esomar

Synthetic Data

Ray Poynter

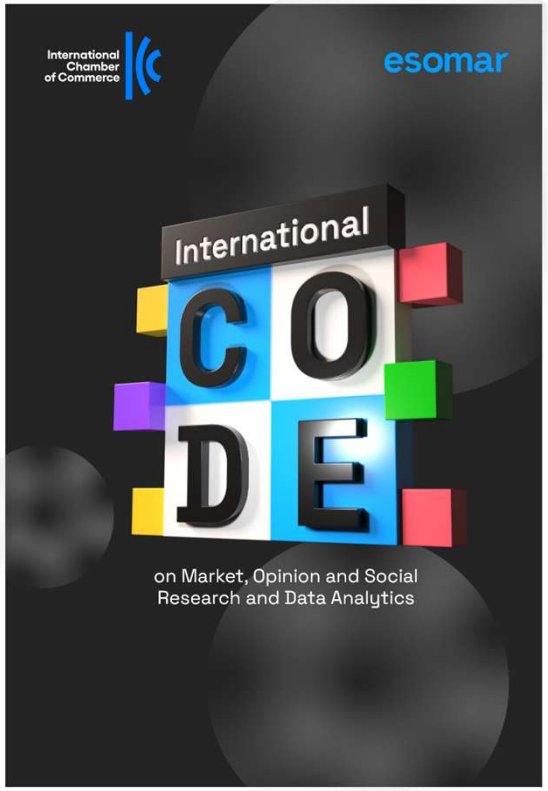
Chair of Esomar Professional Standards Committee

Agenda

- Definitions
- Augmentation
- Other Synthetic Data
- Esomar's Framework
- Slovakia
- Safe Experimentation
- Q & A

esomar

Definitions – Synthetic Data



Synthetic Data

- Information that has been generated to replicate the characteristics of real-world data.

Synthetic Persona

- A digital representation of a person generated to mimic the behaviours, preferences, and characteristics of real people or groups.

