
17		=F5	=profit(C22)		= (F4 - F5) / (F6 - 1)			
18		B8:						
19		=B6+\$F\$7						
20								

FIGURE 2.12: Another shot at pie profit sensitivity.

	A	B	C	D	E	F
1						
2		title	Weekly profit versus price			
3		y-axis	profit			
4		x-axis	table reference: Pie Price			
5		max profit	8672.41			
6		price	7.63			
7		max profit	x	y		
8		a	7.63	0		
9		b	7.63	8672.41		
10						

FIGURE 2.13: Profit-price plot settings.

We locate the plot in the dashboard worksheet. This is familiar to us from the last time we met the model with only weekly profits.

Done.

Well, almost done. What about NPV? We perform the same tasks as we did for profit maximizing price. We use a data table to create the for loop of replacing prices in the npv worksheet with prices from the data table, and for each price calculate a new NPV. Here is the table.

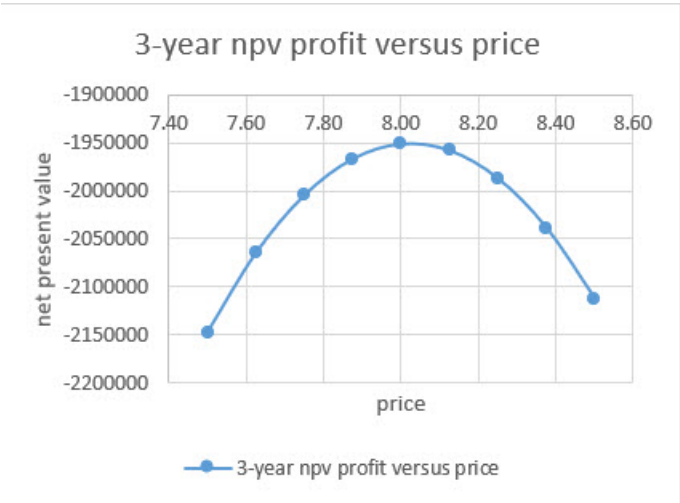
Of course we want to plot it. Plotting produces one more bit of information than the table: it makes very explicitly the non-linear shape of profits, and multiperiod profits here labeled net present value.



FIGURE 2.14: The plot thickens.

	J	K	L	M	N	O	P	Q
1								
2								
3		table reference: Pie Price	8.00					
4		price	net present value		max	8.5		
5			-1963131		min	7.5		
6		7.50	-2163246		nodes	9		
7		7.63	-2079646		width	0.125		$= (O4 - O5) / (O6 - 1)$
8		7.75	-2018427					
9		7.88	-1979589		max profit	-1963131		$= MAX(L6:L14)$
10		8.00	-1963131		optimal price	8		$= INDEX(K6:K14, MATCH(O9, L6:L14, 0))$
11		8.13	-1969055					
12		8.25	-1997359					
13		8.38	-2048044					
14		8.50	-2121110					
15								
16		B27:	C26:					
17		=O5	=npv-sensitivity!C27					
18		B28:						
19		=K6 + \$O\$7						

FIGURE 2.15: NPV price sensitivity using Data Table



Oh, and one more thing, this plot dramatically shows that Make-a-Pie is under water, a quaint trope that more than suggests a uniformly, across the range of prices and for three years, a loss in the value of the enterprise. Simone Tortiere definitely furrows her brow over this rendering. What to do next?

### 2.3.5 One way or the other

We now venture into the world of a 2-way sensitivity analysis.<sup>3</sup> We brace ourselves and strap in. First, of course, we set up a grid. This time two are two lists, here for price, and also for a variable of interest, filling cost per unit. Second, we make a grid and supply the reference for the npv calculation (from a separate npv-sensitivity worksheet to avoid possible calculation conflicts, oh my!). The references for the price and filling cost per unit on in the same worksheet as the data table. The npv sensitivity(2) worksheet does the calculations for the 2-way sensitivity table.

We call up Data > Table > What if, the row-column dialogue box appears. We choose price for the row and filling cost per unit for the column.

**FIGURE 2.16:** NPV price x filling cost 2-way Data Table setup.

We press OK and see this table.

This took 5 seconds to run. That’s too slow for a production model as we are often a bit impatient for results when under the track referee’s gun to start the race! One way around this is to create our own model in Visual Basic for Applications. Here is a picture of such a model in the Developer ribbon.

The user-defined function (UDF) is NPV\_pie. It calculates the same value as the npv worksheet and is a good quality control check on what turns out to be a fairly complex model. The VBA function employs a FOR-NEXT loop.

<sup>3</sup>Winston (2019) Chapter 17 shows us how to construct one-way and two-way data tables in good detail. We must always be sure to have data table reference cells in the same worksheet as the data table. In Chapter 93 he performs a many-way data table. This is a technique that easily lends itself to a spider-plot. Microsoft Office help has this to say about 2-way data tables.



	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
23																	
24					row	column											
25				reference	9.00	3.5								max	8.5		
26														min	7.5		
27														nodes	9		
28														width	0.125		
29																	
30																	
31																	
32																	
33																	
34																	
35																	
36																	
37																	
38																	
39																	
40																	
41																	
42																	
43																	
44																	

FIGURE 2.17: NPV price x filling cost 2-way Data Table results.

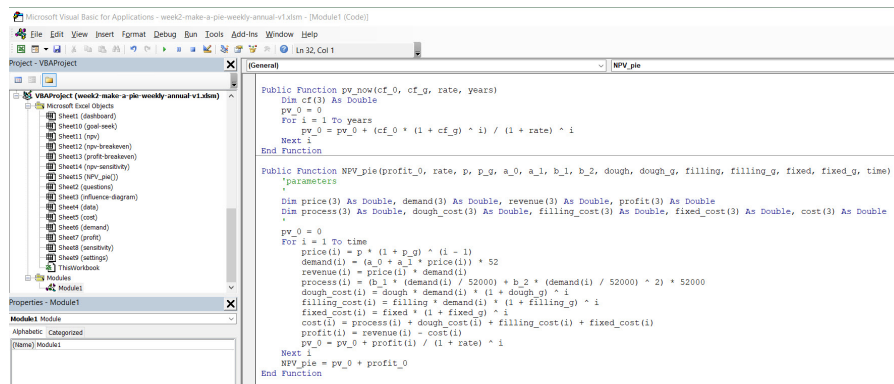


FIGURE 2.18: A faster way? VBA user-defined function (UDF).

When we build grids like price and filling cost per unit we are implicitly using a for loop structure. We have an initial value like the minimum price. We then in the next cell increment the previous cell's value and then move to the next cell until there are no more cells to go to.<sup>4</sup>

Here is an illustration how we can use a function, or for that matter any formula, in a two way table (simply two inputs to the formula or the function). Suppose that we have the normal distribution describe cost behavior. We propose several values of the mean  $\mu$  and standard deviation  $\sigma$ . We want to relate a mean with a standard deviation as if a scenario, just like a price and filling cost per unit combination. The formula for the Gaussian (normal) distribution is this beast.

$$Pr(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

<sup>4</sup>The logical thinking behind the spreadsheet seems more like the do-done logic in macro-assembler languages that underly Excel, VBA, C++, FORTRAN, APL2. Here is an IBM reference for examples. Cell by cell calculations also seem analogous to the addresses in processor memory registers. Here is an example for x86 Intel processors.

where  $d$  is some sort of data we are considering, say,  $d = 9$ . We ask this question: **How plausible is it to observe a price that is \$9.00 per pie when the mean and standard deviation are different values, and, this is crucial, we believe that price observations in general follow a Gaussian (normal) distribution?**

A data table comes to mind, of course! We set up  $\mu$  and  $\sigma$  grids, reference this beast of a formula, copy and paste into cells in Excel’s memory caches, for example, like this rendering.

	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
42																
43																
44																
45																
46																
47																
48																
49																
50																
51																
52																
53																
54																
55																
56																
57																
58																
59																
60																
61																
62																
63																
64																
65																
66																
67																
68																

**FIGURE 2.19:** Gaussian probabilities (aka Noram Distribution)

Instead of the Gaussian beast, we could have used the Excel NORM.DIST() function (with the last parameter set to FALSE for the probability mass calculation for which we clamor.) And instead of that we use our own, bespoke, NPV\_pie() function.

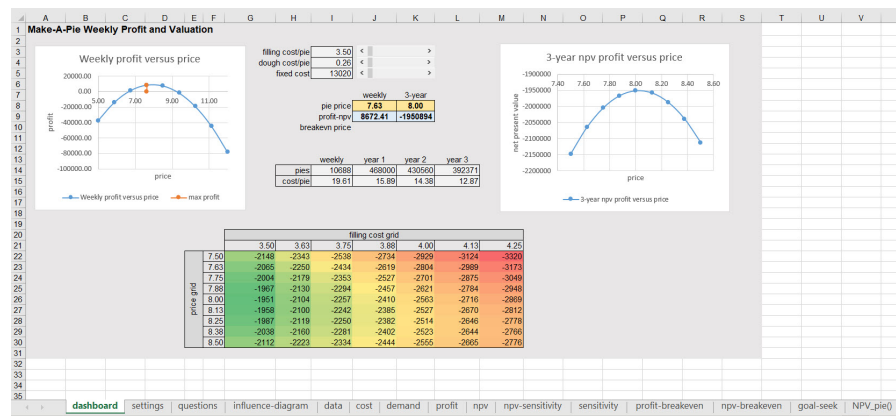
	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	

**FIGURE 2.20:** The wonderful world of color.

And oh yes we selected the table results and at the Home > Conditional Formatting button selected Color Scales > Green-Yellow-Red. We have the notorious heat map. Use of heat maps for performance and heat map discussions are notorious because some believe them to be misleading and not good practice, as in this article by Osama Salah. He references the risk management international standard ISO 31000 for support. We can, and should, have some debate over this topic.

## 2.4 Lo and behold yet again

This is what we have been waiting for, agina. We modified the modular nature of this spreadsheet application extensively. The dashboard worksheet interacts with all but the questions worksheet, at least in this iteration of the application. All Simone Tortiere needs to do, after paying our invoice, is to put the cursor on the slide bars that help her understand how the profit maximizing price changes with changes in the unit and fixed cost assumptions.<sup>5</sup>



Is there too much in this dashboard, a question we should ask Simone Tortiere? Are the colors of the heat map misleading? Only if they do not lead to a rational discussion of the lack of value in the organization, based on all of the assumptions of this analysis.

What price is best? All we have to do is live with whatever assumptions we make and read the dashboard. Using this and our beliefs about the credibility and plausibility of our results we ponder further. If enough political and emotional capital are present for a change, then change we must.

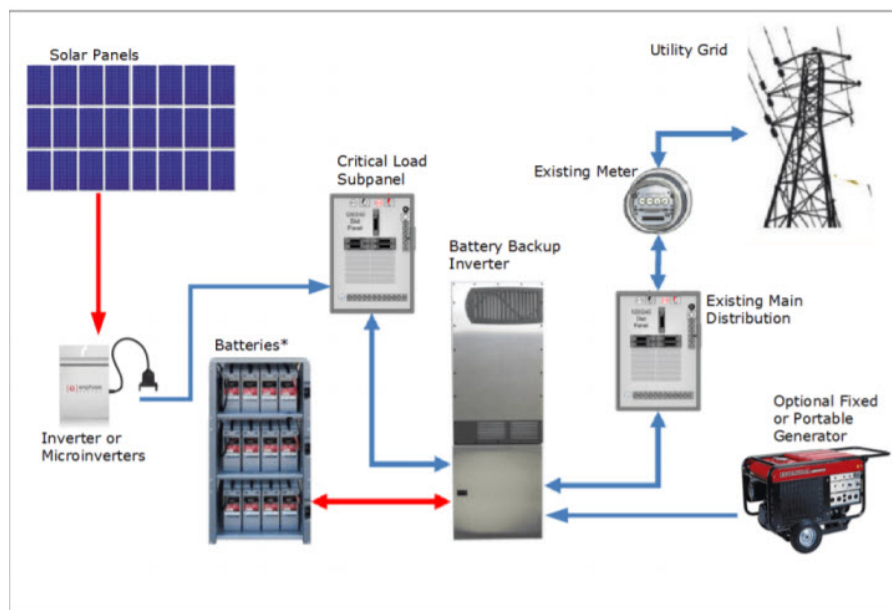
A break-even price is not even possible in this environment. If we went back to high school and solved the equation quadratic in prices that is the net present value equation we would get complex number solutions. Why? Because the NPV iceberg is entirely under water, it is negative. We would have to do something much more radical like change the structures behind the assumptions to float NPV back up to more solar energy.

<sup>5</sup>Winston (2019) Chapter 27 has a discussion on the implementation of user form controls. These are located on the Developer ribbon at the Insert button. Here are directions to show this ribbon in your workbook.



# 3

## *Case: Salmerón Solar Systems LLC*



**FIGURE 3.1:** The solar system - package version.

### 3.1 Photovoltaic systems

With 50 packaged small scale photovoltaic systems already leased to industrial operators, Toni Salmerón would like to offer her US based solar leasing company for sale at a price of \$1,000,000 before the beginning of next year's operations. Tomi and her partners have built up the business over the past five years using their savings and good old-fashioned sweat equity. They hold key patents on battery, control, and manufacturing technology that they intend to retain. Toni wants us to develop a three-year economic analysis to evaluate buyers' tenders for the company.

### 3.2 The particulars

The company's property taxes are \$35,000 per year and are expected to grow at an annual rate of 4%. Toni currently spends \$4,800 per year per system to maintain and administer the fleet. Administrative and maintenance costs are expected to increase by 7% per year.

Currently, Toni leases her systems for \$1,000 per month each. She leases 60% of the systems each month. Toni believes demand for her systems is highly elastic. Her market surveys and pricing experiments indicate that the percentage of the fleet leased each month increases by 7% for each \$100 per system per month reduction in the lease price. For example, at \$600 per month, she expects that 88% of her systems would be rented. She also believes lease prices can be increased 9% in years two and three without affecting the fleet lease percentage established during the first year.

At the end of three years, Toni assumes for the purpose of the assessment that the buyer will resell the business for three times the revenue earned in year three. Until the end of the third year, the fleet size will remain constant so that no systems will be bought or sold.

---

### 3.3 Some definitions are in order

Toni defines cash flow simply as revenue minus expenses, and for this assessment will ignore depreciation and income taxes. After all, that is on the buyer in any case. She assumes that cash flow in year three includes the proceeds from the sale of the business at the end of the year and that year one's cash flow does not include the purchase price as this transaction will have happened prior to the first year in the analysis. She also defines overall investment profit as the net present value of the cash flows over the forecast horizon, assuming a discount rate of 10%.

---

### 3.4 Requirements

1. We construct an influence diagram for our analysis, labeling the decision variables, the key parameters, and the performance measures.
2. We then use the influence diagram diagram to construct an Excel spreadsheet. We label the cells containing the key problem param-

eters and constants. Our design and implementation uses named ranges, and no hard coding of constants anywhere.

3. We use a data table to determine the system lease price that achieves the highest overall investment performance. A graph builds on our ability to communicate results.
4. We then construct a 2-way data table to analyze the sensitivity of overall investment profitability to at least one of following parameters: purchase price, annual maintenance cost/system, annual property taxes, and/or lease price growth rate.
5. We deposit results in a dashboard to summarize and focus a discussion around this organization/s value.





---

## ***Part 2 – Optimization***

---

Getting our arms around

- Designing models with influence diagrams, again
- Identifying the key business questions
- Identifying decisions, formulating objectives, and developing constraints
- Collecting data relevant to the decision framework which helps to answer the business questions
- Implementing model as a linear or non-linear program
- Analyzing sensitivity of results with respect to changes in decision and constraints parameters



## 4

---

### *Bringing Pie to Earth*

---

---

#### 4.1 It all started here

The first centuries of Islam saw the confluence of Hindu, Persian, and Mesopotamian science and mathematics. It is the algorithms that concern us here. The Latinized (translated, that is) name of Persian mathematician Muḥammad ibn Mūsā al-Khwārizmī (Persian: موسا بن محمد خوارزمی, romanized: Moḥammad ben Musā Khwārazmī; 780 – 850 CE) is *Algorithmi*—from the Arabicized al-Khwarizmi and formerly Latinized as Algorithmi. He led the research and teaching programs at Baghdad’s *Beit-al-Hikima*, the House of Wisdom.

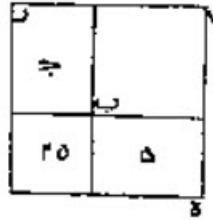
Central to our purposes, Al-Khwarizmi wrote and compiled a procedure in *The Compendious Book on Calculation by Completion and Balancing*, 813–833 CE (ibn Mūsā al Khwārizmī (1881)) which we still use today to systematically solve linear and quadratic equations. We come across both in our work: net present value and profit both are quadratic in price, or volume. He employed both *reduction* and *balancing* (whatever we do to one side of the equation we must do on the other side). The algorithms we use to solve for optimal decisions depend explicitly on the procedures developed and taught. As did al-Khwarizmi, we too will use algebra and the geometrical representation of equations to describe, explain, and interpret our results.

This page from al-Khwarizmi’s work describes his procedure to find the largest square of 8, and of course it is 64. We will use equations, drawings, and graphs to perform much the same tasks for this menagerie of models.

- allocation models
- covering models
- blending models
- network models

As we should always, we begin with a simple, but instructive, and even useful model for a decision maker.

علي تسعة ونلتين ليتم السطح الأعظم الذي هو سطح رة فبلغ ذلك كله أربعة وستين فاحذنا جذرها وهو ثمانية وهو أحد اضلاع السطح الأعظم فإذا نقصنا منه مثل ما زدنا عليه وهو خمسة بقي ثلثة وهو ضلع سطح  $\overline{AB}$  الذي هو المال وهو جذره والمال تسعة وهذه صورته



وأما مال واحد وعشرون درهما يعدل عشرة اجذاره فانا نجعل المال سطحاً مربعاً مجهول الاضلاع وهو سطح  $\overline{AD}$  ثم نضم اليه سطحاً متوازي الاضلاع عرضه مثل أحد اضلاع سطح  $\overline{AD}$  وهو ضلع  $\overline{DN}$  والسطح  $\overline{DB}$  فصار طول السطحين جميعاً ضلع  $\overline{DN}$  وقد علمنا ان طوله عشرة من العدد لان كل سطح مربع معاصري الاضلاع والزوايا فان أحد اضلاعه مضروباً في واحد جذر ذلك السطح وفي اثنين جذراه فلما قال مال واحد وعشرون يعدل عشرة اجذاره علمنا ان طول ضلع  $\overline{AD}$  عشرة اعداد لان ضلع  $\overline{AD}$  جذر المال فتقسمنا ضلع  $\overline{AD}$  بنصفين علي نقطة

FIGURE 4.1: Origins of algebra.

---

## 4.2 From humble beginnings

Allocation models choose mixes and compositions typically of competing decisions to optimize an objective (often some sort of cash flow or profit) subject to less-than or equal to constraints on capacity and availability of a resource. Examples include product and sales mix, vendor mix, and team member compositions. Resource constraints often dictate a trade off among alternatives.

Make-A-Pie (MAP) produces two kinds of pies: savory and fruit. Each product requires some labor in procurement of ingredients, assembly of pies, and baking. MAP sells pies through a local distributor as well as by contract to grocery store chains. These marketing channels estimate maximum potential sales for each product in the coming quarter. The MAP accounting department calculates profit contributions for each product. The decision problem is to determine the product mix which maximizes MAP's profit for the quarter by choosing production quantities for the savory and fruit pies.

We ask three questions.

1. **What are we deciding?** We need to set product mix. These are the volume of pies to be produced. There will be many possible, feasible alternatives from which to choose. For MAP the decisions are just the number of savory pies and number of fruit pies.
2. **What is success?** We next define a performance criteria against which we measure the impact of our decision alternatives. In many organizations, the criterion is profit, or net cash flow, or net present value of benefits, costs, and risks to performance. At the least we will need to know the contribution of decisions to profit. To that end we ask the management accountants to render profit per pie. This number when multiplied by the decision number of pies will yield the total profit of the decision alternative. We are not yet done. Our criterion is **maximize profit**.
3. **How is success constrained?** First of all, the decision variable themselves may have minimums and maximums. Contracting with grocery chains may require minimum volumes of on-time and in-full delivery of pies to maintain the marketing channel. Second, each process will have only so many hours of operation in a given week. The time frame of a week, of the limitations of work force, materials, and even square footage of floor space to produce all present themselves as candidates to constrain the feasible combinations of fruit and savory pie production volumes. Then there is potential demand which requires a number of fruit and savory pies.

We have additional information from Tortiere and her CFO Marta Fazi. Adhering to good spreadsheet engineering practices we store this first report in a separate worksheet called `margin`.

	A	B	C	D	E
3					
4			<b>fruit</b>	<b>savory</b>	
5		<b>Sales</b>	\$61,991.00	\$19,601.00	
6		<b>Variable Costs</b>			
7		Direct materials	\$15,279.63	\$5,096.26	
8		Direct labor	\$12,535.96	\$4,332.17	
9		Variable manufacturing overhead	\$3,822.78	\$1,524.52	
10		Variable selling & administrative	\$6,199.10	\$1,960.10	
11		<b>Total</b>	<b>\$37,837.47</b>	<b>\$12,913.05</b>	
12					
13		<b>Allocated Fixed Costs</b>			
14		Manufacturing overhead	\$9,687.83	\$4,330.69	
15		Selling & Administrative	\$8,577.78	\$2,277.78	
16		<b>Total</b>	<b>\$18,265.61</b>	<b>\$6,608.47</b>	
17					
18		<b>Net Profit</b>	\$5,887.92	\$79.48	
19		<b>Units Sold</b>	8102	2267	
20		<b>Net Profit Per Unit</b>	\$0.7267	\$0.0351	
21					
22		<b>Marginal Revenue Per Unit</b>	\$7.6513	\$8.6462	
23		<b>Variable Cost Per Unit</b>	\$4.6701	\$5.6961	
24		<b>Marginal Profit Per Unit</b>	\$2.9812	\$2.9501	
25					
26		<b>Direct materials per unit</b>			
27		Filling	\$1.13	\$1.12	
28		Dough	\$0.75	\$1.12	
29					

**FIGURE 4.2:** Product margin calculations.

Marta believes that the savory line of pies is very unprofitable from a pure accounting model of value. However, she knows that allocated fixed costs must be incurred given a capacity to produce a range of the volume of pies. These expenses include environmental, health, and safety costs, general maintenance of equipment and work spaces, and even long-term employee benefits. Thus, she produced a marginal revenue, variable cost, and materials report.

We work with Marta to develop a further understanding of the business. We ask about the production process, its sub-processes, resource requirements per pie per pie line, and the availability of resources. Critical to this understanding

is the way the decisions to produce pies winds its way into the underlying production processes and material availability.

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		<b>Availability and capacity</b>										
3		<b>processing</b>	fte	days	hours / shift	shifts / day	minutes / week					
4		staging	3	7	8	2	20160	G3	=C4*D4*E4*F4*60			
5		assembling	3	7	8	2	20160					
6		baking	10	7	8	3	100800					
7		packaging	2	7	8	2	13440					
8						total	154560	G8	=SUM(G4:G7)			
9												
10			unit cost		material proportion							
11		<b>direct materials</b>	fruit	savory	fruit	savory						
12		dough	1.1315453	1.1240097	0.6	0.5	C12	=margin/C27		E12 =C12 / SUM(C512:C513)		
13		filling	0.7543635	1.1240097	0.4	0.5						
14												
15			beginning inventory		addition to inventory		usage		usage	ending inventory	average availability	
16		<b>direct materials</b>	fruit	savory	fruit	savory	fruit	savory	total	fruit	savory	
17		dough	108	42	5000	2000	4861.2	1133.5	5994.7	246.8	908.5	7150
18		filling	98	62	4000	2000	3240.8	1133.5	4374.3	857.2	928.5	6160
19												
20		<b>processing</b>	fruit	savory			G17		I17	J17		L17
21		staging	1.00	1.00			=E12 *margin/C519		=G17+H17		=C17 + E17 - G17	
22		assembling	2.00	1.50								
23		baking	9.00	10.00								
24		packaging	1.00	1.00								
25		minutes / pie	13.00	13.50			=SUM(C21:C24)					
26		gross labor capacity	11889	11449			=G58 / C25					
27		factor	0.80	0.80								
28		maximum labor capacity	9511.3846	9159.1111			=C27 * C26					
29												

**FIGURE 4.3:** Product availability drivers.

Marta likes the documentation using FORMULATEXT() as she trains, and educates, analysts to think through the work flow. We agree that the minutes it takes to procure, assemble, bake, and package pies is a unit of productivity measure of importance. Labor is key to the operation. The available labor minutes by activity constrains the production of pies. Adhering to good spreadsheet engineering practices we store the report in a separate worksheet called **availability**. Labor minutes of availability across pie lines is influenced by the full time equivalent work force, number of shifts (maximum of 3 8 hour shifts), number of days (7 in a week), minutes in a day run through hours.

We model material availability through inventory. We start with beginning of the week inventory at the end of the previous week. During the week we procure ingredients thus adding to the beginning of week inventory. Then we start to use the ingredients in some proportion that is possibly different for each pie line. At the end of the week we tally up additions and subtractions to arrive at the end of week inventory. We then define material availability as beginning of week inventory plus additions to inventory procured during the week, but in advance of a production run in any shift.

#### 4.2.1 1. Decisions, decisions

We attend to our first question, first. **What are we deciding?** We know we need to set the product mix. These are the volume of pies to be produced. There will be many possible, feasible alternatives from which to choose. For

MAP the decisions are just the number of savory pies and number of fruit pies.

Here is a 2 dimension mapping of the decision space of all volumes of fruit and savory pies.



**FIGURE 4.4:** Bounds on product decisions.

First, we notice that we define decisions only in the positive quadrant of all possible combinations of fruit  $X$  and savory  $Y$  pies. Both decisions are continuous on the positive real number plane.

$$X \geq 0 \quad (4.1)$$

$$Y \geq 0 \quad (4.2)$$

Not only this constraint, but we also observe that there are minimum and maximum boundaries for the decisions.

$$6000 \leq X \leq 9511 \quad (4.3)$$

$$300 \leq Y \leq 9159 \quad (4.4)$$

Both a graph and algebraic expression help us describe the decision space of possibilities.

#### 4.2.2 2. Resources they are a'constraining

Perhaps borrowing too liberally from the Bob Dylan verse, we proceed to the second modeling question **What is success?** We now work with Marta



and her MAP analysts to define a performance criteria against which we can measure the impact of our decision alternatives. In many organizations, the criterion is profit, or net cash flow, or net present value of benefits, costs, and risks to performance. In general this measure is multi-faceted for example when removing snow in Montreal PQ, there are competing objectives of cost and environmental impact minimization (Labelle et al. (2002)).

At the least we will need to know the contribution of decisions to profit. To that end we review Marta's profit per pie report. These marginal profits when multiplied by the decision number of pies will yield the total profit of the decision alternative. We are not yet done. Our criterion is then to **maximize profit**.

This graph depicts the family of equal-profits combinations of  $X$  = fruit pie volume and  $Y$  = savory pie volume, known affectionately as **iso-profit** curves. We can compactly express these curves using this equation.

$$aX + bY = c \quad (4.5)$$

We already know what  $X$  and  $Y$  are. Coefficient  $a$  is the profit contribution per unit of decision  $X$  units (fruit pies!). Coefficient  $c$  is the profit contribution per unit of decision  $Y$  units (savory pies!). These add up to  $c$  total profits.

For example if we produce  $X = 10$  fruit pies and  $Y = 5$  savory prices, *and*  $a = \$2.00$  marginal profit per fruit pie produced and sold and  $b = \$3.00$  marginal profit per savory pie produced and sold, then we can compute total profit  $c$ .

$$aX + bY = c \quad (4.6)$$

$$(2.00)(10) + (3.00)(5) = c \quad (4.7)$$

$$20 + 15 = c \quad (4.8)$$

$$35 = c \quad (4.9)$$

Yes, it is as simple as this. The hard thing we had to do was come up with the right decisions and the marginal profits. We will use this formulation throughout our modeling work flow. We will also produce plots that will depend on the solution of  $Y$ , on the vertical axis, the ordinate, in terms of  $X$  on the horizontal axis, the abscissa, of the plot. In homage to al-Khwarizmi, we balance and reduce terms to arrive at the solution.

$$aX + bY = c \quad (4.10)$$

$$aX - aX + bY = c - aX \quad (4.11)$$

$$0 + bY = c - aX \quad (4.12)$$

$$bY = c - aX \quad (4.13)$$

$$\frac{bY}{b} = \frac{c - aX}{b} \quad (4.14)$$

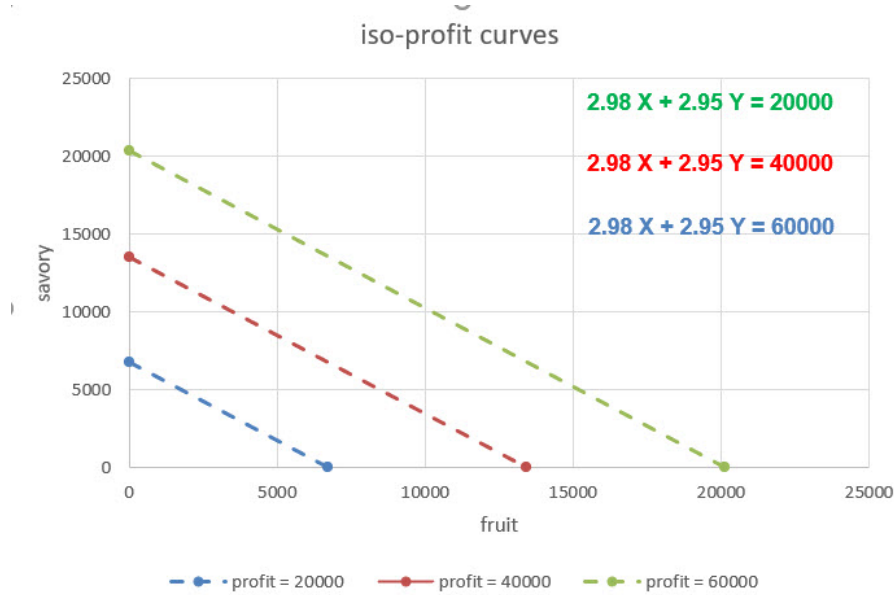
$$(1)Y = \frac{1}{b}(c - aX) \quad (4.15)$$

$$(1)Y = \frac{c}{b} - \left(\frac{a}{b}\right)X \quad (4.16)$$

$$Y = \frac{c}{b} - \left(\frac{a}{b}\right)X \quad (4.17)$$

Where we used the algebraic facts that  $aX - aX = 0$ ,  $b/b = 1$ , and the distributive property of multiplication over addition. The term  $c/b$  is the intercept and the term  $-(a/b)$  is the slope of the straight-line (often called linear) equation of decision  $Y$  in terms of decision  $X$ . In all of its glorious detail.

After all of that algebra in all of its glorious detail, we deserve to see our handiwork.



