The Law of Small Numbers: Accurate Estimating with Limited Data

Christian Smart, Ph.D., CCEA

GALORATH
EXPERIMENT

A DRUG HAS AN 80% SUCCESS RATE

FIVE PEOPLE ARE GIVEN THE DRUG

WHAT IS THE PROBABILITY THAT EXACTLY FOUR OUT OF THE FIVE ARE CURED?
The Law of Large Numbers

STANDARD STATISTICS REQUIRES LARGE DATA SETS

The Law of Large Number is a foundational concept in statistics:

If a sample is large enough, the sample average should be close to the mean

- Cardano proposed the concept
- Jacob Bernoulli established the first mathematical theorem

The key results of classical statistics require large data sets.
THE LAW OF SMALL NUMBERS

ISSUES WITH SMALL DATA SETS

For several decades the psychologists Daniel Kahneman and Amos Tversky studied biases that affect our decision-making processes. In 2002, their research was honored with a Nobel Prize in Economics. One of their key contributions was a paper “Belief in the Law of Small Numbers.” In this paper, Kahneman and Tversky define the Law of Small Numbers as the mistaken belief that a small sample accurately reflects the probabilities of a population.

In small data sets you can find patterns where none exist!
EXAMPLE 1

CONSIDER THE DATA SET ON THE LEFT

ONLY 10 DATA POINTS BUT THERE IS CLEARLY A STRONG TREND!

ACTUALLY, THE DATA WERE RANDOMLY GENERATED!
IN A POPULATION OF 100 DATA POINTS THAT HAVE NO TREND IT IS POSSIBLE TO FIND SMALL SAMPLES THAT DISPLAY A TREND.

THE DATA POINTS IN THE POPULATION HAVE ZERO CORRELATION.

BUT THERE ARE SAMPLES OF 10 DATA POINTS THAT HAVE STRONG TRENDS.
It is easy to find examples of spurious correlations in small data sets. Tyler Vigen has devoted a website and a book to the subject.
FITTING TO NOISE

STUDIES INDICATE THAT 30-70% OF VARIABLES INCLUDED IN STEPWISE REGRESSION ARE PURE NOISE

RULE OF THUMB

50 DATA POINTS PLUS 10 DATA POINTS FOR EVERY ADDITIONAL PARAMETER

BASIS

THE RULE OF THUMB IS BASED ON SIMULATION STUDIES OF FITTING REGRESSIONS TO RANDOMLY GENERATED DATA
WHAT CAN BE DONE?

COLLECT MORE DATA
WORK WITH OTHER GOV'T AGENCIES (E.G., AIR FORCE, COMMERCIAL COMPANIES)
CAN BE DIFFICULT TO DO

IMPUTE
USE IMPUTATION TO FILL IN MISSING DATA
NEED TO BE CAREFUL TO PRESERVE CORRELATION STRUCTURES

GO LOWER
ESTIMATE AT THE COMPONENT LEVEL INSTEAD OF THE SUBSYSTEM LEVEL
MORE NOISE

BAYES
USE TECHNIQUES DESIGNED TO ESTIMATE WITH LIMITED DATA SUBJECTIVE
SHOW ME THE DATA!

COLLECT MORE DATA

THERE ARE AIR FORCE AND COMMERCIAL MISSIONS WHOSE DATA WOULD BE USEFUL TO NASA

STATUS QUO

NASA HAS ALREADY DONE SIGNIFICANT WORK IN THIS AREA WITH THE ESTABLISHMENT AND COLLECTION OF CADREs

CAVEATS

LOW HANGING FRUIT HAS BEEN COLLECTED, THIS OPTION REQUIRES MORE EFFORT
MISSING DATA
Even with significant data sets, there will be missing fields. Cuts down on the number of data points that can be used for modeling. Imputation uses the data you have to fill in missing inputs.

STATUS QUO
Nothing significant has been done in this area.

POTENTIAL
This is a low-intensity effort – time and resources. One of the easiest options.
GO LOWER

ESTIMATE AT A LOWER LEVEL

BREAK DOWN UNIQUE SUBSYSTEMS TO COMPONENTS THAT HAVE COMMONALITY WITH OTHER MISSIONS
INCREASES NUMBER OF AVAILABLE DATA POINTS WHEN UNIQUE MISSIONS ARE MODELED

STATUS QUO

MANY MODELS ARE AT THE SUBSYSTEM LEVEL – PCEC, NICM, EVEN SEER-SPACE

POTENTIAL

WILL REQUIRE NEW APPROACH TO DATA COLLECTION – CADREs ARE TYPICALLY AT THE SUBSYSTEM LEVEL
CONSIDER DEFENSE DEPT.’S FLEXFILE APPROACH
SOME COMMERCIAL MODELS ARE AT THE LOWER LEVEL (E.G. SEER-H)
BAYESIAN PARAMETRICS

