What matters and what doesn’t: building pragmatic and robust simulation models

Dr. Francisco J Zagmutt
Managing Partner
EpiX Analytics
fzagmutt@epixanalytics.com

www.epixanalytics.com
Why do we build models

- Decision support
- Deal with what’s available (data and knowledge) – bosses don’t like data wish lists
- Therefore, must make **assumptions** and **simplifications**
- How do we accomplish this while providing insightful answers?
What we will discuss today

Two common mistakes can impact a model usefulness:

• Methodological errors (doing it wrong)
• Overly complex models (lack of parsimony)

Complex models:

• More data
• Higher error rate
• Can be intractable - -> cloud rather than inform decision

Mistakes affect spread and shape of outputs -> statistics e.g. percentiles and variance.

Impact on decision supported by model from irrelevant to fatal’

Therefore, key to get them right…
ModelAssist for @RISK

Free simulation and @RISK training and reference tool:  
http://www.epixanalytics.com/ModelAssist.html

- Based on EpiX Analytics’ decades of experience in risk analysis consulting, training, and research

- Option to install locally or visit online

- Page numbers are Mxxxx. For example, M0407 is “Selecting the appropriate distributions for your model”

- Keep an eye on these during this talk!
A motivational example

You are in charge of predicting next year’s total borrowed money for the consumer lending division in a bank.

Making lots of assumptions, a very simplistic calculation could be:

\[
\text{Expected total} \_{\text{Borrowed}} = \text{mean # of loans/year} \times \text{mean loan size} + \ldots
\]

…however, the number of individual loans and the loan size is highly variable (and uncertain for the future).

How do we incorporate this variability in our calculation above?
We hire a marketing consultancy, and after 4.2M in fees, they conclude that:

- # loans/year is Pert(450,650,800)
- Loan size is $\text{Lognormal}(1000,4000)$

Therefore, we could use @RISK to “stochastize” prior calculation using:

$$\text{Expected total}_{\text{Borrowed}} = \text{Pert}(450,650,800) \times \text{Lognormal}(1000,4000) + \ldots$$

Let's see how this model works…
Let’s visualize what our model does

<table>
<thead>
<tr>
<th>Iteration</th>
<th># loans</th>
<th>$ per loan</th>
<th>Total borrowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>550</td>
<td>28,126</td>
<td>15,469,300</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between iterations and loan values.](image)
Let’s visualize what our model does

<table>
<thead>
<tr>
<th>Iteration</th>
<th># loans</th>
<th>$ per loan</th>
<th>Total borrowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>550</td>
<td>X</td>
<td>28,126</td>
</tr>
</tbody>
</table>

© EpiX Analytics LLC
Let’s visualize what our model does

<table>
<thead>
<tr>
<th>Iteration</th>
<th># loans</th>
<th>$ per loan</th>
<th>Total borrowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>550</td>
<td>X 28,126</td>
<td>15,469,300</td>
</tr>
<tr>
<td>2</td>
<td>770</td>
<td>X 6,307</td>
<td>4,856,390</td>
</tr>
</tbody>
</table>

Are these calculations correct? Model says that for each scenario, **every client** borrows the same amount!! How would this affect your model?
Technical aspects that we will cover today

1. When a multiplication is a sum
2. Confusing variability and uncertainty
3. Dependencies matter
4. Separate inputs from calculations
5. Starting simple
6. The power of distribution identities
1. When a multiplication is a sum (M0089)

Multiplications can be a shortcut to sum identical numbers
e.g. $10 \times 3 = 10 + 10 + 10$

However, random numbers are not identical.
E.g. the sum of $n \text{Lognormal}(10,4) \neq \text{Logn}(10,4) \times n$

• Impact: gross overestimation of output variance.

• Correct modeling (M0435):
  • Simulate $n$ distributions individually, then sum them e.g. RiskCompound
  • Use identities (e.g. $\text{Binomial}(1,p) + \text{Binomial}(1,p) = \text{Binomial}(2,p)$)
  • Use CLT approximation (e.g. $\text{N}(\mu,\sigma) + \text{N}(\mu,\sigma) = \text{N}(2\mu,\sqrt{2}\sigma)$)
  • Use actuarial methods (FFT, Panjer, DePril)

Let’s see how to solve our banking problem, now also calculating revenue assuming 7.5% IR
Correct approaches

<table>
<thead>
<tr>
<th>Approach 1: Central Limit Theorem:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clients:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>671 n=Pert(450,650,800)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total money borrowed</td>
</tr>
<tr>
<td>$ 6,730,198</td>
</tr>
<tr>
<td>Total $=N(n*μ, σ√n)</td>
</tr>
<tr>
<td>Mean $6,713,667</td>
</tr>
<tr>
<td>St.Dev $103,643</td>
</tr>
<tr>
<td>Total revenues (approach 1):</td>
</tr>
<tr>
<td>$ 504,765</td>
</tr>
<tr>
<td>R=Total $*IR</td>
</tr>
</tbody>
</table>

Approach 1 (M0075): Use the CLT $N(n^*μ, σ√n)$ to calculate total money borrowed, given n random # of clients. Then multiply by IR.

