Where do the (big) data come from
1: introduction

Rod Little

Department of Biostatistics

UNIVERSITY OF MICHIGAN

Outline

- My journey in statistics, and why it’s such a great field for a career
- Statisticians love to toss coins: describe and compare two roles of randomization in

  (a) sample selection (talk 2)

  (b) treatment assignment (talks 3 and 4)
- Maximum Likelihood, a major tool for statistical inference (talk 5)

Early days

1956-1968: Glasgow Academy

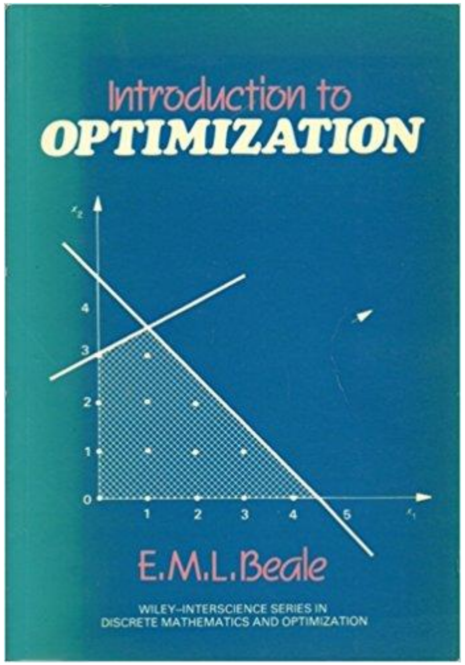
1968-71: BA Mathematics, Gonville and Caius College, Cambridge University

(R.A. Fisher's college)

Postgraduate training

1972-74 MS/PhD in Statistics and Operations Research

D.R. Cox (winner of the 2017 International Prize in Statistics)
Ph.D. with Martin Beale



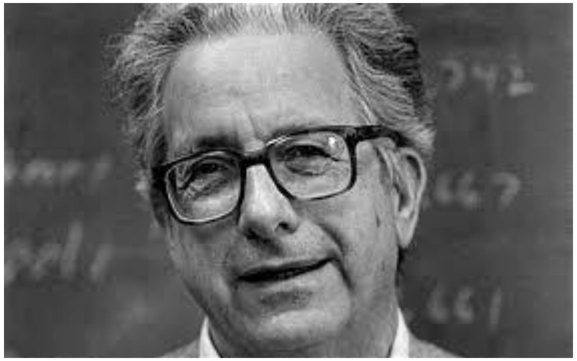
Introduction to
OPTIMIZATION

E.M.L.Beale

WILEY–INTERSCIENCE SERIES IN
DISCRETE MATHEMATICS AND OPTIMIZATION

1975-76 Post-doc, Department of Statistics, University of Chicago. My first chair was Paul Meier, of the “Kaplan-Meier curve” in survival analysis. (One of the most cited papers in science)

Paul Meier – from The Telegraph obituary (2011)



"Paul Meier, who died on August 7 aged 87, was a statistician who championed the idea of testing new medical treatments through randomised trials, so helping to lead a revolution in clinical research and **saving, albeit indirectly, millions of lives…**

… The idea of assigning subjects in medical trials solely on the basis of random selection might now seem obvious. But, like many medical innovations, it did not seem so at the time Meier proposed it in the 1950s…. **Many physicians were horrified at the idea that their selection should be random**, together with an equally randomly-selected "control" group of patients who were given the standard treatment or a placebo… At first Meier's arguments met with incomprehension: "When I said 'randomise' in breast cancer trials," he recalled in 2004, "I was looked at with amazement by my medical colleagues: **'Randomise? We know that this treatment is better than that one.' I said, 'Not really!'".**

3. World Fertility Survey (1976-80)

Recruited by Sir Maurice Kendall, Director of the World Fertility Survey (WFS)

Sir Maurice was a prominent statistician, noted for the treatise with Alan Stuart “The Advanced Theory of Statistics”

Sir Maurice was proud that WFS conducted **probability surveys** in developing countries, a design that he liked to describe as “scientific”…

Also a poet and joker, see “Hiawatha designs an experiment:” http://www.mscs.mu.edu/~paulb/Pomes/hiawatha.pdf

3. World Fertility Survey (1976-80)

Recruited by Sir Maurice Kendall, Director of the World Fertility Survey (WFS)