USE METHODS DESIGNED FOR SMALL DATA SETS

THE BAYESIAN APPROACH TO PARAMETRICS ALLOWS YOU TO LEVERAGE ALL AVAILABLE DATA INCLUDING EXPERIENCE

STATUS QUO

VERY LIMITED USE IN PRACTICE

POTENTIAL

REQUIRES NO ADDITIONAL COLLECTION EFFORTS
REQUIRES A MODERATE AMOUNT OF EFFORT TO CORRECTLY APPLY

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
Imputation
MISSING DATA

Dealing with holes in your data set

When developing multivariate CERs, you may initially have a decent-sized data set

Then you notice a few missions are missing one key independent variable, a few others are missing another, etc.

Pretty soon you have whittled down your data set to only a few missions that have all the independent variables you want to consider

One option would be to go out and spend time and effort on collecting the missing information

But what if you could use statistics of the data set to make reasonable assessments of the missing values?

It turns out you can with statistical imputation
IMPUTATION

Filling in holes in your data

Well-developed statistical theory

Methods have been developed for a variety of missing types, whether at random or not at random

Important to use techniques that do not change the correlation structure of the model (if intent is to use the data for multiple regression)

Imputation techniques use a combination of maximum likelihood methods along with Bayesian techniques
Bayesian Parametrics
BAYES’ THEOREM

CONDITIONAL PROBABILITY

NUMEROUS APPLICATIONS

1. **ENIGMA**
   Bayesian techniques were used to help crack the Enigma code in World War II, shortening the war

2. **PROPERTY and CASUALTY INSURANCE**
   Used for over a century to set premiums when there is limited data

3. **HEDGE FUND MANAGEMENT**
   Also used in election forecasting and game theory

CONDITIONAL PROBABILITY:

$$Pr(A|B) = \frac{Pr(A)Pr(B|A)}{Pr(B)}$$

Bayesian techniques have been successfully used in a variety of applications - their use is no mere academic exercise in fancy statistics. They have skin in the game.
EXAMPLE 1: DRUG TESTING

**LAW OF TOTAL PROBABILITY:**
\[
\Pr(B) = \Pr(B|A) \Pr(A) + \Pr(B|A') \Pr(A')
\]

**BAYES' THEOREM:**
\[
\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B|A) \Pr(A) + \Pr(B|A') \Pr(A')}
\]

**ANSWER:**
\[
\Pr(A|B) = \frac{0.02(0.95)}{0.02(0.95) + 0.99(0.05)} \approx 27.7\%
\]

**QUESTION**
What is the probability that someone who fails a drug test is not a drug user?

**ASSUMPTIONS**
95% of the population are non-users
If someone is a drug user, it returns a positive result 99% of the time
If someone is not a user, it returns a positive result 2% of the time

**DEFINING TERMS**
A = Event that someone is NOT a drug user
B = Event that someone tests positive for drugs

**COMPLEMENT**
A' = Event that someone uses drugs
B' = Event that someone tests negative for drugs
QUESTION
There are three doors. Behind one is a car. Behind the other two cars are goats. You pick a door. Monty opens a different door and shows you a goat, and offers you the chance to switch

DEFINING TERMS
Let $A_i$ denote the event that the car is behind the $i$th door, WLOG assume:
1. You pick door #1
2. You are shown a goat behind door #3, define this as event B

PRIOR PROBABILITY
Initial assumption is that $P(A_1) = P(A_2) = P(A_3) = 1/3$

CONDITIONAL PROBABILITY
$P(B | A_3) = 0$
$P(B | A_2) = 1$
$P(B | A_1) = 1/2$

EXAMPLE 2: MONTY HALL PROBLEM

BAYES’ THEOREM:

$$Pr(A_1|B) = \frac{Pr(A_1) Pr(B|A_1)}{Pr(A_1) Pr(B|A_1) + Pr(A_2) Pr(B|A_2) + Pr(A_3) Pr(B|A_3)}$$

POSTERIOR PROBABILITIES:

$$Pr(A_1|B) = \frac{(1/3)(1/2)}{(1/3)(1/2) + (1/3)(1) + (1/3)(0)} = \frac{1/6}{1/6 + 1/3} = \frac{1}{3}$$

$$Pr(A_2|B) = \frac{(1/3)(1)}{(1/3)(1/2) + (1/3)(1) + (1/3)(0)} = \frac{1/3}{1/6 + 1/3} = \frac{2}{3}$$

$$Pr(A_3|B) = 0$$

ANSWER: YOU SHOULD SWITCH!
RSDO EXAMPLE

NASA’s Rapid Spacecraft Development Office (RSDO) uses streamlined acquisition processes and fixed-price contracts to cut costs for robotic Earth-orbiting satellites.

