RISC

decisions with confidence

Big Data: What is a significant sample size?

Gavin Ward

Devex, Aberdeen

21 June 2023

World Oil

https://www.worldoil.com/magazine/2022/january-2022/features/big-data-what-is-a-significant-sample-size/

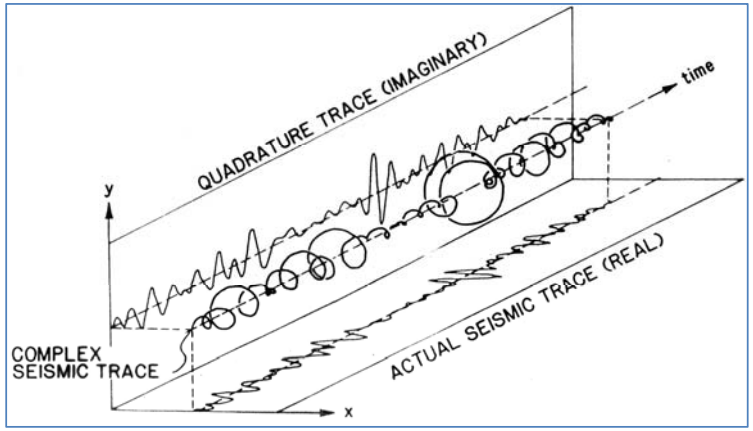
Contents

Introduction: Why do we seem to continually predict poorly?

A famous financier once asked, "*Why is an MBA student who's learned about discounted cash flow, like a baby with a hammer?*"

Answer:

"*Because to a baby with a hammer, everything looks like a nail*".



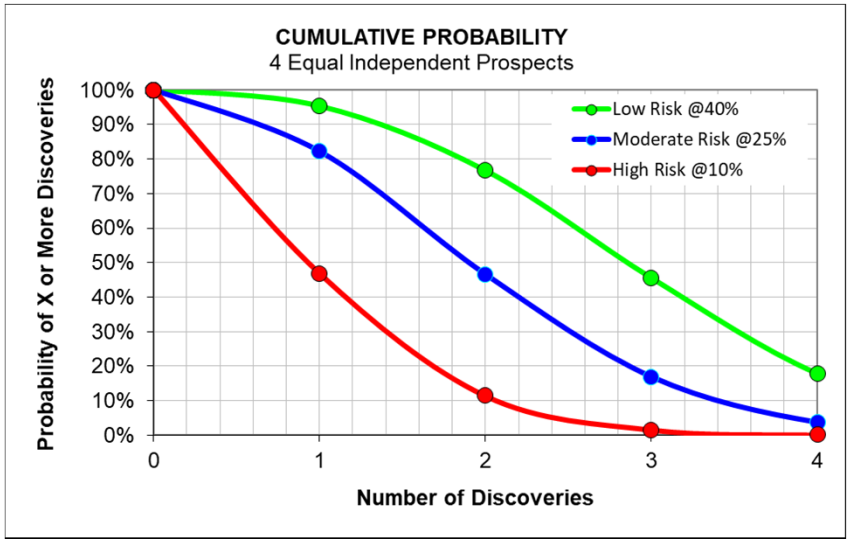
- Decision Makers are continuously bombarded with requests for funding supported with NPV's
- Don't ignore the assumptions of the input forecasts to the discounted cash flow NPV's
- How do you distinguish NPV's that are positive from those that merely result from forecasting errors?