Sir Maurice was a prominent statistician, noted for the treatise with Alan Stuart “The Advanced Theory of Statistics”

Sir Maurice was proud that WFS conducted **probability surveys** in developing countries, a design that he liked to describe as “scientific”…

Also a poet and joker, see “Hiawatha designs an experiment:” http://www.mscs.mu.edu/~paulb/Pomes/hiawatha.pdf

5. UCLA Biomathematics (1983-93)

From Wil Dixon's Retirement Party (Oct 1986)

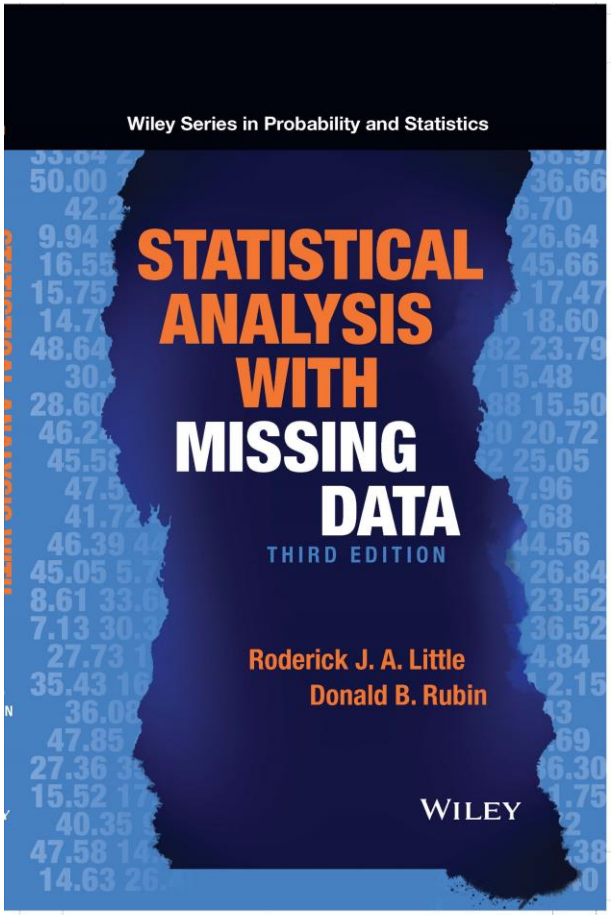
The data of young Sherman Mellinkoff
Had extremes that were knocking his stockings off
He called in Wil Dixon,
Whose trimmed means soon fixed 'em
Dr. M. became Dean, Biomath-took-off (RL)

Wilfred Dixon (developer of trimmed means, and BMDP, an important early statistical software program)

6. University of Michigan Biostatistics (1993-present)

- Fine university in a wonderful town
- Great faculty, staff, students
- No earthquakes, hurricanes, floods (just the odd lost tornado)
- Songs and skits…