**FIGURE 4.5:** Iso-profit analysis.

We parameterize the iso-profit curves with three levels of profits, \$20,000, \$40,000, and \$60,000, color coordinating the equations and the dashed iso-profit lines. An increase in profits shifts the iso-profit lines to the right. Not

only that, but increased profits require, of course, more sales and production of pies. The slopes of each of the iso-profit curves are the same. This means they are parallel to one another, they cannot ever cross in this slice of the decision universe.

We can measure the profit tradeoff between savory and fruit pies directly through the slope. While we might have avoided calculus in our analytics journey, or perhaps not encountered it officially at all, we can use it to think through the marginal profit tradeoff. We start with the total change in profit  $dc = a dX + b dY$ , that is the sum of the change in profit due to a change in volume of fruit pies,  $a dX$ , and the change in profit due to a change in volume of savory pies,  $b dY$ .

These are lovely mathematical and linear operators at work. Without the fuss of knowing exactly what all that entails, but because they are linear (and by the way smooth) operators, we can solve for the ratio of  $dY/dX$ , the vauntedY profit tradeoff, when, and this is key, the change in profit is stable, that is, 0,  $dc = 0$ .

$$dc = a dX + b dY \quad (4.18)$$

$$0 - a dX = a dX + b dY - a dX \quad (4.19)$$

$$-a dX = b dY \quad (4.20)$$

$$\frac{dY}{dX} = -\frac{a}{b} \quad (4.21)$$

In our case  $a = 2.98$  and  $b = 2.95$ . The two pie lines are literally neck-to-neck. The tradeoff is evidenced in the negative slope. A (very small) increase in fruit pie volumes **reduces** profit by  $2.98/2.95 = 1.01$ , just a bit over a dollar. This tradeoff is the same for any level of profit, thus the  $dc = 0$  assumption. Phew!

Let's overlay the minimum and maximum decision constraints on the iso-profit curves to get some perspective and maybe another insight or two.

The low profit regime (\$20,000) just seems to makes it into the feasible rectangle of minimum and maximum pie requirements. The high profit regime (\$60,000) does not at all. The **decision core** will be somewhere within the feasible rectangle. One such profit is a mid profit regime of (\$40,000). Is this the optimal profit? Not yet! By shifting the iso-profit lines we must see that there is one (1) optimal, that is, maximizing profit, fruit and savory volumes solution. This plot depicts the situation.

That's right, the maximizing profit decision is to produce and sell the maximum number of fruit and savory pies. This yields a profit of \$55,378. We simply moved the high profit regime line to the left to land on the northeast vertex (also known as a corner) of the four-sided polytope (also known here as a variant called a rectangle) that describes the region of feasible decision combinations. Again PHEW!

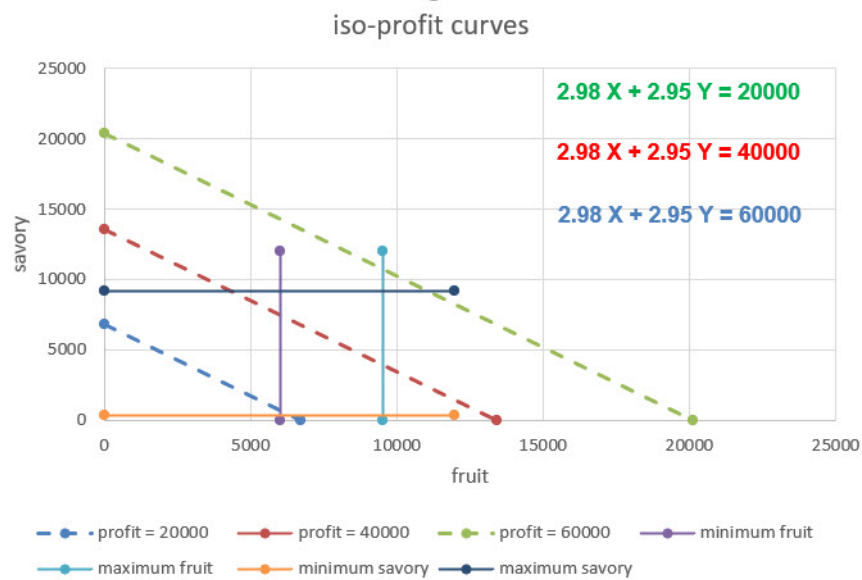


FIGURE 4.6: Bounds meet iso-profits.

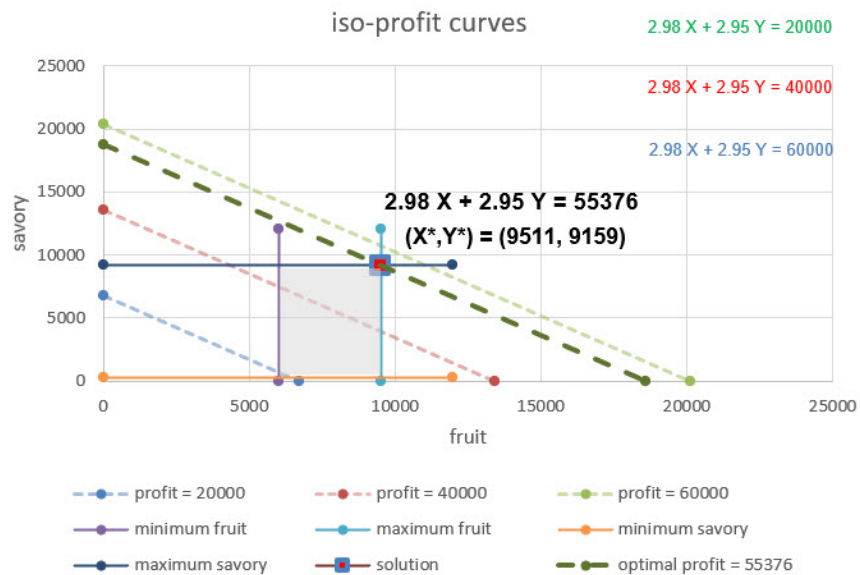


FIGURE 4.7: A solution?

### 4.2.3 3. Not without constraint

We reach our third modeling question **how is success constrained?** First of all, the decision variable themselves may have minimums and maximums. Contracting with grocery chains may require minimum volumes of on-time and in-full delivery of pies to maintain the marketing channel. We know that this is sufficient to get at a notion of an optimized objective.

Second, each of the procuring, assembling, baking, and packaging processes will have only so many possible hours of operation in a given week. The strict time frame of a minute, an hour, a day, a week, of the limitations of work force, materials, and even square footage of floor space to produce all present themselves as candidates to constrain the feasible combinations of fruit and savory pie production volumes. Then there is potential demand which requires a number of fruit and savory pies. Even here markets, competition, communication create further constraints.

Again, we attempt an algebraic description. For the decided upon volume of  $X$  = of fruit pies and  $Y$  of savory pies, we have four (4) process resource constraints.

$$\text{staging} : X + Y \leq 20160 \quad (4.22)$$

$$\text{assembling} : 2X + 1.5Y \leq 20160 \quad (4.23)$$

$$\text{baking} : 9X + 10Y \leq 100800 \quad (4.24)$$

$$\text{packaging} : X + Y \leq 13440 \quad (4.25)$$

Here are two (2) materials constraints. We might call these process input constraints as well.

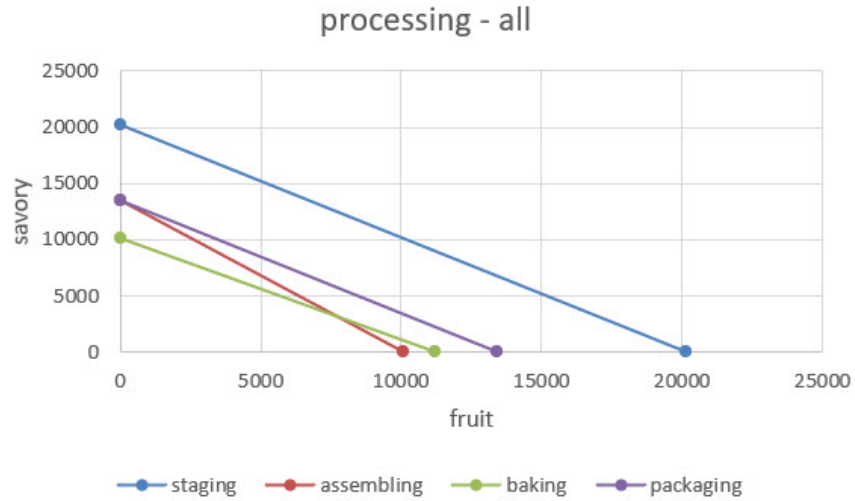
$$\text{filling} : 0.6X + 0.5Y \leq 7150 \quad (4.26)$$

$$\text{dough} : 0.4X + 0.5Y \leq 6160 \quad (4.27)$$

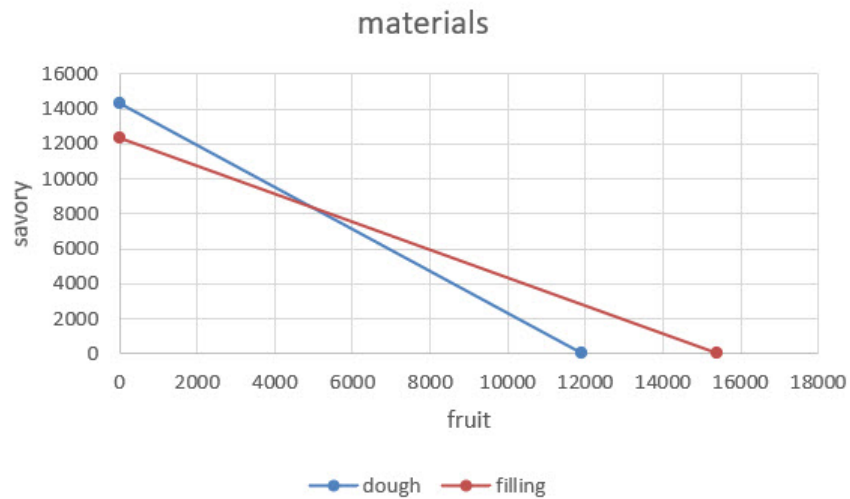
With the same results we developed for profit curves, we can find intercepts and slopes and plot these relationships. Here are the constraints for the four processes. We leave out the profit relationship for now.

We seek only the most constraining of these curves. We immediately see that both staging and packaging outstrip the assembling and baking processes. To the left of the assembling and backing process constraints lies a potential set of feasible decisions. Also we note that the slopes of the staging and packaging process constraints are the same. This adds to their irrelevance and redundancy. It is not that they are unimportant. Not at all, just not helpful in identifying the region within which decisions can be found feasibly. By feasible we then mean the decision space that fits all of the constraints. This criterion is then met only by the most constraining of the processes here.

And here are two ingredients for the material group of constraints.



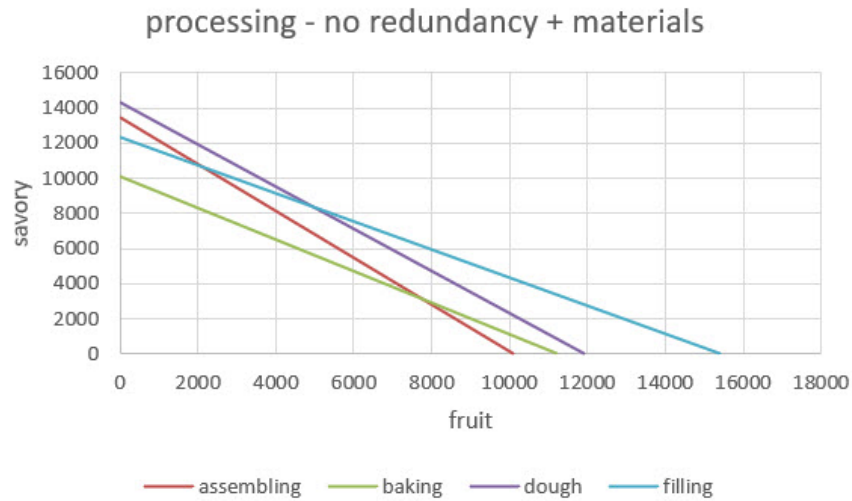
**FIGURE 4.8:** Product and process together at last.



They intersect at around  $X = 5,000$  fruit pies and  $Y =$  a little under 9,000 savory pies. To the left of the kinked boundary at this point we may, just maybe, an optimal decision.

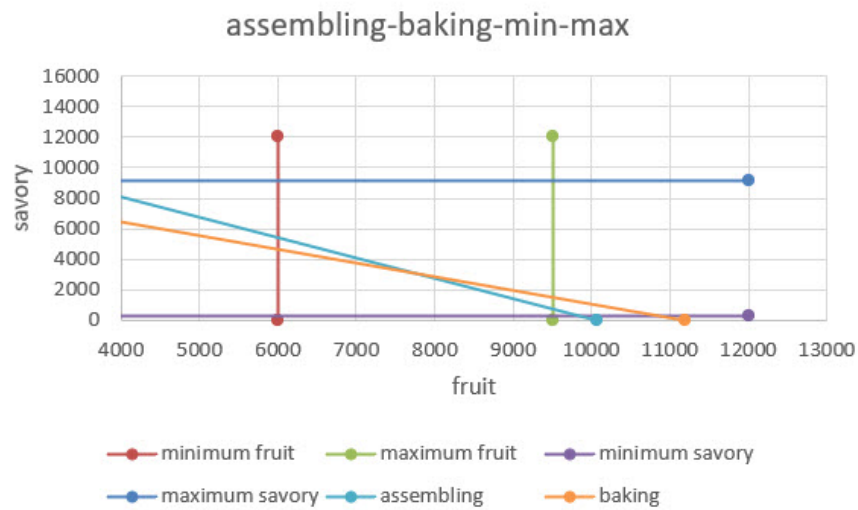
We now put the material constraints together with the relevant assembling and baking constraints.

Wow, both of the material constraints are cut away by just the assembling and baking process constraints. Those two are the most constraining of the



**FIGURE 4.9:** The principle of parsimony at work.

suite of six constraints. They alone define, along with the demand constraints, the feasible set of decisions.



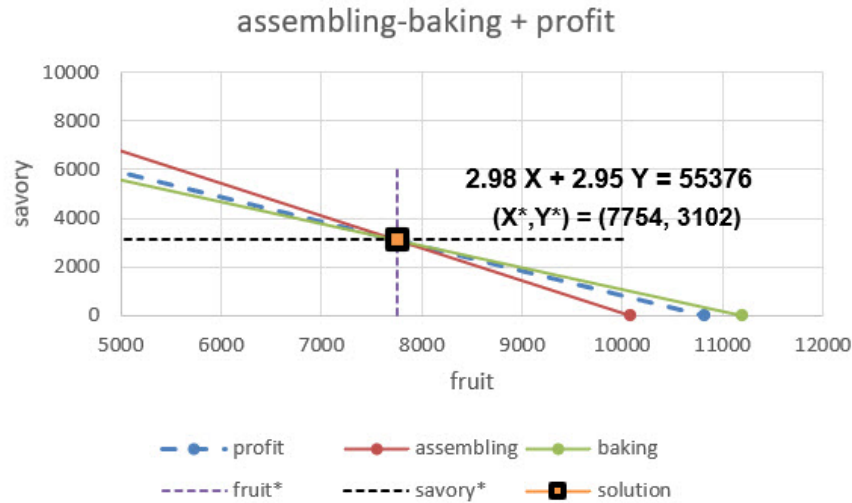
**FIGURE 4.10:** Keeping it simple.

That point at which the assembling and baking processes lies well within the minimum and maximum boundaries of the demand requirements. We arrive

at the feasible decision space. Just one more step will get us to our goal of maximizing profit with pies!

#### 4.2.4 Let us eat pie

And produce and sell it. Our sponsor Simone Tortiere and colleague Marta Fazi have followed us through this intense analysis. We are now ready to uncover the answers which we seek in this case. Here is the the final overlay of an iso-profit line through the intersection of the assembling and baking constraints.



**FIGURE 4.11:** Process yielding profit solution.

The iso-profit line registers a maximizing profit of \$55,376 against optimal 7,764 fruit and 3,102 savory pies. How can we know this is the best choice? First, any pie decision greater than this will be infeasible, will not honor the assembling and baking process constraints. Second, any iso-profit to the right of the dashed line will surely be at a greater profit but then will evoke decisions that are infeasible. Any iso-profit to the left will enlist feasible decisions, but will be less profitable than the profit at the dashed line. We have found an optimizing solution, the hard way!



### 4.3 An easier way?

We can use the algebra of al-Khwarizmi to solve consecutively, much like the process we just followed, by eliminating, balancing, reducing constraints until we get to that one, or two, or three, or more feasible decision points. At those points we can move in the direction of increasing profit, moving the iso-profit curve to match the feasible decision point, all very geometric, in an algebraic sort of way. George Dantzig (1949) developed such an algorithm, the Simplex algorithm of linear programming.<sup>1</sup>

We use the Solver add-in from Excel to help us solve this problem algebraically. We check that Excel has enabled the Solver add-in using the instructions at this site. There are instructions for Windows, Mac-OS, and other operating systems. We find the Solver add-in in the Analysis group in the Data ribbon (way to the right!). Frontline Systems gives us this tutorial for using the Solver add-in, replete with examples and spreadsheets.

We set up a separate worksheet for our decision optimization. Here is the setup.

	A	B	C	D	E	F	G	H
1								
2		<b>Make-A-Pie product mix</b>						
3								
4			<b>fruit</b>	<b>savory</b>	Maximize: C13 By Changing: C5:D5 Subject to: E16:E19<=F16:F19 E22:E23<=F22:F23 C5:D5<=C6:D6 C5:D5<=C7:D7			
5		<b>pies produced</b>	7754	3102				
6		minimum	6000	300				
7		maximum	9511	9159				
8								
9		unit selling price	\$7.65	\$8.65				=margin!C30
10		unit variable cost	\$4.67	\$5.70				=margin!C23
11		unit margin	\$2.98	\$2.95				=C9-C10
12								
13		<b>total margin</b>	\$32,266					=SUMPRODUCT(C11:D11,\$C\$5:\$D\$5)
14								
15		<b>processing</b>	<b>minutes per pie</b>	<b>used</b>	<b>available</b>			
16		staging	1.00	1.00	10855.4	20160		=SUMPRODUCT(C16:D16,\$C\$5:\$D\$5)
17		assembling	2.00	1.50	20160	20160		=SUMPRODUCT(C17:D17,\$C\$5:\$D\$5)
18		baking	9.00	10.00	100800	100800		=SUMPRODUCT(C18:D18,\$C\$5:\$D\$5)
19		packaging	1.00	1.00	10855.4	13440		=SUMPRODUCT(C19:D19,\$C\$5:\$D\$5)
20								
21		<b>direct materials</b>	<b>proportion per pie</b>	<b>used</b>	<b>available</b>			
22		dough	0.60	0.50	6203.08	7150		=SUMPRODUCT(C22:D22,\$C\$5:\$D\$5)
23		filling	0.40	0.50	4652.31	6160		=SUMPRODUCT(C23:D23,\$C\$5:\$D\$5)
24								

**FIGURE 4.12:** Solving with Solver.

<sup>1</sup>Here is the abstract. “Activities (or production processes) are considered as building blocks out of which a technology is constructed. Postulates are developed by which activities may be combined. The main part of the paper is concerned with the discrete type model and the use of a linear maximization function for finding the ‘optimum’ program. The mathematical problem associated with this approach is developed first in general notation and then in terms of a dynamic system of equations expressed in matrix notation. Typical problems from the fields of inter-industry relations, transportation nutrition, warehouse storage, and air transport are given in the last section.”

We color decisions yellow, criteria blue, and constraint usage green. Extensively SUMPRODUCT comes to our aid. We see the technology of pie making through the process and material constraints. The proportions of material filling and dough and the processing of the materials into pies evoke the design of each pie. We express labor and capital supply and demand, customer supply and demand through costs, availability, capacity values and constraints. This worksheet refers to other worksheets like availability and margin for values from reports and analysis.

We are ready to use Solver.

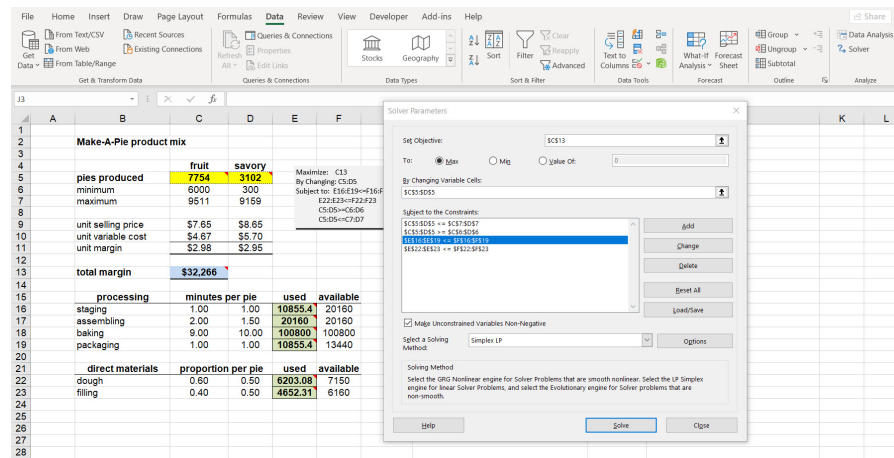


FIGURE 4.13: More from Solver.

Let's look at one of the constraints, actually all four of the processing constraints. Here is how we enter and edit a constraint.

Constraint arrays can be chosen without the need to put each constraint separately. Thus the way we set up the problem with constraints in rows together, decisions in columns together, allow us this succinct way to implement the model. We press OK and go back to the main Solver dialogue. If we think we have loaded all of the data into Solver, we then press Solve.

This dialogue asks if we want to use the new solution. If yes, this will wipe out the previous solution. We also can write other reports to worksheets. Sensitivity analysis is one such report we will need. Press OK to see the solution deposited into our model.

All of this is very mechanical. Our design of worksheets definitely has that character. But design is born of principles and practice. These spreadsheets follow the logic of the algebraic model. The model and its inputs in turn follow the scope of Tortiere's and Fazi's requests, born of the need to understand the

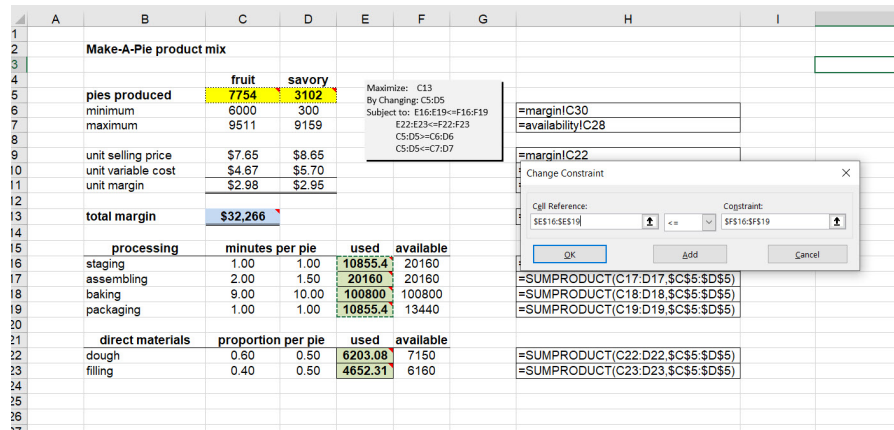


FIGURE 4.14: Solver constraints.

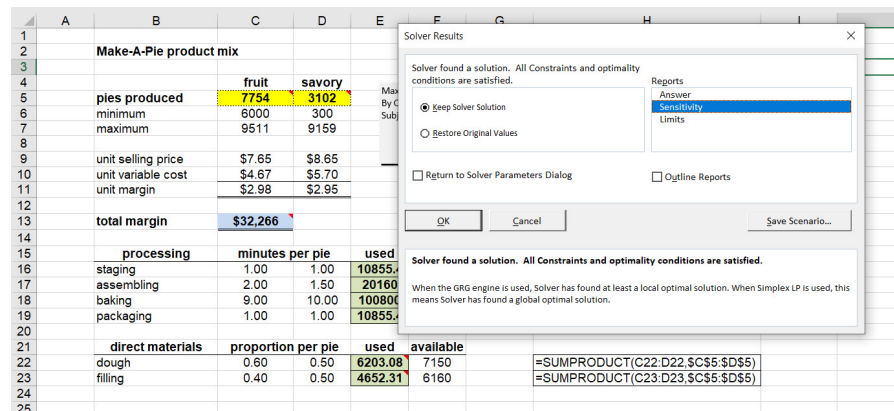


FIGURE 4.15: Solver gets a feasible and optimal solutions!

technology, markets, and competition in the markets all around pie making and feeding others.<sup>2</sup>

<sup>2</sup>We can visit Winston (2019) for several of the tricks up his sleeves. Chapter 29 is an introduction. He follows this with chapters 30-35 for various examples of linear programming of many business problems. Other chapters in his book reveal further examples.



# 5

## *Expanding horizons*

### 5.1 Dido's bullhide



*Dido Purchases Land for the Foundation of Carthage. Engraving by Matthäus Merian the Elder, in *Historische Chronica*, Frankfurt a.M., 1630. Dido's people cut the hide of an ox into thin strips and try to enclose a maximal domain.*

**FIGURE 5.1:** Dido's bullhide; getting the most out of a surface.

Queen Dido fled the tyrannical ruler of Tyre to found the city of Carthage. Rome later razed it to the ground, salted the earth, and to this day, the site in Tunisia is inarable. Virgil at about 21-19 BCE regaled the ruling class of Rome with its origin story, that of Aeneas, Greek veteran of the Trojan War. Aeneas landed at Carthage and into Dido's arms. Star-crossed lovers were not blessed by gods or humans, all ending in tragedy for Dido and Carthage, eventually, and the establishment of Rome by myth.

Wonderful story, but here is a formal optimization problem solved by Dido which resulted in her, and her retinue, founding of Carthage. They solved the problem of finding the figure bounded by a line which has the maximum area for a given perimeter. The solution is a semicircle. Here is a statement of the problem in Virgil's Aeneid, of course, with an engaging back story up front to keep up at the end of our seats.

"The Kingdom you see is Carthage, the Tyrians, the town of Agenor; But the country around is Libya, no folk to meet in war. Dido, who left the city of Tyre to escape her brother, Rules here—a long and labyrinthine tale of wrong Is hers, but I will touch on its salient points in order....Dido, in great disquiet, organised her friends for escape. They met together, all those who harshly hated the tyrant Or keenly feared him: they seized some ships which chanced to be ready... They came to this spot, where to-day you can behold the mighty Battlements and the rising citadel of New Carthage, And purchased a site, which was named 'Bull's Hide' after the bargain By which they should get as much land as they could enclose with a bull's hide."

Well, we suppose they took the largest bull they could find and produced the longest length of leather thong they could and enclosed land area within the semicircle of leather lacing. Well, that's an optimization problem, solved!

Around 300 BCE Euclid wonders about the minimal distance between a point a line, and proves that a square has the greatest area among the rectangles with given total length of edges. About 100 BCE Heron proves that light travels between two points through the path with shortest length when reflecting from a mirror. This last optimization had psychological and metaphysical implications since the reigning theory from Plato was that light emanates from our souls outward into the world and what we see is what is reflected back to us.

Here we go this time with part two (ni in Japanese this time). In mathematical finance, the idea of maximising a profit is, of course, very important, but it's modulated by the need to simultaneously manage the risk of loss. This all boils down to a sophisticated version of the very everyday idea of not putting all your eggs in one basket, also known as diversification.

Cervantes wrote *Don Quixote* in the 1600s. Here's a useful quote.

"It is the part of a wise man to keep himself today for tomorrow and not venture all his eggs in one basket."

Not only is diversification involved here, but also the beginnings of the notion of delayed consumption. The idea is that we might need to save for the future. Inventory is a sort of delayed consumption as well. Cervantes almost gets it. Another quote from Shakespeare expands on Cervantes (although backward in time!) the more sophisticated idea. In the Merchant of Venice, Shakespeare cleverly diversifies the asset structure of a whole estate, the portfolio, across

time, when Antonio, in a conversation with his friends, told his friends he wasn't spending sleepless nights worrying over his commodity investments, and he says:

"My ventures are not in one bottom trusted, nor to one place, nor is my whole estate upon the fortune of this present year, therefore my merchandise makes me not sad."

Hidden in there is this idea of thinking about having a balance between the near future and the rather more distant future in terms of the return on a portfolio of investments. Very nice, but how often should we diversify? Here is a heuristic from Rabbi Isaac bar Aha, in Aramaic. It's the simple equal allocation strategy, and what you should do with your wealth, you should hold a third in land, a third in merchandise, and a third, as he put it, "at hand", which means in cash. That is the so-called *one over n* strategy. If you have  $n$  assets, you put an equal amount into each one. If you keep doing that, you're not necessarily turning over your portfolio very frequently when you move from one allocation to the other over time. Is there a Confucian version? Not necessarily, but Confucius did say that we who will not economise will have to agonise. We might think of that when we refinance, especially in tail events like the current pandemic - recession - whatever.

It is interesting that profit in prices in the NPV calculation we performed with Make-A-Pie is quadratic. When we go from the quadratic total profit to the marginal profit per unit we move to the linear. That was the big idea in optimization introduced in the 1870's by Walras. What is new now? Life is truly not quadratic! Normal (Gaussian) distributions are quadratic. Markets are not. They are not symmetric, nor are our profit decisions. They are skewed, they have outliers. We could use this for inspiration here here as we expand our horizons and dialogue with Robert Hamilton in his Introduction to Merchandize (Hamilton (1777)).



240 **A L G E B R A.** **PART II.**

13. Two men A and B working together, perform a piece of work in 8 days; A could have done it alone in 12 days. In what time could B have done it?
14. Three men can do a piece of work, the first in 50 hours, the second in 60 hours, and the third 75 hours. In what time will they do it all working together?
15. A and B together can do a piece of work in 12 hours, A and C in 20, and B and C together in 15 hours. In what time will all three do it?
16. The same things being given, required in what time each person can do it separately?
17. A gentleman distributing money to some poor people, found that, if he gave each half a crown, he would have 5 sh. over, but, if he gave each 3 sh. he would want 4 sh. Required the number of poor?
18. The paving of a square court at 6 d. a yard cost as much as the inclosing at 5 sh. a yard. Required its extent?
19. An army being drawn up in a square, there were 79 men over; but, in attempting to enlarge each side of the square by one man, there were 89 too few. Required the number of men in a rank, and altogether?
20. Required two numbers whose sum is 63, and difference 27?
21. Required two numbers whose sum is 48, and quotient 3?
22. Required two numbers whose difference is 14, and their proportion as 5 to 7?
23. Required two numbers, such that 5 times the greater, together with 3 times the lesser, may amount to 132, and 4 times the greater together with 6 times the lesser, may amount to 144?
24. Required two numbers, such that 5 times the greater may be 7 more than 8 times the lesser, and 7 times the greater may be 2 more than 12 times the lesser?
25. Required two numbers, such that  $\frac{1}{2}$  of the first, together with  $\frac{1}{3}$  of the second, may amount to 36 and  $\frac{1}{4}$  of the first, together with  $\frac{1}{5}$  of the second, may amount to 34?
26. Required two numbers, whose sum is nine times greater than their difference, and the amount of twice the greater, together with thrice the lesser, is 88?
27. Required two numbers, whose quotient is 5 and product 180?
28. Required two numbers, whose difference is 5, and the difference of their squares 205?
29. Required two numbers, whose proportion is as 2 to 3, and the sum of their squares 325?
30. Required two numbers, whose proportion is as 3 to 5, and the difference of whose squares is  $\frac{1}{2}$  of their product?

