The Scientist Game

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington
Seattle WA

May 14, 2024

RCTdesign.com – Emerson Statistics

Origins

- 1956: Eleusis invented by Robert Abbott
- 1977: Described by Martin Gardner *Scientific American*
- c. 1993: I adapted Scientist Game for enrichment activities in my daughters' elementary school classes
- 1994+: I started seeing how various collaborators fared
  - Statistician, oncologist, physicist, biologist colleagues
  - Graduate students in biostatistics, public health

Overview

- A simplified universe
  - One dimensional universe observed over time
  - Each position in the universe has an object
  - Goal is to discover any rules that might determine which objects are in a given location at a particular time

Objects

- Objects in the universe have only three characteristics, each with only two levels

  - Color: White or Orange
  - Size: BIG or small
  - Letter: A or B

- So only 8 kinds of objects in the universe:

A a B b A a B b

Universal Laws

- The level of each characteristic (color, size, letter) for the object at any position in the universe is either
  - completely determined by the prior sequence of **that** characteristic for objects at **that** position,

    OR
  - is completely random (anything is permissible)
  - (No patterns involving probabilities less than 1)
  - (Adjacent positions have no effect)

Universal Laws

- Furthermore any pattern to the objects at a position over time is "stationary"
  - The exact pattern repeats itself over a finite period of time (the "cycle")
  - The following "pattern" is not considered possible, because the exact same sequence does not re-appear

b A b A A b A A A b A A A A b A A A A A

Examples of Universal Laws

- Color only (a cycle of length 2):

b a A b a a a B A a A A

- The next object in the sequence must be white, but any size or letter will do:

a A b B

Examples of Universal Laws

- Size and letter (a cycle of length 4):

A a B b A a B b A a B b

- The next object in the sequence must be a big A, but any color will do

A A

Examples of Universal Laws

- Size only (a cycle of length 2):

B a B b A b B a B b A a

- The next object in the sequence must be big, but any color or letter will do:

A B A B

Examples of Universal Laws

- No discernible pattern (in available data):

A a b a A b B a B B A a

- If there is truly no deterministic pattern, then any object may appear next:

a A b B a A b B

Scientific Task

- Goal is therefore to decide for some position
  - whether a rule governs the level of each characteristic, and
  - if so, what that rule is (pattern to the sequence)

Hypothesis Generation

- Initially we have observational data gathered over time
  - Amount of available information varies from position to position
  - We want to identify some position that is the most likely to be governed by some deterministic rule

Observational Data

Time

Pstn	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
…														
114.	b	a	A	b	a	a	B	a	A	A	a	b	?	?
115.	A	a	B	b	A	a	b	b	A	a	a	A	?	?
116.							B	b	A	a	a	b	?	?
117.						b	B	A	b	b	b	a	?	?
118.	A	b	B	A	B	b	A	B	B	a	B	B	?	?
119.		B	b	B	b	A	a	A	a	B	A	b	?	?
120.										B	B	b	?	?
121.				B	A	a	B	b	b	a	b	A	?	?
…														

Observational Data

Time

Pstn	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
…														
118.	A	b	B	A	B	b	A	B	B	a	B	B	?	?
…														

Next Step?

- Further observation?
  - Might take too long
  - Won't really establish cause and effect

Experimentation

- You can try to put an object in the position
  - If it cannot come next, it disintegrates and you can try another
  - If it can come next, it stays and you can try a different object to follow it
  - Ultimately, a sequence of experiments can be used

Experimental Goal

- You need to devise a series of experiments to discover
  - whether a deterministic rule governs the sequence of objects at position 118, and
  - if there is such a rule, what it is

Real World

- Problem:
  - You must buy objects to experiment with
    - (apply for a grant)
- Question:
  - What object should you try next in the sequence in order to determine the rule?

Possible Experiments

Time

Pstn	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
118.	A	b	B	A	B	b	A	B	B	a	B	B	?	?

Possible Experiments

a A b B a A b B

- Which experiment do you do first?

Reviewing the Grant Application

- Did you choose a good experiment?

- In order to determine whether your grant application should be funded, we review an ideal scientific approach

  1. Observation

  2. Formulating hypotheses

  3. Devising experiments which discriminate between hypotheses

Results of Observation

Time

Pstn	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
118.	A	b	B	A	B	b	A	B	B	a	B	B	?	?