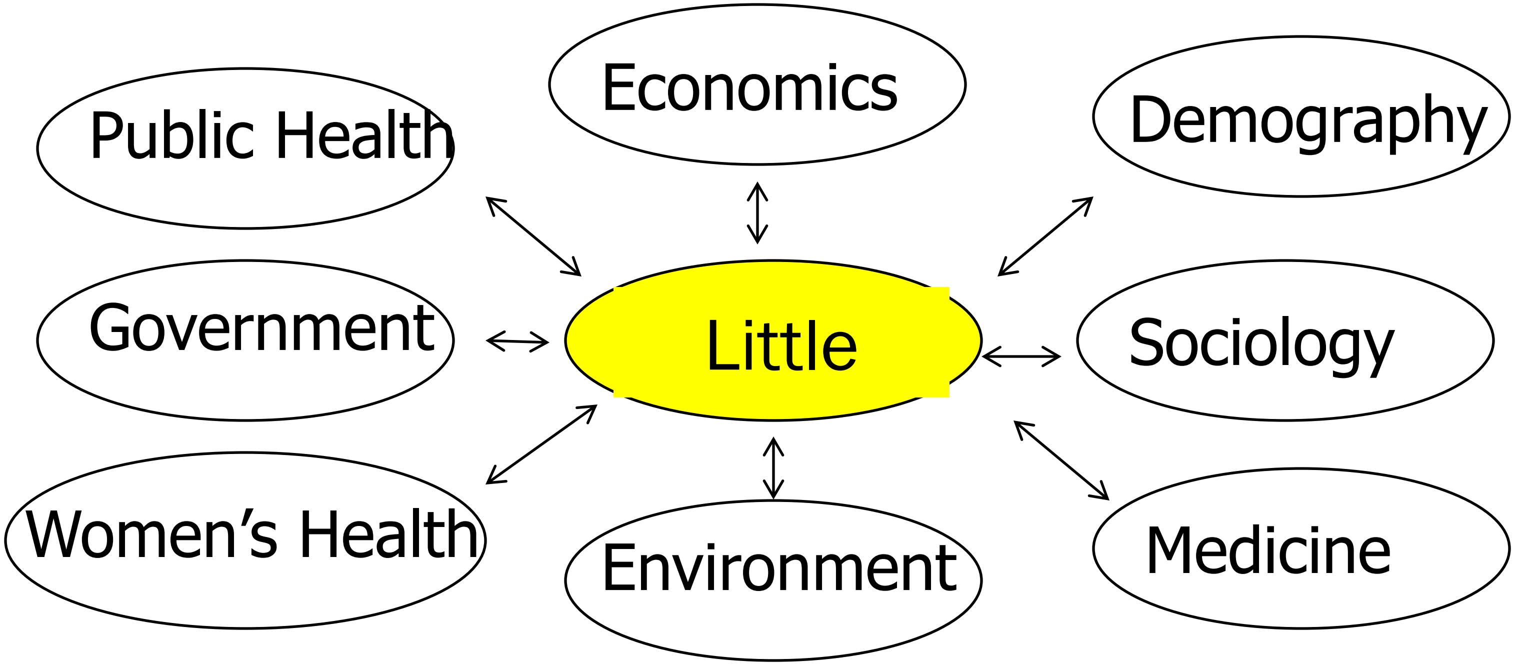
My own work


Wiley Series in Probability and Statistics
STATISTICAL ANALYSIS WITH MISSING DATA
THIRD EDITION
Roderick J. A. Little
Donald B. Rubin
WILEY

Little,R.J. and Rubin, D.B. (2019) *Statistical Analysis with Missing Data*, 3rd edition. Wiley: New York

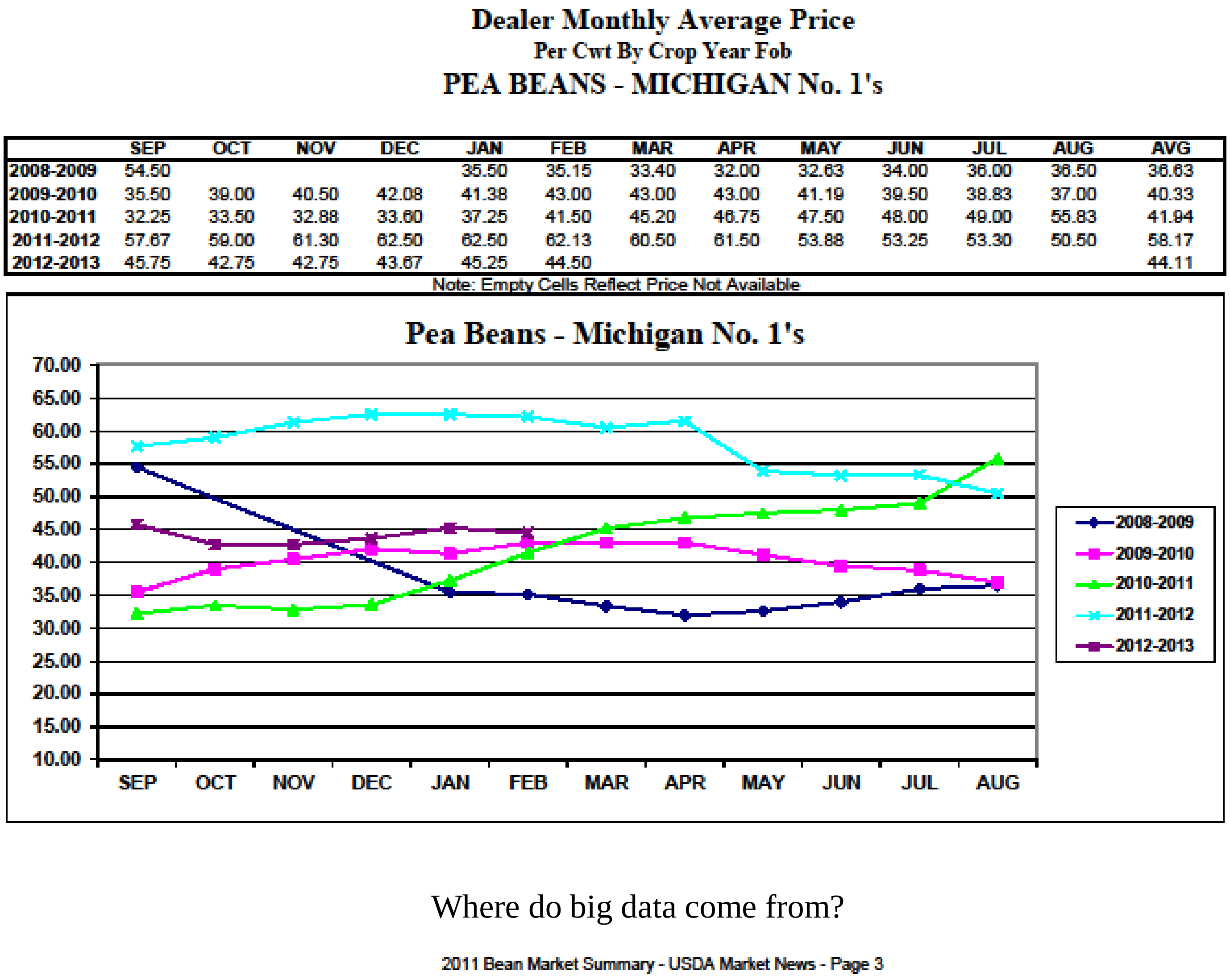
Little, R.J. (2012). Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 3, 309-372