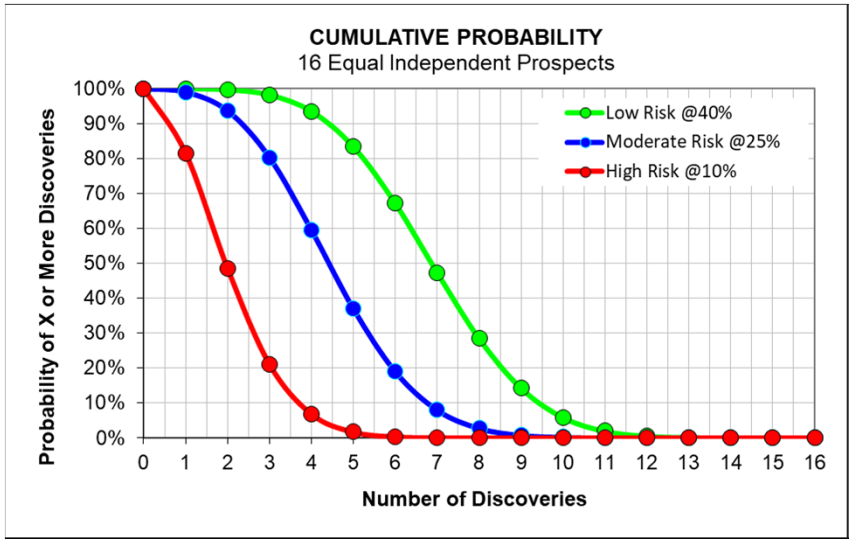
NOTE: Discounted cash flow gives Net Present Value (NPV)

Why do samples matter?

- Why do we sample?
- Why does sample size matter?
  - Improves predictability of outcome

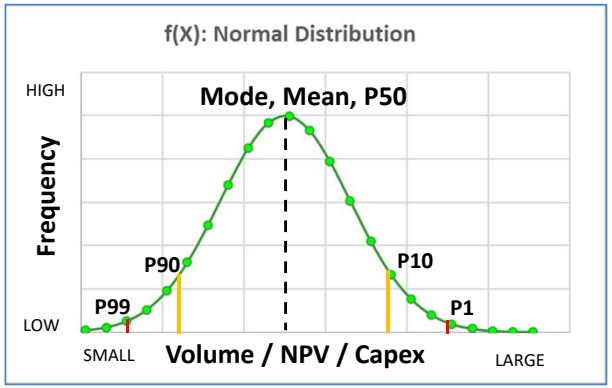
- Resource size forecasts only '*credible*' if portfolio contains a statistically significant number of samples.

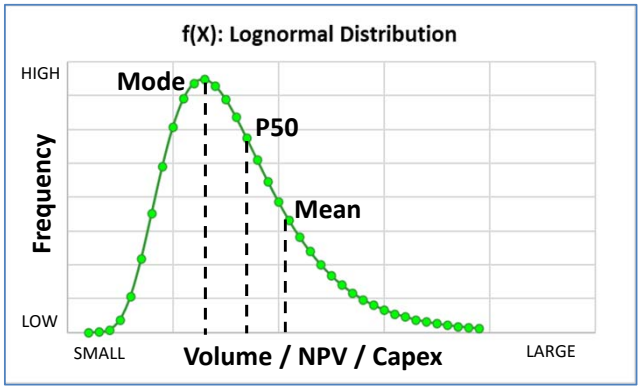


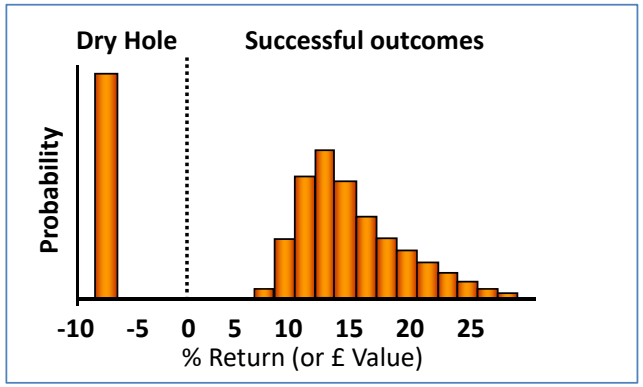


Distributions

- Normal distribution P50 = Mean = Mode (most likely).
- Lognormal distribution P50 ≠ Mean ≠ Mode.
- Descriptive term '*Most Likely*' is misleading as it contains no information about variance.