esomar

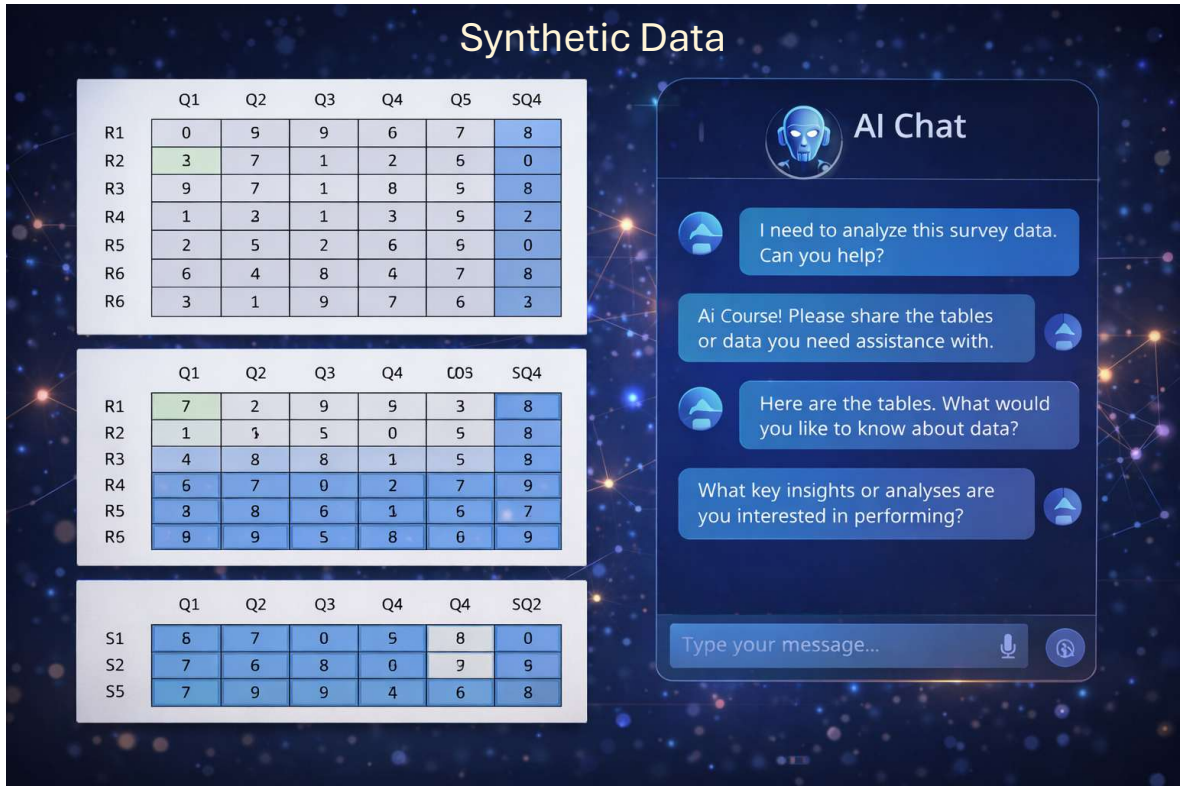
Synthetic Data

ID	Q1	Q2	Q3	Q4	Q5	SQ1
R1	1	4	5	4	9	S
R2	2	3	4	5	2	S
R3	1	S	3	3	7	S
R4	3	4	4	5	8	S
R5	2	5	5	3	6	S
R6	1	3	S	4	4	S
R7	3	4	3	5	9	S
S1	S	S	S	S	S	S
S2	S	S	S	S	S	S

esomar

Synthetic Personas

Digital Twins
Group Personas
Virtual Respondents
Simulacra
Agents



esomar

Synthetic Data Overview

- Augmented Synthetic Data
- Other Synthetic Data
  - Digital Twins
  - Group Personas
  - Synthetic Respondents
- *Anonymised Synthetic Data*
  - *A real dataset tweaked to make it anonymous, retaining key statistical inferences*
- *Randomised Synthetic Data*
  - *Generated to test theories, software, etc*

esomar

Augmented Data

ID	Q1	Q2	Q3	Q4	Q5
R1	1	4	5	4	9
R2	2	3	4	5	2
R3	1	4	3	3	7
R4	3	4	4	5	8
R5	2	5	5	3	6
R6	1	3	5	4	4
R7	3	4	3	5	9
S1	S	S	S	S	S
S2	S	S	S	S	S

Example Problem for Sample Boosting

We wanted to collect

- 250 Young Men
- 250 Young Women
- 250 Older Men
- 250 Older Women
- 1000 in Total

We have collected

- 125 Young Men
- 250 Young Women
- 250 Older Men
- 250 Older Women
- 875 in Total

Solutions:

1. Live with it
2. Weight the data, upweighting the Young Men
3. Use a statistical method to boost the Young Men
4. Use AI/Machine Learning to boost the Young Men

Statistical Methods to Boost Young Men

- There are a family of methods for boosting the data that existed years before Generative AI and other leading AI techniques were available
- This was more common in academic than commercial studies
- Random sampling and nearest neighbour
  - These techniques tend to be based on a combination of re-selecting people at random from the underrepresented group (Young Men in our example) and blending them with other people from the same data set
  - SMOTE is one example of this approach

esomar

SMOTE
(Synthetic Minority Over-sampling Technique)

In simple terms, and in our case of 125 Young Men

- Pick a Young Man at random
- Find the five most similar responses to him
- Pick one of them at random
- Blend the two profiles to make a new one
- Repeat until you have the required number (250 in our case)

esomar

The Consequence of SMOTE and Similar

- The data is more central and less dispersed
- The data is more flexible
  - Can be fed into other processes (e.g. dashboards and what-if models) – many processes do not like (or ignore) weightings
  - It can be analysed in terms of other cross-breaks
- But we can't do significance testing with traditional statistical approaches
- When done well, it is slightly better than weighting
- Most techniques do not boost/replicate the open-ended comments