31.

**CHAP. V. EQUATIONS.** 241

31. Required two numbers, whose proportion is as 4 to 5, and the sum of their cubes 5103?
32. 504 bushels wheat were mixed from a cargo at 21 sh. and another at 17 sh. 6 d. in such proportion that the value was 19 sh. Required how much of each kind?
33. Required to mix 50 lb. tea at 9 sh. with some at 14 sh. and some at 8 sh. in order to make up 100 lb. at 10 sh.?
34. A mixture of gold and silver weighs 170 ounces, and its specific gravity is 15. Required how much of each metal it contains, the specific gravity of gold being 19, and of silver 10 $\frac{1}{2}$ ?
35. A goldsmith mixes 18 ounces of gold with alloy, and forms a mass of 24 ounces, the specific gravity of which is 16. Required the specific gravity of the alloy?
36. A sum of L. 30 was paid in half guineas and half-crowns. The number of pieces paid was 80. Required the number of each?
37. A gentleman has a chaise valued at L. 48, and two pair horses; the value of the chaise and horses when the best pair is yoked is thrice as much as the other pair, but when the worst pair are yoked, only twice as much. Required the value of the horses?
38. A vessel containing 200 cubical inches is filled with a mixture of wine and spirits, the specific gravity of the wine is .950, of the spirits .858, and of the mixture .875. Required how much of each?
39. A and B left together 16 fons; A's fortune was  $\frac{1}{2}$  greater than B's; but each being equally divided among the families, the proportion which fell to A's sons was only  $\frac{1}{5}$  of the portion which fell to B's sons. How many fons had each?
40. A farmer kept a servant for every 40 acres he possessed; but, on taking a lease of 104 more acres, he engages three new servants, and then he had a servant for every 36 acres. Required the number of acres and servants?
41. A mercer bought two kinds of silk, and sold them at 2 sh. a yard profit; he paid as much for 2 yards of the first as for 3 of the second, and received as much for 4 of the first as for 5 of the second. Required how many yards of each kind?
42. A hare, 30 of her leaps before a greyhound, takes 4 leaps for the greyhound's 3, but 2 of the greyhound's leaps are equal to 3 of the hare's. How many leaps must each take before the greyhound catch the hare?
43. A son asked his father how old he was, who answered, when 5 years ago I was thrice as old as you, but, if we live 5 years longer, I shall be only twice as old as you. Required their ages?
44. Two persons discoursing of their money, says A to B, if you will give me 2s. 20, I shall have as much as you; but B replies, if you

We have to remember when this was published! In this book are complete systems of trade, book keeping (double entry is the Italian approach used by Genovese bankers), probability (for annuities, pensions, forecasting), and, of course, solutions of simultaneous equations. Our task remains to build on this foundation. Our first stop will be Make-A-Pie's inventory.

## 5.2 Making dough

There are two outputs from the production of Make-A-Pie pies, fruit and savory pies. Our clients Toriere and Fazi want to be sure they can meet changes in demand, hold down storage and replenishment costs, save on space, and not invest too much in inventory. Quite an ask! Here is the set up.

1. **Decisions.** Managers need to know how many fruit and savory pies to hold on hand to buffer changes in demand per week. If demand booms and production did not make enough pies, then inventory smooths the spike. If demand flags and there is too much inventory on hand, scarce short-term investment, and its financing, erodes costs and threatens profitability. With profitability threatened, plow



back of earnings into the business for improvements and corrective maintenance is impaired. Is it possible to have negative and positive levels of ordering decisions? Why not, if it is possible to accept and not accept pies into inventory.

2. **Criteria.** Executives (Tortiere and Fazi) determine that the appropriate incentive for materials managers is the minimization of cost, without sacrificing quality of course, and numerous other constraints we will soon mention. There are several costs here. They usually are part of the **cost of carry**.
  - **Production cost**, is assessed per unit here simply the number of pies, both fruit and savory. Working backward into the production process there would also be dough is in pounds and fillings in gallons. Make-A-Pie learned long ago to separate materials production from pie production. In fact, they own the copyrights and patents to the production of their proprietary dough and pie fillings. They then designate contract manufacturers to produce the required ingredients. Of course, Make-A-Pie audits quality daily sometimes. For our purposes, and to answer Fazi's question, we will focus on the inventory of pies.
  - **Storage cost**, also called *holding cost*, is also in fruit and savory pie units both per unit of time, a week in this case. What's involved in storage? Space in square footage, walk-in refrigerators, shelving, temperature and humidity control, daily maintenance and cleaning, among others, all contribute to this cost. We also should include insurance and factoring costs, both financial. There is labor too, administrative tracking, maintenance of storage facilities, cleaning, inspection.
  - **Replenishment cost** occurs with any order whatever the size and is fixed per order. When inventory gets down to a **safety stock** level a **\*\*reorder point\*** triggers production of more pies. Demand for pies drives this cost. The demand rate is the ratio of demand to the order size, also known as the lot size.
3. **Constraints** are legion! At first we will not use any explicit constraints to get the classic **economic order quantity (EOQ)** model. But they will exist. One certainly is the inter-holding period constraint where we have beginning inventory, add to the inventory and use the inventory during a holding period, and net out the level of ending of period inventory. Other constraints will arise from receiving (load dock comes to mind) and storage capacities including walk-in refrigerators and freezers, shelf-space, facility availability. Still others might be an exogenously required level of **safety stock** as insurance against forced outages of supplies.

As we should always, we begin with a simple, but instructive, and even useful model for a decision maker.

### 5.2.1 Model me an EOQ

First we draw a picture. Here is an influence diagram to help us model inventory order decisions. We might try this visual art site to make an influence diagram. Barring that here is a hand-drawn diagram on a Microsoft white board.

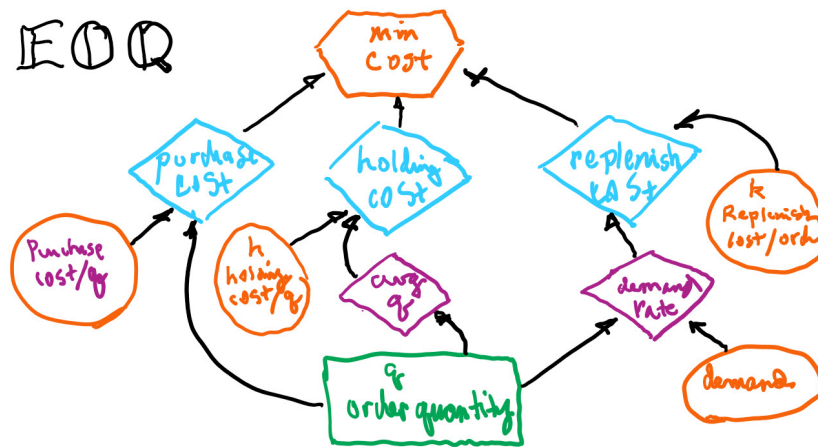


FIGURE 5.2: EOQ on the board.

Aside from the scrawl all roads (arrows, that is) lead to the minimize total cost criterion for the manager. The three costs are at the next level, again, all pointing to the criterion hexagon. In this first exercise we will focus only on the pure inventory costs of storage (holding) and replenishment.

- The decision is the amount of the order in the green box,  $q_i$  for  $i = 1, 2$  where fruit is 1, savory is 2. If  $q_1 = 200$  this means that the correct number of extra fruit pies to be made per week is 200 fruit pies. This is called the **order quantity**.
- Holding cost per unit  $h_i$  times the average order  $q_i/2$  across a holding period of a week feeds directly into the total cost criterion. If the total per pie holding cost of fruit pies is \$50/pie, then the fruit pie holding cost is  $h_i q_i = (0.60)(200) = \$100$ ,
- We also have two demands driving the need for inventory,  $d_i$  for which when combined will eventually build a product mix of the two pie lines. The number of orders is the demand rate  $d_i/q_i$ . If fruit pie demand  $d_1 = 11000$  and fruit pie order level  $q_1 = 200$  then the number of orders is  $d_i/q_i =$

$10000/200 = 50$  replenishments. If replenishment cost is fixed at We multiply the orders by the replenish cost per order  $k_i$ . This creates the replenishment cost which feeds the total cost criterion.

- We leave purchase cost out of this initial analysis and give it over to the production for Make-A-Pie. We will come back to this portion of the analysis since we will eventually need to value this inventory.

The total cost function so far is this expression for each  $q_i$ . Since the cost  $C_i$  for each pie simply adds together across the two pie products we really only need to look at one of the products at a time.

$$C_i = h_i \left( \frac{q_i}{2} \right) + k_i \left( \frac{d_i}{q_i} \right) \quad (5.1)$$

We do not have any tradeoff between inventories of the two products in this version of the model. What determines the approximate location of a minimum cost is in the replenishment, or reordering of lot sizes, the second term. Total cost is the sum of all of the product inventory costs  $C_i$ .

We can find the marginal holding and replenishment costs of  $q_i$  with this optimization. We drop the subscript to ease our notional anxieties, just for a moment. We first find the first derivatives of  $C$  with respect to  $q$ . This gives us the marginal impact of a change in  $q$  on  $C$ . We want to find that level of  $q$  such that the impact of a small change in  $q$  on  $C$ . At this level of  $q$ , in this case,  $C$  will be at its very desirable lowest possible level.

$$C = h \left( \frac{q}{2} \right) + k \left( \frac{d}{q} \right) \quad (5.2)$$

$$\frac{dC}{dq} = \frac{h}{2} - k \left( \frac{d}{q^2} \right) = 0 \quad (5.3)$$

$$\frac{h}{2} = k \left( \frac{d}{q^2} \right) \quad (5.4)$$

$$q^2 = \frac{2kd}{h} \quad (5.5)$$

$$q = \sqrt{\frac{2kd}{h}} \quad (5.6)$$

This is the **economic order quantity (EOQ)** fabled since Ford W. Harris (1913) asked how many parts should be made at one time, the economic lot size. Lot size, also known as the order size, increases with replenishment cost and demand and decreases with holding cost.

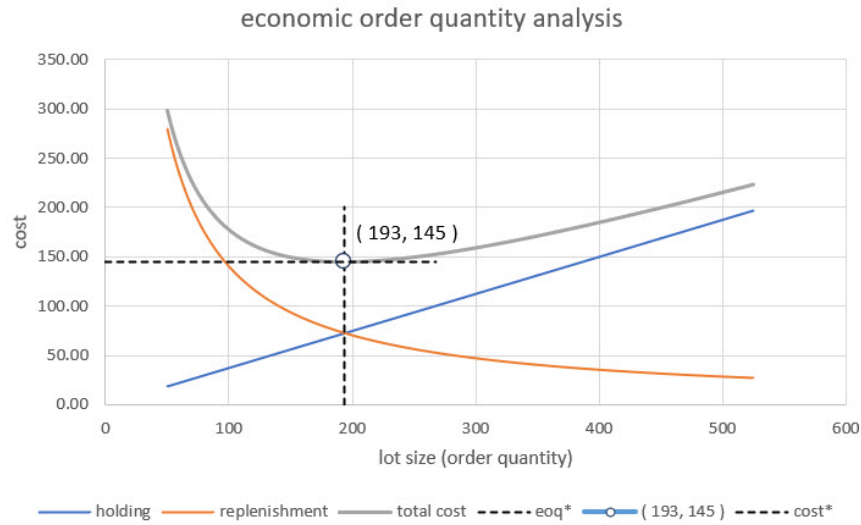
Let's take a look at a plot of holding, replenishment and total cost for  $h = 0.75$ ,  $k = 2.00$  and  $D = 7000$ . The economic order quantity is calculated here.

$$q^* = \sqrt{\frac{2kd}{h}} \quad (5.7)$$

$$= \sqrt{\frac{2(2)(7000)}{0.75}} \quad (5.8)$$

$$= 193 \quad (5.9)$$

And this is with rounding. When we substitute  $q^* = 193$  into the total cost expression we have the minimized total cost of a lot size  $C^* = 145$ .



**FIGURE 5.3:** EOQ neo-classical analysis.

Holding cost rises with order quantity in a nice straightline fashion from 0 so that the marginal holding cost is just  $h/2 = 0.375$  per unit. Replenishment cost is a hyperbola whose slope  $-2/q^2$  is the marginal replenishment cost per unit, a rapidly decreasing amount as  $q$  approaches the holding cost line, thereafter declining at a declining rate. Total cost is not at all linear in  $q$ . The non-linearity derives solely from the need to incorporate a demand per order cost to calculate replenishment.

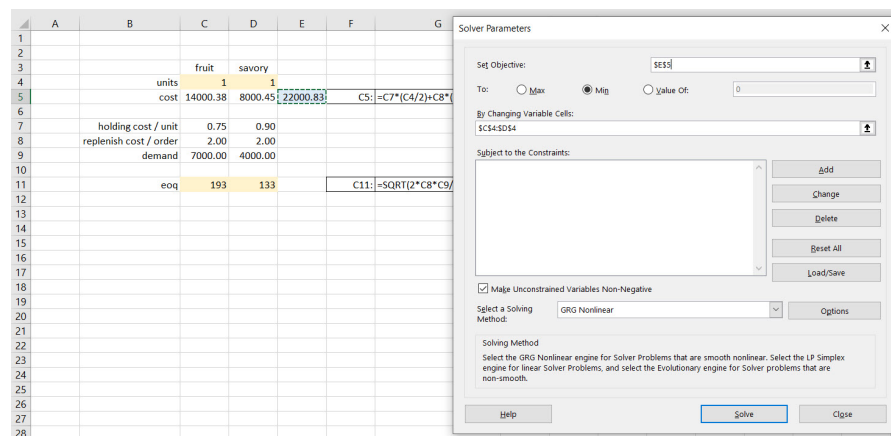
The optimal economic order quantity is the level of lot size where any further decrease in total cost equals any incremental or further increase in total cost, the bottom of the total cost curve. We should note that this optimal quantity occurs exactly where replenishment cost equals holding cost. EOQ implicitly takes into account the timing of reordering, the cost incurred to place and replenish an order, and the cost to store and hold product. If we constantly place small orders to maintain a specific inventory level, the ordering costs are

higher, and there is a need for additional storage space. All of this sense was discovered by Ford and his colleagues.

The EOQ formula assumes that consumer demand is constant. The calculation also assumes that both replenishment and holding costs remain constant. This fact makes it difficult or impossible for the formula to account for business events such as changing consumer demand, seasonal changes in inventory costs, lost sales revenue due to inventory shortages, or purchase discounts a company like Make-A-Pie might realize for buying or making inventory in larger quantities.

### 5.2.2 Implementing the simple EOQ

Here is a Solver setup for the raw EOQ formula we just derived. All of the data we need is built into this setup.



**FIGURE 5.4:** EOQ a la Solver

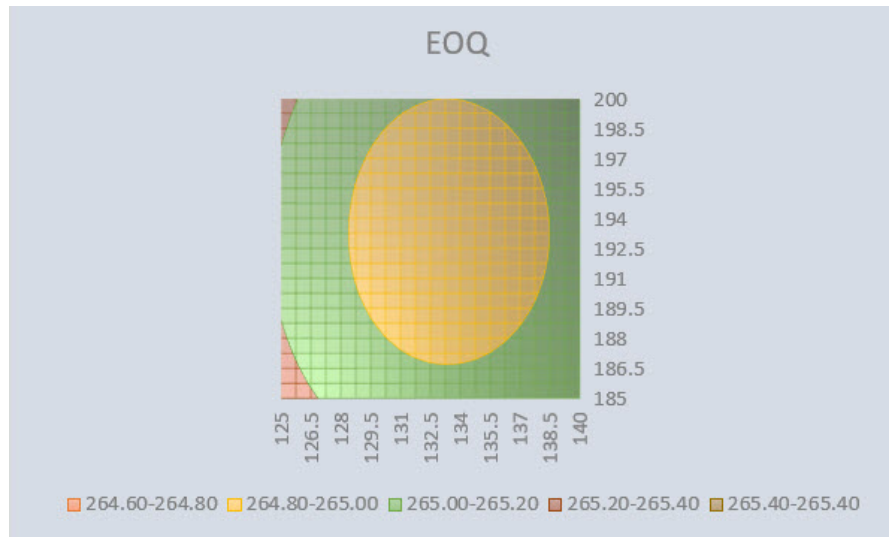
We set total cost to be minimized by choosing the two decision cells. Since this is a non-linear optimization problem, Solver uses a gradient reduction approach. This method effectively uses a smart approximation iteration to get closer and closer to an optimal solution. We need to take care here as there might be other optimal solutions. There are no constraints, yet. We press OK and solve.

Shown in row 11 is the EOQ formula we derived. Close enough for our purposes. We should, each week, be sure there is this amount of inventory available to manage these weekly demands, given this inventory cost structure.

The costs of fruit and savory pie inventories can be visualized with this 2-dimensional grid of contours.

We can locate the optimal lot sizes of fruit and savory pies somewhere in the

	A	B	C	D	E	F	G
1							
2							
3			fruit	savory			
4		units	194	133			
5		cost	144.91	120.00	264.91	C5: =C7*(C4/2)+C8*(C9/C4)	
6							
7		holding cost / item	0.75	0.90			
8		replenish cost / order	2.00	2.00			
9		demand	7000.00	4000.00			
10							
11		eq	193	133		C11: =SQRT(2*C8*C9/C7)	
12							
13							

**FIGURE 5.5:** EQO Solver setup.**FIGURE 5.6:** EOQ cost map.

middle of the yellow ochre blob. A brute force grid like this can be computationally and visually useful for a 2 product inventory. More products than that would be difficult if not impossible to visualize, even with our expansive imaginations.

### 5.2.3 Life constrains our simple model

What happens when we have only so much cubic footage of storage space. Each pie package is about 3"x12"x12". This translates into 0.25 cubic feet. We might only have so much, so we create a constraint like this.

$$a_1 q_1 + a_2 q_2 \leq S$$

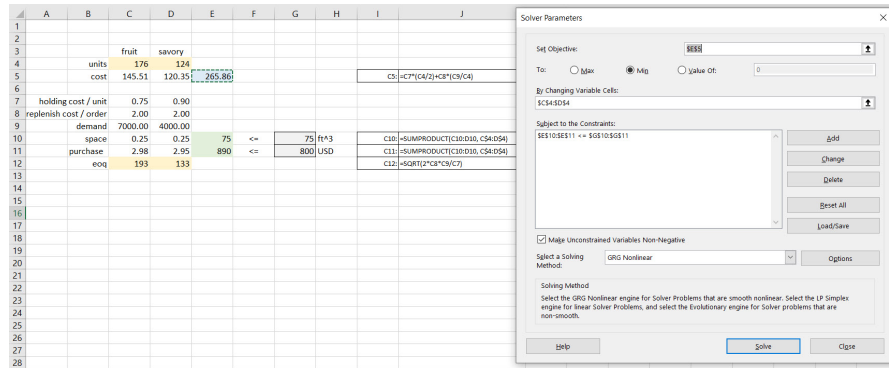
where the coefficients on the  $q$ s are the space occupied per pie and  $S$  is the available space.

Another wrinkle in the time-space continuum is financing of the inventory investment. We might suppose that Tortiere and Fazi set aside \$800 per week to cover short-term, recurring inventory investment. How they come up with that amount is the subject of a financial management course. But here is where the purchase cost enters the inventory. To value the inventory, the inventory management effectively takes in inventory from production management at least at the variable cost per pie. This first-in-first-out FIFO valuation can be assailed on many accounts, again best left for a management accounting course. Here is the constraint.

$$b_1 q_1 + b_2 q_2 \leq I$$

where the coefficients on the  $q$ s are the variable production costs per pie and  $I$  is the desired limit on inventory investment.

In all its glory, here is the Solver setup for the revised EOQ model.



**FIGURE 5.7:** Bounding the EOQ solution.

Again we press OK, but also this time in the next dialogue request the Sensitivity Report. We get this solution.

	A	B	C	D	E	F	G	H	I	J	K
1											
2											
3			fruit	savory							
4		units	158	112							
5		cost	147.94	121.82	269.76					C5: =C7*(C4/2)+C8*(C9/C4)	
6											
7		holding cost / item	0.75	0.90							
8		replenish cost / order	2.00	2.00							
9		demand	7000.00	4000.00							
10		space	0.25	0.25	67	<=	75	ft <sup>3</sup>		C10: =SUMPRODUCT(C10:D10, C\$4:D\$4)	
11		purchase	2.98	2.95	800	<=	800	USD		C11: =SUMPRODUCT(C10:D10, C\$4:D\$4)	
12		eq	193	133						C12: =SQRT(2*C8*C9/C7)	
13											
14											
15											
16											

The constraints make the lot sizes smaller indeed. We should by now expect this. We notice that we have some space to spare, also known as **slack**. But we do use up our allocated investment funds for inventory.

Remember that sensitivity report? Here it is.

	A	B	C	D	E
1	<b>Microsoft Excel 16.0 Sensitivity Report</b>				
2	<b>Worksheet: [week4-inventory-eoq.xlsx]Sheet4</b>				
3	<b>Report Created: 4/11/2021 2:17:11 PM</b>				
4					
5					
6	<b>Variable Cells</b>				
7				<b>Final</b>	<b>Reduced</b>
8		<b>Cell</b>	<b>Name</b>	<b>Value</b>	<b>Gradient</b>
9		\$C\$4	units fruit	157.5363773	0
10		\$D\$4	units savory	112.047998	0
11					
12	<b>Constraints</b>				
13				<b>Final</b>	<b>Lagrange</b>
14		<b>Cell</b>	<b>Name</b>	<b>Value</b>	<b>Multiplier</b>
15		\$E\$10	space	67.39609381	0
16		\$E\$11	purchase	799.9999982	-0.063460456
17					
18					

**FIGURE 5.8:** Marginal values of each constraint.

There is this new idea called the **Lagrange Multiplier** also known as the **shadow price**. It is zero for the space constraint since we do not use all of the space, there is some **slack**. On the other hand we use up all of the



inventory investment funding. Whenever the right-hand side of a constraint is completely used, we can expect a Lagrange Multiplier to show up.

Here is a little bit more than we have bargained for. We formulate a one constraint EOQ model by minimizing the cost function with a further cost (possibly benefit) to avoid exceeding the investment constraint only. We ignore the space constraint, since we already know we have enough, and more, space to go around.

Since we only want to focus on the simplest version of this model for illustration, we train our sights on just one of the  $qs$ , again. Even so, we have introduced a new variable into our midst,  $\lambda$ . We take partial derivatives of cost first with respect to  $q$ , for any level of  $\lambda$ , then with respect to  $\lambda$ , for any level of  $q$ . We end up with two simultaneous equations in two unknowns,  $q$  and  $\lambda$  when we set the partial derivatives to zero. There will be an important insight waiting for us at the end of this tunnel.

$$C = h \left( \frac{q}{2} \right) + k \left( \frac{d}{q} \right) + \lambda(I - bq) \quad (5.10)$$

$$\frac{\partial C}{\partial q} = \frac{h}{2} - k \left( \frac{d}{q^2} \right) - \lambda b = 0 \quad (5.11)$$

$$\frac{\partial C}{\partial \lambda} = I - bq = 0 \quad (5.12)$$

After some serious, and sometimes tedious, algebraic manipulation courtesy of al-Khwarizmi, we arrive at this solution.

$$q^* = \sqrt{\frac{2bkdI^2}{(b-h)I^2 + 2b^2d}} \quad (5.13)$$

$$\lambda^* = \frac{h}{2b} - \frac{bkd}{I^2} \quad (5.14)$$

Yes, these are as beastly as any other equations we might have, or hope never to have, seen. We see that investment  $I$  features prominently, as does demand. These large amounts are honed down by the marginal holding and replenishment costs. An increase in the spread between variable cost  $b$  and holding rate  $h$  will decrease the amount of the lot size needed to manage inventory.

It is hard to say whether an increase in demand  $d$  in the presence of this constraint will or will not increase lot sizes. It certainly decreases the marginal impact of an increase in inventory investment  $I$  on cost,  $\lambda$ . An increase in marginal holding costs also unambiguously increases the impact of investment on total cost. For our joint space and investment constraints, an increase of investment of \$1 decreases costs by \$0.06. This seems true unambiguously as

well.<sup>1</sup> This shadow price of inventory investment also looks like an interest rate (it is!). In fact this number is behind the use of inventory to collateralize short-term borrowing and lending operations.

$$\frac{d\lambda^*}{dI} = \frac{bkd}{I^3}$$

So that the rise over run of inventory impact on cost will be always negative in this model as it is inversely related to changes in  $\lambda^*$ .

We should plug some numbers into these formula. That is a great exercise for the less faint of algebraic heart among us. Toy models like this at least give us some insight into what might or might not impact total inventory cost. More constraints, more products in inventory, seemingly add to the complexity of the model, but these general relationships will continue to hold mostly.

---

### 5.3 Back to the future

Tortiere and Fazi still need to plan for the future. They have figured out with us a way to express product mix, pricing, and inventory as decisions chosen to optimize relevant objective functions subject to various constraints. They still need guidance on longer term planning. They ask now for a multi-period production planning model.

#### 5.3.1 Bottom line up front

We are in luck! We have done this before. Most of the elements of such a model are present in previous models. Here is a template we can work from.

Which layout should we choose?

Here is a source-sink network representation of the model from a previous engagement.

$P$  is production,  $B$  and  $I$  are inventory,  $D$  is demand, the sink. We will definitely need to clean this up. We will use it to scope out the data we need and the client's requirements at the same time.

In a network like this a common, if not also neutral, node is the source. Its links point to the 6 production periods with ranges of potential production at a production cost, or it could be profit. Networks represent flows from and to common and disparate nodes. Each edge is an amount or condition flowing. The arrows indicate the direction of flow and thus by convention the sign of

---

<sup>1</sup>If we are bold, we might calculate the rate of change of cost minimizing  $\lambda$  with respect to inventory investment  $I$ . That operation gives us

A							B	C	D	E	F	G	H	I	J	K	L	
1	TEMPLATE PRODUCTION PLAN																	
2		Period		1	2	3	4	5	6									
3		Beg Inv		2750	5750	4750	2750	1500	1500									
4		Production P(t)		4000	3500	4000	4250	3500	4000									
5		Demand		1000	4500	6000	5500	3500	4000									
6		Ending Inv I = Beg Inv + Prod - Dem		5750	4750	2750	1500	1500	1500	>=								
7										<=								
8		Min Prod(% of max)	50.00%	2000	1750	2000	2250	1750	2000									
9		Max Production		4000	3500	4000	4500	3500	4000									
10		Avg Inventory		4250	5250	3750	2125	1500	1500									
11																		
12		Holding cost/unit	1.50%	\$3.60	\$3.75	\$3.98	\$4.28	\$4.20	\$3.90									
13		Prod Cost/unit		\$240	\$250	\$265	\$285	\$280	\$260									
14																		
15																		
16																		
17																		
18																		
19		Period		Beg Inv	Production	Demand	End Inv I	Avg Inv	Min Prod	Max Prod	Cost/u	h %						
20		1		2750	4000	1000	5750	4250	2000	4000	\$240	\$3.60						
21		2		5750	3500	4500	4750	5250	1750	3500	\$250	\$3.75						
22		3		4750	4000	6000	2750	3750	2000	4000	\$265	\$3.98						
23		4		2750	4250	5500	1500	2125	2250	4500	\$285	\$4.28						
24		5		1500	3500	3500	1500	1500	1750	3500	\$280	\$4.20						
25		6		1500	4000	4000	1500	1500	2000	4000	\$260	\$3.90						
26																		
27																		
28																		
29																		
30																		
31																		
32																		
33																		
34																		
35																		
36																		
37																		
38																		
39																		
40																		
41																		
42																		
43																		
44																		
45																		
46																		
47																		
48																		
49																		
50																		
51																		
52																		
53																		
54																		
55																		
56																		
57																		
58																		
59																		
60																		
61																		
62																		
63																		
64																		
65																		
66																		
67																		
68																		
69																		
70																		
71																		
72																		
73																		
74																		
75																		
76																		
77																		
78																		
79																		
80																		
81																		
82																		
83																		
84																		
85																		
86																		
87																		
88																		
89																		
90																		
91																		
92																		
93																		
94																		
95																		
96																		
97																		
98																		
99																		
100																		

FIGURE 5.9: Multiperiod production template.

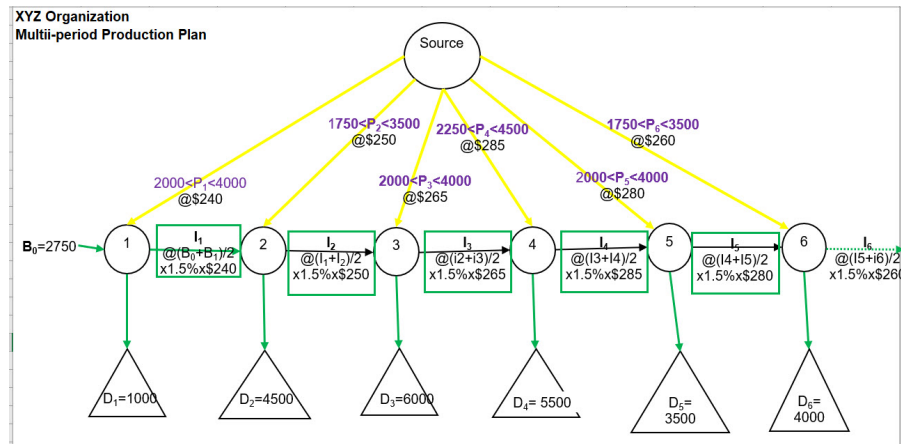


FIGURE 5.10: Where the template came from.

an amount of flow. Graphical representations like these help us understand complex relationships like feed forward and feed back flows. We know these relationships in algebra as simultaneous equations.

At period 1 initial production along with initial inventory feed demand. What is left over from demand leaves period 1 as ending inventory at a holding cost becoming the beginning inventory into period 2. Potential production enters period 2 with beginning inventory to meet outgoing period 2 demand. Again what is left over is inventory and a holding cost to the next period node and so on to the final inventory leaving the period 6 node. The algebra of the node

is arrows into a node will sign positive while arrows from a node will sign negative.

### 5.3.2 Tighter, and tighter

Here is a tighter formulation that may suit us. We forecast that customers demand  $D_1, \dots, D_T$  be the demands for the next  $T$  periods. For our purposes  $T = 6$  weeks. When we get to simulation we will assume that the  $D_t$ 's are independent random variables, and that all stockouts are backordered. We will let  $c_t$  denote the unit purchase cost in period  $t$ , and let  $x_t$  denote the inventory level at the beginning of period  $t$ , where a positive  $x_t$  means that  $x_t$  units of inventory are carried from the previous period, and a negative  $x_t$  is a backlog of  $-x_t$  units is carried from the previous period.