- We identified position 118 which had some regular patterns
  - Color cycle of length 2: (orange, white)
  - Size cycle of length 4: (big, little, big, big)
  - Letter cycle of length 3: (A, B, B)

Define Hypotheses

- Deterministic pattern vs random chance
  - Some (or all) of the observed patterns for each characteristic might be coincidence
- Is coincidence realistic?
  - Assume each level equally likely and a sample size of 12
  - Chance of observing a pattern for a single characteristic
    - 1 out of 1,024 for a cycle of length 2
    - 1 out of 512 for a cycle of length 3
    - 1 out of 256 for a cycle of length 4
    - 1 out of 134,217,728 for all three simultaneously

Possible Hypotheses

- Assuming sufficient data to see any rule

118. A b B A B b A B B a B B ? ?

Hypotheses

Color, Size, and Letter

Color, Size

Color, Letter

Size, Letter

Color

Size

Letter

All coincidence

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Possible Experiments

Hypotheses

A

Color, Size, and Letter

Color, Size

Color, Letter

Size, Letter

Color

Size

Letter

All coincidence

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	
Color, Letter	
Size, Letter	
Color	
Size	
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	
Size, Letter	
Color	
Size	
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Possible Experiments

Hypotheses	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	
Color	
Size	
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	+
Color	
Size	
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	+
Color	+
Size	
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	+
Color	+
Size	+
Letter	
All coincidence	

Most Popular First Choice

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	+
Color	+
Size	+
Letter	+
All coincidence	

Most Popular First Choice

- A noninformative experiment

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments
	A
Color, Size, and Letter	+
Color, Size	+
Color, Letter	+
Size, Letter	+
Color	+
Size	+
Letter	+
All coincidence	+

Next Worse Choice

- If all hypotheses equally likely, a 7-1 split

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments A	b
Color, Size, and Letter	+	-
Color, Size	+	-
Color, Letter	+	-
Size, Letter	+	-
Color	+	-
Size	+	-
Letter	+	-
All coincidence	+	+

If Eliminate Hypotheses 1 by 1

- Guessing a number between 1 and 1,000
  - You can ask Yes-or-No questions
  - Strategy 1: Elimination 1 by 1
    - "Is it 137? (NO) Is it 892? (NO) …"
    - On average it will take 500 questions
  - Strategy 2: Binary search
    - Split remaining hypotheses in half each time
    - "Is it > 500? (NO) Is it > 250? (YES) Is it > 375?..."
    - You can know the answer after 10 questions (2^{10}=1,024)

Other Suboptimal Experiments

- If all hypotheses equally likely, a 6-2 split

118. A b B A B b A B B a B B ? ?

Hypotheses	A	b	b	B	a
	Possible Experiments				
Color, Size, and Letter	+	-	-	-	-
Color, Size	+	-	-	-	-
Color, Letter	+	-	-	-	-
Size, Letter	+	-	-	-	-
Color	+	-	+	-	-
Size	+	-	-	+	-
Letter	+	-	-	-	+
All coincidence	+	+	+	+	+

Optimal Experiments

- Based on a binary search

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments: A	b	b	B	a	B	a	A
Color, Size, and Letter	+	-	-	-	-	-	-	-
Color, Size	+	-	-	-	-	+	-	-
Color, Letter	+	-	-	-	-	-	+	-
Size, Letter	+	-	-	-	-	-	-	+
Color	+	-	+	-	-	+	+	-
Size	+	-	-	+	-	+	-	+
Letter	+	-	-	-	+	-	+	+
All coincidence	+	+	+	+	+	+	+	+

Interpreting Good Experiments

- We can easily describe what we were testing for in the three “best” experiments

  – Is Letter important? **B**
    - We used the size and color that would work regardless

  – Is Size important? **a**
    - We used the letter and color that would work regardless

  – Is Color important? **A**
    - We used the size and letter that would work regardless

Sequence of Experiments

- Separate question into three experiments
  - Address each characteristic separately
  - Avoid “confounding” the question
- Perform these 3 experiments in sequence
  - Results uniquely identify the 8 hypotheses
  - (Eliminating hypotheses 1 at a time would on average take 4 experiments)
- (No other series of 3 will always do this)

Optimal Experiments

- Based on a binary search

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments		
	B	a	A
Color, Size, and Letter	-	-	-
Color, Size	+	-	-
Color, Letter	-	+	-
Size, Letter	-	-	+
Color	+	+	-
Size	+	-	+
Letter	-	+	+
All coincidence	+	+	+

Other Experimental Sequences

- Based on results from first experiment, additional good experiments for the second stage possible, e.g.,
  - Suppose first experiment: **A** <u>and</u> it does not disintegrate
    - Then we know color does not matter
  - Good choices for the next experiment: **B a B a**
    - Choose different letter or size, but not both
  - BUT: If first experiment had disintegrated, only two good choices: **B a**
    - Must use orange, because color matters

Supposing A Works

- Based on a binary search

118. A b B A B b A B B a B B ? ?

Hypotheses	Possible Experiments: B	a	B	a	A
Color, Size, and Letter					
Color, Size					
Color, Letter					
Size, Letter	–	–	–	–	+
Color					
Size	+	–	+	–	+
Letter	–	+	–	+	+
All coincidence	+	+	+	+	+