Give five historical data points of RSDO missions, what is the best way to estimate an RSDO mission?

One option is to develop a CER with the five data points.

However, given our discussion of the law of small numbers, a trend line developed from these five points may not be meaningful.

Alternative – consider a larger set of robotic Earth-Orbiting satellites.
COST ANALYSIS

Applying Bayes’ Theorem to Parametrics

EARTH-ORBITING DATA
There are many more earth-orbiting data points if we do not restrict our attention to only RSDO missions

However, these missions are substantially more expensive than the RSDO analogues

Classical techniques apply here but they will likely overestimate a new RSDO mission by a significant amount

This is like the statistician who lost her car keys in a parking lot at night when someone saw her looking for them under a street light – when asked where she lost her keys she responded that the keys were near her car, far from the street light. When asked why she was looking near the street light, she responded that was where the light was shining.

\[ y = 0.2679x^{0.887} \]
\[ R^2 = 0.6163 \]
COST ANALYSIS

Applying Bayes' Theorem to Parametrics

COMBINING THE DATA

Bayes’ Theorem provides a way to use both sources of data.

The large, generic set of earth-orbiting missions can be used as a prior.

The small specifically applicable data set can be used to update the prior probabilities.

General Set

Specific Subset

To be successful, you have to have your heart in your business, and your business in your heart.

Thomas Watson
Applying Bayes' Theorem to Parametrics

RESULTS

The Bayesian CER coefficients are weighted averages of the coefficients of the CERs based on the two separate data sets.

The Bayesian CER is applicable to a wider range of data than just the RSDO missions.

In testing the CER on an RSDO mission not in the data set, the Bayesian CER was more accurate than either of the RSDO-only CER or the all Earth-orbiting CER.
THE BASIC METHOD AND ITS ISSUES

• Everything we have discussed to this point relies on the use of a basic Bayesian method that has been presented before - for details see:
  • Christian Smart, “Bayesian Parametrics: How to Develop a CER with Limited Data and Even without Data,” 2014 ICEAA Conference
• There are some issues with this basic approach:
  • For a power equation, it requires the use of log-transformed ordinary least squares (biased low)
  • One assumption is that we know that the variance of the CER based on the sample data is known with certainty
  • Another assumption is that the residuals are lognormally distributed (normal in log space)
UNKNOWN VARIANCE

THE MOST PROBLEMATIC ASSUMPTION IS THAT OF KNOW VARIANCE

In our example, the sample data is the small set, the one for which we have the least confidence in knowledge of the population variance.

GOOD NEWS THIS CAN BE HANDLED ANALYTICALLY

By Cochran’s Theorem, the variance can be modeled as a scaled inverse-Chi Square distribution.

ESTIMATING PARAMETERS

Degrees of freedom is the same as the degrees of freedom in the regression of the sample data.

Sample variance is the scaling factor.
(LOG) NORMAL RESIDUALS

GAUSSIAN ASSUMPTION MAKES THE MODEL ANALYTICALLY TRACTABLE
Gaussian likelihood has Gaussian as a conjugate prior
Even with unknown variance model is still tractable
Log-t has been proposed as an alternative to lognormal
Also - nonparametric methods have no distributional assumption

IF RESIDUALS ARE NOT GAUSSIAN, MUST USE SIMULATION
With small samples, residuals may be better modeled with log-t
Markov Chain Monte Carlo simulation can be used to do the Bayesian analysis via simulation

MARKOV CHAIN MONTE CARLO (MCMC)
MCMC uses conditional simulation – each trial depends on the previous
Can implement in R or use WinBUGS
RSDO - RESULTS COMPARISON