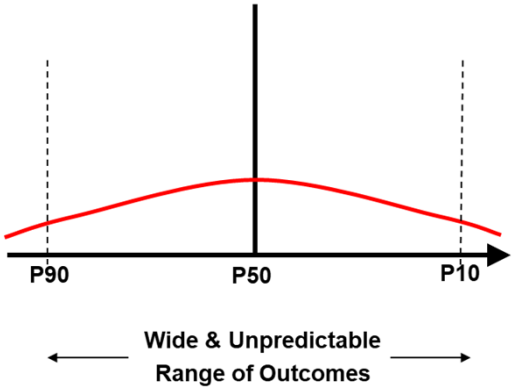


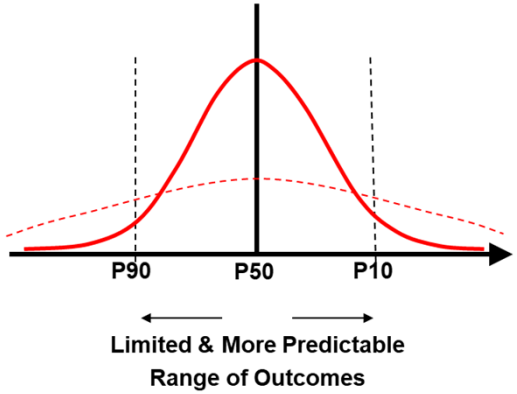




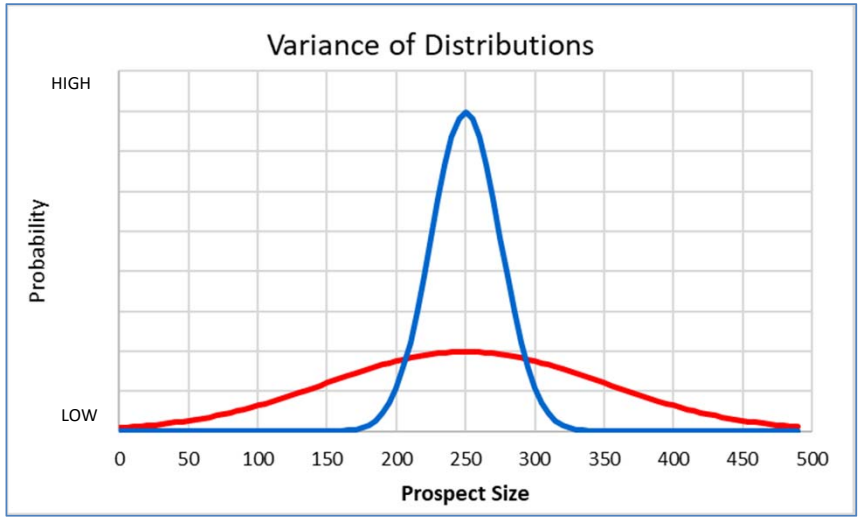
Variance

- Variance is a measure of dispersion / spread of all data points in a data set





- Example: Prospects from two geological basins with same P50/Mean volume of 250 MMstb
  - Red distribution has mean/mode 250 MMstb of oil and variance 10,000
  - Blue distribution has mean/mode 250 MMstb of oil and variance 625



NB: Standard Deviation $(SD) = \sqrt{v}$

Variance is the average of the squared distances from each point to the mean, which is the mid-point (P50) for a normal distribution.

Confidence: When distributions not available

- Forecasts rely on 2 terms: **Confidence Level** & **Level of Accuracy**
- **Confidence Level** tells you how "*sure*" you can be.
  - Represents how often the true percentage of the population who would pick an answer lies within the confidence interval.
  - 95% confidence level means you can be 95% certain;

- **Level of Accuracy** is +/- number (e.g.: US$45 million +/- $5)
- Put Confidence Level together with Confidence Interval
  - 95% sure that the true percentage of population is between US$40 million and US$50 million.
- The wider the confidence interval you're willing to accept, the more certain you can be that the answers from the whole population would be within that range.

