And now for something completely different: Evidential statistics and the analysis of ordinal patterns

Warren Thorngate, Emeritus professor
Psychology Department, Carleton University
Ottawa, Ontario, Canada
warren.thorngate@carleton.ca

Presented at TARDIS, University of North Texas
19 September 2015

Warm-ups

Beware of the man of one method or one instrument, either experimental or theoretical.

John R. Platt (1964). Science, strong inference – proper scientific method (the new Baconians). Science, 146(1642), 347-353. quoted from p. 352.
http://pages.cs.wisc.edu/~markhill/science64_strong_inference.pdf

We do not make intellectual progress by challenging conclusions. We make intellectual progress by challenging assumptions.

Tamostu Shibutani, Lecture in sociology, 1964.

https://en.wikipedia.org/wiki/Tamotsu_Shibutani

Traditional statistical practice

- 95% of all statistical analyses in social science research employ fewer than 5% of available statistical tests
- Almost all of the Chosen Few
  - Are variants of the General Linear Model
    - Rely on normal, parametric assumptions
    - Focus only on differences among means
    - Partition variance into epicyclic, orthogonal components
  - Cling to Neyman-Pearson "significance testing"
    - Concern with generalizing from sample to population (inference) rather than from prediction to observation (evidence)
    - Think of inference as a game of 20 Questions (Newell, 1973)

Why are so few alternatives employed?

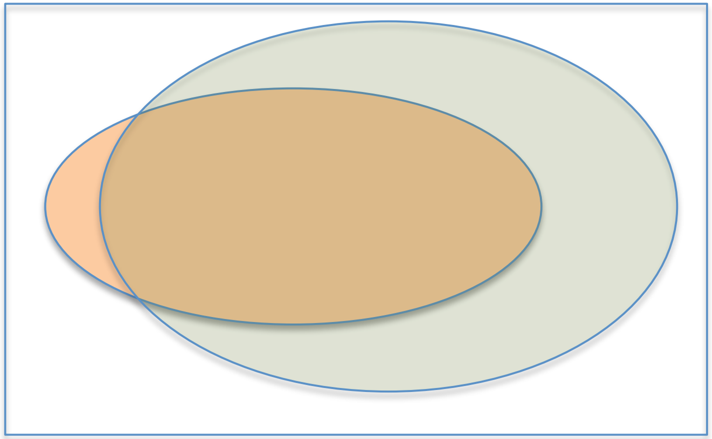
- Typical answer: “Almost all of the important questions in psychology can be answered with traditional statistical methods (GLM variants and N-P inference)!”
  - Is this true?
  - If not true, what alternatives exist and how might they improve our analytical toolkit?

Here I stand

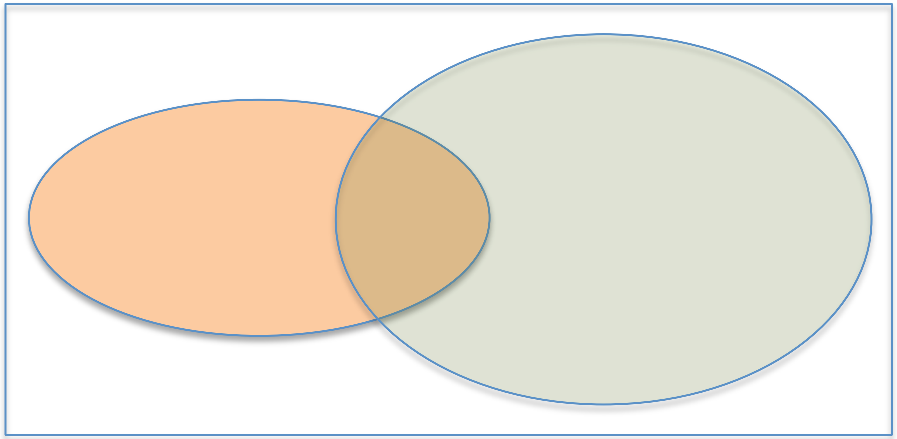
Important questions in psychology

Questions answerable by General Linear Model (think SPSSP menu items)

Current research ideology



My opinion



History might well show our overuse of the GLM has done more to retard the development of social sciences than any other methodological constraint.

What is good about the GLM?

- It is mathematically elegant
  - Everything adds up!
- It is relatively easy and profitable to derive and extend GLM formulas
  - So it can be extended for different research designs, if assumptions are met
    - Consider the proliferation of omnivariate techniques on SPSS menus
- Lots of people know how to use it, and don't want to change
  - But have modified their thinking to fit its Procrustean Bed
  - like the QWERTY keyboard, or English spelling

What is bad about the GLM + NP inference? (A small sample)

- It requires data from a ratio scale, which we rarely have
- It focuses on differences in central tendency, and dismisses differences in variability and shape of data.
- It partitions variance into Ptolemaic cycles and epicycles, tempting us to use statistical models as psychological models
- It is generally unsuitable for examining single cases, and severely limited in examining correlated observations
- It limits inferences to “on average” which are frequently not true “in general”
- Its inference engine is unsuited for asking many of our most important research questions:
  - Not “Can we generalize from samples to populations?”
  - But “Can we generalize from theories to observations?”

What to do?