Bayesian Considerations

- With a binary search, we want to eliminate 50% of the hypotheses based on their “prior probability”
- If we had a prior belief that most of our universe was random chance, then perhaps **b** might be a good first experiment
- We can cook up an example to try to make that a good Bayesian choice when we have seen (perhaps spurious) patterns like 118

Toy Example

- Assume 99.85% of positions truly have no pattern
- Others have “independent, equiprobable patterns” with cycle length < 12
- Expect to see 0.000000002% with patterns like 118
  - Maybe we examined millions of positions
- Best approach may have been to try: **b**
  - Discriminates between no pattern (like 49% of positions) and some pattern (like the other 51% of positions)
  - On average, 2.52 experiments (1 expt 49%, 4 expt 51%)

Details of Toy Example

- Not for the faint of heart, but…
- Suppose color, letter, size independent
- For each factor
  - 99.5% of sites have no pattern
  - Rest equally likely to have cycles of 2, 3, or 4
  - For every length of cycle, all patterns equally likely
    - E.g., for big white letters
      - Cycle length 2: AA, AB, BB each 1/3
      - Cycle length 3: AAB, ABB each 1/2
      - Cycle length 4: AAAB, AABB, ABBB each 1/3

Posterior Probabilities for 118

- We chose a site with observed patterns of color cycle=2, letter cycle=3, size cycle=4

- Of all such patterns, the truth will be
  - All coincidence 49.5%
  - Letter only 16.9%
  - Size only 11.3%
  - Color only 11.3%
  - Letter, size only 3.8%
  - Color, letter only 3.8%
  - Color, size only 2.6%
  - Color, letter, size 0.9%

What If Data Insufficient?

- Suppose deterministic cycle length > 12

118. A b B A B b A B B a B B ? ?

Hypotheses	A	b	b	B	a	B	a	A
	Possible Experiments							
Color, Size, and Letter	+	-	-	-	-	-	-	-
Color, Size	+	-	-	-	-	+	-	-
Color, Letter	+	-	-	-	-	-	+	-
Size, Letter	+	-	-	-	-	-	-	+
Color	+	-	+	-	-	+	+	-
Size	+	-	-	+	-	+	-	+
Letter	+	-	-	-	+	-	+	+
All coincidence	+	+	+	+	+	+	+	+
Cycle length > 12	?	?	?	?	?	?	?	?

If Cycle Length > 12

- We would have no information to be able to guess the true pattern

- BUT, we might gain some information from A as a first experiment (and thus **partial** vindication)

  – If A disintegrated we would know that there was some deterministic pattern with cycle length > 12
    - But we would have no clue about the pattern

  – And we have to question the efficiency of this strategy
    - Deterministic patterns of length > 12 might demand A and thus we cannot presume one of the other hypotheses is established

Moral: Hypotheses

- The goal of the experiment should be to “decide which” not “prove that”
- A well designed experiment discriminates between hypotheses
  – The hypotheses should be the most important, viable hypotheses

Moral: Experiment

- All other things being equal, an experiment should be equally informative for all possible outcomes
- In the presence of a binary outcome, use a binary search
  - (using prior probability of being true)
- But need to consider simplicity of experiments, time, cost
  - (What lessons can be learned from Master Mind?)

In the Presence of Variability

- We use statistics to quantify the precision of our inference
- We will describe our confidence/belief in our conclusions using frequentist or Bayesian probability statements
- Discriminating between hypotheses will be based on a frequentist confidence interval or a Bayesian credible interval

Interval Estimates

- Frequentist confidence intervals
  - The set of all hypotheses for which the observed data are "typical"
  - There is more than a negligible probability of obtaining such results when those hypotheses are true
- Bayesian credible intervals
  - The set of hypotheses that are most probable given the observed data
  - Must incorporate our prior belief in the hypotheses

Frequentist Evidence

- Does frequentist evidence provide evidence?
  - Is it relevant to calculate the probability of data that you know you observed?
  - Relevance especially questionable if calculated on a hypothesis that is unlikely *a priori*

- My answer in experimental design: Yes
  - Design an experiment that has results that are not consistent with one of the viable, important hypotheses

Statistical Experimental Design

- I believe a scientific approach to the use of statistics is to
  - Decide a level of confidence used to construct frequentist confidence intervals or Bayesian credible intervals
  - Ensure adequate statistical precision (sample size) to discriminate between relevant scientific hypotheses
    - The intervals should not contain two hypotheses that were to be discriminated between

Impact on Statistical Power

- In evaluating a design, I always examine the alternative hypothesis for which we have I equal one-sided type I and type II errors
  - E.g., 97.5% power to detect the alternative in a one-sided level 0.025 hypothesis test
- In this way, at the end of the study, the 95% CI will not contain both the null and alternative hypotheses
  - I will have discriminated between the hypotheses with high confidence