We can produce, or contract with a product to procure,  $y_t$ . With these conventions then  $y_t - x_t \geq 0$  is the size of the order, the lot size in in period  $t$  with procurement cost  $c_t(y_t - x_t)$  and an increase of the inventory level to  $y_t$ . Since customers demand  $D_t$  units during the period, the inventory level at the beginning of period  $t + 1$  is then this expression.

$$x_{t+1} = y_t - D_t.$$

If  $y_t = y$  the loss function in period  $t$  is given by

$$G_t(y) = h_t E(y - D_t)_+ + k_t E(D_t - y)_+$$

We use subscripted  $+$  strokes to avoid writing this expression repeatedly.

$$(b - a)_+ = \max(0, b - a)$$

If we take expectations as perfectly fulfilled, then our model has no so-called randomness, becomes nicely deterministic, and perhaps a little rosier, for now.

Let's bring money to the table. Economic losses include the costs of holding and setting up, replenishing, back ordering, whatever vocabulary we use, of inventory. This situation is just like the myopic, period by period, static approach we used for the economic order quantity of Harris (1913). We have finally introduced here  $h_t$  as the holding cost per unit of any positive levels of inventory on hand and  $k_t$  as the replenishment, or some might say, the back order cost. In contracts  $k_t$  could be in a penalty or breakage clause as well. We attempt to find the policy, that is, the time path, for managing inventory from time  $t = 1$  to time  $t = T$  in our formulation. If we extinguish inventory, or at least factor it out and sell it at time  $t = T + 1$ , then we would in that time produce more or reimburse customers for funding any backlogs.

We now formulate our multi-period production plan.

**Decisions** include the path of production, or contract production known as

procurement and even tolling,  $y_t$  from time  $t = 1$  to the finite horizon  $t = T$ , which in our problem is 6 weeks.

**Criterion** is just to minimize costs. We must set out our criterion equation. It gets a bit messy, but says it all. For the moment we will inventory management's life on the job be very rosy indeed by the gift of  $k_t = 0$ . We will see that this artifice will straighten any quadratic lines.

$$\min_{y_t \geq x_t} \sum_{t=1}^{t=T} \left[ c_t y_t + h_t \left( \frac{x_t + x_{t-1}}{2} \right) \right]$$

**Constraints** (and oops we already have at least one buried in the criterion) include any upper and lower bounds on production, inventory (for space and funding reasons at the least), as well as the relationship between current and past inventory levels.

$$y_{min} \leq y_t \leq y_{max} \quad (C1) \quad (5.15)$$

$$x_{min} \leq x_t \leq x_{max} \quad (C2) \quad (5.16)$$

$$x_t = x_{t-1} + y_{t-1} - D_{t-1} \quad (C3) \quad (5.17)$$

$$x_0 = x \quad (C4) \quad (5.18)$$

The first two constraints, C1 and C2, just set the stage for bounding a decision in practical life. We have only so much production capacity and so much space to store inventory. These are easy ways to express much more difficult to formulate ideas, including space and capacity decisions.

The third constraint C3 is the kicker. It allows decisions in one week to affect a decision in another week. This moves us into the much more complex realm of far-sighted rather than the myopia of the static EOQ.

The fourth constraint C4 means we have to seed the process of decision making with a beginning inventory, an initial value. If this sounds like differential equations, it is, except we use the discretization of derivatives known as differences. We will also in our implementation hold the holding cost  $h_t$  and the variable cost of production  $c_t$  constant.

### 5.3.3 Implementing the model

We use the template, repurpose it, and input the data from reports which CFO Fazi provide. Here is a summary of her report.

We have 6 weeks in the planning cycle. A large beginning inventory seems to consume funding, space, and the attention of management disproportionately to production. Both Tortiere and Fazi realize this and want us to help pare this level down. The problem is how low? They contend that the fragility of their products requires special handling, storage, and then there is shrinkage

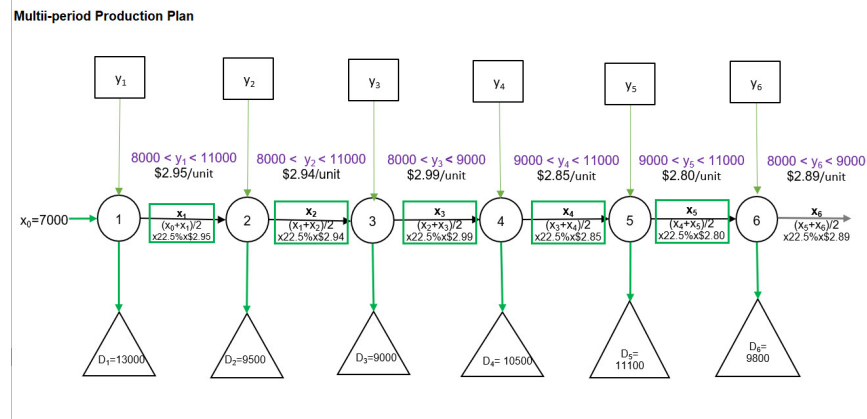
	A	B	C	D	E	F	G	H	I
1									
2			Period	1	2	3	4	5	6
3			beginning Inventory $x(0)$	7000					
4			demand $D(t)$	13000	12500	9500	10500	11100	9800
5			min production	8000	8000	8000	9000	9000	8000
6			max production	11000	11000	9000	11000	11000	9000
7			variable production cost/unit	\$2.95	\$2.94	\$2.99	\$2.85	\$2.80	\$2.89
8			holding cost/unit (%variable)	22.50%					
9									
10									

**FIGURE 5.11:** Some data is in order.

of product in inventory. So the only way to contend with high storage costs per unit is to constrict the volume, while still meeting demand and within the constraints imposed by production. In week 3 they forecast a partial shut-down of production facilities to refit and upgrade equipment.

### 5.3.4 A revised rendering

In conversations with Tortiere, Fazi and their colleagues we draw this map of the model.



**FIGURE 5.12:** A network view.

Sources are from production and appear in squares. Demands are sinks in triangles. They flow into and out from period nodes in circles. Each period node links to past and future adjacent nodes. They follow the algebraic model. We notice from the discussion to use the average inventory across a period to assess holding period cost. Fazi wants, for the time being, to prescind from replenishment lot sizes.

We use the report and the flow network to develop this worksheet model. Here is the Excel Solver setup.

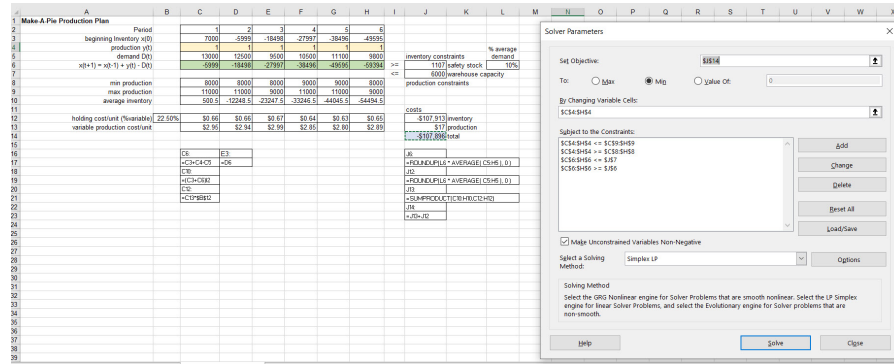


FIGURE 5.13: Setting up Solver version.

We also notice a late request to require a  $b = 10\%$  of demand per period safety stock. This is the constraint. We use a  $\geq$  relationship since a requirement is a matter of have at least a certain level of inventory on hand.

$$x_t \geq bD_t$$

The request for this constraint may be result of Fazi's recent conversations with Make-A-Pie's insurance broker about the coverage for the high level of inventory investment. If Make-A-Pie internally reserves some production, the insurance policy might be able to reduce, or at least not increase, the deductible for inventory losses. In this way safety stock can act as self-insurance.

Now for the final model, and this only the beginning of our analysis.

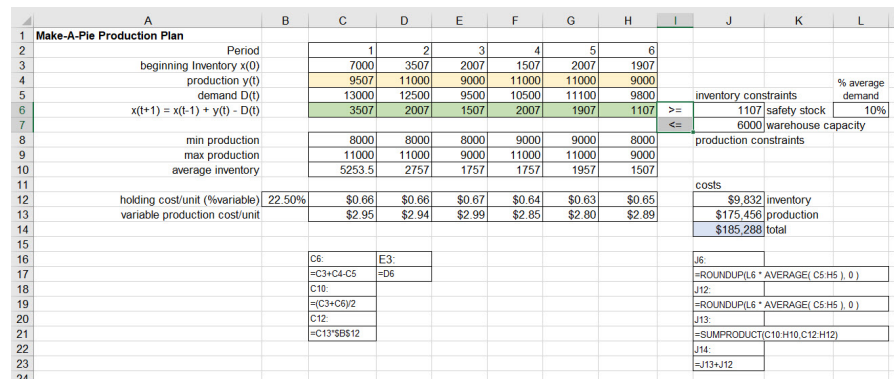


FIGURE 5.14: A bang-bang solution presents itself.

We do not violate any constraints. We have reduced inventory and inventory

cost. We have a multiperiod production plan whose path depends on past and future demand. We also notice we have almost double the warehouse space needed to fulfill demand, production, and inventory requirements.

---

#### 5.4 Next steps?

Of course we will find the breaking points of this model. We can render a sensitivity analysis report, attempt to understand the range of values production, and by deduction inventory, policy can evolve. What would happen if Make-A-Pie expands to all five boroughs of NYC and Westchester and Long Island? What would happen if Make-A-Pie grows new lines of business? How would we model these questionss



# 6

---

## *Case: Pricing Production at Make-A-Pie*

---

### 6.1 Yet another ask!

Tortiere and Fazi now want us to do the seemingly impossible? Price pie. Our job is to insert a pricing decision into the multi-period production planning model. We will use the existing demand framework from the previous chapter. We could alternatively set up a similar Salmeron Solar model? Try it for practice of course! We will try to sow up our discussions on production, demand, inventory, in short a good hunk of the Make-A-Pie supply chain. That's is and that's certainly enough! But, you know, the Salmeron multiperiod planning model might be an interesting final project. Hmmm. We will see.

### 6.2 A process unfolds

We will soon hop into another problem that Tortiere faces. We will need to use simulation to help her solve it. For now let's try to

1. Watch this walkthrough video. You can access this video [here](#)
2. Build an influence diagram with production , inventory, demand, and now pricing. This could follow the approach for the production process network.
3. Start from the Make-a-Pie cost minimization multi-period production model. Adding more weekly periods to the model.
4. Use a Make-A-Pie demand worksheet as the base for price based demand.
5. Make a new pricing decision row. Now be sure all of the cell calculations flow through the model cleanly, especially for price drive demand. Remember we solve for price.
6. Check Solver's settings to be use we have the right objective cell,

decisions including price, just add price cells next to the production sequence with a comma

7. Run GRG Non-linear as cost is now quadratic in prices, just like the NPV and weekly profit analysis of weeks 1 and 2.
8. Remember this is cost minimization. You should get a not too surprising result: the model does the absolute minimum production and finds a price to support that policy.
9. Override the cost minimization assumption by creating a revenue and then a profit cell. Maximize profits. We might see different price and production possibilities? Or not?
10. Always perform sensitivity analysis, especially with respect to the right hand-side constraint values. For GRG-non-linear you will be examining the Lagrange multiplier values interpreted as shadow prices.

---

## ***Part 3 – Simulation***

---

Getting our arms around

- Developing a waiting time archetypal model to answer specific business questions
- Simulating correlated variates
- Applying variates to the waiting time model
- Collecting data relevant to the decision framework inherent to the simulation model
- Exploring simulated data
- Analyzing sensitivity of results with respect to changes in in key decision parameters



# 7

---

## *Probability for Real People*

---

### 7.1 Can we rationally reason?

Some might wonder if this section header even makes sense! CFO Fazi looks at it and knows everyone has a reason, some good and some not so good. But sometimes our reasons are not founded in any data that can be observed either by ourselves or by others. Our reasoning can be founded on falsified data, delusions, unfounded opinions, beliefs with no authoritative grounds. Her questions move from the various relationships in production planning in the previous part of this project to the forecasting of what might happen at all. What would be the rationale for a forecast? She begins her thought experiment with the weather in her hometown of Albany, New York.

Rationality in the context we have been discussing at least means that we, as decision makers, would tend to act based on the consistency of observed reality with imagined and through ideas about the world in which data are collected. We attempt to infer claims about the world based on our beliefs about the world. When confronting ideas, imbued with beliefs, with observed reality we might find ourselves in the position to *update our beliefs*, even those, and sometimes especially those, we so dearly hold.

In our thinking about anything we would venture candidate hypotheses  $h$  about the world, say the world of meeting demand, providing services, marshalling resources in specific markets. Of course the whole point is that we do not know which hypothesis  $h$  is more plausible, or not. We then collect some data  $d$ . When we perform this task, we move from the mental realm of the possibility of hypotheses, theories, surmises, and model to the realm of observed reality. We may well have to revise our original beliefs about the data.

To implement our maintained hypothesis of rationality, we begin our search for **potential consistencies of the collected data with our hypotheses** that are fed by the data. In our quest we might find that some one of the hypotheses has more ways of being consistent with the data than others. When the data is consistent with a hypothesis, that is, when the hypothesis is reasonable

logically, then our belief in that hypothesis strengthens,<sup>1</sup> and becomes more plausible. If the data is less consistent with the hypothesis, our belief in that hypothesis weakens. So far we have performed this set of tasks with conjectures about virus testing and voter alliance in zip codes. Let's switch up our program and consider the following very simplified question about the weather.

CFO Fazi's ask is this question about winter consumer behavior in the Albany-area market:

---

*We see people carrying snow shovels. Will it snow?*

---

What is the data  $d$ ? We have recorded a simple observation about the state of the weather so that single piece of data ( $d =$  We see people carrying snow shovels). Here is where our beliefs enter. We have two *hypotheses*,  $h$ : either it snows today or it does not.

Let's figure out how to solve this problem? We have three *desiderata*:

1. We should include our experiences with snow in our analysis.
2. We should collect data about carrying snow shovels in January as well.
3. We prefer more consistency of data with hypotheses to less consistency.

Here we go, let's strap ourselves in.

### 7.1.1 Priors: what we think might happen

Our observation is about the weather: clouds, wind, cold. But we want to know about the snow! That is our objective and we have definite ideas about whether (don't pardon the pun!) it will snow or not. We will identify our beliefs, ever before we make our observations, about snow. The analytical profession and custom is to label these beliefs as *a priori*,<sup>2</sup> and thus the ellipsis *prior*,

---

<sup>1</sup>The core idea of *strengthen* is to take us from a more vulnerable to a less vulnerable place or state. Synonyms for strength include *confirm* and *validate*.

<sup>2</sup>The *a priori* elements of any argument include just about everything you and I know, including the kitchen sink! We can't help but to have these antecedent thoughts, experiences, shared and not-so-shared histories. They tend to persist in most humans, including us. At least that is what we will maintain. Thus it is a *necessity* to include these beliefs in our discussion. Without their consideration we most plausibly will introduce unsaid and denied

contentions we hold when we walk into the data story we create with the question of *will it snow?*

Our prior contentions are just the plausibilities of each hypothesis whatever data we eventually collect. After all we have to admit to everyone what we believe to be true as the antecedent to the consequent of observations and the plausibility of snow. This move allows us to learn, to revise, to update our dearly held beliefs. We thus can grow and develop. This is in a phrase a *sine qua non*, a *categorical imperative*, a *virtually unconditioned* requirement for change.

What might we believe about whether it will snow (today)? If you come from Malone, New York, north of the Adirondack mountains, you will have a different belief than if you come from Daytona, Florida, on the matter of how many ways snow might happen in a given month. So let's take as our benchmark Albany, the capital of the state of New York.

We will refer to some data to form hypotheses and their plausibility using this weather statistics site. The site reports the average number of days of snowfall in January, when there is at least a 0.25 cm accumulation in a day. It is 10.3 days. These are the number of ways (days) in January, in Albany, NY, that it is true, on average and thus some notion of expected, or believed to be, that it snows. The total number of ways snow could possibly fall in any January (defined by calendar standards) is 31. While the formation of the hypotheses *snowy* and *nice* days is informed by data, we are asking a question about snow because we have yet to observe if will snow. We cannot observe something that has not yet happened. We can thus characterize hypotheses and conjectures as **unobserved data**.

Thus we might conclude that we believe that it is plausible (probable) that snow *can* fall  $10.3/31 = 30\%$  of the different ways snow can fall. Note very well we will talk about *priors* as *potentials* and *conjectures* and *hypotheticals*, and thus used the modal verbs *can* or *might*. Thus we believe it might not snow, because it is possible, with plausibility  $1 - 0.30 = 0.70$ , or, multiplying by 100, 70%, according to the law of total probability of all supposed (hypothesized) events. We only have two such events: *snow* and *not snow*. Probabilities must, by definition, add up to 1 and must, again by definition be a number between 0 and 1.

Nice ideas, nice beliefs, are ours as yet to be observed, but projected notions of a snowy day. But how real, how plausible, how rational, that is, how consistent are they with any *observed data*? Is there any *observed data* we can use to

---

bias, let blindspots have the same focus as clearly understood experiences, and produce time and resource consuming blind alleys. But we should hang on here: even blind alleys and blind spots are extremely important bits of knowledge that help us understand what does not work, an *inverse insight* as exposed by Bernard Lonergan (1970).

**TABLE 7.1:** Priors by hypotheses

hypotheses	priors
snowy day	0.3
nice day	0.7

**TABLE 7.2:** data meets hypotheses

hypotheses	shovels	hands
snow day	0.7	0.3
nice day	0.1	0.9

help us project which of our unobserved data, our hypotheses, is more or less reasonable?

### 7.1.2 Likelihoods: thinking about the data

Life in the Northeast United States in January much revolves around the number of snow days, also known as days off from school. A prediction of snow meets with overtime for snow plow drivers, school shut downs, kids at home when they normally are in school. On some snowy days we see people carrying snow shovels, on others we don't. On some nice days we see people with snow shovels, on others we don't. Confusing? Confounding? A bit.

Now we link our observations of shovels with our unobserved, but through about and hypothesized, prediction of snow. We then suppose we observe that people carry snow shovels about 7 of the 10 snowy days in January or about 70%. On nice days we observe that people carry shovels at most 2 days in the 21 nice days or about 10%.

This table records our thinking using data we observe in Januaries about weather conditions.

First of all these probabilities register yet another set of beliefs, this time about whether we see shovels or not, *given, conditioned by*, the truth of each hypothesis  $h$ . We write the conditional probability  $\Pr(d|h)$ , which you can read as “the probability of  $d$  given  $h$ ”. Also here we will follow the convention that this set of results of our assessment of the relationship of shovels to snowy days as a likelihood.<sup>3</sup>

<sup>3</sup>For Pierre Simon de Laplace et al. (1902), likelihood also has the idea of  $\Pr(h|d)$ . Let's stick to our knitting, and tolerance for ambiguity, with using the rows of this table as our entries for likelihood.



### 7.1.3 Altogether now

Do we have everything to fulfill our *desiderata*? Let's check where we are now.

1. We should include our experiences with snow in our analysis.

---

Yes! We put our best beliefs forward. We even (sometimes this is a courageous analytical step) quantified the ways in which snow and not snow would occur, we believe, in Albany NY in an average January.<sup>4</sup>

---

2. We should collect data about carrying snow shovels in January as well.

---

Yes we did! Again we elicited yet another opinion, belief, whatever we want to colloquially call it. That belief is what we register and document based on observation of shovels and just hands in the presence of snowy and nice days in a January.

---

3. We prefer more consistency of data with hypotheses to less consistency.

---

Not yet! We will impose our definition of rationality here.

---

Let's start out with one of the rules of probability theory. The rule in question is the one that talks about the probability that *two* things are true. In our example, we will calculate the probability that today is snowy (i.e., hypothesis *h* is true) *and* people carry shovels (i.e., data *d* is observed). The *joint*

---

<sup>4</sup>we really need to think further about our notions of an average or centrally located anything. This means more consideration later, including deviations from these locations measured by scale.

**probability** of the hypothesis and the data is written  $\Pr(d, h) = \Pr(d \wedge h)$ , and you can calculate it by multiplying the prior  $\Pr(h)$  by the likelihood  $\Pr(d|h)$ . The conjunction is a *both-and* statement. We express conjunctions using the wedge  $\wedge$  symbol. Logically, when the statement that both  $d$  and  $h$  is true, then the plausibility, now grown into probability is:

$$\Pr(d \wedge h) = \Pr(d|h) \Pr(h)$$

When we divide both sides by  $\Pr(h)$  we get the definition, some say derivation, of condition probability. If we count  $\#()$  the ways  $d \wedge h$  are true and the ways that  $h$  are true then

$$\#(d|h) = \frac{\#(d \wedge h)}{\#(h)}$$

Then the number of ways the data  $d$  are true, given  $h$  is true, equals the total number of ways that  $d$  and  $h$  per each way that  $h$  is true. We have thus normed our approach to understanding a conditional statement like if  $h$ , then  $d$ . Even more so, when we combine the law of conditional probability with the law of total probability we get Bayes Theorem. This allows us to recognize the dialectical principle that, yes, we recognize  $h = \text{snowy}$ , but we also know that every cloud has its silver lining and that there is a non-snowy day and thus a

$$\text{not } h = \neg h = \text{nice}$$

lurking in our analysis.

Here it in in all its glory.

$$\Pr(h | d) = \frac{\Pr(d | h) \Pr(h)}{\Pr(d | h) \Pr(h) + \Pr(d | \neg h) \Pr(\neg h)} \quad (7.1)$$

$$= \frac{\Pr(d \wedge h)}{\Pr(d | h) \Pr(h) + \Pr(d | \neg h) \Pr(\neg h)} \quad (7.2)$$

The numerator is the same as the conjunction both  $d$  and  $h$ . The denominator is the probability that either both  $d$  and  $h$  or both  $d$  and  $\neg h$  are true. While the build up to this point is both instructive, and thus may at first be *confusing*, it is useful as it will highlight the roles these probabilities perform in the drama that is our analysis.

We had better get back to the data or get lost in the weeds of the maths. So, what is the probability it is true that today is a snowy day *and* we observed people to bring a shovel?

Let's see what we already have. Our prior tells us that the probability of a snowy day in any January is about 30%. Thus  $\Pr(h) = 0.30$ . The probability that we observe people carrying shovels is true given it is a snowy day is

**TABLE 7.3:** Both data and hypotheses

hypotheses	shovels	hands	sum
snow day	0.21	0.09	0.3
nice day	0.07	0.63	0.7
sum	0.28	0.72	1.0

**TABLE 7.4:** both data and hypotheses in days in January

hypotheses	shovels	hands	sum
snowy day	6.51	2.79	9.3
nice day	2.17	19.53	21.7
sum	8.68	22.32	31.0

70%. So the probability that both of these things are true is calculated by multiplying the two to get 0.21. We can make this

$$l\Pr(\text{snowy}, \text{shovels}) = \Pr(\text{shovels} \mid \text{snowy}) \times \Pr(\text{snowy}) \quad (7.3)$$

$$= 0.70 \times 0.30 \quad (7.4)$$

$$= 0.21 \quad (7.5)$$

This is an interesting result, something odds makers intuitively know when punters put skin in the game. There will be a 21% chance of a snowy day when we see shovels in people's hands. However, there are of course *four* possible pairings of hypotheses and data that could happen. We then repeat this calculation for all four possibilities. We then have the following table.

Just to put this into perspective, we have for the 31 days in a January this table.

We have four logical possibilities for the interaction of observed data and unobserved hypotheses. We arrange these possibilities in two stacked rows. We recall that visualization is everything, even in tables! Here is the first row.

1. Snowy and shovels

$$l\Pr(\text{snowy}, \text{shovels}) = \Pr(\text{shovels} \mid \text{snowy}) \times \Pr(\text{snowy}) \quad (7.6)$$

$$= 0.70 \times 0.30 \quad (7.7)$$

$$= 0.21 \quad (7.8)$$

2. Snowy and just hands

$$l\Pr(\textit{snowy}, \textit{hands}) = \Pr(\textit{hands} \mid \textit{snowy}) \times \Pr(\textit{snowy}) \quad (7.9)$$

$$= 0.30 \times 0.30 \quad (7.10)$$

$$= 0.09 \quad (7.11)$$

In this row the prior probability about snow is 0.30.

Here is the second row.

1. Nice and shovels

$$l\Pr(\textit{nice}, \textit{shovels}) = \Pr(\textit{shovels} \mid \textit{nice}) \times \Pr(\textit{nice}) \quad (7.12)$$

$$= 0.10 \times 0.70 \quad (7.13)$$

$$= 0.07 \quad (7.14)$$

2. Nice and just hands

$$l\Pr(\textit{nice}, \textit{hands}) = \Pr(\textit{hands} \mid \textit{nice}) \times \Pr(\textit{nice}) \quad (7.15)$$

$$= 0.90 \times 0.70 \quad (7.16)$$

$$= 0.63 \quad (7.17)$$

In this row the prior probability about nice days is 0.70.

An insightful exercise is to carry these calculations from the number of ways snow with and without shovels occurs given we think we know something about snow. The same with the number of ways a nice day might occur with and without shovels, given what we think about nice days.

Let's put one calculation together with a not so surprising requirement. When we conjoin snow with shovels, how many possible ways can these logical statements occur? It is just the 31 days.

We now have all of the derived information to carry our investigation further. We also total the rows and, of course, the columns. We will see why very soon.

The row sums just tell us as a check that we got all of the ways in which snow occurs in 31 days. What is brand new are the column sums. They add up the ways that data occurs across the two ways we hypothesize that data can occur: snow, no snow (nice day). They tell us the probability of carrying a shovel or not, across the two hypotheses. Another way of thinking about the  $p(d)$  column sums is that they are the expectation of finding snow or hands in the data. The consistency of all of these calculations is that column sums equal row sums, 100%. All regular, all present and correct, probability-wise.

### 7.1.4 Updating beliefs

The table lays out each of the four logically possible combinations of data and hypotheses. So what happens to our beliefs when they confront data? In the problem, we are told that we really see shovels, just like the picture from Albany, NY at the turn of the 20th century. Is surprising? Not necessarily in Albany and in January, so you might expect this behavior out of habit during a rough Winter. The point is that whatever our beliefs have been about shovel behavior, we should still subject them to the possibility of accomodating the fact of seeing shovels in hands in Albany in January, a winter month in the Northern Hemisphere.

We should recall this formula about the probability of seeing both an hypothesis and data:

$$\Pr(h \mid d) = \frac{\Pr(d \wedge h)}{\Pr(d)} = \frac{\Pr(d \mid h) \Pr(h)}{\Pr(d)}$$

Now we can trawl through about our intuitions and some arithmetic. We worked out that the joint probability of *both **snowy day** and **shovel*** is 21%, a rate reasonable given the circumstances. In our formula, this is the product of the likelihood  $\Pr(d = \text{shovels} \mid h = \text{snow}) = 0.70$  and the prior probability we registered that snow might occur  $\Pr(h = \text{snow}) = 0.30$ .

Relative to the product of the likelihood of shovels given a nice day and the chance that snow might occur is the the joint probability of *both **nice day** and **shovel*** at 10%, or  $\Pr(d = \text{shovels} \mid h = \text{nice}) \Pr(h = \text{nice}) = 0.10 \times 0.70 = 0.07$ , again a reasonable idea, since we plausibly wouldn't see much shovel handling on that nice day in January..

Both of these estimates are consistent with actually seeing shovels in people's hands. But what are the chances of just seeing shovels at all? This is an ***either or*** question. We see shovels 21% of the time on snowy days or we see shovels 7% of the total days in January on nice days. We then add them up to get 28% of the time we see shovels in all of January, whether it snows or not.

So back to the question: if we do see shovels in the hands of those folk, will it snow? The hypothesis is  $h = \text{snow}$  and the data is  $d = \text{shovels}$ . The joint probability of both snow and shovels is  $\Pr(d, h) = 0.21$ . But just focusing on the data we just observed, namely that we see shovels, we now know that the chances of seeing shovels on any day in January in Albany, NY is  $\Pr(d) = 0.28$ . Out of all of the ways that shovels can be seen in January then we would anticipate that the probability of snow, upon seeing shovels, must be  $\Pr(h \mid d) = \Pr(d, h) / \Pr(d) = 0.21 / 0.28 = 0.75$ .

What is the chance of a nice day given we see shovels? It would be again likelihood times prior or  $0.10 \times 0.7 = 0.07$  divided by the probability of seeing shovels any day in January 28%. We then calculate  $0.07 / 0.28 = 0.25$ . We now

**TABLE 7.5:** Unobserved belief tempered by observed data = posteriors.

hypotheses	shovels	hands	priors	posterior shovels	posterior hands
snow day	0.7	0.3	0.3	0.75	0.125
nice day	0.1	0.9	0.7	0.25	0.875
sum	0.8	0.2	1.0	1.00	1.000

have the posterior distribution of the two hypotheses, snow or nice, in the face of data, shovels. So what are the odds in favor of snow when we see shovels?

$$OR(h \mid d) = \frac{\Pr(h = \text{snow} \mid d = \text{shovels})}{\Pr(h = \text{nice} \mid d = \text{shovels})} \quad (7.18)$$

$$= \frac{0.75}{0.25} \quad (7.19)$$

$$= 3 \quad (7.20)$$

We can read this as: when we see people with shovels in January in Albany, NY, then it is 3 times more plausible to have a snowy day than a nice day. The ratio of two posteriors gives us some notion of the plausible divergence in likely outcomes of snowy versus nice days. Again we must append the circumstances of time and place: in a January and in Albany, NY.

Here is table that summarizes all of our work to date.

---

## 7.2 Enough with the shovels!

Let's apply the probability analysis we developed so far to understanding the plausibility of the demand for pies at one of the smaller restaurants Make-A-Pie vends to. Here is a day by day count of the number of pies sold (the *count* variable) and the weather (clear coded as *C* or rainy coded as *R*). The original data only has the date, count, and the weather. The rest is analysis based on this data.

After carefully verifying and validating the data and its sources, the next step in spreadsheet engineering is always to create named ranges. Here they are the names of the columns.

To derive the days of the week in a potential analysis, we can create a table of the day number, with 1 as Sunday. Then the day of the week would be an INDEX() lookup of the date's corresponding day of the week in the table. The result is the day name. With this we can possibly sort out all Saturdays and Sundays to isolate activity on weekdays and weekends.