Hints from other sciences and music

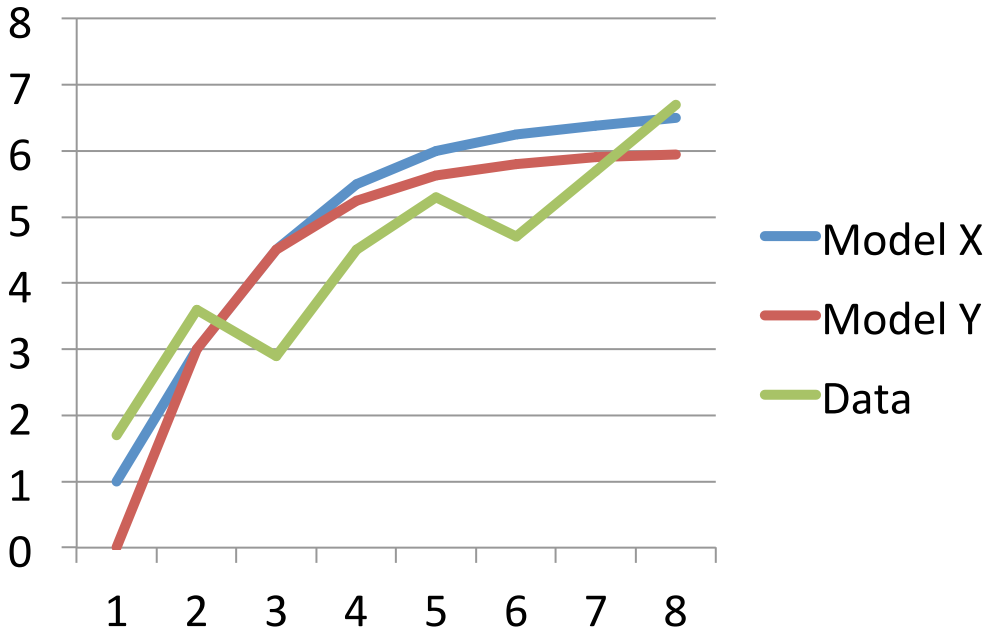
A bit of personal history

- Background in math, physics, music
- Specialized in mathematical modeling and computer simulation of social behaviour addressing research questions in untraditional ways
- Hung out with evolutionary biologists, ecologists, and other scientists who think and do science differently than we do
- Learned important sylistic difference between
  - Empiricists versus rationalists
  - "Variance Splitters" versus "Goodness of Fitters"
  - Inferential versus evidential statistics
- Began to wonder if alternative approaches might benefit psychological research

Example: Assessing mathematical models

- How close is the fit between predictions and observations?
  - Inter-ocular trauma test
  - Goodness of fit measures (evidential statistics)

– Praying for insignificant prediction-observation differences!
  - Because significant differences invalidate the model



Limits of mathematical modelling

- Most theories in psychology make only ordinal predictions, not point predictions required by mathematical models

- Most data collected in psychology have at best ordered metric properties, and most only ordinal properties

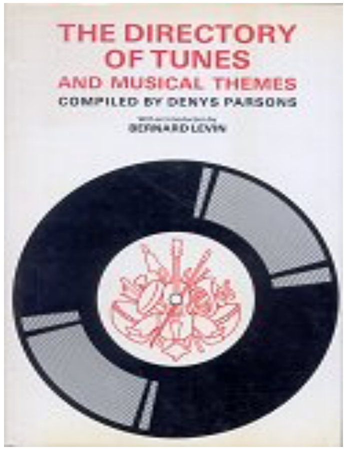
- Most goodness-of-fit assessments employ aggregated data (usually averages) under the often-mistaken assumption that all participants are using the same mental processes

What the world needs

- A simple means to evaluate the fit between ordinal predictions and ordinal properties of observations
  - Men will score lower than women
  - Scores will improve with practice (a predicted order, more quickly sooner than later (a predicted order of differences)
  - Stress will be higher on Mondays and Fridays than on Tuesdays and Thursdays, and will be lowest on Wednesdays (a semi-order)
- A way to test prediction-observation fit with both individuals and aggregates
- Simple, easily calculated, interpretable indices of fit
  - That can be aggregated across people and studies

Enter the CBC, Barbara Frum and Denys Parsons

- CBC started the radio show *As it Happens* in 1968
- Hosted by Barbara Frum for many years
  - Son = David Frum, G.W. Bush speechwriter
- Frum frequently interviewed British eccentrics to end shows on a light note
- In 1975, she interviewed a British Library press officer *Denys Parsons* (1914-?), author of his Directory of Tunes and Musical Themes
- Eureka! An idea was born



Denys Parsons (1975)
Directory of Tunes and Musical Themes

- Coded the adjacent ordinal relations of the first 16 notes of popular tunes: Up, Down, Repeat
- Examples:
  - Oh Canada: URDUUUUUDUURUUUR
  - Ode to Joy: RUURDDDDRUURDRUR
  - Doe a deer: UUDUDUDUURDDUDUU
- No code repeated in 14,000 tunes!
  - Like testing goodness of fit of 14,000 theories!
- Extend Parsons' Code = ordinal pattern of adjacent notes
  - Generalize beyond adjacent notes to all pairs of notes
  - Replace notes with psychological data
- **Voila! Ordinal Pattern Analysis (OPA)** is born

The basics of OPA

- Gather at least two measurements you think are relevant to testing a theory
- Ask yourself "How often do the predictions derived from my theory match the sample of measures I have?"
- Count, a la Kendall's Tau
  - **#hits** = the number of times the order of each possible pair of predictions is matched by the order of relevant measurements
  - **#misses** = the number of times the order of each possible pair of predictions is mismatched by the order of relevant measurements
  - **Ignore ties** (unless predicting weak orders, ≥ or ≤) and **missing data**
- Calculate
  - **pH** = Probability of a Hit = #hits / (#hits + #misses)
    - or if you prefer…
  - **IOF** = Index of Observed Fit = (#hits - #misses) / (#hits + #misses)
- Save #hits and #misses for future meta analyses.

