UCC Mathematical Sciences Seminar Series
2025.11.13

Lyrics to accompany the slides

Thank you. Good afternoon. I am going to talk about two historical articles that still have lessons – and surprises – for us today. They also show a side of Ronald Fisher's personality that few got to see. And you will get to see how one of them was peer-reviewed. 59/59

==========

So here is the **plan**. • There were two investigations. The first ruled out the father's age, and the second ruled out the birth order. Both retellings are now in print, and I have put also put the supplemental information on two of my webpages. So I won't give you all the details in this presentation. 56/115

==========

• The story is a **mix** of epidemiology, genetics, statistics, optimization, history – and archaeology!. It's a chance to reflect on **how far we have come** in statistical methods/com but also on what we have **forgotten/missed** along the way. Teaming up with Supratik to recreate the data for the first study taught me **a lot** about **sufficient statistics**, and about **data privacy/disclosure**. The **concept** of sufficiency was covered in the formal math-stat courses I had taken and taught in my 55 years in statistics, but its full **implications** were not not all that real for me until two years ago.

And, for me, and for one of the Biometril reviewers, getting to go behind the scenes of the second study 'felt like the statistical equivalent of finding and opening an unknown Pharoah's tomb.' (Those are the reviewer's words). 140/255

==========

• In the late 1950s, what we today call Down syndrome or Down's syndrome was found to be a **genetic disorder** caused by the presence of all or part of a **third copy of chromosome 21**. Its scientific name is **trisomy 21.** It is typically associated with **physical growth delays**, characteristic **facial** features, and a range of **intellectual disability**. 59/314

==========

• This Wikipedia page gives some of its important **epidemiology** features, including the **large role of maternal age**. But when and how did it become clear that it's the mother's age that matters? 33/347

==========

• It goes back to the **statistical work of the physician-scientist Lioned Penrose**, and his behind-the-scenes collaborator Ronald Fisher. Penrose's main work was on genetics of intellectual deficit, but he had wide ranging interests. As a Quaker, he opposed war, and spent the World War II years working in Canada. After the war, he returned and became the chair of genetics at UCL when Fisher moved to Cambridge University. His wife Margaret and his four children, Oliver, Roger, Jonathan and Shirley, all had notable careers too. 86/433

==========

• This is the title of his 1933 paper 9/442

==========

• Penrose starts out with the tricky statistical issue of what today we call **confounding**: the strong correlation of the parent's ages means that that if you look at EITHER parent, the probability of Downs will seem to be strongly related to **that** parent's age. So, how to disentangle them and find which is the driver and which is the passenger? 61/503

==========

• He borrowed an idea from the eminent American geneticist. In research on guinea pigs, Sewall Wright had used **partial correlations** to show that once one **removed (eliminated)** the effect of the mother's age, there was no effect of the father's age. 42/545

==========

• Today the idea of computing a Pearson correlation between a **binary** (0/1) Y and a **quantitative** X seems strange, but it was common back then, and so Penrose used it too. Here is all he says about his data. 40/585

==========

• From what we can tell, Penrose did **all** of his analyses based on the frequencies in this one table. We will come back to it soon. There are 154 cases of Down syndrome and 573 Normal children from the **same** 150 families. Today you would keep, and share, the raw data in a data frame with 727 rows, 1 per child, and 3 columns, D(0/1), M and F. 69/654

==========

• These are the 'CRUDE' correlations, where he **collapsed over the ages of the other parent**. They are both strong. And of course, no surprise, the correlation between the parents' ages is very high: 0.829. But what happens when he, as the social scientists like to say, **partials out** the effect of the age of that other parent? The effect of the mother's age remains, but the effect of the father's age disappears. As he went on to say in his paper, paternal age is not a significant factor, while maternal age is to be regarded as very important. 99/753

==========

• This is that table again, highlighting the frequencies he used to calculate each of the 3 pairwise correlations: a 42 x 2 frequency table of fathers' ages x Down/Normal children, the corresponding 31 x 2 table involving mothers' ages, and the 42 x

31 one for the 727 pairs of parental ages. 53/806

==========

• To be fair to Penrose, he did anticipate your discomfort (and reviewers' discomfort) with correlations involving a BINARY variable, so he also used a **crisper alternative** 27/833