For our purposes we would like to know the odds of selling more than 10

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2														
3		date	count	clear	weekday	day_of_week		day_index	day					
4		2021-12-27	11	R	2	Monday		1	Sunday					
5		2021-12-28	32	C	3	Tuesday		2	Monday					
6		2021-12-29	21	C	4	Wednesday		3	Tuesday					
7		2021-12-30	19	C	5	Thursday		4	Wednesday					
8		2021-12-31	13	C	6	Friday		5	Thursday					
9		2022-01-01	2	R	7	Saturday		6	Friday					
10		2022-01-02	1	R	1	Sunday		7	Saturday					
11		2022-01-03	10	C	2	Monday								
12		2022-01-04	9	C	3	Tuesday		E4:						
13		2022-01-05	3	R	4	Wednesday		=WEEKDAY( B4 )						
14		2022-01-06	6	C	5	Thursday		F4:						
15		2022-01-07	19	C	6	Friday		=INDEX( day, MATCH( E4, day_index, 0 ), )						
16		2022-01-08	25	C	7	Saturday								
17		2022-01-09	26	C	1	Sunday								
18		2022-01-10	10	C	2	Monday								
19		2022-01-11	2	C	3	Tuesday								
20		2022-01-12	1	C	4	Wednesday								
21		2022-01-13	22	R	5	Thursday								
22		2022-01-14	21	C	6	Friday								
23														

FIGURE 7.1: Daily Pie demand and the weather.

pies when the weather is clear. We can answer this query by making a 2x2 contingency table of the conditional data counts of pie demand. This panel details the computations.

	N	O	P	Q	R	S	T	U	V
2									
3		threshold	10						
4									
5			weather						
6		count	R	C					
7		>10	2	8	10				
8		<=10	3	6	9				
9			5	14	19				
10									
11			Pr( >10   C ) = 57.14%						
12			OR = 1.33x						
13									
14		O7:							
15		= ">" & threshold							
16		Q7:							
17		=COUNTIFS( count, \$O7, clear, "=" & Q\$6 )							
18		Q9:							
19		=SUM(Q7:Q8)							
20		P11:							
21		="Pr( "&O7&"   "&Q6&" ) = "&TEXT( Q7/Q9 * 100, "##.00")&"%"							
22		="OR = "&TEXT( Q7 / Q8, "##0.###")&"x"							
23									

FIGURE 7.2: What are the odds?

The key is the associative counting of pies by *both* a count greater than a

threshold of 10 *and* clear weather. We accomplish this, for example in cell Q7, with =COUNTIFS( count, \$O7, clear, “=”&Q\$6 ). We can read this statement as count the pies (count) greater than 10 (\$O7) *and* with weather (clear) which is indeed clear (Q\$6). This yields 8 pies. When the weather is clear we sell a total of 14 pies.

The probability of selling more than 10 pies when the weather is clear is about  $8/14 \times 100 = 57\%$ . We sold 8 pies greater than the threshold and 6 pies less than the threshold when the weather was clear. The ratio of  $8 : 6 = 1.33$  reports the odds of selling greater than 10 pies. We interpret this number with the phrase: it is 1.33x more likely to sell greater than 10 pies, than not, when the weather is clear.<sup>5</sup>

---

### 7.3 What did we accomplish?

We have travelled through the complete model of probabilistic reasoning.

1. We started with a question. The question at least bifurcates into the dialectical *is it?* or *is it not?*.
2. We then began to think about beliefs inherent in the question for each of the hypotheses buried in the question.
3. We then collected data that is relevant to attempting an answer to the question relative to each hypothesis.
4. Then we conditioned the data with the hypotheses inside the question. It is always about the question!
5. Finally we derived plausible answers to the question.
6. We then illustrated the example with data collected, threshold set, table constructed, odds computed.

What is next? We continue to use this recurring scheme of heuristic thinking, sometimes using algorithms to count more efficiently, applied to questions of ever greater complexity. In the end our goal will be to learn, and learning is inference.

---

<sup>5</sup>Some folks might want to aggrandize this statement by saying that is is 33% more likely to sell more than 10 pies when clear. We might caution ourselves as we might be falling into a distortion by magnifying an apparent difference. The probability of selling more than 10 pies when clear is  $8/14 = 4/7$ . on the other hand the complementary probability of selling up to 10 pies when clear is  $6/14 = 3/7$ . We get to sell only 1 pie in crossing the 10 pie threshold given this data set. An inflated way of expressing the results of this analysis? Isocrates would agree.



---

## 7.4 Try this out

1. Start with a question for analysis using a indicative-interrogative statement format, for example “We observe X. Will Y occur?” Based on this statement identify the unobserved data of the hypothesis and the observed data. Use binary hypotheses and observations.
2. Rework the Albany NY example using your hometown or city. Develop initial distribution of hypotheses, distributions of data given a hypothesis, joint distributions of hypotheses and data. Find the probability that a particular hypothesis might occur given a specific piece of data.
3. Rework the spreadsheet example using various thresholds. Use the Data Table sensitivity analysis to calculate results. Plot the Data Table. Lead a discussion with some skeptical people about your findings.



# 8

---

## *Waiting on a Simulation*

---

---

### 8.1 A story of chance

Stan Ulam got sick one day, and it lasted for quite a while. The year was about 1945 . The place, a secret facility in a not so secret desert in the Southwest United States. To wile away the time he played, as many might do, solitaire, a game of self-imposed patience. He was a mathematician and soon became impatient with not winning several hands in a row.

The combinations of potential hands and their rule by rule repositioning in an ace-high order, and Stan knew this, results in  $52! \approx 8.07 \times 10^{67}$  possible games. If he could play 4 games per hour, non-stop, it would take him about  $3.29 \times 10^{62}$  years. That wasn't happening! He discussed this idea with his colleague John von Neuman who had just been involved in building the ENIAC computer (see Thomas Haigh (2014) for a history). He supposed he could generate  $N$  games.

#### 8.1.1 A short digression on volatility

For each game Stan Ulam determines which proposed game is, or is not, winnable. The number of winnable games is  $W$ . The relative frequency  $f$  of winnable, acceptable games is then  $f = w/N$ . Next he would let  $N$  get larger thinking that the accuracy of this thought experiment should improve. There are after all just two possibilities win or lose, a binary outcome. If there are just two outcomes, and like the flipping of a coin, equally likely or nearly so, then solitaire wins and losses can be modeled by a binomial distribution with  $N$  games a probability of a win  $p$ . and thus a loss  $1 - p$ . The standard deviation of  $p = f = w/N$  wins over losses is the sampled  $f$  from a binomial distribution of a 1 shot sample (also known as a geometric distribution, but neither here nor there.

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Here's a sketch of the derivation. The sketch includes a manipulation of expectations. Now, we do remember a bit of information about expectations. They are ultimately weighted averages of outcomes, like the  $f/N$  we, includ-

ing Stan, sampled, where the weights are assigned probabilities of occurrence of outcomes. We let  $E()$  stand in for this aggregation. Then we remember that the standard deviation is the square root of the variance. The variance in turn is the expectation (weighted probability average) of squared deviations of outcomes from means (which are in turn expectations). If  $f$  is a binomial outcome, then  $Ef = np$  and  $Var(f) = E(f - Ef)^2 = np(1 - p)$ . Here is the sketch without much, in any, commentary.

$$Var(p) = E(p - Ep)^2 \quad (8.1)$$

$$= Ep^2 - (Ep)^2 \quad (8.2)$$

$$p = \frac{f}{N} \quad (8.3)$$

$$Var(p) = E(f/N)^2 - (E(f/N))^2 \quad (8.4)$$

$$= (1/n^2)(Ef^2 - (Ef)^2) \quad (8.5)$$

$$= (1/n^2)var(f) \quad (8.6)$$

$$= (1/n^2)np(1 - p) \quad (8.7)$$

$$= \frac{p(1 - p)}{n} \quad (8.8)$$

The standard deviation  $\sigma_p$  of  $Var(p)$  is then

$$\sigma_p = \sqrt{\frac{p(1 - p)}{n}}$$

Phew! Only a little waving of the hands required.

### 8.1.2 Back Solitaire

This means that as the number of sampled solitaire (Klondike style) hands climbs, the standard deviation of relative frequencies declines and thus the process produces more accurate results. The win rate for recreational play has been apocryphally estimated at  $p = 0.43$ . This is almost a random walk.

John von Neumann, his *computer operator* wife Klara von Neumann, and others worked with Ulam's idea, code-named Monte Carlo (it was secret until about 1948) on the ENIAC computer with its 17,000+ vacuum tubes. They were able to find solutions to problems they had no idea even existed. They also built thermo-nuclear weapons with designs enabled by Monte Carlo methods on the ENIAC. Haigh et al. document the use of the stored program modern computing paradigm installed in the ENIAC as well as the programs themselves (Thomas Haigh (2014)),

We might have three or so takeaways from all of this story-telling. First, we should always keep a notebook, the original stored program facility. Second, we can use Monte Carlo to generate new insights through a random brute force

method. Third, advances in technology provide a touchstone for technique: always be on the look-out for off trend movements in a domain of knowledge.

The second takeaway has an important implication for us. Generative models build joint probabilities of highly interrelated groupings of data. In our working example in this chapter we will need the capability of letting multiple eateries communicate with different times of day. With 20 eateries and 2 shifts there are over  $4 \times 10^{18}$  possible interactions. That's a lot of potential communication. Conditioning the data will help us pare this down to size. Monte Carlo will help us understand customer experience at the eateries.

---

## 8.2 A piece of pie, please

Tortiere and Fazi started baking vegan pies because they really wanted to open a chain of small eateries where the pies, with salads of course, would be the main attraction. They would also offer coffee, near-coffee as they call products from some chains from the Northwest United States, perhaps liquor and beer, very cosmopolitan. They hire some restaurant experts from a popular TV network to help them understand some of the business side of an eatery. The so-called experts build a robot. This is a pie-eating-robot, not a person, but a machine that can sense, observe, even build updateable probability tables for how long it takes to get served.

Our model (we are the so-called experts) should be able to go from one eatery to another with memory. and we We do remember that last time we ordered a cup of coffee and a piece of pie at a lunch counter eatery. The model should also be able to the same eatery or another in the morning or the afternoon. Time of day traffic at eateries could be different, exhibiting different customer experiences. The primary metric we elicit from our clients is their concern over how long a customer remains in line, gets to order, and receives the order. Too slow and the customer might bolt next door and eat meat pies. Too fast and service quality, including basic courtesy might jump out of the back window.

We will program a robot to visit two eateries, order coffee and a piece of pie, and estimate the waiting times at each. The robot enters the first eatery, either in the morning or in the afternoon. The pie-eating-robot begins with a vague idea for the waiting times, say with a mean of 5 minutes and a standard deviation of 1. After ordering a cup of coffee and a piece of pie at the first eatery, the robot observes a waiting time of 4 minutes, an observation below the mean, about 1 standard deviation's worth. It updates its probability of being at this eatery, the so-called entry, just getting in the queue, prior distribution, using Bayes' theorem of course, with this information. This gives

it an exit, resulting, otherwise known as a posterior distribution for the waiting time at the first eatery.

Now the robot moves on to a second eatery. It might be morning or afternoon. When this robot arrives at the next eatery, what is its expectation upon entering, the so-called prior, the probability that the next eatery is to be entered? It could just use the resulting distribution, also known as a posterior, from the first eatery as its entry distribution, also known as a prior, for the second eatery. But that implicitly assumes that the two pie-eateries have the same average waiting time. Eateries are all pretty much the same, but they aren't identical.

On the other hand, it doesn't make much sense to ignore the observation, the experience of waiting, from the first eatery. That would make the robot have amnesia and we want the robot to remember, and possibly compare. So how can the eatery customer-robot do better? It needs to represent the population of eaterys and learn about that population. The distribution of waiting times in the population becomes the prior, the probability, for each eatery, until, and if, the pie-eating-robot encounters a new customer experience, which will be compared with and conditioned by prior experiences.

Whatever the eatery, the robot has a simple model in its robotic cortex.

$$\mu_i = \alpha_i + \beta_i A_i$$

where  $\mu_i$  is the average waiting time in minutes at eatery  $i$ ,  $\alpha_i$  is the average morning waiting time,  $\beta_i$  is the average difference in morning and afternoon waiting times, and  $A_i$  is a zero/one indicator of whether we are in the afternoon shift, 1, or present ourselves to the morning shift, 0.

Eateries covary in their intercepts and slopes. Why? At a popular eatery, wait times are on average long in the morning, because staff are very busy, in a word, slammed. The eatery might be the only place open for blocks, but the same eatery might be much less busy in the afternoon, leading to a large difference between morning and afternoon wait times. At such a popular eatery, the intercept is high and the slope is far from zero, because the difference between morning and afternoon waits is large. But at a less popular eatery, the difference will be small. Such a less than popular eatery would make us wait less in the morning, because it's not busy. and there is not much of a change in the afternoon. In the entire population of eateries, including both the popular and the unpopular, intercepts and slopes covary. This covariation is information that the robot can use.

### 8.3 A brief interlude with Cholesky

The information structure we impose on the customer-robot relates morning and afternoon wait times. We model this as covariance, equivalently as correlation, which is covariance scaled (divided) by the product of the standard deviations of morning and afternoon. Scaling forces a covariance that can possibly range from negative to positive infinity into a far more useful range between -1 and +1. If we (through a scanner darkly<sup>1</sup> with our robot, that is)

We have two random parameters, so our problem is like this one. Given a  $2 \times 2$  standardized variance-covariance matrix, also known as a correlation matrix,  $R$ , where  $\rho$  is the coefficient of correlation,

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

We start with uncorrelated variates  $x$  from no particular distribution. Our job is to transform the uncorrelated variates to produce correlated variates  $z$  with the same expected variance-covariance matrix as  $R$ . This will put informational structure into the otherwise independently occurring  $x$ .

When we say transform we mean this mathematically.

$$z = Lx$$

such that

$$R = LL^T$$

This may be a tall order if it weren't for a trick we might have learned when solving simultaneous equations called triangularization.

### 8.4 Solution

We can decompose  $R$  into the product of upper and lower triangular symmetrical matrices. This is the standard trick of matrix algebra noted above.

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix} \begin{bmatrix} \ell_{11} & \ell_{21} \\ 0 & \ell_{22} \end{bmatrix}$$

<sup>1</sup>Pilfered directly from Philip K. Dick's memorable novel, *A Scanner Darkly*. Our aim is not at all as dystopic as Dick's.

When we multiply the upper and lower triangular matrices we get this interesting matrix.

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} \\ \ell_{11}\ell_{21} & \ell_{21}^2 + \ell_{22}^2 \end{bmatrix}$$

Now we use another trick by matching elements of  $R$  with the elements in our new matrix. Here we go.

- $\ell_{11}^2 = 1$  implies  $\ell_{11} = 1$
- $\ell_{11}\ell_{21} = \rho$  then implies  $\ell_{21} = \rho$
- $\ell_{21}^2 + \ell_{22}^2 = 1$  then implies that  $\ell_{22} = \sqrt{1 - \rho^2}$

That wasn't as bad as we might have thought when we started. We then let  $L$  be the new mashed up  $R$  matrix. Here we already know that  $\ell_{12} = 0$  from the original definition of  $L$ .

$$L = \begin{bmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{bmatrix} \quad (8.9)$$

This matrix can be expanded to more than 2 dimensions. The procedure to make that happen is called a **Cholesky** factorization, really an algorithm, well beyond these proceedings, but on a path for Monte Carlo generation.

## 8.5 Now we can simulate

We generate correlated  $z = Lx$  building on the uncorrelated  $x$ . We first generate a random  $x_1$  and, independently, a random  $x_2$ . We can use the `=RAND()` function in Excel to perform this task.

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Then we generate a  $z_1$  and a  $z_2$  using the  $x$  vector of random numbers, but transformed by pre-multiplying  $x$  with  $L$ .

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1 - \rho^2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

We do remember that By definition, if  $x_1$  is not correlated with  $x_2$ , then  $\rho_{12} = 0$ . We can check our maths with this calculation.

$$xx^T = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1x_2 \\ x_1x_2 & x_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The multiplication of a column vector of independently drawn  $x$  with its transpose, the row vector of those same  $x$  random numbers will always return the



identity matrix  $I$ , a matrix of 1's on the diagonal and 0's elsewhere. This shows that variates are perfectly correlated with themselves but not each other.

Back to the main business at hand. We now calculate

$$zz^T = (Lx)(Lx)^T = Lxx^TL^T = LIL^T = LL^T$$

But,  $R = LL^T$  so that  $R = zz^T$ .

Thus we have sketched out these steps.

- Generate uncorrelated  $x$ .
- Generate  $z = Lx$ , where  $L$  reflects the desired correlation structure.
- Find that  $L$  generates the same correlation structure as the correlations in  $z$ .

### 8.5.1 Excel or bust!

Here is a demonstration in Excel. It will use VBA without apology.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1																					
2		x <sub>1</sub>	0.596711		R	1	-0.07		L	1	0		z <sub>1</sub>	0.596710686			x <sub>1</sub>	x <sub>2</sub>	z <sub>1</sub>	z <sub>2</sub>	
3		x <sub>2</sub>	0.899758			-0.07	1			-0.07	0.49755		z <sub>2</sub>	0.405904968			0.596711	0.899758	0.596711	0.405905	
4																					
5			C3:			F3:	G2:			J3:	K3:			N2:N3		runs	x <sub>1</sub>	x <sub>2</sub>	z <sub>1</sub>	z <sub>2</sub>	
6			=RAND()			=-0.07	=rho			=rho	=(1-rho*2)*(1/2)			=MMULT(L,x)			1	0.613314	0.311155	0.613314	-0.34998
7																	2	0.475564	0.789996	0.475564	-0.08325
8																	3	0.347664	0.65543	0.347664	-0.07623
9																	4	0.419936	0.595977	0.419936	-0.14198
10																	5	0.45506	0.415174	0.45506	-0.21267
11																	6	0.489882	0.139332	0.489882	-0.30739
12																	7	0.02145	0.733992	0.02145	0.172153
13																	8	0.881076	0.626292	0.881076	-0.45705
14																	9	0.127158	0.374143	0.127158	0.006396
15																	10	0.819525	0.882067	0.819525	-0.34874
16																	11	0.265359	0.068601	0.265359	-0.16826
17																	12	0.4505	0.448675	0.4505	-0.30774

Vectors of  $x$  and  $z$  flank the correlation matrix  $R$  and the transformation matrix  $L$ , just like the maths. We name the region B2:N3 **calculation**. The  $x$  vector is populated with `RAND()` functions. These will always return a uniformly distributed random number between 0 and 1. This will act like a probability if we would like to shape the draws into something more useful to our purposes. We use the `MMULT()` array function to matrix multiply  $L$  and  $x$  to get  $z$ . We then simulate this act 10,000 times.

To simulate we designate the Q3:T3 range by naming it **interface**. This is a go-between region that mediates the calculations with the presentation of a single run to another named range called **simulation** in cells Q5:T5.

We then (cleverly) write a routine that replaces our fingers. They would otherwise press F9 10,000 times to recalculate the sheet, drawing new  $x$  values and transforming them into new  $z$  values. Each time the fingers would drop a copy of the interface value row into the next available row below the **simulation** range Q5:T5, and do it all over again, shall we say it again? Yes, 10,000 times. We replace this labor intensive work with this Visual Basic for Applications

(VBA) code which may be viewed by pressing the Visual Basic button in the Developer ribbon.

```
Sub cmdMonteCarlo_Click()
    Dim R1 As Range, R2 As Range, R3 As Range
    Dim iRun As Long
    Dim ltk As Double
    Application.ScreenUpdating = False
    Application.Calculation = xlCalculationManual
    Set R1 = Range("calculation")
    Set R2 = Range("interface")
    Set R3 = Range("simulation")
    For iRun = 1 To 10000
        R1.Calculate
        R2.Calculate
        R3.Offset(iRun, 0).Value = R2.Value
    Next iRun
    Calculate 'update statistical analysis
    Application.Calculation = xlCalculationAutomatic
    Application.ScreenUpdating = True
    MsgBox "Congratulations! Pat yourself on the back! ", vbOKOnly + vbInformation, "Monte Carlo simulation"
End Sub
```

**FIGURE 8.1:** Cholesky does a demo.

We might some appropriate pithy remark in the MsgBox. We can also change the number of simulations to some other assumption of the analysis by editing the FOR-NEXT loop. Otherwise, this code is self-sufficient and ready to reuse. We simply specify the named ranges. The rest is nicely automated. Assigning this subroutine to a button finishes the job.

## 8.6 Back to the eatery robot

The robot enters an eatery and observes waiting times. We will have this robot do this many times in the morning and / or the afternoon for 20 eateries, thus the simulation.

Here is the setup, calculation and graph for the random  $\alpha_i$  waiting time intercepts correlated with  $\beta_i$  afternoon waiting time differentials. This table is the starter for the rest of the waiting time simulation. It represents our prior expectations about the range and shape of morning intercept times and afternoon slope times. We will use this table in the main simulation in the next section.

The intercepts and slopes for 20 eateries plots a negative relationship. We shape the random variates into Gaussian (normal) distributions. In an x\_1 cell in column F we make a normally distributed random variable with a 0 mean and 1 standard deviation. This is none other than the centered and scaled normal z-score. NORM.INV() will report the number of standard deviations away from 0 corresponding to a given probability, here drawn as a number from 0 to 1 with RAND(). But we need a probability version of this draw that transforms z's that may positive or negative and very large or very small

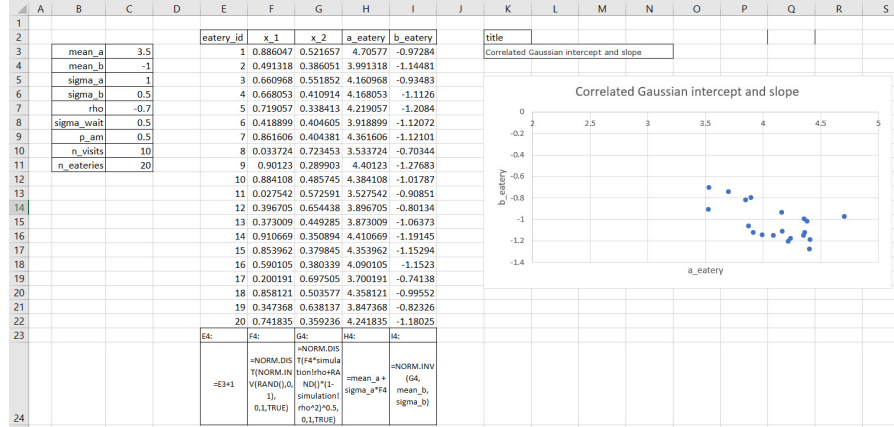


FIGURE 8.2: Simulating wait time parameters.

into a probability. That magic we perform with NORM.DIST set to TRUE, the cumulative normal density distribution. In the end we get a normally distributed random number with mean zero and standard deviation one.

We move to an  $x_2$  cell in column G where we *Cholesky* (a new verb, perhaps) the  $x_1$  normal random number into a new normal (pun intended!) but correlated random number. We use the second row of the Cholesky matrix, which when multiplied by another randomly drawn  $x_2$  in Excel gives us in Excel this algebraic formula.

$$z_2 = \rho x_1 + \sqrt{1 - \rho^2} x_2$$

We remember that  $z_1 = x_1$  in the Cholesky transformation. Maybe this is too much information, but it does come in handy when numbers blow up or do not at all align with anyone's common sense.

We then get to the random intercepts and slopes in columns H and I. The hard work has already been done. We draw normal variates with mean and standard deviations already specified for a and b. The graph concurs with our expectations that they are random and negatively correlated.

### 8.6.1 The main course

The robot randomly enters eatery number one and orders again, and possibly, again, or goes to eatery number two to do the same. The robot may go in the morning, or possibly in the afternoon. We and the robot do not expend any calories, get heartburn, or even have a good time. The robot only cares about how long it takes to get served. This means the robot tracks parameters for each eatery and time of day.

	A	B	C	D	E	F	G	H	I	J	K
1									E3:	F3:	F4:
2		key	run	eatory_id	A	mu	wait		=IF( RAND() > p_am, 0, 1 )	=INDEX( a_eatory, MATCH( D3, eatory_id, 0 ) ) + INDEX( b_eatory, MATCH( D3, eatory_id, 0 ) ) * E3	=F3 + sigma_wait *NORM.INV (RAND(), 0,1)
3		1-1-0	1	1	0	3.508972	4.156685				
4		2-2-1	2	2	1	2.862296	2.583731				
5		3-3-1	3	3	0	3.849713	3.962056				
6		4-4-1	4	4	1	2.949318	3.145916				
7		5-5-0	5	5	1	2.692138	3.165851				
8		6-6-0	6	6	1	2.807911	3.277582				
9		7-7-1	7	7	1	2.834218	3.549979				
10		8-8-0	8	8	1	2.927171	2.684472				
11		9-9-0	9	9	0	3.744481	3.264105				
12		10-10-0	10	10	0	4.16716	3.307825				
13		11-11-1	11	11	0	4.002217	4.385343				
14		12-12-1	12	12	1	3.06329	2.661684				
15		13-13-0	13	13	0	3.754629	3.43123				
16		14-14-0	14	14	0	3.606935	4.400666				
17		15-15-1	15	15	0	3.505498	2.946349				

**FIGURE 8.3:** Simulating wait time setups.

We introduce more, uncorrelated, uncertainty by allowing the customer-robot to choose whether to sample an eatery in the morning  $A = 0$  or afternoon  $A = 1$ . We use the IF() statement to make a choice with a threshold of p\_am. Basically this is Jakob Bernoulli's probability. The robot jumps into the eatery in the morning if a uniformly drawn RAND() number exceeds this threshold, otherwise wait for the afternoon shift.

We calculate the average wait time using each eatery\_id in an INDEX(MATCH()) statement to lift the correct intercept  $\alpha_i$  and slope  $\beta_i$  for eatery  $i$ . We then add intercept to the slope times the 0 or 1 afternoon marker. This is the mean, along with an assumed standard deviation that samples a waiting time from a normal distribution in the wait column. We do this lots and lots of times.

With all of this data it behooves us to study it closely.

### 8.6.2 Exploring wait times

We now have several hundred logically related versions of how 20 eateries with two shifts, morning and afternoon produce wait times for wary and unwary customers. We did this by building a logically endowed robot to perform the calculations all in Excel. Here is the workup in a separate worksheet called eda (for Tukey (1977) exploratory data analysis).

There are three tables. The one on the left at the top builds the parameters for an equally spaced set of intervals from which we can count the number of wait time results occur in the interval. All of that action occurs in the table to the right. Just below the grid parameter table is a summary of results.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2							interval	begin	end	mid	frequency	relative frequency	cumulative relative frequency	gaussian cdf
3		mu_grid	wait				1	-0.3928	-0.0985	-0.2457	2	0.0010	0.0010	1.9E-05
4		max_wait	5.78714				2	-0.0985	0.19574	0.04859	3	0.0015	0.0025	7.6E-05
5		min_wait	-0.3928				3	0.19574	0.49002	0.34288	7	0.0035	0.0060	0.00027
6		intervals_wait	21				4	0.49002	0.78431	0.63716	13	0.0065	0.0125	0.00089
7		width_wait	0.29428				5	0.78431	1.07859	0.93145	17	0.0085	0.0210	0.00261
8		count_wait	1999				6	1.07859	1.37287	1.22573	12	0.0060	0.0270	0.00694
9							7	1.37287	1.66716	1.52002	17	0.0085	0.0355	0.01664
10		wait					8	1.66716	1.96144	1.8143	28	0.0140	0.0495	0.03619
11		min	-0.3928				9	1.96144	2.25573	2.10859	69	0.0345	0.0840	0.07149
12		0.25	2.8970				10	2.25573	2.55001	2.40287	122	0.0610	0.1451	0.12864
13		0.50	3.4321				11	2.55001	2.8443	2.69715	184	0.0920	0.2371	0.2116
14		0.75	4.0124				12	2.8443	3.13858	2.99144	247	0.1236	0.3607	0.31955
15		max	5.7871				13	3.13858	3.43286	3.28572	280	0.1401	0.5008	0.44551
16		IQR	1.1154				14	3.43286	3.72715	3.58001	250	0.1251	0.6258	0.57727
17		mu_wait	3.4072				15	3.72715	4.02143	3.87429	261	0.1306	0.7564	0.70086
18		sig_wait	0.8866				16	4.02143	4.31572	4.16858	194	0.0970	0.8534	0.80478
19		mad	0.6891				17	4.31572	4.61	4.46286	143	0.0715	0.9250	0.88312
20		skewness	-0.5629				18	4.61	4.90429	4.75714	93	0.0465	0.9715	0.93608
21		kurtosis (+3)	3.9781				19	4.90429	5.19857	5.05143	37	0.0185	0.9900	0.96817
22							20	5.19857	5.49286	5.34571	15	0.0075	0.9975	0.98561
23							21	5.49286	5.78714	5.64	5	0.0025	1.0000	0.99411
24							G3:	H3:	I3:	J3:	K3:	L3:	M3:	N3:
25							=1	=min_mu	=H3 + width_mu	=(H3 + I3)/2	=COUNTIFS(INDIRECT(C\$3), ">="&H3, INDIRECT(C\$3), "<"&I3)	=K3/count_wait	=L3	=NORM.DIST(J3, mu_wait, sig_wait, TRUE)
26							G4:	H4:	I4:		K23:			
27							=G3+1	=I3	=H4 + width_mu		=COUNTIFS(INDIRECT(C\$3), ">="&H23, INDIRECT(C\$3),			
28														

FIGURE 8.4: Building EDA

We compose a typical data summary with these items:

1. Maximum, minimum, and the 25th, 50th, and 75th quartiles, since they break the data into four parts. All four are measures of location. We measure central tendency with the median. Generally these are all we need to make a box plot. We also record the Interquartile Range (IQR), a scale metric useful for helping us determine outliers.
2. Mean and standard deviation to get other measures of location and scale. We also compare with median with the mean and see they are very close.
3. Skewness measures how lopsided, asymmetric the distribution of data is. Kurtosis is very nearly a scaled standard deviation of the standard deviation. It measures the thickness, or thinness, of the tail where we find less frequent outcomes. This data is nearly symmetric with a meso-kurtic, a medium-thick, tail. It so looks like the normal distribution, and it should, since we sampled from this distribution in the first place.

4. Other measures such as the data can be included.

Now off to the right is a table where we calculate, using 21 intervals, the frequency, relative frequency, and cumulative relative frequency of the data occurring in intervals. We use enough intervals with interval mid-point spacing designed for the use of managers and other decision makers. The documented cells at the bottom of the table document the way we construct intervals and their mid-points. We use midpoints in graphing. We count the integer number of times that a wait time we just sample happens to occur in a an interval.

The Excel function COUNTIFS(array, criterion, array, criterion) does the heavy lifting for us. An array here is the vector of wait times. All criteria are logical statements. The syntax for a finding the number of wait time samples in an interval with beginning of interval value in a cell and an ending of interval value in another cell is “>=&begin\_cell and “<=&end\_cell. The very last interval, to ensure we count all of the data has a “<=&end-cell logical statement. We check each of the beginning and ending interval cell formulae and the logical criteria to count wait time observations.

In the last column of the table we compute the normal (Gaussian) cumulative distribution function (CDF) version of cumulative relative frequency. We use mid-points from our interval estimations and the mean and standard deviation from the data summary table. We set the last argument to TRUE to calculate the CDF. Our mechanical work is done. Here is a graph of the results.