Example 1

- **Theory**
  - The more neural connections we have, the more irony we perceive. Neural connections increase with age. Ergo, people more frequently detect irony as they age
- **Method**
  - Fred takes the Irony Questionnaire (IQ) in the following years of age (aN)
    - a10, a15, a22, a25, a31, a32, and a49
  - Irony Questionnaire is scored out of 100 = highest irony
- **Predicted Ordered Pairs (POP set) of IQ scores**
  - a10<a15, a10<a22, a10<a25, a10<a31, a10<a32, a10<a49
  - a15<a22, a15<a25, a15<a31, a15<a32, a15<a49
  - a22<a25, a22<a31, a22<a32, a22<a49
  - a25<a31, a25<a32, a25<a49
  - a31<a32, a31<a49
  - A32<a49

Example 1 continued

- **Results = IQ scores on Irony Questionnaire at different ages (aN)**
  - a10 = 39
  - a15 = 32
  - a22 = 41
  - a31 = 67
  - a32 = 58
  - a49 = 67

- **Scores = evidential statistics**
  - #hits = 4+4+3+0+1 = **12**
  - #misses = 1+0+0+1+0 = **2**
  - #ties = 0+0+0+1+0 = **1**
  - #NA= 0+0+0+0+0 = **0**
  - **pH** = 12 / (12+2) = **+0.86** (86% correct predictions)
  - **IOF** = (12-2) / (12+2) = **+0.71** (71% better than chance)

- **Significance** = inferential statistics
  - Use resampling or bootstrap techniques. But who cares?

Example 2

- No-Name Resilience Theory derivation
  - Depression is highest immediately following a tragedy, then steadily declines
  - Decline is more likely to happen for women than for men
  - Decline is more likely to happen for younger people than for older people

Depression scores following a tornado (0-10)

Person	Age	Day d1	Day d2	Day d3	Day d4	Day d5
F1	19	9	6	6	3	4
F2	21	NA	5	7	NA	3
F3	34	5	7	3	3	9
F4	52	2	5	4	7	9
M1	20	7	6	7	5	2
M2	22	4	4	3	4	3
M3	31	4	3	8	9	8
M4	47	1	NA	4	7	7
	Average =	4.6	5.1	5.2	5.4	5.2

- POP set for daily decline of depression
- d1>d2, d1>d3, d1>d4, d1>d5
- d2>d3, d2>d4, d3>d5
- d3>d4, d3>d5
- d4>d5

Results

Person	Age	Hits	Misses	Ties	NA	IOF
F1	19	8	1	1	0	+0.78
F2	21	2	1	0	7	+0.33
F3	34	4	5	1	0	-0.11
F4	52	1	9	0	0	-0.80
M1	20	8	1	1	0	+0.78
M2	22	5	1	4	0	+0.67
M3	31	2	7	1	0	-0.56
M4	47	0	5	1	4	-1.00
Females		15	16	2	7	**-0.03**
Males		15	14	7	4	**+0.03**
	< 30 years	23	4	6	7	**+0.70**
	>30 years	7	26	3	4	**-0.58**

Semi-orders and scope

- Semi-orders occur when predictions do not address all pairs of relevant data
  - Example: Productivity is higher on Tuesdays than on any other workday
    - Predicts Tu>Mo, Tu>We, Tu>Th, Tu>Fr
    - Does not address ordinal relations among Mo, We, Th, Fr
- Scope = #pairs addressed by predictions / #pairs in data
  - Example:
    - There are 5x4 / 2 = 10 possible pair of weekdays
    - Above prediction about Tuesdays addresses 4 of them
    - So Scope of prediction = 4 / 10 = 0.40
  - Scope is important when comparing fits of different predictions

Useful subsets of predicted ordered pairs (POP subsets)

- When comparing POPs from 2+ different theories (say, theories X and Y), consider separately
  - 1. Ordered pairs addressed by X but not by Y (X-unique)
  - 2. Ordered pairs addressed by Y but not by X (Y-unique)
  - Ordered pairs addressed by both
    - 3. Those predicting the same ordinal relations = **convergent** predictions. Example: X predicts A>B; Y predicts A>B)
    - 4. Those predicting opposite ordinal relations = **divergent** predictions. Example: X predicts A>B; Y predicts B>A)
- Divergent predictions are classic tests of competing theories, but fits of the three other subsets can also be illuminating