==========

• one that you are probably more comfortable with. It also fits with how a lot of people (still) view case-control studies. For now, ignore the 2 equations and just stay with the 4 mean numbers in black (the observed ages, the 39.3 vs 33.8 and the 37.2 vs 31.2), which show that the parents of the cases are decidedly older. But it still leaves open whether there it is 1 culprit or 2 joint culprits. To settle this you need to look at the 2 fitted equations, which allowed him to work out the expected (predicted) age of fathers, given the mothers ages, and vice versa. So plug the observed mothers' ages (here 35.7 and 31.6) into the first equation to get the expected ages of the fathers, shown in magenta. They are very close to the observed ones. Conversely, when we go in the other direction, and plug the observed fathers ages into the second equation (in cyan) we get 2 expected mothers' ages that don't match very well with those we observe. There is almost a 2 year discrepancy. So to Penrose there could be no doubt which parent's age matters. 194/1027

==========

• One of the delights of studying old material is that some of what went on behind the scenes **beforehand** is now available to us through archives, and digital archives at that! In the Penrose papers held at UCL, we can read that **early on** Penrose consulted an eminent statistician. BTW Wherever you see a redacted term, substitute today's term Down Syndrome or Down's Syndrome. 65/1092

==========

• As often happens, it was not the **original question he put to the consultant** that mattered so much, as it was the **alternative way of looking at the data** that he got in reply. Penrose's question had to do sample size and standard errors when one has **correlated (family)** data. But look at the answer he got from Fisher – and all within **2 days**. October 29, 1932 was a Saturday, and October 31, Halloween, a Monday.

78/1170

==========

• A lot as been written about Fisher's brusque manner, and there is a hint of it in the first sentence of his reply. But then, look at what follows. He starts by seeing the data in the frequency tables I showed earlier as simply two separate 2-way tables of mothers' and fathers' ages, 1 for the cases, and one for the non-cases. And when he says just compare the mothers ages (x's) of the cases and non-cases using a straight $t$-test, he doesn't fuss in the least that the data are family-based (probably because paired tests would have even smaller standard errors?) You will object that this does not **settle the issue** of whether it is the father's or mother's age that matters. But now look at Fisher's **second** piece of advice on how to measure any separate contribution of the father: use the fathers' ages **corrected for mothers' ages**. 151/1321

==========

• That was **then**, but what might Penrose been able to do **today**? 13/133

==========

• In 2014 we were keen to try out some modern methods that don't require these (2-step, indirect) adjustments. Unlike in 1933, we now have methods for studying determinants of **'imaginary'** variates **directly**. The probit and logit models were mostly applied to toxicology data. Cornfield saw a way to go from the dis-

criminant score to a logistic curve for epidemiology. Nelder and Wedderburn's groundbreaking 1972 paper broug much greater generality. McFadden's work took place in a parallel universe, and wasn't immediately noticed by epidemiologists or statisticians. 86/1420

==========

• But, to do so, I thought we needed to have those two separate 2-way tables of mothers and fathers ages, 1 for the cases, and one for the non-

==========

• I would have been happy to find **even one** solution. So, on my website, I issued a challenge to people with stronger Sudoku and computing skills than I had. I also shared this problem with a visiting speaker; he thought an MCMC approach might work, but he didn't take it any further. 53/1576

==========

• I visited my alma mater University College Cork in mid 2023 12/1588

cases. As best we can tell, Penrose NEVER SEPARATED them, and instead worked with the three 2-D (margin frequency tables you have seen earlier. I wrote to the Penrose family asking whether they could find the raw data, but they could not. His daughter Shirley (a geneticist) did send me a book with lots of other sibships data. His son Oliver (a mathematician) also replied and told me there could be lots of solutions. 103/152

==========

• and I happened to mention it to Supratik Roy. And within a few weeks, I had an answer back from him. Not only had he corrected some annoying entries in the table, he had used LINEAR PROGRAMMING to come up with solutions. There is an abbreviated account of this in our CHANCE article. My webpage has a fuller appendix, along with a link to the video of Supratik's Nov. 2023 presentation on it, and on some further ex-

plorations he made. 81/1669

==========

• We now had a full 3-D frequency table, where the total frequency in each mother-father cell was split into the number of children unaffected and the number of children affected. 31/170(

==========

• Here is one of the first solutions, with the colours inside each cell showing the split. 17/1717

==========

• Once Supratik had cracked the Sudoku, I was keen to apply some modern methods to the various solutions, and I wondered how much the results would vary from solution to solution. I was expecting I might have to use the Rubin variance formula that is used for multiple imputation, when one generates multiple (perturbed) copies of a dataset. 59/1776

==========

• I started with this oldest technique – still a favourite today – which classifies the children on the basis of a **linear discriminant (LD) score**. You see that it is made up mostly of the mother's age, so the **dividing line** is almost **vertical**. But how much do you think the dividing line changes from solution to solution? I was quite surprised that the dividing line (and the fitted linear combination) did **not** vary from solution to solution. 79/1855