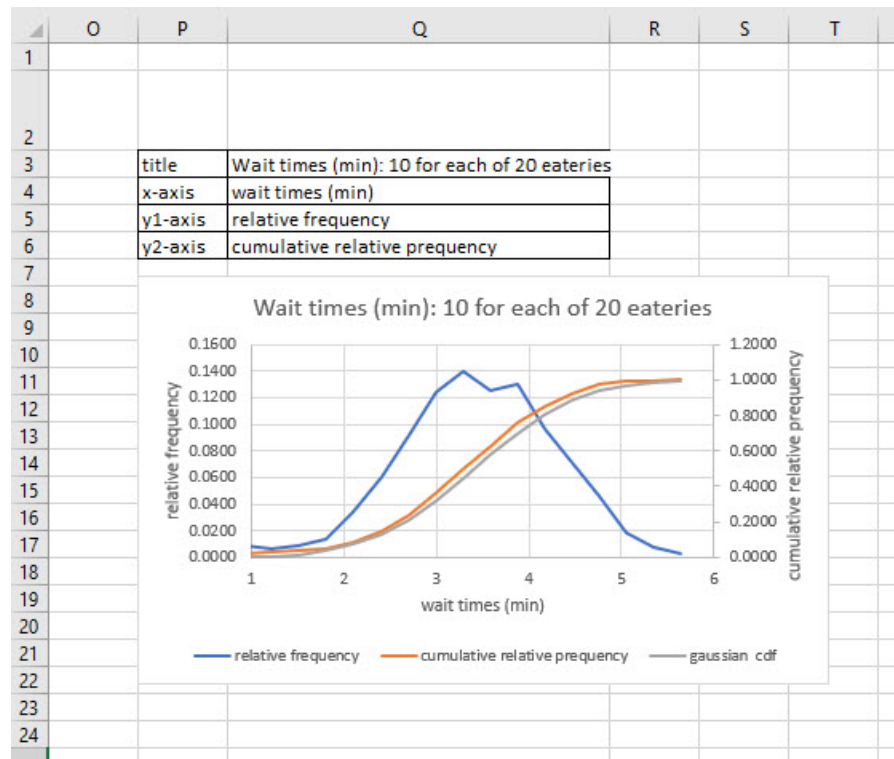
We see that the simulation and normal CDF’s nearly collide into one another. The utility of a distribution like this is to help us ask how certain are we about average wait times? Probability intervals with lower and upper bounds will help us allocate scarce resources in a more principled fashion. We continue to remind ourselves that models are abstractions from reality, they are not customer experience, or the staff and facilities needed to enhance the experience.

---

## 8.7 Any next steps?

We certainly should take a paste-values frozen copy of the simulation, split it into morning and afternoon sections and run the same exploratory data analysis on the splits. This will enable us to determine whether the eateries have much the same or radically different customer experiences.

We can also ask the question about which eateries are more or less popular, if we think popular is longer wait times for both morning and afternoon shifts. If popular, the conjecture is that longer wait times are due to customer congestion. Unfortunately they could be due to staff shortages, relatively less trained

**FIGURE 8.5:** Finally, a picture.

staff or even the physical configuration of an eatery. But at least we have a start on beginning to advise Make-A-Pie.

Can we use this model in other settings? We might try to map the parameters of this model to other processes. For example the model might apply to analyze queues at vaccination sites. We might also conceive of the use of the model for tracking time to correct faults in geographically distributed electric and gas equipment. We might also use this sort of model in any environment where the number of steps, the number movements, the distance to and from, varies by some category in time and space (physical and imaginary) and where the categories are correlated. As always one size will probably only fit one. Customization will require further scoping, designing, and implementing.





# 9

---

## *The Outer Limits*

---

### 9.1 Tales of tails

Rod Serling portrayed the edge of human dire straits and decisions in his classic TV series, *The Twilight Zone*.<sup>1</sup> We will go to the edge in this round of simulation, but we hope, not too far over the edge of our capabilities to analyze. The edges and horizons of our work will be peering into the tails of distributions of decision outcomes. One of these outcomes is workers compensation claims experience.

Let's set the scene. Make-a-Pie owner Simone Tortiere meets regularly with other vegan food company owners. This week their agenda covers the rising incidence of and cost of covering workers compensation claims. A claim arises when a worker, on the job, becomes disabled for any reason. The New York State Insurance Fund (NYSIF) offers several plans and products to employer policy-holders. The vegan company owners around the table at host Candle Cafe try to get a handle on claims, rejection of claims, required drug formularies and treatment codes, increasing cost of disability, loss of key personnel, among other things.

Many of the vegan businesses are NYSIF policy-holders. For example, if a covered worker falls, has extended health issues, the fund will cover expenses for a specific time frame, using specified drug and other therapy treatments, all under the direction of NYSIF medical associates. Premiums will undoubtedly rise with more claims experience. Safety Group plans, while paying dividends to policy-holders with relatively low claims experiences, often groups all food-related workers into one class. This may, or may not, disadvantage the vegan food industry.

The group decides on this course of action.

- Gather workers compensation claims experience across the group

---

<sup>1</sup>Local lore places the zone at a now decommissioned train station in Dryden, NY. Serling ran the entire series with his company Cayuga Production, named for the lake where Serling's family owned a summer home Ithaca, NY, not far from Dryden on, yes, Route 13.

- Attempt to model future claims to understand the range and shape of the distribution of claims, all based on existing claims experience
- Use the future claims model to simulate a self-insurance portfolio

All of this is a very tall order for experts in vegan food production and service. Tortiere has already talked to us about the issues. She recommends that our analytics service can at least start to structure the group's next conversation in a month with some provisional results.

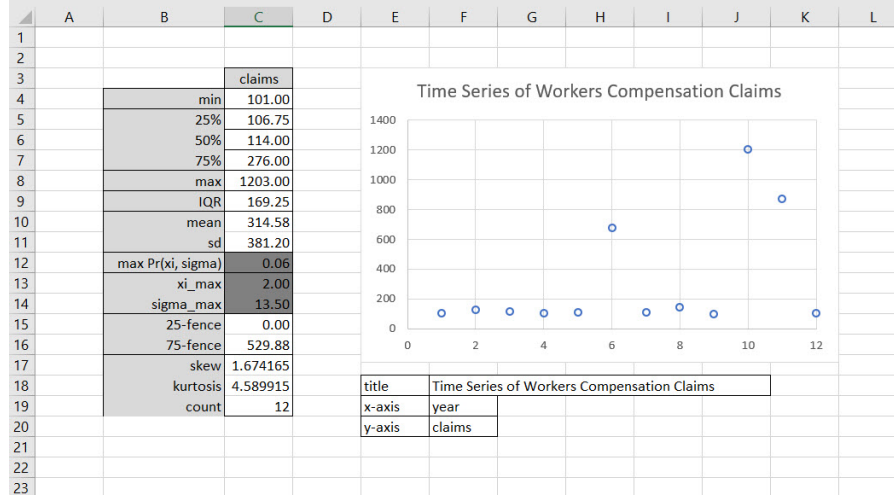
## 9.2 It always starts with data

Not really, because we actually begin with Tortiere's question: can we insure ourselves and be better off? Yes, then we begin our analysis by gathering 12 years of annual workers claim experience from several of the vegan establishments.

	A	B	C
1			
2		year	gross_cla
3		1	107
4		2	126
5		3	118
6		4	106
7		5	108
8		6	675
9		7	110
10		8	143
11		9	101
12		10	1203
13		11	875
14		12	103
15			

**FIGURE 9.1:** Some workers compensation claims data.

This is only a small sample of owners experience, all in hundreds of thousands of dollars. We visualize the times series.



Two patterns emerge. There seems to be a base-line of claims experiences. But there are several very large experiences as well. We ask is there some sort of threshold that splits the analysis of claims experience?

Well a well-known exploratory technique exists to examine the behavior of data, especially extreme values in this claims data. We see that Tukey (1977)'s outlier fences detect outliers above a threshold claims level of \$529,000. Another technique develops an excess of high threshold series. We generate a series of possible thresholds  $u$ . Then we sort the claims from lowest to highest. For each threshold and claim  $x$  we calculate the excess of threshold metric. Souvik Ghosh and Sidney Resnick (Ghosh and Resnick, 2010) develop just how reasonable a mean excess plot (MEP) is to explore extreme tail behaviors.<sup>2</sup>

We use this tool to evaluate this metric.

$$\max(x, y) = \frac{x + y + |x - y|}{2}$$

Let's try this with  $x = 0$  and  $y = 5$ . We remember that we are trying to measure claims 5 in excess of a threshold of 0.

<sup>2</sup>This is a highly technical discussion including the topology (normed vector spaces) of the region beyond a threshold. Some of the nuggets they reveal include: - MEP is linear when  $0 < \xi < 1$ . In this region, while there will much variability,  $\sigma$ , uncertainty measured by  $\xi$  seems more at bay. The mean is well defined in this region. - When  $\xi \geq 1$ , the MEP can mislead the analyst. A slowly varying tail process, such as with water seepage and ozone layering in urban areas, can fool any attempt to provide a systematic diagnostic. Perhaps this is simply because a tail with a value of  $\xi \geq 1$  cannot be discerned from the data, a distribution, or any other systematization except one: we know it is unknown, the inverse insight itself.

$$e(u) = \max(0, x - u) \quad (9.1)$$

$$= \frac{x + y + |x - y|}{2} \quad (9.2)$$

$$= \frac{0 + 5 + |0 - 5|}{2} \quad (9.3)$$

$$= \frac{5 + 5}{2} \quad (9.4)$$

$$= 10/2 \quad (9.5)$$

$$= 5 \quad (9.6)$$

$$= (5 - 0)_+ \quad (9.7)$$

This case seems to work. This works only because when  $x > y$ , then  $x - y < 0$  and  $|x - y| = y - x > 0$ . Here we flex our algebraic muscles to verify this relationship.

$$e(u) = \frac{0 + (x - u) + |0 - (x - u)|}{2} \quad (9.8)$$

$$= \frac{x - u + |-(x - u)|}{2} \quad (9.9)$$

$$\text{if } x > u \quad (9.10)$$

$$e(u) = \frac{x - u + x - u}{2} \quad (9.11)$$

$$= \frac{2x - 2u}{2} \quad (9.12)$$

$$= x - u \quad (9.13)$$

$$\text{if } x \leq u \quad (9.14)$$

$$e(u) = 0 \quad (9.15)$$

We can use this analysis to program spreadsheet cells. We use the subscript  $+$  to economize on notation. Next we average  $e(u)$  for each threshold. We remember from somewhere, maybe an elementary statistics course, the for  $u$  a constant that  $E u = u$ .

$$E e(u) = \sum_i^{N_u} \pi_i (x_i - u)_+ \quad (9.16)$$

$$= \pi_1 (x_1 - u)_+ \dots \pi_{N_u} (x_{N_u} - u)_+ \quad (9.17)$$

$$= E x - u \quad (9.18)$$

On the other hand does the variance of  $e(u)$  have anything at all to do with  $u$ ?

$$\text{if } x > u \quad (9.19)$$

$$\text{then} \quad (9.20)$$

$$\text{Var } e(u) = E[e(u) - E e(u)]^2 \quad (9.21)$$

$$= E[(x - u)^2 + (E x - u)^2 - 2(x - u)(E x - u)] \quad (9.22)$$

$$= E[x^2 - 2ux + u^2 + (E x)^2 - 2u E x + u^2 - 2x E x + 2u E x - u^2] \quad (9.23)$$

$$= E x^2 - (E x)^2 + u^2 - 2u E x \quad (9.24)$$

$$= \sigma_x^2 + u^2 - 2u \mu_x \quad (9.25)$$

We can use these calculations to build probability intervals around  $e(u)$  should the desire ever percolate to the surface.

Expectations build around a number  $N_u$  of non-zero  $e(u)$  values. It is possible that there are less than  $N$  observations for which a claim  $x$  only exceeded a threshold  $u$ . It is very possible that given a stream of claims, a very few set of claims which exceeds a high threshold will occur and thus we can state another rule of the excess over threshold story that  $N_u \leq N$ .

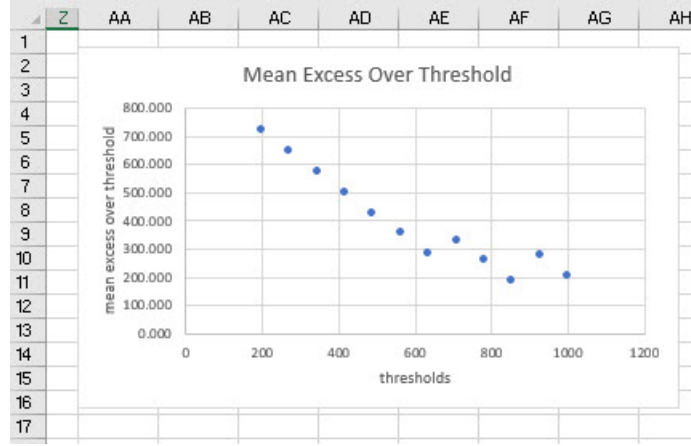
Here we calculate the mean excess of claims over a series of potential thresholds  $u$ .

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2		year	claims	claims	x - 200	x - 273	x - 345	x - 418	x - 491	x - 564	x - 636	x - 709	x - 782	x - 855	x - 927	x - 1000	threshold u grid				u	
3		9	101	101.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	max	1000			200.00
4		12	103	103.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	min	200			272.73
5		4	106	106.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	intervals	11			345.45
6		1	107	107.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	width	72.7273			418.18
7		5	108	108.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					490.91
8		7	110	110.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					563.64
9		3	118	118.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					636.36
10		2	126	126.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					709.09
11		8	143	143.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					781.82
12		6	675	675.000	475.000	402.273	329.545	256.818	184.091	111.364	38.636	0.000	0.000	0.000	0.000	0.000	0.000					854.55
13		11	875	875.000	675.000	602.273	529.545	456.818	384.091	311.364	238.636	165.909	93.182	20.445	0.000	0.000	0.000					927.27
14		10	1203	1203.000	1003.000	930.273	857.545	784.818	712.091	639.364	566.636	493.909	421.182	348.455	275.727	203.000						1000.00
15				u	200	272.727	345.455	418.182	490.909	563.636	636.364	709.091	781.818	854.545	927.273	1000						
16				mean	717.667	644.339	572.212	499.485	426.758	354.030	281.303	229.909	187.182	144.455	101.727	203.000						
17																						
18			claims		F4	G16																V16
19			lowest to		"x"	-MAX(0,																A*10 + 1710
20			highest		ROUND(F	#F4																
21					5,0)	G16)																
22						G17																
23						TRANSPOSE(u)																
24						G18																
						BF																
						SD-GA																
						"%6")																

FIGURE 9.2: Mean excess plot calculations.

We notice that there are a lot of zeros in this table. That is a direct result from just considering only positive differences of  $x - u$ . We use this idea to calculate a conditional mean with the AVERAGEIF() function. We generate the thresholds  $u$  on an equally spaced grid. What is  $N_u$ ? In column L after all of the zeros we are left with  $N_u = 3$ , and in column N  $N_u = 2$ . We look at Now for the plot of all of this handiwork.

The plot is a straight line until some turbulence occurs as the threshold approaches the maximum value of claims in the data. This usually indicates that the following distribution will work well to match this data.



**FIGURE 9.3:** MEP claims.

### 9.3 A distribution to remember

We need an observational model with these qualities:

1. It generates a regularly occurring ranges of outcomes.
2. On occasion the distribution generates very high outcomes.
3. It will be relatively easy to implement.

Well, we always hope for 3) but will get by with 1) and 2). Such a distribution is the Generalized Pareto Distribution (GPD) built on ideas by the Italian civil engineer, sociologist and economist Vilfredo Pareto at the turn of the last century. The very high outcomes he observed were wealth in the hands of an elite. For us, it's just a lot of bad work days in a year, no elites are involved. The GPD has been used in quality control for manufacturing, epidemic outbreaks, severe weather, finance collapse, and political violence. We use it for workers compensation claims experience.

Here is the probability distribution function. There are three parameters, location  $\mu$  is a threshold of claims  $x$ ,  $\sigma$  is the scale or dispersion of claims,  $\xi$  is the shape of claims that allow thick or thin tails in the distribution. That last quality is crucial. It marks a way to measure kurtosis in our observational model.

$$Pr(x | \mu, \sigma, \xi) = (1/\sigma)[1 + \xi(x - \mu)/\sigma]_+^{-(1+\xi)/\xi}$$

where  $\mu > 0$  is the location parameter (a known value, the threshold),  $\sigma > 0$

is Again we use the short hand notation  $h_+ = \max(h, 0)$ . If  $\xi < 0$ , then we are taking a root of a negative number in the square brackets. We must not! At least not here as this will entail complex numbers, a fate we do not want to tempt, again, at least here. If  $\xi$  should wander into a negative range, then this condition must hold true, or else, we have a singular model, and we will get a !NUM error in Excel.<sup>3</sup>

Here we isolate the  $x$  to get an idea of what range (mathematicians sometimes call this a support)  $x$  can take on before the model blows up.

$$1 - \xi(x - \mu)/\sigma < 0 \quad (9.26)$$

$$1 < \xi(x - \mu)/\sigma < 0 \quad (9.27)$$

$$\frac{\sigma}{\xi} + \mu < x \quad (9.28)$$

$$x > \frac{\sigma}{\xi} + \mu \quad (9.29)$$

In other words,  $x$  definitely needs to be greater than the  $\mu$  threshold. In general, the support is  $x > \mu$  for  $\xi > 0$ , and  $\mu < x < \mu - \sigma/\xi$  for  $\xi < 0$ . This will be our observational model. Such a model takes a  $\xi, \sigma$  combination as a conjecture, for a given level of threshold  $\mu$ , a hypothesis as a given, a condition.

The model computes the probability that a  $x$  occurs given this conjecture. Of course, we conjecture many combinations of the shape and scale parameters. Which one do we pick? The most plausible one. It turns out that the most plausible hypothesis is the hypothesis with the highest probability of choosing a particular hypothesis of  $\xi, \sigma$  given all of the data we have available, here 12 annual observations of claims.

How do we get to that probability? We simply find the probability that **both** the data occurs, given a hypothesis, **and** the probability that the hypothesis itself occurs. We calculate the probability both of the data in a model and of the probability of the hypothesis as the product of the probabilities. This is justly called the **Law of Conditional Probability**.

$$\Pr(x \wedge (\sigma, \xi)) = \Pr(x \mid \sigma, \xi) \Pr(\sigma) \Pr(\xi)$$

The little wedge  $\wedge$  means **and** and the  $\Pr(\mu) = 1$ , means we know  $\mu$  with certainty, after all we are setting it ourselves. But that does mean there will be some scrutiny to assure ourselves of the level of the threshold.

Yes, this a beast of a table. But while there is a lot going on here, we can start from left to right and top to bottom. This calculation machine is our claims-robot's cerebral cortex in its separate grid-approximation worksheet.<sup>4</sup> A grid on the left has all of the conjectured  $\xi$  shape and  $\sigma$  parameter combinations

<sup>3</sup>What distribution to choose? Taleb (2019) has a bit to say about the matter, and very technically, and practically, so.

<sup>4</sup>Grids, and techniques to approximate functions like our GPD, are popular in many fields,

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1																									
2																									
3																									
4																									
5																									
6																									
7																									
8																									
9																									
10																									
11																									
12																									
13																									
14																									
15																									
16																									
17																									
18																									
19																									
20																									
21																									
22																									
23																									
24																									
25																									
26																									
27																									
28																									
29																									
30																									
31																									
32																									
33																									
34																									
35																									
36																									
37																									
38																									

FIGURE 9.4: A grid to approximate claims distribution plausibility.

from the grid-setup worksheet. This table is pivotable, a task we perform in the next worksheet.

We can unpack the IF() statements in cells C6 and D6 by realizing that we must stay at the same  $\xi_i$  only while we loop through the list of  $\sigma$ 's, otherwise move to the next  $\xi_i$ . Cells C5 and D5 start the parade with the beginning entries of the  $\xi_i$  and  $\sigma$  lists.

Cell C6 tests whether the previous D5 is the end of the  $\sigma$  list by using the MAX() function. If true then the INDEX(..., MATCH()) retrieves the next  $\xi_i$ , otherwise stay at the same  $\xi_i$  in cell C5.

At the same time, the IF() statement in D6 tests whether or not the  $\sigma$  in D5 is the last  $\sigma$  in the list. If true, then go back to the beginning of the  $\sigma$ , otherwise go to the next  $\sigma$  in the list.

We end up With 25 nodes, that is,  $5 \times 5 = 25$  hypotheses, in a grid. We can then proceed to use the GPD observational model, one fit for use with location,  $\mu$  is the threshold, shape, approximated by  $\xi_i$ , and scale,  $\sigma$ , approximated by  $\sigma$ .

Our next stop on the magical mystery tour is the mashing together of observed data with the unobserved data of hypotheses, all 25 combinations of set  $\mu$  and approximated  $\sigma$  and  $\xi$ . These hypothetical parameters turn up in the GPD observational model again, here for reference.

including option pricing in finance, queue measurement in operations, customer sentiment analysis in marketing, reservoir flow in hydrological engineering, and beam stress analysis in civil engineering. We already used grids to optimize Simone Tortiere's pie price in our first outing with decision models.



$$Pr(x | \xi, \sigma) = \frac{1}{\sigma} \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-(1+\xi)/\xi}$$

This distribution is less beastly than the Gaussian (normal) distribution, after all it does not have  $\pi$  in it! So we just drop in one of the claims observations for  $x$ , and one of the  $\xi, \sigma$  combinations from the 25 node grid and compute. There will be  $12 \times 25 = 300$  such calculations, much more effectively and efficiently performed in the spreadsheet.

For example, the probability of observing  $x = 101.00$  given a hypothesis that  $\xi = 1.00$  and  $\sigma = 12.00$  is 0.0710, found in the first calculation cell F5 of the table. Then we calculate the probability that the next claim occurs under the same condition, and so one until we get to the end of the 12 claims observations. We now have 12 probabilities that observations occur, all conditional, in this row, on the same hypothesis  $\xi$  and  $\sigma$ .

To calculate the probability that we see both observation 1 and observation 2 and, ..., observation 12, we multiply all of the probabilities together with the PRODUCT() function. We now find the probability both of the data,  $Pr(x | \xi, \sigma)$  and of the hypotheses,  $Pr(\xi)$  and  $Pr(\sigma)$ . the hypotheses probabilities are in columns E and F, We assume all hypotheses are equally likely until we happen to update our assumptions. We multiply the three probabilities together and now we have this expression all up and down column T.

But the next column U tells the story. We take the column T joint probabilities, sum them up to get the grand total probability both of all of the data and all of the hypotheses. We use this grand total probability to normalize each of the column T joint probabilities. How do we do that? In column U we divide each cell in column T by the grand total probability. This ends up computing the one thing we have been looking for all along, the probability of a particular hypothesis given the data. Column U is the contribution of each joint probability to the medley of mashing together data and conjectures about the data.

## 9.4 What does it all mean, so far?

Does our grid tell us anything useful? On its own it is not in a form easy to interpret. We have the raw  $Pr(\mu, \sigma | wages)$  in column U. We did build a key in column W above. Now is the time to put it to good use. We need to calculate the total probability of any particular  $\mu$  or  $\sigma$ . Here is the ultimate grid that relates each hypothesized  $\mu$  with each hypothesized  $\sigma$ . The link between them is the probability both of  $\mu$  and  $\sigma$ , that is,  $Pr(\mu, \sigma | wages)$  in column U.

Right in the cross-hairs is the maximum joint probability of  $\xi$  and  $\sigma$ . We

	A	B	C	D	E	F	G	H	I	J
1										
2										
3						sigma				
4				12	12.75	13.5	14.25	15	Pr( xi )	
5			1.00	0.013703	0.015093	0.016329	0.017399	0.018297	0.080820	
6			1.50	0.049081	0.050835	0.052005	0.052645	0.052817	0.257383	
7			2.00	0.058409	0.058799	0.058632	0.057997	0.056977	0.290814	
8			2.50	0.046253	0.045850	0.045101	0.044074	0.042833	0.224112	
9			3.00	0.030789	0.030245	0.029517	0.028647	0.027674	0.146872	
10			Pr(sigma)	0.1982	0.2008	0.2016	0.2008	0.1986		
11			C4:	D3:						
12			=grid-setup!E3	{=TRANSPOSE( 'grid-setup'!E10:E14 )}						
13				D5:						
14				=INDEX( Pr_h_d, MATCH( \$C5&"-"&D\$3, key, 0) )						
15						F9:			I5:	
16						=SUM(F4:F8)			=SUM(D5:H5)	
17										
18										

**FIGURE 9.5:** In the cross-hairs: variability and uncertainty.

might interpret these parameters as uncertainty and variability. In any case, they indicate the maximum probabilities of hypotheses and the value of those best hypotheses  $\xi = 2$  and  $\sigma = 13.5$ .

The marginal probability of  $\Pr(\xi = 2.00)$  is the highest density in the I column. We calculate it realizing that this probability must take into account any of the ways in which  $\xi = 12.00$  interacts jointly with each of the hypothesized  $\sigma$ 's. The key word in the last sentence is the indefinite pronoun *any*. This pronoun denotes an *either-or* proposition: either  $\sigma = 12$  or  $12.75$  or, ...,  $15$ . Either-or situations have probabilities that add up and thus the **SUM()** in cell I5.

Similarly the marginal probability of  $\Pr(\sigma = 13.5)$  is the highest density for the hypothesized  $\sigma$ s. This probability is also the sum of the either-or probabilities of  $\sigma = 10$  interacting jointly with any of the hypothesized  $\mu$ s. We often refer to this calculation as integrating out, in this case, the  $\mu$ s for each  $\sigma$ , and *vice-versa* for integrating out the  $\sigma$ s for each  $\mu$ .

Thus we end our estimation of the hypotheses which are most compatible with the claims data. We now have the ability to simulate the claims experience in a principled way.

## 9.5 The joint's ajumpin'

We know that  $\xi$  and  $\sigma$  are joined at the hip. The estimation table is a joint probability table. Let's investigate further. We will calculate the mean, variance (standard deviation), and covariance (correlation). In this way we can

generate claims with correlated shape and scale parameters exactly in the manner we did with waiting times using Cholesky Factorization.

In this table we first calculate the expectations of  $\xi$  and  $\sigma$  for means, standard deviations and covariances. These are weighted averages with probabilities and grid outcomes ported directly from the  $\xi \times \sigma$  table. The probabilities for the covariance calculation are the diagonal joint probabilities in that table. Since they do not add up to one we follow our standard practice by normalizing them using their sum.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2															
3		Pr(xi)	xi	dev	dev^2		Pr(sigma)	sigma	dev	dev^2		Pr(xi_i,sigma_j)	normalized	dev_xi*dev_sigma	
4		0.080820	1.00	-1.05	1.10		0.1982	12.0000	-1.5005	2.2515		0.0137	0.0703	1.5746	
5		0.257383	1.50	-0.55	0.30		0.2008	12.7500	-0.7505	0.5633		0.0508	0.2608	0.4123	
6		0.290814	2.00	-0.05	0.00		0.2016	13.5000	-0.0005	0.0000		0.0586	0.3008	0.0000	
7		0.224112	2.50	0.45	0.20		0.2008	14.2500	0.7495	0.5617		0.0441	0.2261	0.3377	
8		0.146872	3.00	0.95	0.90		0.1986	15.0000	1.4995	2.2485		0.0277	0.1420	1.4254	
9												sum_pr_xi_sigma	0.194919		
10															
11		B4:	C4:	D4:	E4:		G4:	H4:	I4:	J4:		L4:	M4:	N4:	
12		=xi-sigma-grid*I4	=xi-sigma-grid*I4	=C4 - mu_xi	=D4^2		[=TRANSPOSE(xi-sigma-grid*I4:H9)]	[=TRANSPOSE(xi-sigma-grid*I4:H9)]	=H4 - mu_sigma	=I4^2		=xi-sigma-grid*I4	=L4 / sum_pr_xi_sigma	=D4 *I4	
13															
14															

**FIGURE 9.6:** Simulating Xi and sigma jointly.

Using this calculation apparatus we can then calculate the means, standard deviations, covariance, and correlation that summarize the joint estimation of claims shape  $\xi$  and scale  $\sigma$ . Using these parameters we can then simulate correlated versions of  $\xi$  and  $\sigma$  for each of 6 funding years we will forecast for the self-insurance discussion.

We now have the parameters and the one run simulation of  $\xi$  and  $\sigma$  values for each of the forward years we will include in our burgeoning fund analysis. We also set a maximum claim size to keep our model from exploding.

It turns out we can generate GPD variates with this formula for  $\xi \neq 0$ .

$$Y = \mu + \frac{\sigma(U^{-\xi} - 1)}{\xi} \sim GPD(\mu, \sigma, \xi \neq 0)$$

If  $U$  is uniformly distributed on  $(0, 1]$ , we can use `RAND()` in our spreadsheet. Here is a revised one run simulation with simulated claims. Finally!

In fact this run looks a little like the experience of claims in the small sample. There are several threshold level claims and a nearly 2x claim to thicken the tails of this kurtotic distribution.

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1																	
2																	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	

FIGURE 9.7: Further setting up the xi and sigma simulation.

## 9.6 A sinking feeling?

The Vegan Workers Insurance Fund (VWIF, Fund) would be a monoline insurance fund Federally chartered, operating in New York State, with strategic partners chartered in Bermuda. The fund would specialize in vegan food industry worker compensation risk. As an insurance fund it considers the matching of claims exposures with investments needed to fund claims over a several year period. Premiums charged will depend on the riskiness of the funding, and the variability and size of claims over time. In this way the Fund operates similarly to any sinking fund where cash flow obligations occurring over several time periods must be met with a slate of investment returns.

As a first step, Fund investment analysts would propose a slate of potential funding opportunities while claims analysts determine potential exposures over time. The slate of potential investments reflect the monthly timing of expected cash flow in along with annual annual returns. At the same time, claims are cash flows determined by analysis of various risk drivers and exposures. Our model takes the more naive approach to jump-start the analysis of claims using the GPD to model exposures.

The fund would have a 7 year mandate to meet claims. A 7th year is required to manage overlaps in coverage among participants and restart the investment allocations. Here is the expected claims schedule based on data collected from a small sample of vegan establishments and projected for years 2 to 6. The first year is the funding period for this tranche of exposures.

The main constraint is to have investment surplus from proceeds equal claims requirements. The calculation of investment surplus  $S_t$  for each year  $t = 2, \dots, 7$ . A return  $r_{it}$  from investment  $i$  at year  $t$  accrues according to the schedule of years of cash flow for each investment. These abstractions become a bit clearer in this next panel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
3	Vegan Workers Insurance Fund																						
4																							
5																							
6																							
7																							
8																							
9																							
10																							
11																							
12																							
13																							
14																							
15																							
16																							
17																							
18																							
19																							
20																							
21																							
22																							
23																							
24																							
25																							
26																							
27																							
28																							
29																							
30																							
31																							
32																							
33																							
34																							
35																							
36																							
37																							
38																							
39																							
40																							
41																							

FIGURE 9.8: A model to design a monoline insurance fund.