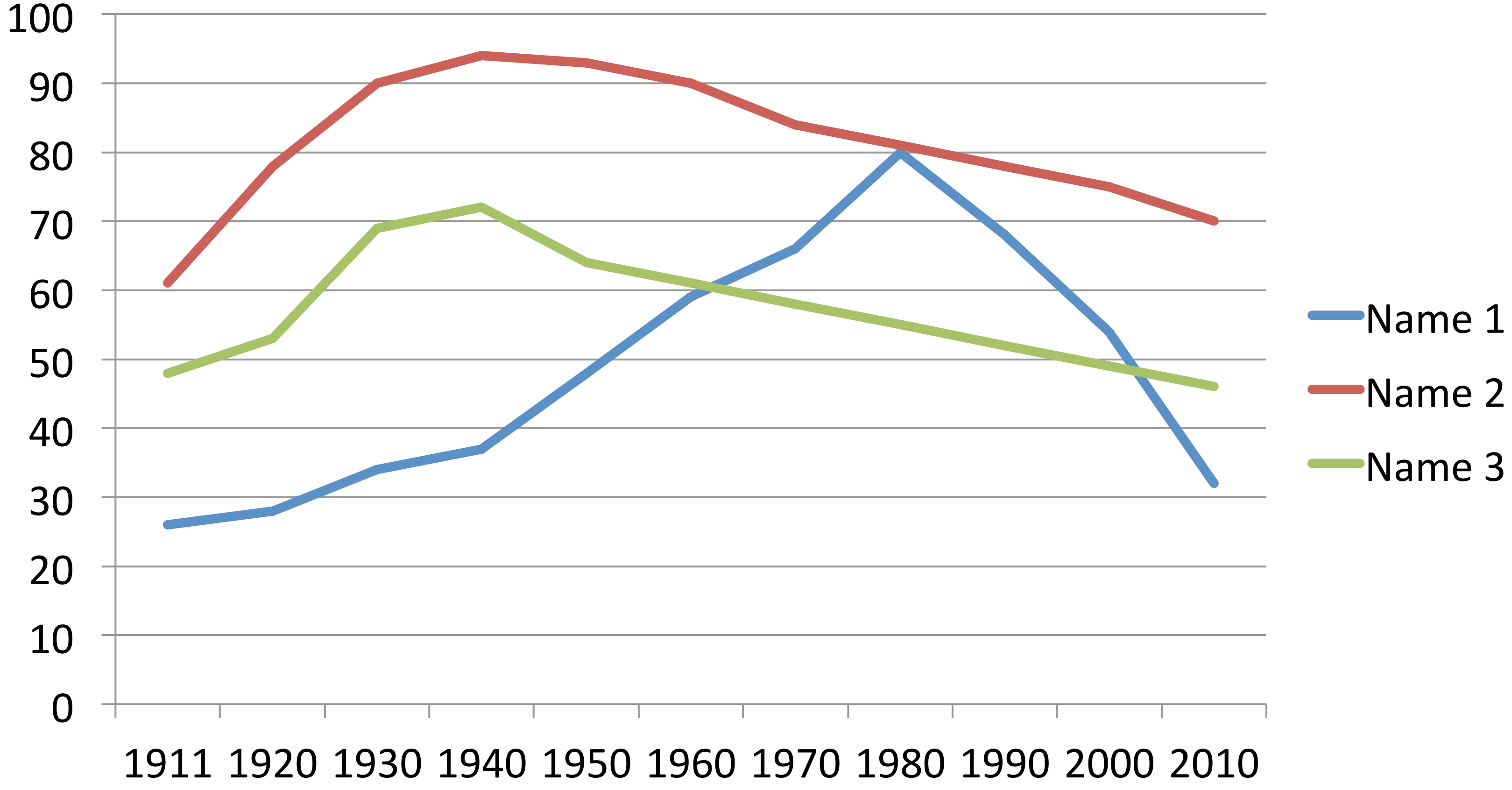
Example 3
The rise and fall of baby names

http://www.ssa.gov/oact/babynames/

- Theory/simulation A
  - predicts names will increase in popularity until a single peak, then fall
- Theory/simulation B
  - predicts same as A
  - Also predicts that fall will be faster than rise
- Theory/simulation C
  - Predicts faster recovery from nadir than decline to nadir