Factors that Affect Confidence Intervals

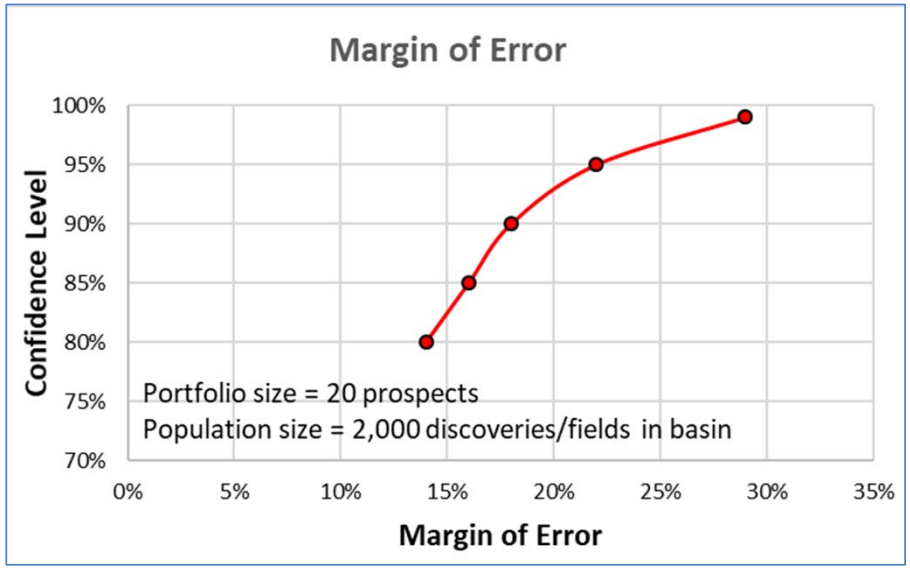
- Size of a significant sample of a population depends on what level of confidence we want in our prediction (e.g.: Low < 50%, High >90% etc.) and the size of the total population of data.

- We don't always know the exact size of the total population of data, but we can estimate this, and precision is not required.

- There are three factors that determine the size of the confidence interval for a given confidence level:

1) Sample size
2) Population size
3) Margin of error

Confidence level	Population Size	Sample Size	Margin of error
99%	2,000	20	29%
95%	2,000	20	22%
90%	2,000	20	18%
85%	2,000	20	16%
80%	2,000	20	14%



What is a good sample size?

- Before you can calculate a good sample size, you need some idea about the degree of precision you require or, the degree of uncertainty you are prepared to tolerate

- Many sample size calculators and statistical guides available but as a guide……..

- Good maximum sample size is usually around 10% of the population, as long as this does not exceed 1000.
  - Population of 5,000 North Sea wells, 10% would be 500.
  - Population of 200,000 onshore North American wells sampling 1,000 (0.5%) will usually give a fairly accurate result.
  - Sampling > 1,000 wells won't add much to the accuracy regardless of Big Data processing power & dataset size

		Size of Population					
		200	500	1,000	2,500	5,000	> 5,000
Marin of Error	+/- 10%	65	81	88	93	94	96
	+/- 7.5%	92	127	146	160	165	171
	+/- 5%	132	217	278	333	357	384
	+/- 3%	169	341	516	748	880	1,067