Our objective is to minimize the front-end investment in year 1. In calculating premia for each policy-holder we can levelize payments according to the capital needed by year. This process is not shown here and would be quite a good exercise. For now we content ourselves with understanding this version of a funding model. In this model we add all of the investments that start with year 1 using the SUMIF() function.

We choose levels of capital to invest in each of the A, B, C, D funds. These funds have starting and ending years. Each then earns a different time-based return. We flow a \$1 dollar (-1) in a starting year across the holding period of the fund to earn at the end of the holding period the invested capital \$1 plus a return. The nested IF() statements establish these flows successfully. They only took a few fevered hours to concoct!

The **constraints** for each year calculate the investment surplus (insurance-speak for dollar returns) by summing up the products of the one plus returns and the amount of capital investment by fund. Investment surplus must equal claims. Claims in turn are the simulated claims from much previous and laborious analysis. To calculate investment surplus we multiply the amount invested with the return in each investment return year from years 2 through 7. This effort requires the use of the SUMPRODUCT() function.

The right-hand side (the bottom here) is the projected claims experience from the simulation of claims in another worksheet. Because Solver badly interacts (as would any interactive algorithm!) with volatile functions like OFFSET() and RAND() we must cut and paste special with values the projection into the claims constraint. Failure to do this will result in Solver errors like non-linear or binary constraints detected, when there aren't any in the model.<sup>5</sup>

<sup>5</sup>Solver linear programming uses a modified simplex algorithm, along with a generalized

We build several headers and data transpositions for input into the fund-sim worksheet all starting at row 28. Two buttons adorn this model. One executes a single run of the model, the other 100 runs of the model. Here is the Visual Basic for Applications script for one run of Solver.

```
Sub solve_capital_paste_special()
' solve_capital_paste_special Macro
'
'
    Sheets("xi-sigma-sim").Select
    Range("Z15").Select
    Range(Selection, Selection.End(xlToRight)).Select
    Selection.Copy
    Sheets("fund-allocation").Select
    Range("H20").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    SolverOk SetCell:="$E$19", MaxMinVal:=2, ValueOf:=0, ByChange:="$E$7:$E$18", _
        Engine:=2, EngineDesc:="Simplex LP"
    SolverOk SetCell:="$E$19", MaxMinVal:=2, ValueOf:=0, ByChange:="$E$7:$E$18", _
        Engine:=2, EngineDesc:="Simplex LP"
    SolverSolve (True)
End Sub
```

**FIGURE 9.9:** On run of the monoline fund design.

We record the key strokes in this macro by turning on the Record Macro feature in the Developer ribbon. The copy of the selection in the xi-sigma-sim worksheet can be optimized simply by naming the range. But it is instructive to see how Excel seems as if to think about all of the steps needed for a selection of a range. The paste utility works on a selection of the first cell, H20, in the fund-allocation sheet. The script will always need to have some way of identifying in which sheet a range resides. We can now invoke Solver and its settings. We can also add and delete constraints as well. All of this points to the way we often need to build VBA scripts: first and foremost just record key strokes, and second, review and revise the script as needed.

## 9.7 Simulate and stimulate

We accomplish multiple runs, with storage, of the funding simulation itself through this subroutine. We simulate several runs, 100 here, of the fund-allocation model in a separate spreadsheet. Because we need to automate the manual cut and copy paste special with values claims projections from

reduced gradient and genetic algorithms, all of which iterate to an approximation solution. The routines expect a stable numerical input, one that does not change with spreadsheet recalculations. Even if we set recalculation to manual, Solver will still react to underlying changes and produce the error.

the claims simulation into the fund allocation constraints, we must resort to VBA.

```
Sub fundMonteCarlo_Click()
    Dim R1 As Range, R2 As Range
    Dim iRun As Long
    Application.ScreenUpdating = False
    'Application.Calculation = xlCalculationManual
    Set R1 = Range("interface")
    Set R2 = Range("simulation")
    For iRun = 1 To 100
        solve capital_paste_special
        R2.Offset(iRun, 0).Value = R1.Value
    Next iRun
    'Application.Calculation = xlCalculationAutomatic
    Application.ScreenUpdating = True
    MsgBox "Congratulations! Fund allocation simulation completed ", vbOKOnly + vbInformation, "VWIFund simulation"
End Sub
```

**FIGURE 9.10:** A modified simulation automated flow.

Here is a snapshot of the simulator itself in its own worksheet.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2																						
3		A.1.2	B.1.3	C.1.4	D.1.7	A.2.3	A.3.4	B.3.5	A.4.5	C.4.7	A.5.6	B.5.7	A.6.7	capital_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr		
4		110.256	0	0	0	593.266	251.45	251.148	0	97.0468	151.406	0	0	110.256	112.241	101.346	158.929	108.532	154.132	102.67553		-fund-allocation/W33
5																						-fund-allocation/IJ31
6	run	A.1.2	B.1.3	C.1.4	D.1.7	A.2.3	A.3.4	B.3.5	A.4.5	C.4.7	A.5.6	B.5.7	A.6.7	capital_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr		T3
7	1	105.762	0	0	0	624.629	265.151	223.351	0	96.6511	130.284	0	0	105.762	107.666	147.371	173.273	100.883	132.63	102.25686		
8	2	161.41	0	0	0	604.591	300.435	209.81	0	95.0375	100.453	0	0	161.41	164.316	105.228	210.805	116.7	102.261	100.54963		
9	3	119.399	0	0	0	2860.86	2473.12	234.95	0	111.636	138.717	0	0	119.399	121.548	204.289	2406	104.456	141.214	118.11062		
10	4	131.047	0	0	0	660.086	211.097	354.67	0	113.095	229.219	0	0	131.047	133.406	106.201	101.801	137.864	233.345	119.655		
11	5	2363.46	0	0	0	3753.97	208.588	2656.14	0	105.482	357.58	0	0	2363.46	2406	956.815	106.86	2391.52	364.017	111.60024		
12	6	99.8711	0	0	0	2772.73	2479.2	242.434	0	117.821	132.554	0	0	99.8711	101.669	101.007	2406	118.365	134.94	124.65471		
13	7	101.769	0	0	0	2822.62	348.38	2421.2	0	117.203	99.9468	0	0	101.769	103.601	103.845	237.448	2406	101.746	124.00084		
14	8	171.596	0	0	0	2505.5	2125.24	322.232	0	472.308	113.109	0	0	171.596	174.685	103.128	1691.18	220.4	115.145	499.7018		
15	9	132.687	0	0	0	2699.33	222.198	2419.75	0	125.777	98.4405	0	0	132.687	135.076	105.97	100.421	2406	100.212	133.07168		
16	10	125.603	0	0	0	2784.66	2459.38	216.967	0	97.6483	102.182	0	0	125.603	127.864	158.44	2406	122.38	104.021	103.1119		
17	11	112.608	0	0	0	4257.1	1783.88	2447.78	0	1666.22	127.45	0	0	112.608	114.635	102.074	149.77	2406	129.744	1762.8585		
18	12	113.596	0	0	0	3325.36	3058.13	203.913	0	707.172	108.949	0	0	113.596	115.641	123.174	2406	102.101	110.91	748.18817		
19	13	117.346	0	0	0	2850.26	269.704	225.858	0	120.155	106.01	0	0	117.346	119.458	2406	154.404	127.753	107.918	127.12424		
20	14	394.726	0	0	0	3134.58	2691.78	342.416	0	2129.97	146.077	0	0	394.726	401.831	156.8	610.27	206.324	150.743	2253.505		
21	15	114.311	0	0	0	2740.44	249.319	2422.2	0	139.893	100.974	0	0	114.311	116.369	118.252	113.913	2406	102.792	148.00679		
22	16	109.694	0	0	0	1381.26	234.91	380.309	0	121.669	236.752	0	0	109.694	111.669	790.899	117.469	156.869	241.013	128.72581		
23	17	117.182	0	0	0	3028.36	336.134	340.734	0	241.076	101.687	0	0	117.182	119.291	2406	101.108	250.973	103.517	255.05886		
24	18	124.624	0	0	0	4027.32	2339.24	2494.51	0	2274.1	2363.46	0	0	124.624	126.867	182.268	107.242	218.359	2406	2406		
25	19	546.829	0	0	0	4814.21	2336.49	2395.23	0	2274.1	2363.46	0	0	546.829	556.672	169.153	104.442	115.604	2406	2406		
26	20	1176.10	0	0	0	608.006	136.436	1773.007	0	134.04	160.331	0	0	1176.10	1184.36	110.637	106.65	103.731	163.677	141.81453		

We generated quite a bit of data here! We named the range B4:T4 **interface** and the range header B5:T6 as **simulation**. The Offset() function in VBA will advance the row from 0 to the value of iRun to record each run's worth of results from the fund allocation model. We will additionally name the ranges of the header and data to perform the analysis for which we have all been not so patiently waiting.

To see the sausage being made in the fund-simulation worksheet we could place a ' single quote in front of the Application.ScreenUpdating = False line in the VBA code to comment this instruction out. We should have a beverage on hand for the viewing, or perhaps popcorn, as this action will take even more precious time away from other activities in the busy analyst's life.

Each run can takes up to 20 seconds or 3 runs per minute depending on available CPU availability and thus we only perform 100 runs where, perhaps, 10,000 might be preferable by some analysts. On a Lenovo IdeaPad from the factory, 10,000 runs could take up to 2.5 hours. A GPU multi-core processor would run such routines nearly in the blink of an eye.

### 9.8 Finally, exhaustively, where are the results?

With all of that derived data we generate, we owe ourselves results. We name each column of simulated data with the cell at the head of the vectors of length 100 we generated. We then reuse the 21 interval (yet another grid!) frequency analysis from the waiting time analysis.

First, here is the grid setup for the frequency table, along with a summary of the simulated data. We build a list box from the data validation feature in the Data ribbon to facilitate choices of variable to analyze. We show here the year 1 capital required to meet simulated claims.

	A	B	C	D	E	F
1						
2						
3		mu_grid	capital_yr1			
4		max_capital_yr1	2363.46		=MAX( INDIRECT(C\$3) )	
5		min_capital_yr1	98.44		=MIN( INDIRECT(C\$3) )	
6		intervals_capital_yr1	21.00		=21	
7		width_capital_yr1	107.86		=(C4 - C5) / C6	
8		count_capital_yr1	100.00		=SUM(K3:K23)	
9						
10			capital_yr1		=MAX( INDIRECT(C\$3) )	
11		min	98.44		=MIN( INDIRECT(C\$10))	
12		0.25	105.92		=PERCENTILE( INDIRECT(C\$3), \$B12 )	
13		0.50	117.98		=PERCENTILE( INDIRECT(C\$3), \$B13 )	
14		0.75	171.21		=PERCENTILE( INDIRECT(C\$3), \$B14 )	
15		max	2363.46		=MAX( INDIRECT(C\$10))	
16		IQR	65.30		=C14 - C12	
17		mu_capital_yr1	300.15		=AVERAGE( INDIRECT(C\$10) )	
18		sig_capital_yr1	511.34		=STDEV.S( INDIRECT(C\$10) )	
19		mad	284.62		=AVEDEV( INDIRECT(C\$10) )	
20		skewness	3.45		=SKEW( INDIRECT(C\$10) )	
21		kurtosis (+3)	14.29		=KURT( INDIRECT(C\$10) ) + 3	
22						
23						
24						

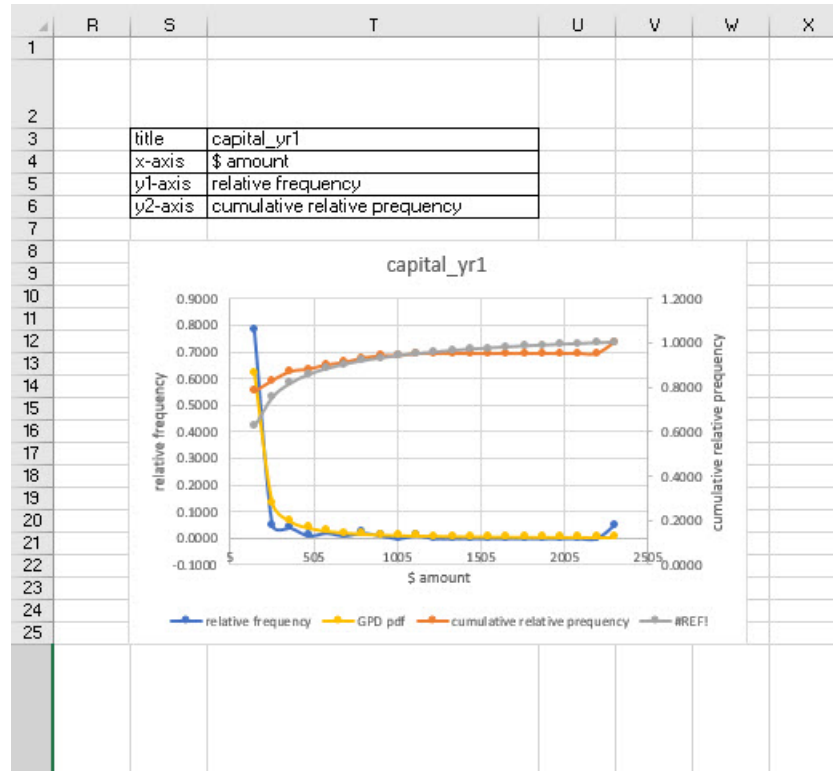
**FIGURE 9.11:** EDA for the claims simulation.

We cannot help but notice the whopping 14+ kurtosis. The volatility of volatility of simulated capital requirement is impressively large, as advertised by the GPD approach to tails.

Next, the table of frequencies of occurrence of capital values across 21 intervals. From these counts we calculate relative and cumulative relative frequencies. For comparison purposes we also compute the estimated GPD probability distribution function (from the mass function) and cumulative probability functions.



Last, and most importantly we have a plot.



Estimated are not exactly, but only generally following the simulation. We might call sampling error here, as we only simulated 100 variations of potential future decisions. We do have the characteristic GPD shape and scale inside gamma distributed exponential rates. What does it all mean? We have a highly variable capital requirement further necessitating risk management for extreme claims experience. An insurance analyst might remark that we need an excess of loss reinsurance treaty along with our general insurance coverage. Then there is the preventive maintenance and capital to support safer and healthier worker conditions to consider.



# 10

---

## *Case: Forecasting Workers Compensation Claims*

---

---

### 10.1 Some background

Make-A-Pie CEO and owner Simone Tortiere helped organize a group of like-minded owners of vegan food providers. Many of the vegan businesses are New York State Insurance Fund workers compensations policy-holders. For example, if a covered worker falls, has extended health issues, the fund will cover expenses for a specific time frame, using specified drug and other therapy treatments, all under the direction of NYSIF medical associates. Premiums will undoubtedly rise with more claims experience. Safety Group plans, while paying dividends to policy-holders with relatively low claims experiences, often groups all food-related workers into one class. This may, or may not, disadvantage the vegan food industry. The owners seek an alternative solution, at least to get a baseline view of the expense and exposure they face altogether.

---

### 10.2 The ask

The group decides on this course of action.

- Gather workers compensation claims experience across the group
- Attempt to model future claims to understand the range and shape of the distribution of claims, all based on existing claims experience
- Use the future claims model to simulate a self-insurance portfolio

All of this is a very tall order for experts in vegan food production and service. Tortiere has already talked to us about the issues. She recommends that our analytics service can at least start to structure the group's next conversation in a month with some provisional results.

### 10.3 Some requirements

The ask only includes the simulation of potential claims experience, not to simulate the self-insurance fund itself. Tortiere's colleagues offer some new data arrives for several years of workers compensation claims experience.

year	gross_claims
1	107
2	126
3	118
4	106
5	108
6	675
7	110
8	143
9	101
10	1203
11	875
12	103

1. Explore the data using summaries, Tukey's fences, and a time series plot. Observe possible thresholds.
2. Perform a mean excess of threshold plot. Identify possible threshold candidates. Be able to communicate the need for such an analysis to a decision maker.
3. Propose an appropriate distributional model that respects the thick tail nature of this data, including the threshold, scale, and shape of the distribution. Discuss its relevance and parameterization. Make a plot of the proposal.
4. Use a grid to approximate the joint probability of conjectured parameters and claims data of the proposed distribution.
5. Calculate summary statistics of the conjectured parameters including mean, standard deviation, and correlation using the distribution of conjectured parameters. Use these statistics to generate 12 years of correlated parameter values.
6. Simulate 10,000 runs of 12 years each run of claims data.
7. Analyze the sensitivity of anticipated claims using a plot of claims relative and cumulative relative frequency of occurrence in the simulation, along with a data summary including quantiles, mean, stan-

dard deviation, skewness, and kurtosis. Discuss the impact of these graphical and tabular results on projected claims experience.



---

## ***Part 4 – Optimal Decision Simulation***

---

Putting together

- Spreadsheet engineering practice
- Native Excel functions
- Probabilistic simulation on a grid
- Multi-variable optimization by simplex algorithm
- Plotting to communicate with a simple dashboard





# 11

---

## *Vegan Workers Insurance Fund: case and notes*

---

### 11.1 Tales of tails

Let's set the scene. Make-a-Pie CEO and owner Simone Tortiere meets regularly with other vegan food company owners. This month their agenda covers the rising incidence of and cost of covering workers compensation claims. A claim arises when a worker, on the job, becomes disabled for any reason. The New York State Insurance Fund (NYSIF) offers several plans and products to employer policy-holders. The vegan company owners around the table at host Candle Cafe try to get a handle on claims, rejection of claims, required drug formularies and treatment codes, increasing cost of disability, loss of key personnel, among other things. The COVID-19 pandemic alone has raised costs of production and service by over 30%. Workers comp claims have kept pace with this rise.

Many of the vegan businesses are NYSIF policy-holders. For example, if a covered worker falls, has extended health issues, the fund will cover expenses for a specific time frame, using specified drug and other therapy treatments, all under the direction of NYSIF medical associates. Premiums will undoubtedly rise with more claims experience. Safety Group plans, while paying dividends to policy-holders with relatively low claims experiences, often groups all food-related workers into one class. This may, or may not, disadvantage the vegan food industry.

The group decides on this course of action.

- Gather workers compensation claims experience across the group
- Attempt to model future claims to understand the range and shape of the distribution of claims, all based on existing claims experience
- Use the future claims model to simulate a self-insurance portfolio
- Suggest and implement two scenarios to sensitize the group to alternative approaches to funding

All of this is a very tall order for experts in vegan food production and service. Tortiere has already talked to her Bronx consultants Rose Mascetti and Marie Ortiz (Mascetti-Ortiz Analytics) about the issues. They have already provided

valuable advice and plans for production, pricing, and strategies for expansion of the business. Tortiere recommends to her colleagues that Mascetti-Ortiz Analytics service can at least start to structure the group's next conversation in a month with some provisional results. The vegan food employers appoint a sub-committee to commission the work. They call themselves the Vegan Workers Group.

## 11.2 It always starts with data

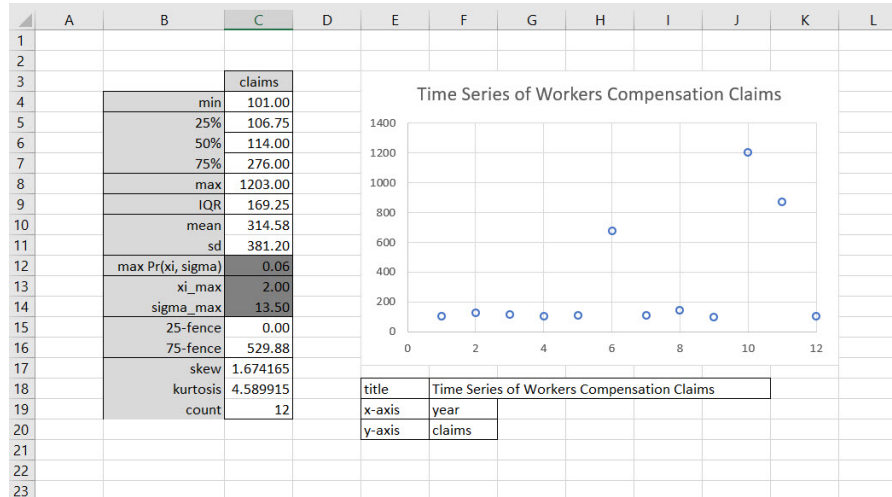
We work with Rose and Marie on their analytics team. Does the analysis always start with data? Not really, because we actually begin with Tortiere's question: can we insure ourselves and be better off? To help answer this question we begin our analysis by gathering 12 years of annual workers claim experience from several of the vegan establishments.

	A	B	C
1			
2		year	gross_claims
3		1	107
4		2	126
5		3	118
6		4	106
7		5	108
8		6	675
9		7	110
10		8	143
11		9	101
12		10	1203
13		11	875
14		12	103
15			
16			

**FIGURE 11.1:** The new claims data.

This is only a small sample of owners experience, all in hundreds of thousands of dollars. We visualize the times series.

Two patterns emerge. There seems to be a base-line of claims experiences. But

**FIGURE 11.2:** EDA for claims - again.

there are several very large experiences as well. We ask is there some sort of threshold that splits the analysis of claims experience?

We know from our business analytics experience that a well-known exploratory technique exists to examine the behavior of data, especially extreme values in this claims data. We see that Tukey (1977)'s outlier fences detect outliers above a threshold claims level of \$529,000. Another technique develops an excess of high threshold series. We generate a series of possible thresholds  $\mu$ . Then we sort the claims from lowest to highest. For each threshold and claim  $x$  we calculate the excess of threshold metric.

We use this tool to evaluate the threshold metric  $\mu$ .

$$\max(x, y) = \frac{x + y + |x - y|}{2}$$

We have some interns from a local Lasallian college. To help them understand the analysis we frequently use simplified examples. Let's try this tool out with  $x = 0$  and  $y = 5$ . We remember that we are trying to measure claims 5 in excess of a threshold of 0.

$$e(\mu) = \max(0, x - \mu) \quad (11.1)$$

$$= \frac{x + y + |x - y|}{2} \quad (11.2)$$

$$= \frac{0 + 5 + |0 - 5|}{2} \quad (11.3)$$

$$= \frac{5 + 5}{2} \quad (11.4)$$

$$= 10/2 \quad (11.5)$$

$$= 5 \quad (11.6)$$

$$= (5 - 0)_+ \quad (11.7)$$

This case seems to work. This works only because when  $x > y$ , then  $x - y < 0$  and  $|x - y| = y - x > 0$ . Here we flex our algebraic muscles to verify this relationship.

$$e(\mu) = \frac{0 + (x - \mu) + |0 - (x - \mu)|}{2} \quad (11.8)$$

$$= \frac{x - \mu + |-(x - \mu)|}{2} \quad (11.9)$$

$$\text{if } x > \mu \quad (11.10)$$

$$e(\mu) = \frac{x - \mu + x - \mu}{2} \quad (11.11)$$

$$= \frac{2x - 2\mu}{2} \quad (11.12)$$

$$= x - \mu \quad (11.13)$$

$$\text{if } x \leq \mu \quad (11.14)$$

$$e(\mu) = 0 \quad (11.15)$$

We can use this analysis to program spreadsheet cells. We use the subscript  $+$  to economize on notation. Next we average  $e(\mu)$  for each threshold. We remember from somewhere, maybe an elementary statistics course, the for  $u$  a constant that  $E\mu = \mu$ .

$$E e(\mu) = \sum_i^{N_\mu} \pi_i (x_i - \mu)_+ \quad (11.16)$$

$$= \pi_1 (x_1 - \mu)_+ \dots \pi_{N_\mu} (x_{N_\mu} - \mu)_+ \quad (11.17)$$

$$= E x - \mu \quad (11.18)$$

$$= \mu_x - \mu \quad (11.19)$$

We must remember not to confuse the  $\mu$  which is the threshold, and the expectation of an outcome (here claims)  $x$  which is  $\mu_x$ . On the other hand does the variance of  $e(\mu)$  have anything at all to do with  $\mu$ ?

First, if  $x > \mu$ , then,

$$\text{Var } e(\mu) = E[e(\mu) - E e(\mu)]^2 \quad (11.20)$$

$$= E[(x - \mu)^2 + (E x - \mu)^2 - 2(x - \mu)(E x - \mu)] \quad (11.21)$$

$$= E[x^2 - 2\mu x + \mu^2 + (E x)^2 - 2u E x + u^2 - 2x E x + 2\mu E x - \mu^2] \quad (11.22)$$

$$= E x^2 - (E x)^2 + \mu^2 - 2\mu E x \quad (11.23)$$

$$= \sigma_x^2 + \mu^2 - 2\mu \mu_x \quad (11.24)$$

We can use these calculations to build probability intervals around  $e(\mu)$  should the desire ever percolate to the surface.

Expectations build around a number  $N_u$  of non-zero  $e(\mu)$  values. It is possible that there are less than  $N$  observations for which a claim  $x$  only exceeded a threshold  $u$ . It is very possible that given a stream of claims, a very few set of claims which exceeds a high threshold will occur and thus we can state another rule of the excess over threshold story that  $N_u \leq N$ .

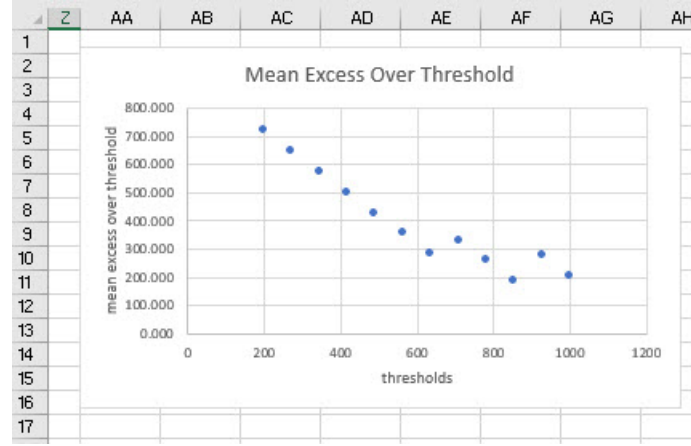
Here we calculate the mean excess of claims over a series of potential thresholds  $u$ .

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2		year	claims	claims	x - 200	x - 273	x - 345	x - 418	x - 491	x - 564	x - 636	x - 709	x - 782	x - 855	x - 927	x - 1000	threshold u grid				u	
3		9	101	101.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	max	1000		200.00
4		12	103	103.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	min	200		272.73
5		4	106	106.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	intervals	11		345.45
6		1	107	107.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	width	72.7273		418.18
7		5	108	108.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				490.91
8		7	110	110.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				563.64
9		3	118	118.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				636.36
10		2	126	126.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				709.09
11		8	143	143.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				781.82
12		6	675	675.000	475.000	402.273	329.545	256.818	184.091	111.364	38.636	0.000	0.000	0.000	0.000	0.000	0.000	0.000				854.55
13		11	875	875.000	675.000	602.273	529.545	456.818	384.091	311.364	238.636	165.909	93.182	20.455	0.000	0.000	0.000	0.000				927.27
14		10	1203	1203.000	1003.000	930.273	857.545	784.818	712.091	639.364	566.636	493.909	421.182	348.455	275.727	203.000						1000.00
15					u	200	272.727	345.455	418.182	490.909	563.636	636.364	709.091	781.818	854.545	927.273	1000					
16					mean	717.667	644.939	572.212	499.485	426.758	354.030	281.303	208.566	135.839	63.112	0.000						
17																						
18			claims		F4		G16															V16
19			lowest to		+*% "SPROUND" (%,0)		+MAX(0, +H4- G16)															+V15+1716
20			highest				G17															
21							IFERROR( POSSE(u,j)															
22							G18															
23							EF( G18,G19, "%0.0															
24																						

**FIGURE 11.3:** MEP for claims - again.

We notice that there are a lot of zeros in this table. That is a direct result from just considering only positive differences  $x - u$ . We use this idea to calculate a conditional mean with the AVERAGEIF() function. We generate the thresholds  $u$  on an equally spaced grid. What is  $N_u$ ? In column L after all of the zeros we are left with  $N_u = 3$ , and in column N  $N_u = 2$ . We look at Now for the plot of all of this handiwork.

The plot is a straight line until some turbulence occurs as the threshold approaches the maximum value of claims in the data. This usually indicates that the following distribution will work well to match this data.



**FIGURE 11.4:** MEP for claims- the eplot this time - again.

### 11.3 A distribution to remember

We need an observational model with these qualities:

1. It generates a regularly occurring ranges of outcomes.
2. On occasion the distribution generates very high outcomes.
3. It will be relatively easy to implement.

Well, we always hope for 3) but will get by with 1) and 2). Such a distribution is the Generalized Pareto Distribution (GPD) built on ideas by the Italian civil engineer, sociologist and economist Vilfredo Pareto at the turn of the last century. The very high outcomes he observed were wealth in the hands of an elite. For us, its just a lot of bad work days in a year, no elites are involved. The GPD has been used in quality control for manufacturing, epidemic outbreaks, severe weather, finance collapse, and political violence. We use it for workers compensation claims experience

Here is the probability distribution function. There are three parameters, location  $\mu$  is a threshold of claims  $x$ ,  $\sigma$  is the scale or dispersion of claims,  $\xi$  is the shape of claims that allow thick or thin tails in the distribution. That last quality is crucial. It marks a way to measure kurtosis in our observational model.

$$Pr(x | \mu, \sigma, \xi) = (1/\sigma)[1 + \xi(x - \mu)/\sigma]_+^{-(1+\xi)/\xi}$$

where  $\mu > 0$  is the location parameter (a known value, the threshold),  $\sigma > 0$

is Again we use the short hand notation  $h_+ = \max(h, 0)$ . If  $\xi < 0$ , then we are taking a root of a negative number in the square brackets. We must not! At least not here as this will entail complex numbers, a fate we do not want to tempt, again, at least here. If  $\xi$  should wander into a negative range, then this condition must hold true, or else, we have a singular model, and we will get a !NUM error in Excel.<sup>1</sup>

Here we isolate the  $x$  to get an idea of what range (mathematicians sometimes call this a support)  $x$  can take on before the model blows up.

$$1 - \xi(x - \mu)/\sigma < 0 \quad (11.25)$$

$$1 < \xi(x - \mu)/\sigma < 0 \quad (11.26)$$

$$\frac{\sigma}{\xi} + \mu < x \quad (11.27)$$

$$x > \frac{\sigma}{\xi} + \mu \quad (11.28)$$

In other words,  $x$  definitely needs to be greater than the  $\mu$  threshold. In general, the support is  $x > \mu$  for  $\xi > 0$ , and  $\mu < x < \mu - \sigma/\xi$  for  $\xi < 0$ . This will be our observational model. Such a model takes a  $\xi, \sigma$  combination as a conjecture, for a given level of threshold  $\mu$ , a hypothesis as a given, a condition.

The model computes the probability that a  $x$  occurs given this conjecture. Of course, we conjecture many combinations of the shape and scale parameters. Which one do we pick? The most plausible one. It turns out that the most plausible hypothesis is the hypothesis with the highest probability of choosing a particular hypothesis of  $\xi, \sigma$  given all of the data we have available, here 12 annual observations of claims.