Collaborative Work



The Joy of Stats: So Much More than Bean Counting!

Statisticians *do* count beans…



Dealer Monthly Average Price
Per Cwt By Crop Year Fob
PEA BEANS - MICHIGAN No. 1's

	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	AVG
2008-2009	54.50				35.50	35.15	33.40	32.00	32.63	34.00	36.00	36.50	36.63
2009-2010	35.50	39.00	40.50	42.08	41.38	43.00	43.00	43.00	41.19	39.50	38.83	37.00	40.33
2010-2011	32.25	33.50	32.88	33.60	37.25	41.50	45.20	46.75	47.50	48.00	49.00	55.83	41.94
2011-2012	57.67	59.00	61.30	62.50	62.50	62.13	60.50	61.50	53.88	53.25	53.30	50.50	58.17
2012-2013	45.75	42.75	42.75	43.67	45.25	44.50							44.11

Note: Empty Cells Reflect Price Not Available

Pea Beans - Michigan No. 1's

2011 Bean Market Summary - USDA Market News - Page 3

How many have died from COVID-19?

- Numbers are tragic – and are also a hot political topic … some countries clearly underreport
- Some politicians have argued that CDC overstates the number to exaggerate the scope of the pandemic
- Death certificates are not reliable -- often there are multiple causes. Excess deaths over historic death rates are a more reliable source
  - These suggest that early COVID 19 death rates were understated

It’s what you do with them that matters…

“…now we really do have essentially free and ubiquitous data… so the complementary scarce factor is the ability to understand that data and extract value from it.”