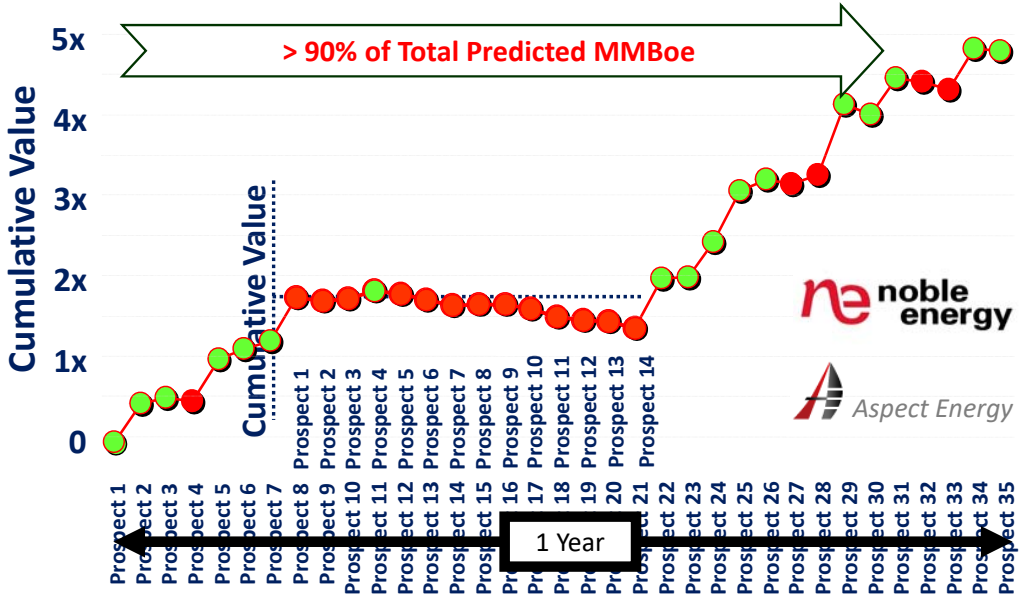
Human Bias in Sampling

- All humans are subject to biases, regardless of technical competency, or level of education

- '*Belief in the law of small numbers*', Tverskey[1], Kahneman[1]
  - Humans regard a sample randomly drawn from a population as highly representative.
  - 'Sample size neglect' is tendency to underestimate how variability of average estimates changes with sample size.

- '*The Difficulty of Assessing Uncertainty*' by Ed Capen[2] ARCO

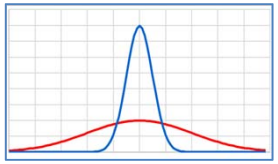
- Glenn McMaster[3] & Peter Carragher[3] bp



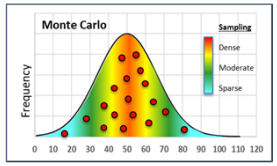
[1]Tverskey, Kahneman, Psychological Bulletin, 1971 [2]Ed Capen, Journal of Petroleum Technology, 1976 [3]Glenn McMaster & Peter Carragher, BP, 2003
Chart: AAPG 2006 (Basics are Boring – The Essentials of Good Portfolio Management at Independent Oil and Gas Companies)

Closing Remarks

- Shaky ground: Using P50 blindly
- Variance is a vital value for describing an estimate



- Monte Carlo simulations are not just for geoscientists



- Monte Carlo simulation most beneficial to fully understand and appreciate the variance when the sample size is at its smallest.
- A level of accuracy provides a useful index of variability, and it is precisely this variability that we tend to underestimate.
- The associated confidence is implicit in the P90/P50/P10 figures, but many upstream documents typically only report one of these (e.g.: Accountants) and therefore lose all information about variability ……**this is not good for making decisions, or decision makers!**

www.riscadvisory.com

Gavin Ward

Director, Europe, Africa, Middle East & South America

gavin.ward@riscadvisory.com

Perth

Level 2
1138 Hay Street
WEST PERTH WA 6005
P. +61 8 9420 6660
E. admin@riscadvisory.com

Brisbane

Level 10
95 North Quay
BRISBANE QLD 4000
P. +61 7 3025 3397
E. admin@riscadvisory.com

London

Level 2
20 St Dunstan's Hill
LONDON UK EC3R 8HL
P. +44 (0)203 795 2900
E. admin@riscadvisory.com

South East Asia

Jakarta
Indonesia
P. +61 8 9420 6660
E. admin@riscadvisory.com