How do we get to that probability? We simply find the probability that **both** the data occurs, given a hypothesis, **and** the probability that the hypothesis itself occurs. We calculate the probability both of the data in a model and of the probability of the hypothesis as the product of the probabilities. This is justly called the **Law of Conditional Probability**.

$$\Pr(x \wedge (\sigma, \xi)) = \Pr(x \mid \sigma, \xi) \Pr(\sigma) \Pr(\xi)$$

The little wedge  $\wedge$  means **and** and the  $\Pr(\mu) = 1$ , means we know  $\mu$  with certainty, after all we are setting it ourselves. But that does mean there will be some scrutiny to assure ourselves of the level of the threshold.

Yes, this a beast of a table. But while there is a lot going on here, we can start

---

<sup>1</sup>What distribution to choose? Taleb (2019) has a bit to say about the matter, and very technically, and practically, so as we thought about before. Here we will continue to work our way through a very thick tailed distribution. An interesting comparison presents itself: how would the thick-tailed approach compare with the conventional approach. The so-called conventional approach is, of course, to use a distribution which converges to a central location, namely a mean-reverting Gaussian distribution.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1																									
2																									
3																									
4																									
5																									
6																									
7																									
8																									
9																									
10																									
11																									
12																									
13																									
14																									
15																									
16																									
17																									
18																									
19																									
20																									
21																									
22																									
23																									
24																									
25																									
26																									
27																									
28																									
29																									
30																									
31																									
32																									
33																									
34																									
35																									
36																									
37																									
38																									

FIGURE 11.5: Xi and sigma grid approximation.

from left to right and top to bottom. This calculation machine is our claims-robot's cerebral cortex in its separate grid-approximation worksheet.<sup>2</sup> A grid on the left has all of the conjectured  $\xi$  shape and  $\sigma$  parameter combinations from the grid-setup worksheet. This table is pivotable, a task we perform in the next worksheet.

We can unpack the IF() statements in cells C6 and D6 by realizing that we must stay at the same  $\xi$  only while we loop through the list of  $\sigma$ s, otherwise move to the next  $\xi$ . Cells C5 and D5 start the parade with the beginning entries of the  $\xi$  and  $\sigma$  lists.

Cell C6 tests whether the previous D5 is the end of the  $\sigma$  list by using the MAX() function. If true then the INDEX(..., MATCH()) retrieves the next  $\xi$ , otherwise stay at the same  $\xi$  in cell C5.

At the same time, the IF() statement in D6 tests whether or not the  $\sigma$  in D5 is the last  $\sigma$  in the list. If true, then go back to the beginning of the  $\sigma$ , otherwise go to the next  $\sigma$  in the list.

We end up With 25 nodes, that is,  $5 \times 5 = 25$  hypotheses, in a grid. We can then proceed to use the GPD observational model, one fit for use with location,  $\mu$  is the threshold, shape, approximated by  $\xi$ , and scale,  $\sigma$ , approximated by  $\sigma$ .

Our next stop on the magical mystery tour is the mashing together of observed

<sup>2</sup>Grids, and techniques to approximate functions like our GPD, are popular in many fields, including option pricing in finance, queue measurement in operations, customer sentiment analysis in marketing, reservoir flow in hydrological engineering, and beam stress analysis in civil engineering. We already used grids to optimize Simone Tortiere's pie price in our first outing with decision models.



data with the unobserved data of hypotheses, all 25 combinations of set  $\mu$  and approximated  $\sigma$  and  $\xi$ . These hypothetical parameters turn up in the GPD observational model again, here for reference.

$$Pr(x | \xi, \sigma) = \frac{1}{\sigma} \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-(1+\xi)/\xi}$$

This distribution is less beastly than the Gaussian (normal) distribution, after all it does not have  $\pi$  in it! So we just drop in one of the claims observations for  $x$ , and one of the  $\xi, \sigma$  combinations from the 25 node grid and compute. There will be  $12 \times 25 = 300$  such calculations, much more effectively and efficiently performed in the spreadsheet.

For example, the probability of observing  $x = 101.00$  given a hypothesis that  $\xi = 1.00$  and  $\sigma = 12.00$  is 0.0710, found in the first calculation cell F5 of the table. Then we calculate the probability that the next claim occurs under the same condition, and so one until we get to the end of the 12 claims observations. We now have 12 probabilities that observations occur, all conditional, in this row, on the same hypothesis  $\xi$  and  $\sigma$ .

To calculate the probability that we see both observation 1 and observation 2 and, ..., observation 12, we multiply all of the probabilities together with the PRODUCT() function. We now find the probability both of the data,  $Pr(x | \xi, \sigma)$  and of the hypotheses,  $Pr(\xi)$  and  $Pr(\sigma)$ . the hypotheses probabilities are in columns E and F, We assume all hypotheses are equally likely until we happen to update our assumptions. We multiply the three probabilities together and now we have this expression all up and down column T.

But the next column U tells the story. We take the column T joint probabilities, sum them up to get the grand total probability both of all of the data and all of the hypotheses. We use this grand total probability to normalize each of the column T joint probabilities. How do we do that? In column U we divide each cell in column T by the grand total probability. This ends up computing the one thing we have been looking for all along, the probability of a particular hypothesis given the data. Column U is the contribution of each joint probability to the medley of mashing together data and conjectures about the data.

---

## 11.4 What does it all mean, so far?

Does our grid tell us anything useful? On its own it is not in a form easy to interpret. We have the raw  $Pr(\mu, \sigma | wages)$  in column U. We did build a key in column W above. Now is the time to put it to good use. We need to calculate the total probability of any particular  $\mu$  or  $\sigma$ . Here is the ultimate grid that

relates each hypothesized  $\mu$  with each hypothesized  $\sigma$ . The link between them is the probability both of  $\mu$  and  $\sigma$ , that is,  $\Pr(\mu, \sigma \mid \text{wages})$  in column U.

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										

**FIGURE 11.6:** Xi x sigma joint distribution and optimal values.

Right in the cross-hairs is the maximum joint probability of  $\xi$  and  $\sigma$ . These indicate the maximum probabilities of hypotheses and the value of those best hypotheses  $\xi = 2$  and  $\sigma = 13.5$ .

The marginal probability of  $\Pr(\xi = 2.00)$  is the highest density in the I column. We calculate it realizing that this probability must take into account any of the ways in which  $\xi = 12.00$  interacts jointly with each of the hypothesized  $\sigma$ 's. The key word in the last sentence is the indefinite pronoun *any*. This pronoun denotes an *either-or* proposition: either  $\sigma = 12$  or  $12.75$  or, ...,  $15$ . Either-or situations have probabilities that add up and thus the **SUM()** in cell I5.

Similarly the marginal probability of  $\Pr(\sigma = 13.5)$  is the highest density for the hypothesized  $\sigma$ s. This probability is also the sum of the either-or probabilities of  $\sigma = 10$  interacting jointly with any of the hypothesized  $\mu$ s. We often refer to this calculation as integrating out, in this case, the  $\mu$ s for each  $\sigma$ , and *vice-versa* for integrating out the  $\sigma$ s for each  $\mu$ .

Thus we end our estimation of the hypotheses which are most compatible with the claims data. We now have the ability to simulate the claims experience in a principled way.

## 11.5 The joint continues to jump

We reprise the forecasting work we already developed for claims experience. The estimation table is a joint probability table. Let's investigate further. We will calculate the mean, variance (standard deviation), and covariance (correlation). In this way we can generate claims with correlated shape and scale parameters exactly in the manner we did with waiting times using Cholesky Factorization.

In this table we first calculate the expectations of  $\xi$  and  $\sigma$  for means, standard deviations and covariances. These are weighted averages with probabilities and grid outcomes ported directly from the  $\xi \times \sigma$  table. The probabilities for the covariance calculation are the diagonal joint probabilities in that table. Since they do not add up to one we follow our standard practice by normalizing them using their sum.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2															
3		Pr(xi)	xi	dev	dev^2		Pr(sigma)	sigma	dev	dev^2		Pr(xi,sigma_ij)	normalized	dev_xi*dev_sigma	
4		0.080820	1.00	-1.05	1.10		0.1982	12.0000	-1.5005	2.2515		0.0137	0.0703	1.5746	
5		0.257383	1.50	-0.55	0.30		0.2008	12.7500	-0.7505	0.5633		0.0508	0.2608	0.4123	
6		0.290814	2.00	-0.05	0.00		0.2016	13.5000	-0.0005	0.0000		0.0586	0.3008	0.0000	
7		0.224112	2.50	0.45	0.20		0.2008	14.2500	0.7495	0.5617		0.0441	0.2261	0.3377	
8		0.146872	3.00	0.95	0.90		0.1986	15.0000	1.4995	2.2485		0.0277	0.1420	1.4254	
9											sum_pr_xi_sigma	0.194919			
10															
11		B4:	C4:	D4:	E4:		G4:	H4:	I4:	J4:		L4:	M4:	N4:	
12		=xi-sigma-grid!I4	=xi-sigma-grid!C4	=C4 - mu_xi	=D4^2		=TRANSPOSE(xi-sigma-grid!D9:H9)	=TRANSPOSE(xi-sigma-grid!D3:H3)	=H4 - mu_sigma	=I4^2		=xi-sigma-grid!L4	=L4 / sum_pr_xi_sigma	=D4 * I4	
13															
14															

**FIGURE 11.7:** Begins the xi-sigma simulation.

Using this calculation apparatus we can then calculate the means, standard deviations, covariance, and correlation that summarize the joint estimation of claims shape  $\xi$  and scale  $\sigma$ . Using these parameters we can then simulate correlated versions of  $\xi$  and  $\sigma$  for each of 6 funding years we will forecast for the self-insurance discussion.

We now have the parameters and the one run simulation of  $\xi$  and  $\sigma$  values for each of the forward years we will include in our burgeoning fund analysis. We also set a maximum claim size to keep our model from exploding.

It turns out we can generate GPD variates with this formula for  $\xi \neq 0$ .

$$Y = \mu + \frac{\sigma(U^{-\xi} - 1)}{\xi} \sim GPD(\mu, \sigma, \xi \neq 0)$$

[illegible]

In fact this run looks a little like the experience of claims in the small sample. There are several threshold level claims and a nearly 2x claim to thicken the tails of this kurtotic distribution.

The Vegan Workers Insurance Fund (VWIF, Fund) would be a monoline insurance fund Federally chartered, operating in New York State, with strategic partners chartered in Bermuda. The fund would specialize in vegan food industry worker compensation risk. As an insurance fund it considers the matching of claims exposures with investments needed to fund claims over a several year period. Premiums charged will depend on the riskiness of the funding, and the variability and size of claims over time. In this way the Fund operates similarly to any sinking fund where cash flow obligations occurring over several time periods must be met with a slate of investment returns.

The fund would have a 7 year mandate to meet claims. A 7th year is required to manage overlaps in coverage among participants and restart the investment allocations. Here is the expected claims schedule based on data collected from

a small sample of vegan establishments and projected for years 2 to 6. The first year is the funding period for this tranche of exposures.

The main constraint is to have investment surplus from proceeds equal claims requirements. The calculation of investment surplus  $S_t$  for each year  $t = 2, \dots, 7$ . A return  $r_{it}$  from investment  $i$  at year  $t$  accrues according to the schedule of years of cash flow for each investment. These abstractions become a bit clearer in this next panel. Here we repurpose a linear programming model proposed by ? (p. 94) to match cash flow requirements with funding.

**FIGURE 11.9:** Claims forecast as input to the monoline design.

Our objective is to minimize the front-end investment in year 1. In calculating premia for each policy-holder we can levelize payments according to the capital needed by year. This process is not shown here and would be quite a good exercise. For now we content ourselves with understanding this version of a funding model. In this model we add all of the investments that start with year 1 using the SUMIF() function.

We choose levels of capital to invest in each of the A, B, C, D funds. These funds have starting and ending years. Each then earns a different time-based return. We flow a \$1 dollar (-1) in a starting year across the holding period of the fund to earn at the end of the holding period the invested capital \$1 plus a return. The nested IF() statements establish these flows successfully. They only took a few fevered hours to concoct!

The **constraints** for each year calculate the investment surplus (insurance-speak for dollar returns) by summing up the products of the one plus returns and the amount of capital investment by fund. Investment surplus must equal claims. Claims in turn are the simulated claims from much previous and laborious analysis. To calculate investment surplus we multiply the amount invested

with the return in each investment return year from years 2 through 7. This effort requires the use of the SUMPRODUCT() function.

The right-hand side (the bottom here) is the projected claims experience from the simulation of claims in another worksheet. Because Solver badly interacts (as would any interactive algorithm!) with volatile functions like OFFSET() and RAND() we must cut and paste special with values the projection into the claims constraint. Failure to do this will result in Solver errors like non-linear or binary constraints detected, when there aren't any in the model.<sup>3</sup>

We build several headers and data transpositions for input into the fund-sim worksheet all starting at row 28. Two buttons adorn this model. One executes a single run of the model, the other 100 runs of the model. Here is the Visual Basic for Applications script for one run of Solver.

```
Sub solve_capital_paste_special()
' solve_capital_paste_special Macro
'
'
    Sheets("xi-sigma-sim").Select
    Range("Z15").Select
    Range(Selection, Selection.End(xlToRight)).Select
    Selection.Copy
    Sheets("fund-allocation").Select
    Range("H20").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    SolverOk SetCell:="$E$19", MaxMinVal:=2, ValueOf:=0, ByChange:="$E$7:$E$18", _
        Engine:=2, EngineDesc:="Simplex LP"
    SolverOk SetCell:="$E$19", MaxMinVal:=2, ValueOf:=0, ByChange:="$E$7:$E$18", _
        Engine:=2, EngineDesc:="Simplex LP"
    SolverSolve (True)
End Sub
```

**FIGURE 11.10:** A dry run of fund design.

We record the key strokes in this macro by turning on the Record Macro feature in the Developer ribbon. The copy of the selection in the xi-sigma-sim worksheet can be optimized simply by naming the range. But it is instructive to see how Excel seems as if to think about all of the steps needed for a selection of a range. The paste utility works on a selection of the first cell, H20, in the fund-allocation sheet. The script will always need to have some way of identifying in which sheet a range resides. We can now invoke Solver and its settings. We can also add and delete constraints as well. All of this points to the way we often need to build VBA scripts: first and foremost just record key strokes, and second, review and revise the script as needed.

<sup>3</sup>The Solver linear programming add-in uses a modified simplex algorithm, along with generalized reduced gradient and genetic algorithms, all of which iterate to an approximation solution. The routines expect a stable numerical input, one that does not change with spreadsheet recalculations. Even if we set recalculation to manual, Solver will still react to underlying changes and produce an error.

## 11.7 Simulate and stimulate

We accomplish multiple runs, with storage, of the funding simulation itself through this subroutine. We simulate several runs, just 100 here to get an idea of the direction of computation, of the fund-allocation model in a separate spreadsheet. Because we need to automate the manual cut and copy paste special with values claims projections from the claims simulation into the fund allocation constraints, we must resort to VBA.

```
Sub fundMonteCarlo_Click()
    Dim R1 As Range, R2 As Range
    Dim iRun As Long
    Application.ScreenUpdating = False
    'Application.Calculation = xlCalculationManual
    Set R1 = Range("interface")
    Set R2 = Range("simulation")
    For iRun = 1 To 100
        solve capital_paste_special
        R2.Offset(iRun, 0).Value = R1.Value
    Next iRun
    'Application.Calculation = xlCalculationAutomatic
    Application.ScreenUpdating = True
    MsgBox "Congratulations! Fund allocation simulation completed ", vbOKOnly + vbInformation, "VWIFund simulation"
End Sub
```

FIGURE 11.11: Automating the simulation work flow for fund design.

Here is a snapshot of the simulator itself in its own worksheet.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2																						
3		A.1.2	B.1.3	C.1.4	D.1.7	A.2.3	A.3.4	B.3.5	A.4.5	C.4.7	A.5.6	B.5.7	A.6.7	capital_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	=fund-allocation!W33
4		110.256	0	0	0	593.266	251.45	251.148	0	97.0468	151.406	0	0	110.256	112.241	101.346	158.929	108.532	154.132	102.67553		=fund-allocation!V31
5																						
6	run	A.1.2	B.1.3	C.1.4	D.1.7	A.2.3	A.3.4	B.3.5	A.4.5	C.4.7	A.5.6	B.5.7	A.6.7	capital_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	claims_vr	=T3
7	1	105.762	0	0	0	624.629	265.151	223.351	0	96.6511	130.284	0	0	105.762	107.666	147.371	173.273	100.883	132.63	102.25686		
8	2	161.41	0	0	0	604.591	300.425	209.81	0	95.0375	100.453	0	0	161.41	164.316	105.238	210.805	116.7	102.261	100.54963		
9	3	119.399	0	0	0	2860.86	2473.12	234.95	0	111.636	138.717	0	0	119.399	121.548	204.289	2406	104.456	141.214	118.11062		
10	4	131.047	0	0	0	660.086	211.097	354.67	0	113.095	229.219	0	0	131.047	133.406	106.201	101.801	137.864	233.345	119.655		
11	5	2362.46	0	0	0	3753.97	208.588	2656.14	0	105.482	357.58	0	0	2362.46	2406	956.815	106.86	2391.52	964.017	111.60024		
12	6	99.8711	0	0	0	2772.73	2479.2	242.434	0	117.821	132.554	0	0	99.8711	101.669	101.007	2406	118.363	134.94	124.65471		
13	7	101.769	0	0	0	2822.62	348.38	2421.2	0	117.203	99.9468	0	0	101.769	103.601	103.845	237.448	2406	101.746	124.0084		
14	8	171.596	0	0	0	2505.5	2125.24	322.232	0	472.308	113.109	0	0	171.596	174.685	103.128	1691.18	220.4	115.145	499.7018		
15	9	132.687	0	0	0	2699.33	222.198	2419.75	0	125.777	98.4405	0	0	132.687	135.076	105.97	100.421	2406	100.212	233.07168		
16	10	125.603	0	0	0	2784.66	2459.38	216.967	0	97.6483	102.182	0	0	125.603	127.864	158.44	2406	122.38	104.021	103.3119		
17	11	112.608	0	0	0	4257.1	1783.88	2447.78	0	1666.22	127.45	0	0	112.608	114.635	102.074	149.77	2406	129.744	1762.8585		
18	12	113.596	0	0	0	3125.36	3058.13	203.913	0	707.172	108.949	0	0	113.596	115.641	123.174	2406	102.101	110.91	748.18817		
19	13	117.346	0	0	0	2850.26	269.704	225.858	0	120.155	106.01	0	0	117.346	119.458	2406	154.404	127.753	107.918	277.12424		
20	14	394.726	0	0	0	3134.58	2691.78	342.416	0	2129.97	148.077	0	0	394.726	401.831	156.8	610.27	206.324	150.743	2253.505		
21	15	114.311	0	0	0	2740.44	249.319	2422.2	0	139.893	100.974	0	0	114.311	116.369	118.252	113.913	2406	102.792	148.00679		
22	16	109.694	0	0	0	1381.26	234.91	380.309	0	121.669	236.752	0	0	109.694	111.669	790.899	117.469	156.869	241.013	128.72581		
23	17 <td>117.182</td> <td>0</td> <td>0</td> <td>0</td> <td>3028.36</td> <td>336.134</td> <td>340.734</td> <td>0</td> <td>241.076</td> <td>101.687</td> <td>0</td> <td>0</td> <td>117.182</td> <td>119.291</td> <td>2406</td> <td>101.108</td> <td>250.973</td> <td>105.517</td> <td>255.05886</td> <td></td> <td></td>	117.182	0	0	0	3028.36	336.134	340.734	0	241.076	101.687	0	0	117.182	119.291	2406	101.108	250.973	105.517	255.05886		
24	18	124.624	0	0	0	4927.32	2339.24	2494.51	0	2274.1	2363.46	0	0	124.624	126.867	182.268	107.242	218.359	2406	2406		
25	19	546.829	0	0	0	4814.21	2336.49	2395.23	0	2274.1	2363.46	0	0	546.829	556.672	169.153	104.442	115.604	2406	2406		
26	20	1176.10	0	0	0	600.096	736.636	772.007	0	134.64	140.321	0	0	1176.10	1196.46	119.637	106.45	103.721	142.91453			

FIGURE 11.12: Alas! The simulation of so many fund designs.

We generated quite a bit of data here! We named the range B4:T4 **interface** and the range header B5:T6 as **simulation**. The Offset() function in VBA will advance the row from 0 to the value of iRun to record each run's worth of results from the fund allocation model. We will additionally name the ranges of the header and data to perform the analysis for which we have all been not so patiently waiting.

To see the (vegan) sausage being made in the fund-simulation worksheet we could place a ' single quote in front of the Application.ScreenUpdating = False line in the VBA code to comment this instruction out. We should have

a beverage on hand for the viewing, or perhaps popcorn as needed, as this action will take even more precious time away from other activities in the busy analyst's life.

Each run can takes up to 20 seconds or 3 runs per minute depending on available CPU availability and thus we only perform 100 runs where, perhaps, 10,000 might be preferable by some analysts. On a Lenovo IdeaPad from the factory, 10,000 runs could take up to 2.5 hours. A GPU multi-core processor would run such routines nearly in the blink of an eye.

---

### 11.8 Finally, exhaustively, where are the results?

With all of that derived data we generate, we owe ourselves results. We name each column of simulated data with the cell at the head of the vectors of length 100 we generated. We then reuse the 21 interval (yet another grid!) frequency analysis from the waiting time analysis.

First, here is the grid setup for the frequency table, along with a summary of the simulated data. We build a list box from the data validation feature in the Data ribbon to facilitate choices of variable to analyze. We show here the year 1 capital required to meet simulated claims.

We cannot help but notice the whopping 14+ kurtosis. The volatility of volatility of simulated capital requirement is impressively large, as advertised by the GPD approach to tails.

Next, the table of frequencies of occurrence of capital values across 21 intervals. From these counts we calculate relative and cumulative relative frequencies. For comparison purposes we also compute the estimated GPD probability distribution function (from the mass function) and cumulative probability functions.

Last, and most importantly we have a plot.

Estimated first year capital levels, again using the SUMIF() function, are not exactly, but only generally following the simulation. We might call this an example of sampling error here, as we only simulated 100 variations of potential future decisions, just as Stanislaw Ulam did in simulating hands of solitaire. We do have the characteristic GPD shape and scale inside gamma distributed exponential rates. What does it all mean? We have a highly variable capital requirement further necessitating risk management for extreme claims experience. An insurance analyst might remark that we need an excess of loss reinsurance treaty along with our general insurance coverage. Then there is the preventive maintenance and capital to support safer and healthier worker conditions to consider.



	A	B	C	D	E	F
1						
2						
3		mu_grid	capital_yr1			
4		max_capital_yr1	2363.46		=MAX( INDIRECT(C\$3) )	
5		min_capital_yr1	98.44		=MIN( INDIRECT(C\$3) )	
6		intervals_capital_yr1	21.00		=21	
7		width_capital_yr1	107.86		=(C4 - C5) / C6	
8		count_capital_yr1	100.00		=SUM(K3:K23)	
9						
10			capital_yr1		=MAX( INDIRECT(C\$3) )	
11		min	98.44		=MIN( INDIRECT(C\$10) )	
12		0.25	105.92		=PERCENTILE( INDIRECT(C\$3), \$B12 )	
13		0.50	117.98		=PERCENTILE( INDIRECT(C\$3), \$B13 )	
14		0.75	171.21		=PERCENTILE( INDIRECT(C\$3), \$B14 )	
15		max	2363.46		=MAX( INDIRECT(C\$10) )	
16		IQR	65.30		=C14 - C12	
17		mu_capital_yr1	300.15		=AVERAGE( INDIRECT(C\$10) )	
18		sig_capital_yr1	511.34		=STDEV.S( INDIRECT(C\$10) )	
19		mad	284.62		=AVEDEV( INDIRECT(C\$10) )	
20		skewness	3.45		=SKEW( INDIRECT(C\$10) )	
21		kurtosis (+3)	14.29		=KURT( INDIRECT(C\$10) ) + 3	
22						
23						
24						

FIGURE 11.13: Making some initial sense of the simulations.

## 11.9 Another scenario emerges

Some of the members of the Vegan Workers Group wonder, and so do we, if the funding tenors matter. A funding tenor is the length of time between investment and return. Some of the tenors are short-term, such as the A fund. Others, like the D fund are longer term. What if we constrained funding to only short term tenors and use just the A fund? We can then compare the all-in (A through D) scenario with the short-term financing (only one year terms) scenario. Then we might at least have a true decision alternative.

One tack we might take is to set all but the one year funding vehicles to start and end with year 8. This will allow the model effectively to set these funds to zero. Another way way we could use is to set the decision cells equal to zero in the solver formulation. Here is a snapshot of the first approach.

The main advantage of this approach is that we can create constraints without reprogramming the solver model. The three-nested IF() statements do all of the work to detail the investment program over the life of claims.

We then run the model once to have this solution.

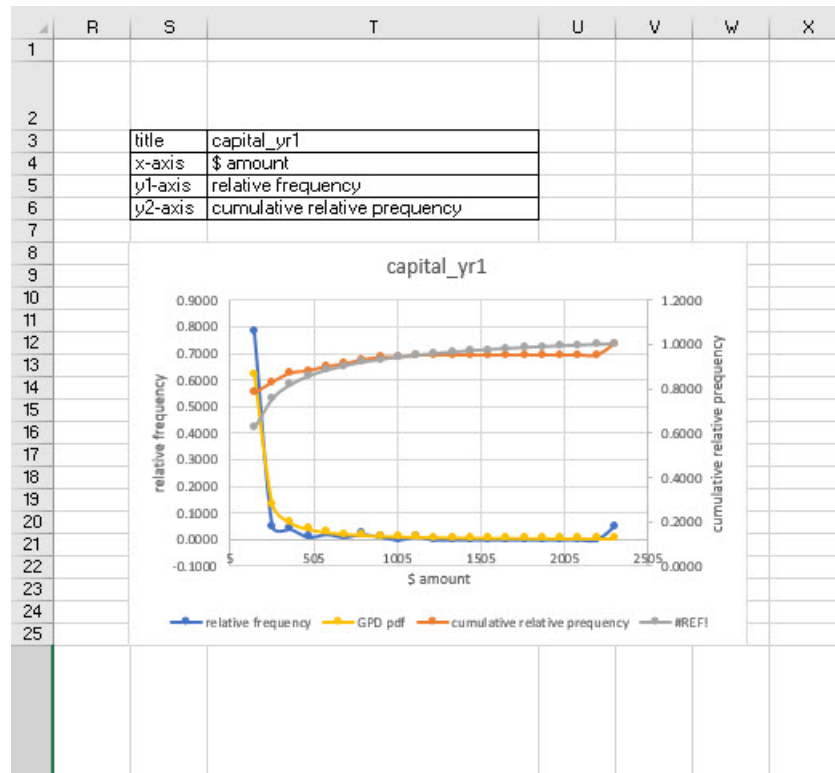


FIGURE 11.14: Setting up a more useful plot.

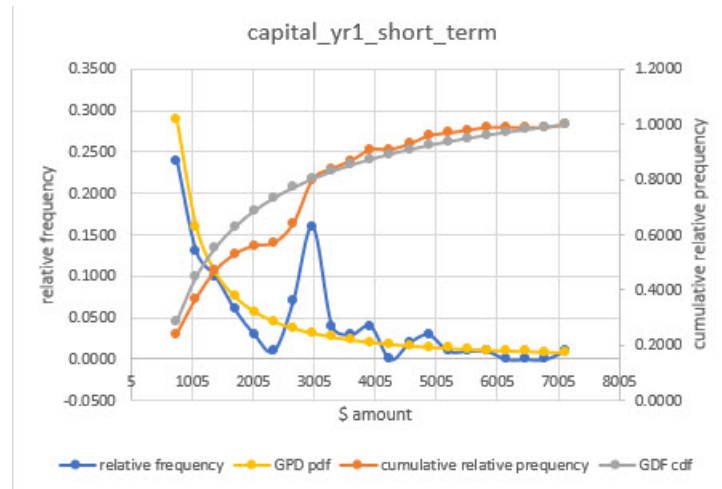
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
2																					
3																					
4																					
5																					
6																					
7																					
8																					
9																					
10																					
11																					
12																					
13																					
14																					
15																					
16																					
17																					
18																					
19																					
20																					
21																					
22																					
23																					
24																					
25																					
26																					

FIGURE 11.15: Trying another fund design scenario for comparison and sensitivity.



	capital_yr1_short_term	capital_yr1_all
min	588.67	98.44
25%	588.67	98.44
50%	588.67	98.44
75%	588.67	98.44
max	7269.73	2363.46
IQR	0.00	0.00
mean	2178.06	300.15
sd	1437.93	511.34
skew	1.016008512	3.446786701
kurtosis	3.765472038	14.28672228
count	100	100

**FIGURE 11.18:** An EDA comparison of the two scenarios.



**FIGURE 11.19:** Even better: a plot emerges.

term for volatile. Literally, the short-term vehicle only scenario also irons out the tail. It reduces the volatility of the standard deviation of the funding required in year one. But it does so at the expense of increasing the standard deviation of funding and the average funding required.

Multiple objectives seem to collide here. Only the preferences for risk and return of the Vegan Workers Group can resolve the deadlock. That determination is beyond the capabilities of this model. Perhaps a goal-programming approach might help the group discern which path to follow?

---

## ***Bibliography***

---

- Dantzig, G. B. (1949). Programming of interdependent activities: Ii mathematical model. *Econometrica*, 17(3/4):200–211.
- de Laplace, P., Truscott, F., and Emory, F. (1902). *A Philosophical Essay on Probabilities*. A Philosophical Essay on Probabilities. Wiley.
- Ghosh, S. and Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications*, 120(8):1492–1517.
- Hamilton, R. (1777). *Introduction to Merchandize*.
- Harris, F. W. (1913). How many parts to make at once. *Factory, The Magazine of Management*, 10(2):135–136, 152.
- ibn Mūsā al Khwārizmī, M. (1881). *The Compendious Book on Calculation by Completion and Balancing, translated as Algebra of Mohammed ben Musa by Frederic Rosen*.
- Labelle, A., Langevin, A., and Campbell, J. (2002). Sector design for snow removal and disposal in urban areas. *Socio-Economic Planning Sciences*, 36:183–202.
- Lonergan, B. J. (1970). *Insight: A Study of Human Understanding*. Philosophical Library.
- Taleb, N. N. (2019). How much data do you need? an operational, pre-asymptotic metric for fat-tailedness. *International Journal of Forecasting*, 35(2):677–686.
- Thomas Haigh, Mark Priestley, C. R. (2014). Los alamos bets on eniac: Nuclear monte carlo simulations, 1947–1948. *IEEE Annals of the History of Computing*, (July-August).
- Tukey, J. (1977). *Exploratory Data Analysis*.
- Winston, W. L. (2019). *Microsoft Excel 2019: Data Analysis and Business Modeling, 6th edition*. Microsoft Press.