Hal Varian, Chief Economist, Google

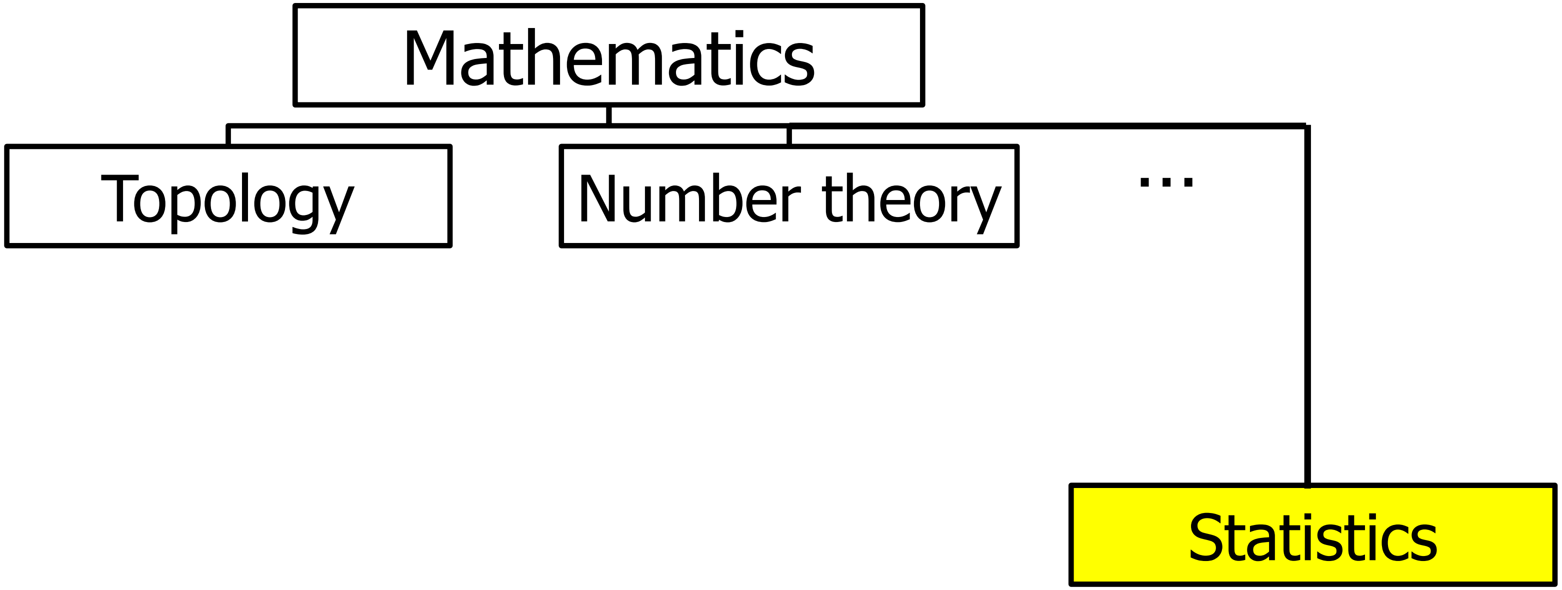
Hal Varian

What is statistics?

Not just facts ...

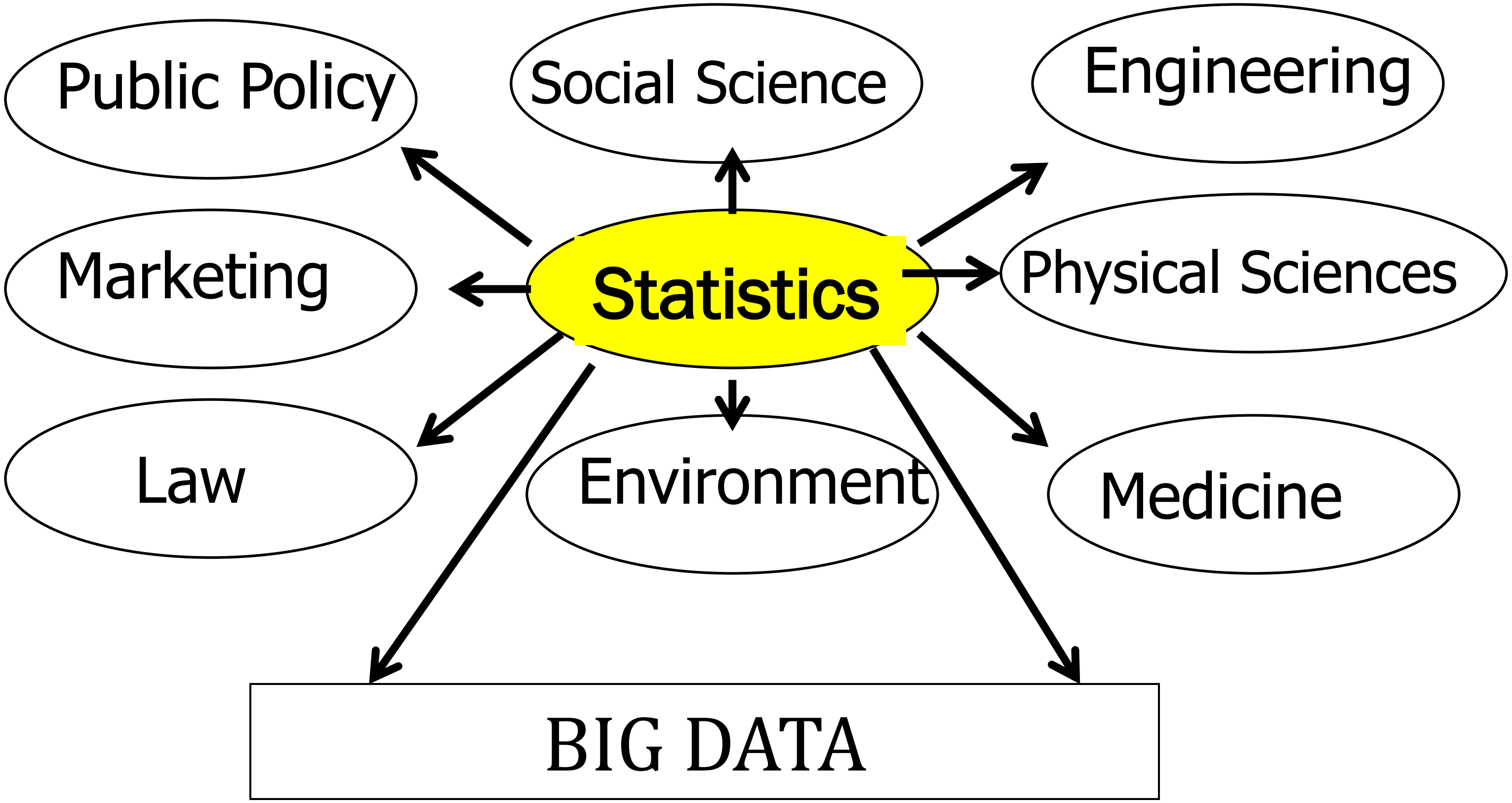
or a (rather pedestrian) subfield of math:

…



What is statistics?

But **data science**:



Statisticians Impacting Science

20/25 of most-cited mathematicians in science in 2002 were statisticians (Science Watch 2002)

Statisticians Impacting Society #1

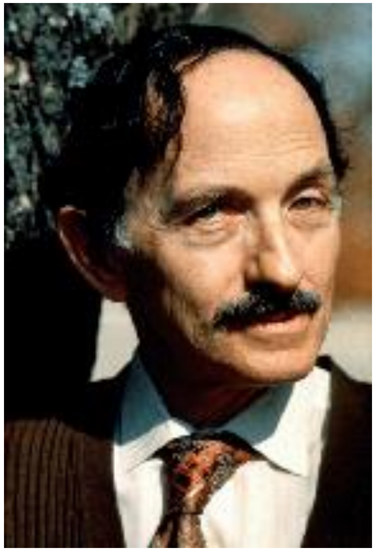
- Sir Ronald Fisher’s experimental designs and analysis of variance have greatly increased the world food supply

Sir Ronald Fisher

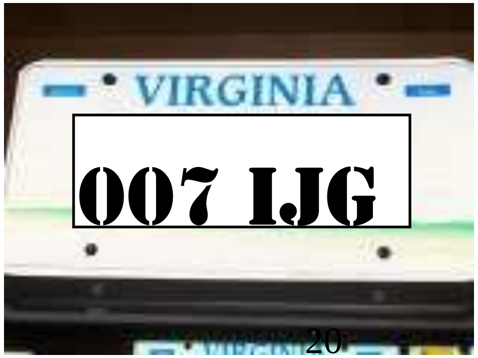
Statisticians Impacting Society #2

Alan Turing sculpture by Stephen Kettle, Bletchley Park. Photo by Jon Callas

Alan Turing and Jack Good's statistical methods helped decode German naval ciphers, arguably reducing the length of World War II by two years or more, saving millions of lives. See e.g. "The Imitation Game"



I. Jack Good (IJG)



Statisticians Impacting Society #3

- Randomization, a strange but clever idea for
  - valid answers about populations from surprisingly small sample surveys
  - randomized clinical trials, pioneered by Sir Bradford Hill, now the gold standard in evidence-based medicine
  - This is the focus of my first two talks

Statisticians Impacting Society #4

- Official government statisticians … not just bean counters, **guardians of our democracy**.
- http://www.huffingtonpost.com/rod-little/decennial-census_b_3046611.html

What is statistics?

- Statistical Design
- Data collection
- Data description
  - Graphs, tables etc.
- Statistical inference
  - Inferring about a broader population based on a sample
  - Hypothesis testing, confidence intervals, etc.
- I’ll focus on design of data collection … but I’ll say a bit about statistical inference