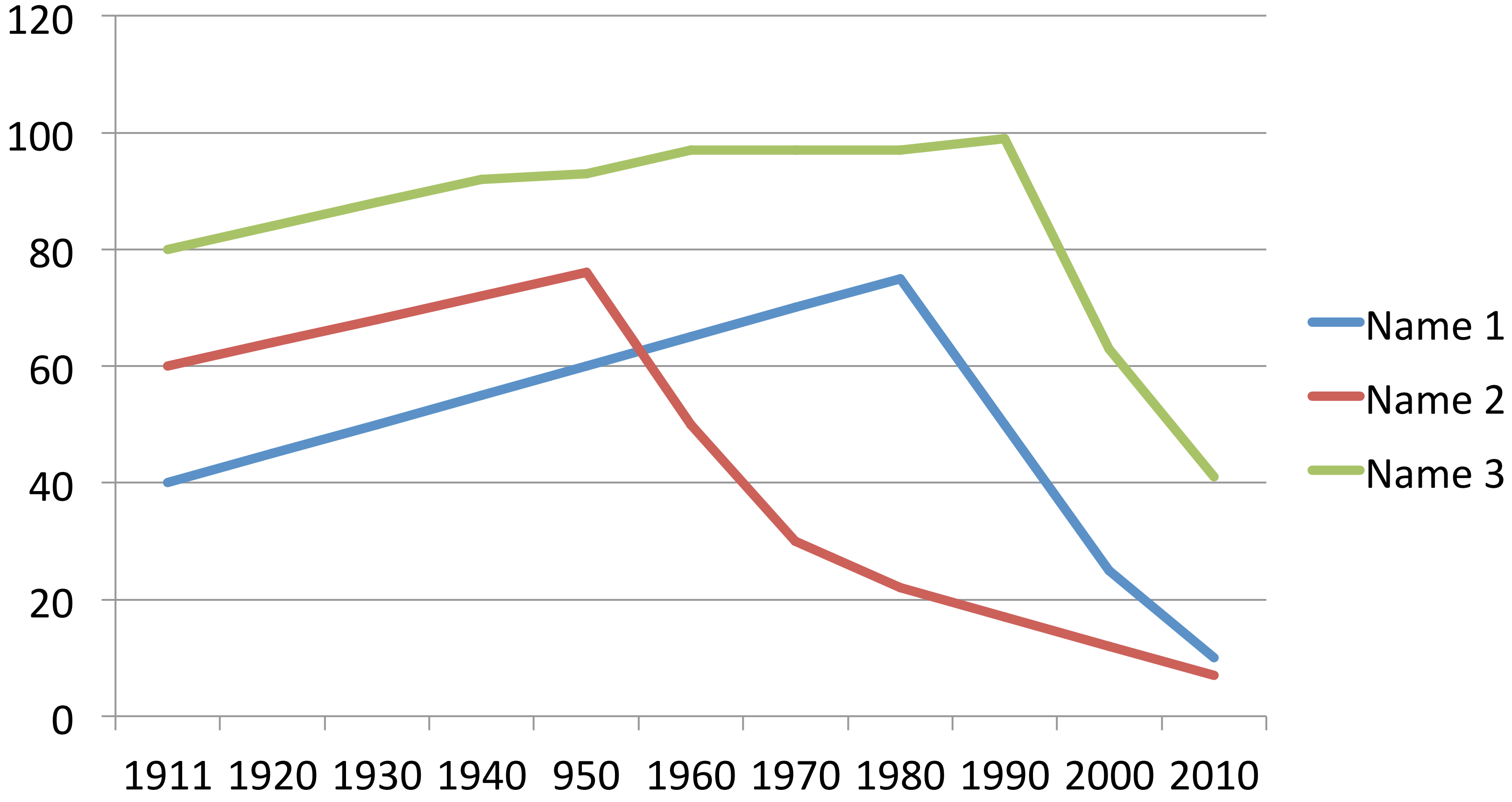
Typical outputs of Simulation A

Names will increase steadily in popularity until a single peak, then fall steadily.



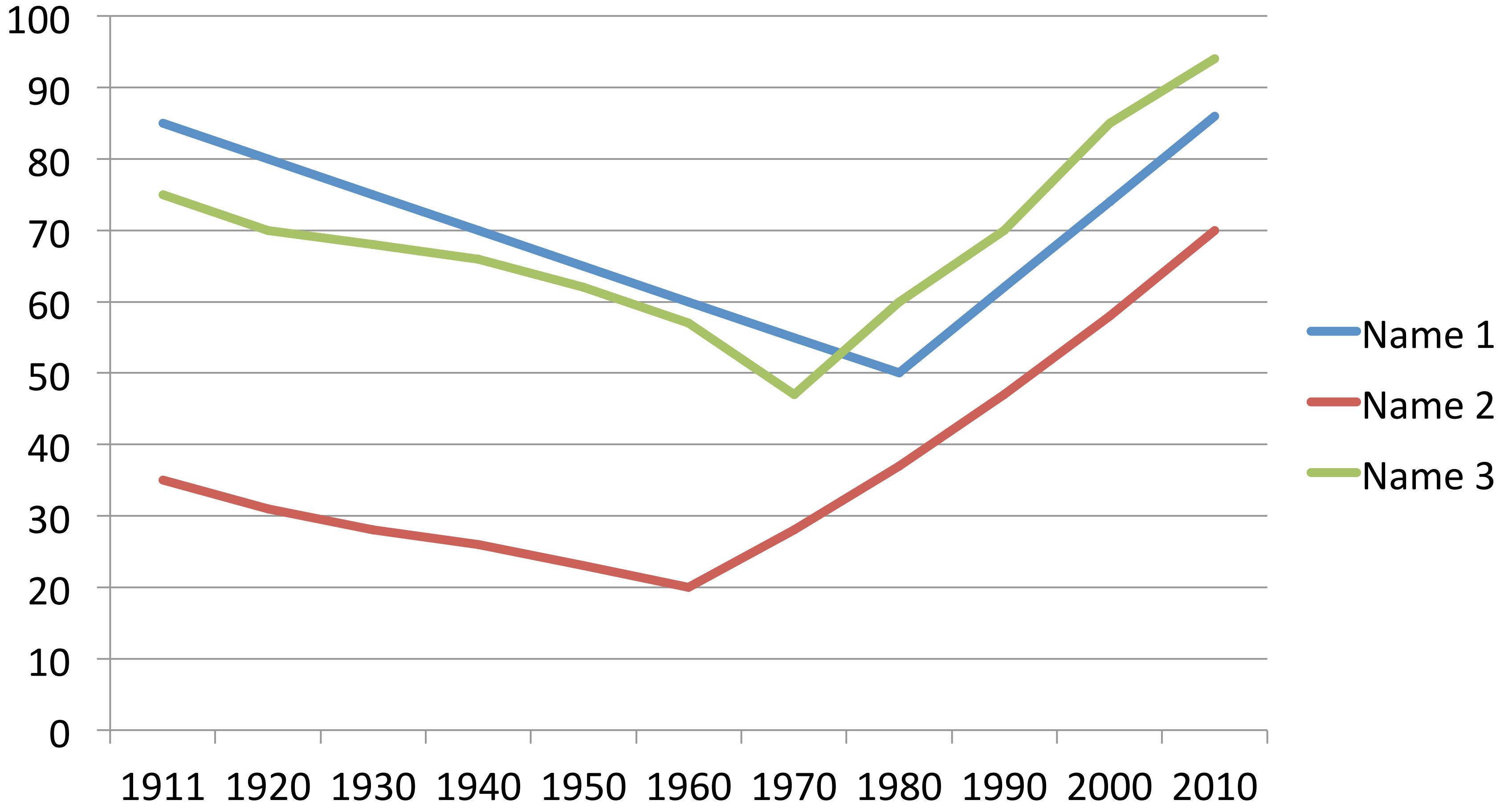
Typical outputs of Simulation B

Names will increase steadily in popularity until a single peak, then fall steadily. Fall will be faster than rise