==========

• And here are the 3 logistic regression fits to one solution. You see that they again support Penrose's conclusion that once you have accounted for the mother age, the father's age does not matter. But I was **not** expecting the fitted coefficients to **stay the same no matter the Sudoku solution**. 52/1907

==========

• **So, the takeaway** The fact that all the solutions give the same values of the statistics that Penrose calculated means that the three 2-D marginal

tables that he published were **sufficient** to derive his statistics. So, why, instinctively, we do feel cheated by not being able to see the full 3-D data? I think it goes back to something that Fisher said quite formally when he came up with the very idea of sufficiency: sufficient statistics are **conditional on the model under consideration**, and there is information in the **remaining data** as to how well the model **fits the data**. 101/2008

==========

**• I am including these 2 slides just to say that we go though this math in the appendix on my website** 22/2030

==========

**• The 3 sufficient statistics in the case of the 2-regressors mode are (1) the 154 cases (ii) the total of their father's ages and (iii) the total of their mother's ages – and once we know the X (design) matrix. These statistics are all calculable from the 3 sets of marginal frequencies that Penrose published!** 56/2086

==========

• I am including these 2 slides just to say that we go though this math in the appendix on my website. 22/2108

==========

• Our piece in CHANCE is due to be published soon. 11/2119

==========

• Now to the much more complicated and laborious study that 'ruled' out another fellow traveler, **birth order**. It featured a much data-intensive effort. Here is Penrose's **Results** paper 29/2148

==========

• And here is the **Methods** paper where he lays out in detail the technical details of the new statistical method suggested by Ronald Fisher. 25/2173

==========

• Penrose's papers in 1934 work are probably the first example of a con-

ditional regression model, and fitted by hand by Maximum Likelihood no less! This was some 40 years before the approach was given the name **conditional logit analysis** by Daniel McFadden. His use of it for qualitative behaviour earned him the Nobel Prize in Economics in 2000. He was born in 1937, just after Fisher had published the method of Discriminant Analysis. 74/2247

==========

• It can be difficult to re-tell history by working from the past up to the present, especially if the topic itself is a bit esoteric. Since some of you won't have personally dealt with conditional logistic regression, here is a modern example I used to use to introduce it to epidemiology and biostatistics students. 55/2302

==========

• Pardoe updates his dataset and analysis each year, but when he first began looking at them, these were the 5 important predictors. The last 3 predictors come from competitions that don't go all the way back to when the Oscars started. 42/2344

==========

• Here are some reasons why 'regular' (unconditional) logistic regression is not appropriate. The dataset is structured into competitions (we call them **matched sets** in epidemiology), with (usually) just one winner per set (in survival analysis Cox called them **risksets**, and 'failures'). Each year provided one realization of a multinomial with $n = 1$. 52/2396

==========

• Let's start with the DATA, which I show in black. For the personal awards (Director/actor/Actress) there tend to be 5 nominees each year, but for the best picture nowadays there are often more than 5. The set of Xs (what I call the **profile** are the predictors, There is 1 winner per com-

petition. So you see why the regular logistic regression won't work (You could get it to run, but it doesn't align with the data structure).
Now let's look at the PARAMETER MODEL and the FITTING, all shown in red. Take the 2025 competition. If we didn't know who was who or what their profiles were, ahead of time each would have 1 chance of 5 of winning. But suppose the $K$ weights (betas, parameters with unknown values) for

the $K$ elements in the profiles produced relative probabilities of omega 1 to omega 5. Then the fitted probabilities would be the 5 $P$'s shown. The 2025 winner (which one has $Y = 1$) is now known, so we can write down the log likelihood contribution from 2025. It will be a function of the betas. Doing the same for each earlier year and adding them up gives us an overall LogLik function which we can maximize with respect to the betas.
Each year at McGill, professor Andrea Benedetti gets teams of our students to update the X matrix, refit the model, and make their own predictions for the upcoming Oscars. Then at our post-Oscars biostatistics seminar, each team describes their model and how it performed that year, and Andrea hands out her own awards to the team that did best. 272/2668
==========
• Going back now to 1934. Penrose's

dataset of 217 families is probably a superset of his earlier 150 families dataset . The analysis again involved two highly correlated suspects, so you might be expecting the same types of analyses. But because of who he got as a reviewer, this disentanglement project involved a much more elegant analysis. 58/2726
==========
• Less than **two weeks** after he submitted the first version, he got a 6-page letter. Here is the first para-

graph. 21/2747

==========

• And here is the second one, getting to the core issue of SELECTION. The only families selected are those with an effected child. [In fact, this criticism also applied to his first study]. 34/2781