[Multiplication.xlsx](Multiplication.xlsx)
• Both correct approaches give the same answer
• The incorrect approach greatly overestimates the total risk ($\sigma_{\text{incorrect}} = 4 \times \text{correct method}$!)

Does this matter?

Comparison of three methods...

<table>
<thead>
<tr>
<th>Values in Millions...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
</tr>
<tr>
<td>0.3168</td>
</tr>
<tr>
<td>0.6335</td>
</tr>
<tr>
<td>0.9503</td>
</tr>
<tr>
<td>1.2670</td>
</tr>
<tr>
<td>1.5838</td>
</tr>
</tbody>
</table>

Total revenues (approach 2):
- Minimum: $345,395.95
- Maximum: $603,077.49
- Mean: $481,214.02
- Std Dev: $50,053.98
- Values: 3000

Yearly revenue (incorrect)
- Minimum: $112,348.24
- Maximum: $1,663,013.58
- Mean: $481,167.55
- Std Dev: $198,910.88
- Values: 3000

Total revenues (approach 1):
- Minimum: $342,991.07
- Maximum: $610,876.51
- Mean: $481,246.24
- Std Dev: $49,951.25
- Values: 3000

© EpiX Analytics LLC
Variability and uncertainty

The difference is subtle, but they can greatly impact a model

Variability: heterogeneity (H) among individuals, or randomness (R) (due to chance, or random samples)

A function of the system being modeled. Can’t be reduced.

- Distribution of household income (H)
- Number of daily coffee breaks (H)
- Catastrophic events (R)
- Quality control sampling (R)

Uncertainty: lack of knowledge about the value of a parameter (e.g. imperfect data) or about the right model to use.

A function of the analyst. Thus, can be reduced with more data

- Probability of loan defaults
- Post-launch sales
- Poverty rates in a country
- % of people bored with this talk (Beta(99,1)?)
2. Confusing variability and uncertainty

• Impact: **under** or **over** estimation of output variance.

• Correct modeling:
  • Variability: **should be repeated** in model to represent heterogeneity or randomness (M0247).
  
• Uncertainty:
  • It’s one value, we just don’t know it: thus, **show only once in model** (M0088)
  • For this reason, can usually be treated the same way as a point estimate (multiply, divide, etc)
What happens if I replicate uncertainties?

Uncertainty: what the observer doesn’t know.

- Impact: underestimation of output spread (e.g. variance).
- Correct modeling: represent uncertainty distribution only once in the model.

Example: We randomly sample 100 individuals in a region and 8 have a disease. How many total infected individuals will there be next week, given the below populations for areas A-E within the region?

<table>
<thead>
<tr>
<th>Area ID</th>
<th>Population in the region</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10,000</td>
</tr>
<tr>
<td>B</td>
<td>5,000</td>
</tr>
<tr>
<td>C</td>
<td>6,000</td>
</tr>
<tr>
<td>D</td>
<td>1,000</td>
</tr>
<tr>
<td>E</td>
<td>9,000</td>
</tr>
</tbody>
</table>
What happens if I replicate uncertainties?

Correct modeling: uncertainty on proportion of disease $p = \text{Beta}(8+1, 100-8+1)$ once. Then simulate infected/area using Binomial(population, $p$) and sum areas.

Incorrect modeling of uncertainty: sample $p$ for each area, then simulate and sum areas as above.
Does this matter?

Certainly! Replicating uncertainties grossly underestimated the risk, even in this basic example.

Likewise (and to add to the confusion!), if I don’t replicate variability, risk will be overestimated.
3. Dependencies matter! (M0242)

**Important rule**: every iteration (sample) in a simulation model has to be a possible scenario.

Correlated (dependent) random variables can’t be sampled independently from each other as this would create impossible scenarios e.g.

- Sampling a **high S&P500** and a very **low Dow Jones**.
- Cost estimation of a project, independent of schedule risks.

- Impact: typically **underestimation** of output variance (but can also result in overestimate).
Correlation modeling options

- **Linear correlation**: Rank order correlation. Most commonly used in @RISK, but limited.
- **Non-linear**:
  - Bootstrap: flexible and easy to implement (M0264)
  - Bayesian MC or MCMC: flexible, but harder to implement (M0052)
  - Conditional model logic (e.g. IF statements): to model causal flow/conditionality (M0097)
  - Copulas: more complex but restricted shapes
4. Separate inputs from calculations (M0088)

**NEVER** hardcode an input variable within a calculation

**NEVER**

**REALLY, DON’T DO IT!**

Instead, keep a separate sheet with inputs and refer to them within your calculations. Named ranges even better.