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Normal likelihood with known variance</th>
<th>Normal likelihood with unknown variance</th>
<th>Student t likelihood, unknown variance, normal error</th>
<th>Student t likelihood, unknown variance, Student’s t error</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>$83</td>
<td>$44</td>
<td>$15</td>
<td>$23</td>
</tr>
<tr>
<td>10%</td>
<td>$94</td>
<td>$59</td>
<td>$25</td>
<td>$41</td>
</tr>
<tr>
<td>20%</td>
<td>$109</td>
<td>$83</td>
<td>$47</td>
<td>$72</td>
</tr>
<tr>
<td>30%</td>
<td>$121</td>
<td>$107</td>
<td>$76</td>
<td>$100</td>
</tr>
<tr>
<td>40%</td>
<td>$132</td>
<td>$133</td>
<td>$112</td>
<td>$129</td>
</tr>
<tr>
<td>50%</td>
<td>$144</td>
<td>$163</td>
<td>$163</td>
<td>$163</td>
</tr>
<tr>
<td>60%</td>
<td>$157</td>
<td>$199</td>
<td>$236</td>
<td>$205</td>
</tr>
<tr>
<td>70%</td>
<td>$171</td>
<td>$248</td>
<td>$352</td>
<td>$266</td>
</tr>
<tr>
<td>80%</td>
<td>$190</td>
<td>$319</td>
<td>$560</td>
<td>$370</td>
</tr>
<tr>
<td>90%</td>
<td>$220</td>
<td>$453</td>
<td>$1,068</td>
<td>$642</td>
</tr>
<tr>
<td>95%</td>
<td>$248</td>
<td>$606</td>
<td>$1,819</td>
<td>$1,169</td>
</tr>
<tr>
<td>99%</td>
<td>$311</td>
<td>$1,044</td>
<td>$4,942</td>
<td>$7,294</td>
</tr>
<tr>
<td>99.5%</td>
<td>$338</td>
<td>$1,275</td>
<td>$7,125</td>
<td>$21,659</td>
</tr>
<tr>
<td>99.9%</td>
<td>$401</td>
<td>$1,922</td>
<td>$15,151</td>
<td>$842,673</td>
</tr>
</tbody>
</table>

FOUR CASES CONSIDERED
1. Base case – Gaussian model (lognormal)
2. (Log) Gaussian residuals with unknown variance
3. Unknown variance, log-t for likelihood, lognormal predictive
4. Unknown variance, log-t for likelihood, log-t predictive

IMPLEMENTED IN BOTH R AND WINBUGS

Unknown variance changes the point estimate

The use of different uncertainty distributions does not change the point estimate (50th percentile)

THE DEVIL IS IN THE TAILS

The use of log-t dramatically effects the tails of the distribution

1 in 1,000 chance that a $163 million cost may grow to tens and hundreds of billions!
• Changing the assumption results in higher estimates
  • The 50th percentile for the Gaussian model is $144 million
  • Relaxing the variance assumption increases the 50th percentile to $163 million
  • Changing the CER residuals assumption does not change the 50th percentile
  • Improves the accuracy – actual cost was $180 million is 10% higher than the estimate with unknown variance
• There is also an increase in the heaviness of the right tail when relaxing the Gaussian assumptions
  • The unknown variance case has a 99th percentile equal to $1 billion
  • Extreme but it happens
COMPARISON COMMENTS (2 OF 2)

• When going to the log-t, the extreme right tail explodes, says that a relatively simple earth-orbiting spacecraft could cost as much as the James Webb Space Telescope (or more)!

• The relaxation of the assumption of known variance is important and the results are logical

• The change of the modeling of the residuals to a Student’s t distribution in log space (log-t) is not logical

• Need to balance common sense with mathematical correctness – cost modeling is both an art and a science
SUMMARY - BAYES
AN IDEAL METHOD WHEN YOU HAVE LIMITED DATA

MYRIAD REAL-WORLD APPLICATIONS
- Cracking Enigma code
- Search-and-Rescue
- Property and Casualty Insurance Premium Setting

USES ALL YOUR DATA
- Can be objective or subjective
- Allows you to use all your data

GAUSSIAN (BASIC) MODEL
- Analytically tractable
- Can be done in Excel

ADVANCED TECHNIQUES
- The Gaussian model has issues with variance assumptions
- If residuals are not lognormal, may require the use of Markov Chain Monte Carlo

FILLS A NEED
- Ideal for small data sets

LOW-HANGING FRUIT
- Currently limited use
- High potential
HOW TO DEAL WITH SMALL DATA

**SUMMARY**

**OVERALL**

**ADMIT THERE IS AN ISSUE**
Be aware that classical statistics does not work well for small data sets (<50 data points)

**DO SOMETHING ABOUT IT**
- Collect data for more missions
- Collect lower-level data
- Imputation
- Bayesian Methods

**DATA COLLECTION**
Time consuming and expensive
Low-hanging fruit has been collected (ONCE/CADRE)

**IMPUTE**
Use more of the data you already have by filling in gaps
Variety of statistical methods, well-developed

**BAYES**
Use Bayesian methods to leverage all your knowledge

**LESSON**
Avoid the abuse of large-sample methods on small data sets and use imputation and Bayesian methods
REFERENCES