==========

• Even as he says it's a lot of work (and it was!), he is VERY encouraging. 17/2798

==========

• The remaining pages of the letter are quite technical, explaining how to do the family by family calculations at each iteration. Here is one page. This is just like the 'omega of the winner over a sum of the omegas of all the nominess' in the Oscar's example. And, yes, this is how (even early on in my time) one included mathematical material in a typed letter. 68/2866

==========

• There are several more back and forth letters over the next 4-5 months, and Penrose came in to UCL a few times to see Fisher in person. The revised version is quite upfront about the 'new and improved' approach, which meant fitting an age-only model, then grouping the predictions by birth order, and comparing the observed with the fitted frequencies. 61/2927

==========

• This table is from the appendix to the Biometrika paper. All but the last column are from the 1934 paper, where Penrose showed that the 'residuals' (the 'O-E's) in the various birth ranks, based on the age-only model, were all within 1 SE of 0. Calculating the residuals was the nastiest part, and the hardest for me to understand, but my simulated-based SE's agree with theirs. Today, we would likely fit an age + birth order model and look at the birth order coefficients, and at fancier GoF statistics. 90/3017

==========

• But how, in the first place did they fit the age-only model? here it is; they categorized age as **7 age-bins**, each 5 years wide. So their model had **6 free** age-effect parameters (piecev constant risks) 36/3053

==========

• The black rectangle has the observed numbers, and you can see how 'artificial' the proportions of affected children were. These became his starting values for the relative probabilities, but the fitted no.s of children with Down Syndrome (in the old days they used the word 'calculated' values ) quickly change as the iterations (improvements) proceed. My paper goes in. some detail into how they did this – without formally writing out the log-likelihood – just by searching for the parameter values that make the fitted frequencies match the observed ones. It looks easy now, but each iteration meant very detailed sibship by sibship calculations of where in each sibship you would expect the affected child to be. 118/3171

==========

• Penrose listed the raw data, sibship by sibship, in both of the 1934 papers. Here are the (head) and (tail) of that file. 24/3195

==========

• This is page one of a more visual representation I made. I just show the first 44 sibships. The red dots are the cases. You can again see, but at a sibship level, the very large (structural) collinearity between age and birth order. In the Biometrika supplement, I address this, and the implications, in a bit more detail. 59/3254

==========

• Two days ago, I asked chatGPT if it could find any photograph of Penrose and Fisher together. It couldn't, but to its credit, it found the full referee's report on the revised version. I will be curious to hear the reactions

of today's Editors. 45/3299

==========

• Fisher was knighted in 1952 and received honorary degrees from (at least) nine universities. **Penrose** got honourary degrees from four. McGill was the first of these, in 1958, when he attended the 10th International Genetics Conference here. 38/3337

==========

• Starting from the right in this photo, you see the McGill principal, then the two other recipients, then Penrose. And at the extreme left is Sewall Wright, who chaired the Congress. Remember the 1926 author who showed that it was the age of the mother than mattered for defects in guinea pigs; that was the same metric Penrose started out with in 1933, before he consulted Fisher about standard errors for correlated data! 74/3411

==========

• I have put on my website Penrose's fascinating diary account of his North American 2 month tour that summer before the congress. He and his wife came and returned home by ship, and had their 13 y. old daughter Shirley with them. 43/3454

==========

In 1964, he and his wife came on a short trip to the USA, where he received this Award from the Joseph P. Kennedy Jr. Foundation. It was set up to study what was at that time called mental retardation. 40/3494

==========

• I have also posted his diary account of this trip, 11/3505

==========

• as well as the program, photos and video from the inaugural awards ceremony. 14/3519

==========

• To wrap up, .... 5/3524

==========

• There is a lot we can take from this story. In the late 1950s, it became known that it involved chromosome 21, and soon geneticists found sub types. But the 1930s finding have

largely held up. But we should no longer speak of a single age-curve, but of a mix, each with its own shape and determinants. And despite the fact that the 1933 analysis ignored the selection, its simplicity still had a several lessons on models, sufficiency, and model-checking – all so relevant today. In our piece in Chance, we suggest that various data-reconstruction challenges, like our so-called Sudoku, might make the concept of sufficiency more relatable-to in math-stat classes. I wasn't able to cover it all today, but the Biometrika piece shows that many dots get connected when one carefully goes though the derivation of the *form* of the conditional logistic model in the 1934 paper, and the *calculations* involved in fitting it to the family data. 160/3684

==========

• I would like to acknowledge two archives, and many many people, going all the way back to Tadgh Carey at UCC in Cork. He awakened my interest in statistical history by (?deliberately) keeping the journals behind locked doors. 39/3723

==========