Let’s see if I can convince you…

Client hired a consultancy to develop a MC forecasting model. Previously, model had predicted spot-on an unexpected growth

However, lately the model has been getting it pretty wrong

Hence, client called us to bring the model to its past glory
Separate inputs from calculations

As standard practice, before working on somebody else’s work, we audit it …

And in this case we found a little “surprise”. Below is one of the key calculations for the forecast

\[
=\text{IF(OR(ISBLANK($E4),ISBLANK($G4),$T4="",\text{ISBLANK($W4)},\text{ISBLANK($AR4="")},") },\text{IF(Summary!$H$6="X","VLOOKUP($E4,Actuals,24,FALSE),((Mar!AZ4*(1+(\text{IF($T4=12,}\text{IF(Mar!AZ4<0.83,\text{RiskDuniform(Inputs!$S$266:$U$290),IF(Mar!AZ4<1.19,\text{RiskDuniform(Inputs!$AL$266:$AN$290),RiskDuniform(Inputs!$BE$266:$BG$290))}),IF($CN4="AFO ",\text{CO4,IF($CN4="EMEA",CP4,(IF($CN4="APFO",CQ4,CR4))))))))))+(VLOOKUP($G4,Adjust,7,FALSE))))))))}}
\]

Those were percentiles calculated from the data and then used to select from two very different calculations…

the thing is, the data was updated monthly but these numbers were static! Therefore, the initial “right” prediction was pure luck.
Simplicity (a.k.a. Parsimony)

“With four parameters I can fit an elephant, and with five I can make him wiggle his trunk”

John von Neumann
5. Start simple (at least at the beginning)

We have all been there: you start modeling and want to be realistic
…but, realistic doesn’t always means insightful. Why?
• More parameters-> more data
• Harder to update (with relevant data)
• Complexity =1/transparency
• Mistakes! (see other slides)
• Slow
• …

Key: create model as simple as possible. Then consider adding complexity
An example from our work

Large mine asked us to increase accuracy of current production forecast

Current model required:
- Information from multiple departments and a dozen of (highly paid) mine engineers
- Two dedicated (again, very well paid!) analysts
- Lots of manual effort and disparate data sources, but…
- Monthly forecast was still largely off-target

Conclusions of our analysis of basic data:
- Despite intuition, geology was immaterial for *uncertainty* in forecast (Engineers already knew it well and planned for it!)
- Main driver of forecast uncertainty was *unplanned* time machines down (broken)
- New model relied on this fact, included a *fraction of the parameters*, and was now on target

Yes, that’s me!
6. Distribution identities can significantly simplify your model (M0288)

**Definition:** Identity means equal to, not approximate e.g. 2+2=4, different from 2+2.001≈4

Many identities available…

But some are more useful for our work:

- Binomial\(n_1, p\) + Binomial\(n_2, p\) = Binomial\(n_1+n_2, p\)
- Poisson\(a\) + Poisson\(b\) = Poisson\(a + b\)
- Binomial(Poisson\(\lambda\), p) = Poisson\(\lambda p\)
- Gamma\(\alpha, \beta\) + Gamma\(\delta, \beta\) = Gamma\(\alpha+\delta, \beta\)
- Poisson(Gamma\(\alpha, \beta\)) = NegBin\(\alpha, 1/(1+\beta)\)
- \(\sum_n\)(Bernoulli(Beta\(\alpha, \beta\))) = Binomial\(n, \alpha/(\alpha+\beta)\)

Confused much? Let’s see an example

Let’s revisit our infection example

Assume that now we know that the probability of infection *varies per individual*.

In other words, a (different) random value from the Beta(9,93) would apply to each individual, then their infection is Bernoulli distributed.

But since we had 31k people that’s **62K distributions** (1 Bernoulli and 1 Beta per person). What if we wanted to model China? We need a super-computer, right? Or….

could we apply that obscure

\[
\text{Sum}_n(Bernoulli(Beta(\alpha, \beta))) = \text{Binomial}(n, \frac{\alpha}{\alpha+\beta})
\]

thingy?

Using identities.xlsx
Yes!

The sum of 62K cells is $= \text{Binomial}(31000, 9/102)$

What would have happened if I instead used $\text{Binomial}(31000, \text{Beta}(9,93))$?

Yes indeed, a **fatal error**
We went through a list of technical issues to keep an eye on.

Now is a good time to go over a **checklist** of procedures (rather than techniques) to keep in mind before, during and after your model is developed.
A reference checklist*

1. Engage your decision makers/executives
   • Let them help you scope the problem statement

2. Let the problem drive the analysis
   • Fancier not always better.

3. Make the analysis as simple as possible, but no simpler
   • Easier to communicate and parameterize.
   • **ALWAYS start simple and later add complexity (if needed)!**

4. Identify all significant assumptions and uncertainties
   • Honesty will pay off

5. Perform sensitivity and scenario analysis
   • To identify key parameters and data gaps

6. Iteratively refine the problem statement and the analysis
   • Adapt model to new evidence and/or needs

7. Present and document results clearly
   • Communication is key!

Thanks for your time!

Please contact me with any questions, or for a free copy of the models used during this talk.

Dr. Francisco J Zagmutt  
Managing Partner  
EpiX Analytics  
fzagmutt@epixanalytics.com