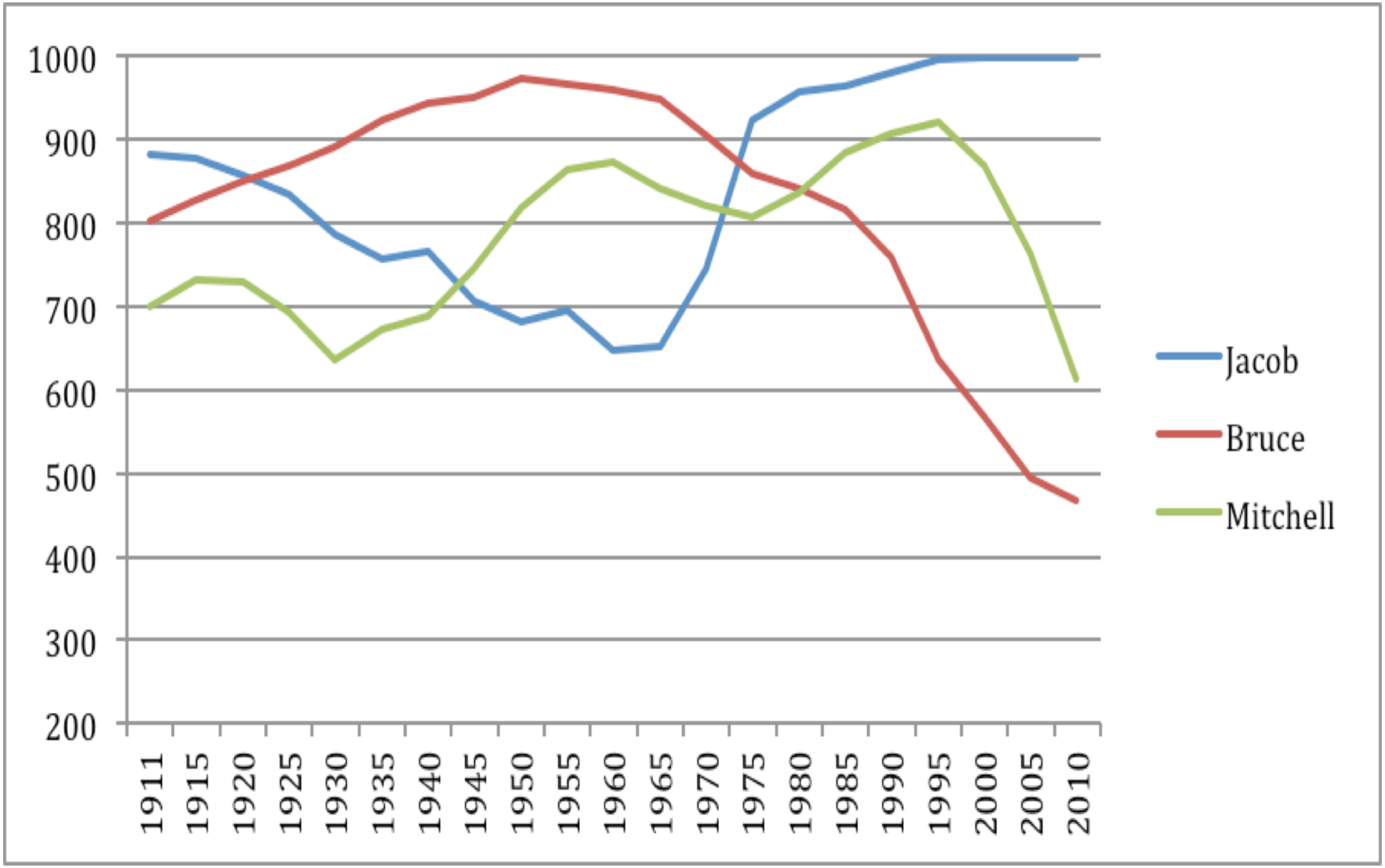
Typical outputs of Simulation C

Faster recovery from nadir than decline to nadir



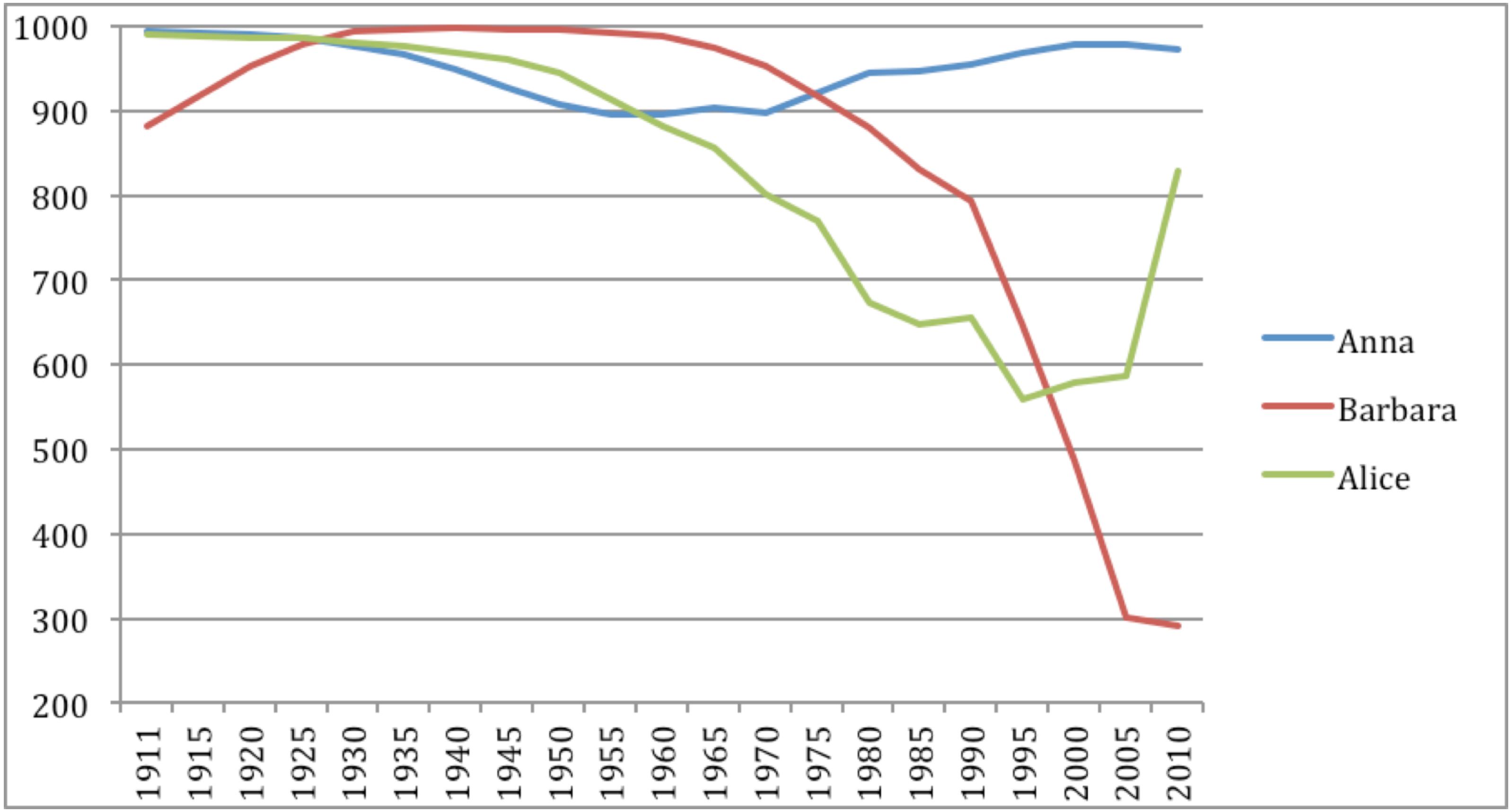
US popularity of 3 male baby names

(inverse rank from top 1,000 names)



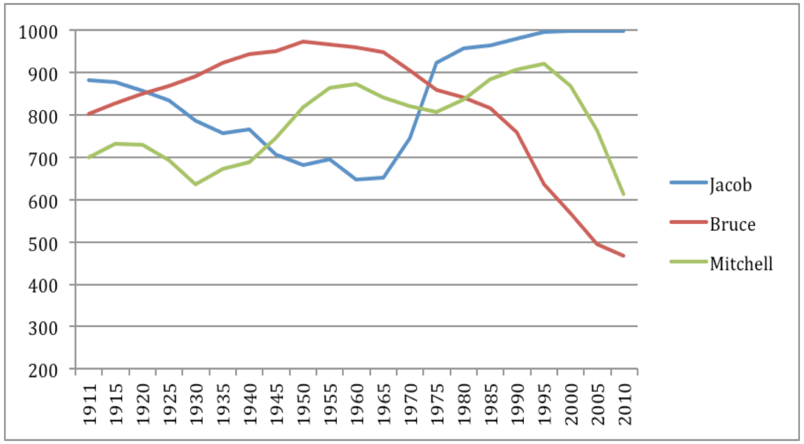
US popularity of 3 female baby names

(inverse rank from top 1,000 names)



Testing fit of data to Theory A
Example = Bruce

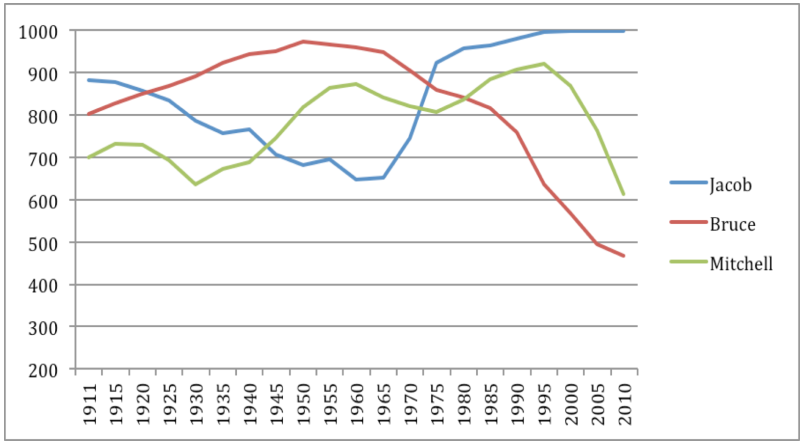
- Step 1: look for peak (1950), then make predictions on either side of it
- Step 2: generate predictions either side of peak
  - 1945 > 1940, 1935, 1930,…, 1911
  - 1955 > 1960, 1965, 1970,…, 2010
  - 1940 > 1935, 1930,…, 1911
  - 1960 > 1965, 1970,…, 2010
  - …
  - 1920 > 1911
  - 2005 > 2010



- Step 3: Count hits and misses
  - 1945 > 1940? Yes. 1945 > 1935? Yes…
  - 1955 > 1960? Yes. 1955 > 1965? Yes…
  - …
  - 2005 > 2010? Yes.
- Step 4: calculate index of fit (ignoring ties)
  - pH = #hits/ (#hits+ #misses) 153 /(153 + 0) = +1.00
  - IOF = (#hits- #misses) / (#hits+ #misses) = 153/153 = +1.00