All of these are true of AI Augmented Synthetic Data

esomar

AI Augmented Synthetic Data

- Many vendors (large, small and new companies)
- Many end clients are making extensive use of it
  - Many are not touching it, some are experimenting
- Most vendors do NOT use LLMs (e.g. ChatGPT) for this
  - Some use it to help estimate open-ended responses
- The techniques are typically a blend of statistics, Machine Learning and other advanced AI approaches
- Inputs
  - Only the study that is being augmented
  - The study and reference data (e.g. census data)
  - The study and historical data
  - The study and reference data and historical data

esomar

AI Augmented Synthetic Data – Use Cases

- To deal with hard-to-reach cases
  - Such as the young men example earlier
- To allow analysis of niches
  - In a tracking study, there may be small brands which do not get enough ratings. Augmentation can allow these to be examined and reported.
- To allow earlier assessments of data
  - In a tracking study, the latest data may look different, but traditionally, we would wait for the next wave before assessing the difference. By augmenting the data, we may be able to spot real changes quicker

Testing the Quality of the Boost

Two cases

- Does it tend to work with your type of data?
- Does it work in this particular case?

Does it work in general with your type of data?

- Take old studies
- Remove some of the data
- Boost the sample to replace the removed data
- Compare the removed data with the boosted data

Does it work in a live case?

- Remove some of the data from the category that needs boosting (e.g. Young Men in our case) – this is a Holdout Sample
- Boost the sample, not using the holdout sample
- Compare the boosted sample with the holdout sample

esomar

Increased Risk of Type 1 Errors

- Running the total data set (real plus the boost) through standard significance testing tends to make P values too optimistic
- The standard error of the mean is the standard deviation divided by the square root of the sample size $\frac{\sigma}{\sqrt{n}}$
  - Synthetic boosts tend to reduce the standard deviation, because they trim the tails, and they appear to increase the sample size
  - The standard error is artificially smaller
- We are more likely to say there is a difference in cases where there is no difference
- New statistics are required
  - Normal CI, MOE & ESS are unsuitable

esomar

Other Synthetic Data

- Digital Twins
- Group Personas
- Synthetic Respondents

esomar

Digital Twins

- Qual and Quant
- Less common / more experimental than Augmented boosts
  - But there are plenty of vendors and presumably clients
- From collected data, create one digital replicant for each original participant
  - Qual data, quant data, and online community data
- Uses
  - A chat interface – the user can ask the twins specific questions, in the same was as people chat with an LLM
  - The digital twins are given a survey and supply answers
  - The digital twins are given a set of qual prompts and give discursive answers

esomar

A Key Paper to Consider

- Research Paper
  "Generative Agent Simulations of 1,000 People"
  - Published November 2024
  - Authors from Google, Stanford & NorthWestern
- Created 1000 Agents (synthetic twins) after a two-hour voice, semi-structured interview with each person
- Asked a wide variety of questions to the Agents and the real humans
  - 85% accuracy between the real answers and the Digital Twins

https://arxiv.org/abs/2411.10109

esomar

Generative Agent Simulations of 1,000 People - Review

- They created models for 1000 people (Agents)
- They asked these Agents 316 questions
  - Not asked to the real people
- They then asked the 316 questions with the real people and got a net, relative, average accuracy of 85%
- They could have asked these Agents another 1000 questions, with no wear and tear
  - And relatively instant responses
- If this proves to be generalisable & reliable in use, it is a game changer
  - There is some opposing evidence

esomar

The difference between Boosting/Augmenting and Digital Twins

Boosting

- There is always some human data in the file
- We know what questions are being asked / answered
- We can't ask new questions, without new human data, which takes time

Digital Twins

- We deal with 100% agents, no people
- The questions can be new questions
- There is potentially no time lag between thinking of a question, asking the question, and getting an answer

esomar

Digital Twins Issues

- Most Twins are powered by Generative AI
- General agreement that Quant uses are ‘directional’ not definitive
- Testing with historical data is possible
- But testing with current projects is difficult
  - Possibly impossible
- Twins trade-off speed/cost against confidence
  - Early screening is a current use case
  - Cases where research would not otherwise be conducted

esomar

Personas

esomar

Personas

- The term is used in different ways by different providers, including:
  - Qual only uses
  - Quant and qual uses
  - Personas representing a single synthetic person
  - Personas representing group
- In this section, I will focus on Personas representing a group of people
  - Other uses are covered in Boosting, Twins, and Wholly Synthetic
- Personas utilise data and AI to bring segments to life