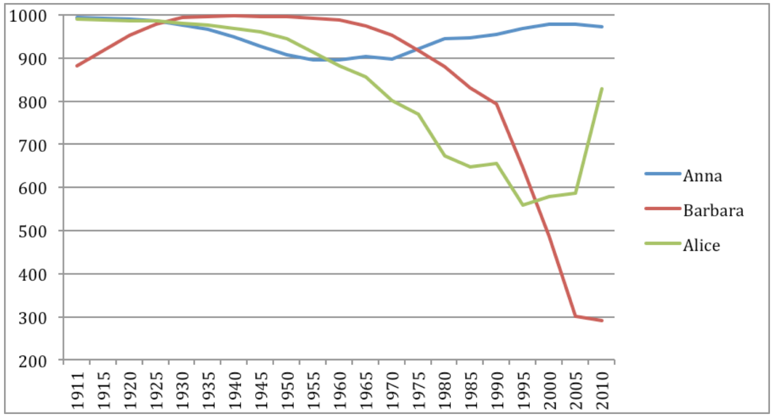
Testing fit of data to Theory A
Example = Jacob

- Step 1: look for peak (2000) , then make predictions on either side of it
- Step 2: generate predictions either side of peak
  - 1995 > 1990, 1985, 1980,…, 1911
  - 2005 > 2010
  - 1990 > 1985, 1980, 1975,…, 1911
  - …
  - 1915 > 1911
- Step 3: Count hits and misses
  - 1995 > 1990? Yes. 1990 > 1980? Yes.
  - 2005 > 2010? Tied (ignore)
  - 1960 > 1955? No. 1960 >1950? No.
  - …
  - 1915 > 1910? No.
- Step 4: calculate index of fit (ignoring ties)
  - pH = 83 / (83+70) = 0.54
  - IOF = (#hits- #misses) / (#hits+ #misses) = 0.08



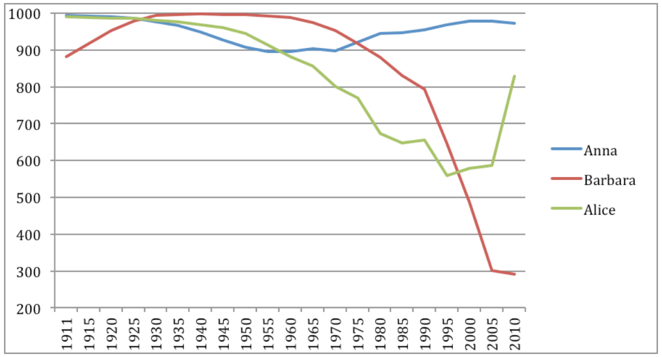
Testing fit of data to Theory B
Example = Barbara

- Step 1: look for peak (1940) , then make predictions on either side of it
- Step 2: generate predictions on either side of nadir
  - 1935 > 1945, 1950, 1955,…, 2010
  - 1930 > 1950, 1955, 1960,…, 2010
  - 1925 > 1955, 1960,…, 2010
  - 1915 > 1965, 1970, 1975,…, 2010
- Step 3: Count hits and misses
  - 1935 > 1945? Tied. 1935 > 1950? Yes.
  - 1930 > 1950? No.
  - 1925 > 1955? No.
  - 1920 > 1960 No.
  - 1965 > 1915? No.
- Step 4: calculate index of fit (ignoring ties)
  - P(Match) = 0.88
  - Index of Observed Fit = 43/57 = 0.75



Testing fit of data to Theory C
Example = Alice

- Step 1: look for nadir (1995) , then make predictions on either side of it
- Step 2: generate predictions on either side of nadir
  - 2000 > 1990
  - 2005 > 1985, 1990
  - 2010 > 1980, 1985, 1990
- Step 3: Count hits and misses
  - 2000 > 1990? No.
  - 2005 > 1985? No. > 1990? No.
  - 2010 > 1980? Yes. > 1985? Yes. > 1990? Yes.
- Step 4: calculate index of fit (ignoring ties)
  - pH = 3 / (3 + 3) = 0.50
  - IOF = 0.00



Now aggregate after analysing

Simulation:	Theory A	Theory B (A-part & extra)	Theory C
Jacob	0.54	0.54 & NA	0.36
Bruce	1.00	1.00 & 0.71	NA
Mitchell	0.67	0.67 & 0.92	NA
Anna	0.63	0.63 & 0.48	0.21
Barbara	1.00	1.00 & 0.88	NA
Alice	0.72	0.72 & NA	0.50
Scope:	TBC	> Scope of Theory A	TBC
Best of the lot?		X	

Further topics

- Combining data across studies
  - Remember, we are allowed to test any theory with any samples of people, tasks, times, etc. that are not part of a theory's scope conditions
  - So simply keep track of hits and misses from each study, then add them up when combining studies. Example:
    - Study 1: 32 hits, 46 misses; IOF = -14/78 = -0.18
    - Study 2: 5 hits, 3 misses; IOF = 2/8 = +0.25
    - Combined: 37 hits, 49 misses; IOF = -12/86 = -0.14
- Delineating prototypes
  - For each pair of research conditions, X and Y, count how often X>Y and Y>X. The most common becomes part of the prototype description
    - Example: If 17 of 20 participants show X>Y, 16 of 20 show Y>Z, then any theory that predicts X>Y, Y>Z and X>Z will have maximum possible pH or IOF

Last words

- Using pH scores to determine domains of validity.
  - Cluster analyses of pH scores to separate groups of participants who support one set of predictions versus those who support other sets.
  - Subdivisions of time
- Is one theory significantly more valid than another?
  - If you cannot break the habit, you can do a one-way ANOVA on sets of pH values where each pH element in a N-element set represents how well one person's data fits each of N theories.
  - In this way, pH and IOF values become (derived) dependent variables
- Confounds
  - As always, high pH or IOF values do not confirm a theory, because the ordinal patterns they match might be caused by one or more confounds. Test situations must be created or selected to separate theoretical predictions from confounding predictions

Enough!

Thank you for your attention

warren.thorngate@carleton.ca
warren.thorngate@rogers.com