esomar

Signoi Personas

UK-based agency – offer Digital Personas via AIBODS

- Bespoke solutions for clients
- Zephyr – 20 Genz Personas

Cameron, 27 – A mobile app developer living in a fashionable neighborhood in London.
Aaliyah, 26 – An architect living in a high-rise flat in Manchester city center.
Ajay, 20 – A film studies student at Goldsmiths in London who plays guitar for an indie band.
Phoebe, 24 – An aspiring painter and barista working at a funky coffee shop in Brighton.

Diego, 21 – A sociology major at the University of Manchester who interns on weekends for a local non-profit focused on diversifying the tech industry.
Gabby, 20 – An environmental science major at the University of Edinburgh who belongs to climate justice and zero waste groups on campus.
Leila, 17- A high school student in Manchester passionate about sci-fi, gaming, and cosplay.

esomar

Lots of other vendors

Evaluating Quant Personas

This falls into the same two questions as earlier

1. Do they work in general in this field?
   - Use past studies and compare the results
2. Do they work for this project?
   - Difficult, it will largely come down to plausibility, and post event testing. For example, running the questions and answers from the personas against omnibus questions.

esomar

Evaluating Qual Personas

- Most quant methods of assessing are not suitable for qual.
- Not fully relevant
  - The number of word & the number of ideas
  - Whether the same ideas or explanations were found
- Relevant
  - Plausibility
  - Was the research helpful and/or enlightening
  - Can the findings be evaluated over time?

esomar

Synthetic Respondents

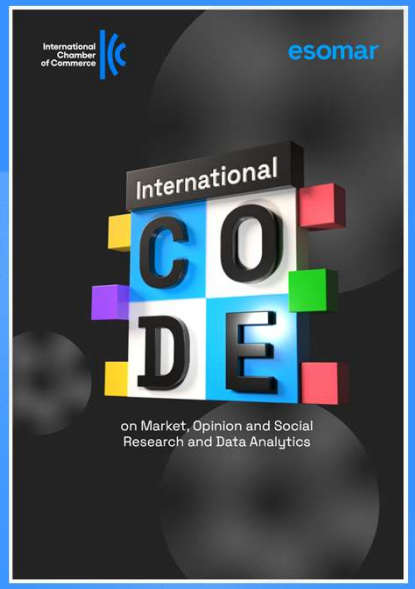
esomar

Synthetic Respondents

- The least developed of these Synthetic Options
- Personas where each case acts as an individual respondent
- The focus is quantitative research
  - For qual, group personas are an easier option
- The input tends to be large amounts of historical data

esomar

Esomar’s AI Framework



esomar

Esomar's Framework

- AI Taskforce – chaired by Pravin Shekar
- Guides and Codes
  - ICC/Esomar Code
  - Esomar 20 Questions for AI Vendors
  - 5 Topics for Discussions about Augmented Data
- Community Circles
  - Synthetic Data and other AI Topics
  - 13 April Group Personas for Qual Research
- New Guidance and Papers
  - Applying the ICC/Esomar Code to all aspects of AI and Synthetic Data
  - Answers to questions, training, and methodology papers

esomar

AI and Slovakia

esomar

AI and Slovakia

- Foundation models (e.g. ChatGPT) are more typical of the USA and less typical of other countries
- Synthetic Data approaches that require large data sets (100s of 1000s of responses) or large AI spend are harder in smaller countries
  - Free-standing AI synthetic data is harder for Slovakia
- Augmented Data, Digital Twins, & Group Personas are as valid for Slovakia
  - They are built from the data you collect

esomar

Safe Experimentation

- Augmentation
  - Give vendors an old dataset with say 25% removed
  - Ask them to create the full data set
  - Compare the total datasets, and compare the missing data

- Digital Twins
  - Take an old data set, remove some of the columns (i.e. answers)
  - Ask vendor to create their twins
  - Ask the questions not shared
  - Evaluate the answers

Q & A

esomar