R.A. Fisher:
Statistics as key tool for scientific research

Daniel Peña

Rector

Universidad Carlos III de Madrid

Facultat de Matemàtiques i Estadística,
Universitat Poltècnica de Catalunya, 29 Septiembre 2011

Outline

1. Introduction
2. Youth and Education (1890-1919)
3. Rothamsted and London (1920-1942)
4. Cambridge and Adelaide (1943-1962)
5. Personality
6. The heritage from Fisher
7. Conclusion

1. Introduction

- **H. Hotelling**: Laplace, Gauss, Pearson and Fisher are the co-founders of statistics and probability.
- **M. Frechet**: K. Pearson and RA Fisher are the founders of Statistics.
- **CR Rao**: Fisher is the founder of modern statistics.
- **B. Efron**: Fisher is the single most important figure in 20th century statistics.

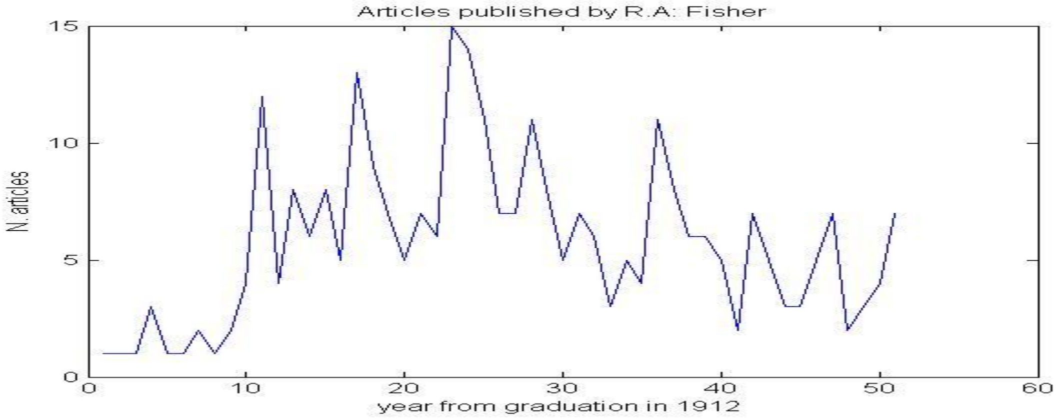
1. Introduction

In the Kotzs and Johnson book, *Breakthroughs in Statistics* I and II, Number of contributions:

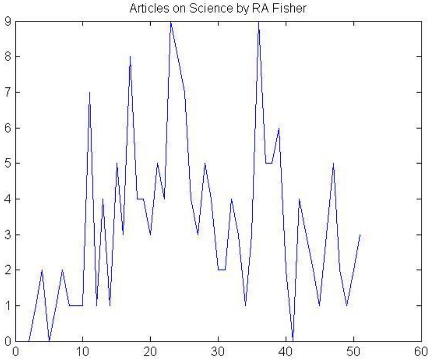
- **Fisher**: **3** (Foundation of MS,22; SMRW,25; DE, 26)
- Hotelling 2 (T, 31; CC, 36)
- Neyman: 2 (HT, 33; Sampling, 34)
- Wald 2 (Sequential, 45; SD,49)
- Box 2 (EO,51; TS, 62)

Note: After Vol III, Rao, Tukey and Smith are added with 2 contributions

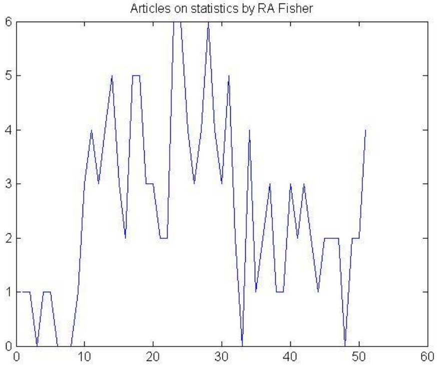
1. Introduction (Fisher's articles)



- 56% on science



- 44% in Statistics



2. First 29 years (1890-1919)

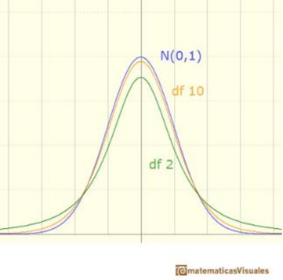
- Born   17/2/1890 in London the 7th child
- Poor eyesight since schoolboy
- Scholarship in Caius College, Cambridge U., 1909.
- Graduate in mathematics from Cambridge U. (1912)
- Work as clerk and school teacher (1913-1919)
- Married in 1917.
- Reject work in K. Pearson’s lab to work in Rothamsted (1919)

Youth Contributions

Inspired by Gosset (Student), works on exact sampling distribution statistics

Proposed a general method of finding sampling distributions by n-dimensional geometry and found the exact distribution of the correlation coefficient

Published 12 papers on Statistics and Genetics



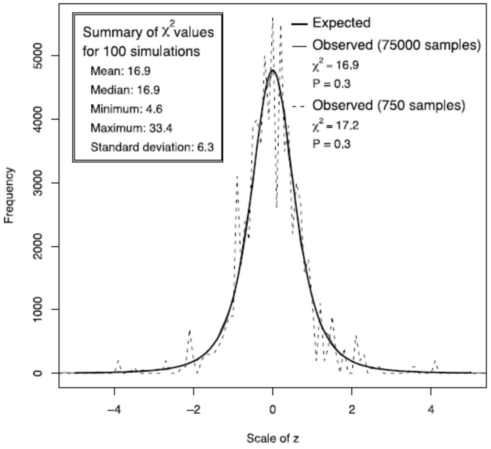
Gosset and the t distribution

The Probable Error of a Mean (Student 1908) that led to today's *t* distribution was the lead article in the March 1908 issue of *Biometrika*.

"Although it is well known that the method of using the normal curve is only trustworthy when the sample is "large," no one has yet told us very clearly where the limit between "large" and "small" samples is to be drawn. The aim of the present paper is to determine the point at which we may use the tables of the (Normal) probability integral in judging of the significance of the mean of a series of experiments, and to furnish alternative tables for use when the number of experiments is too few."

See *The American Statistician, February 2008, Vol. 62, No. 1 1*

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3,000 criminals, from a paper by W. R. Macdonell (*Biometrika*, Vol. I, p. 219). The measurements were written out on 3,000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book, which thus contains the measurements of 3,000 criminals in a random order. Finally, each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation of each sample determined.



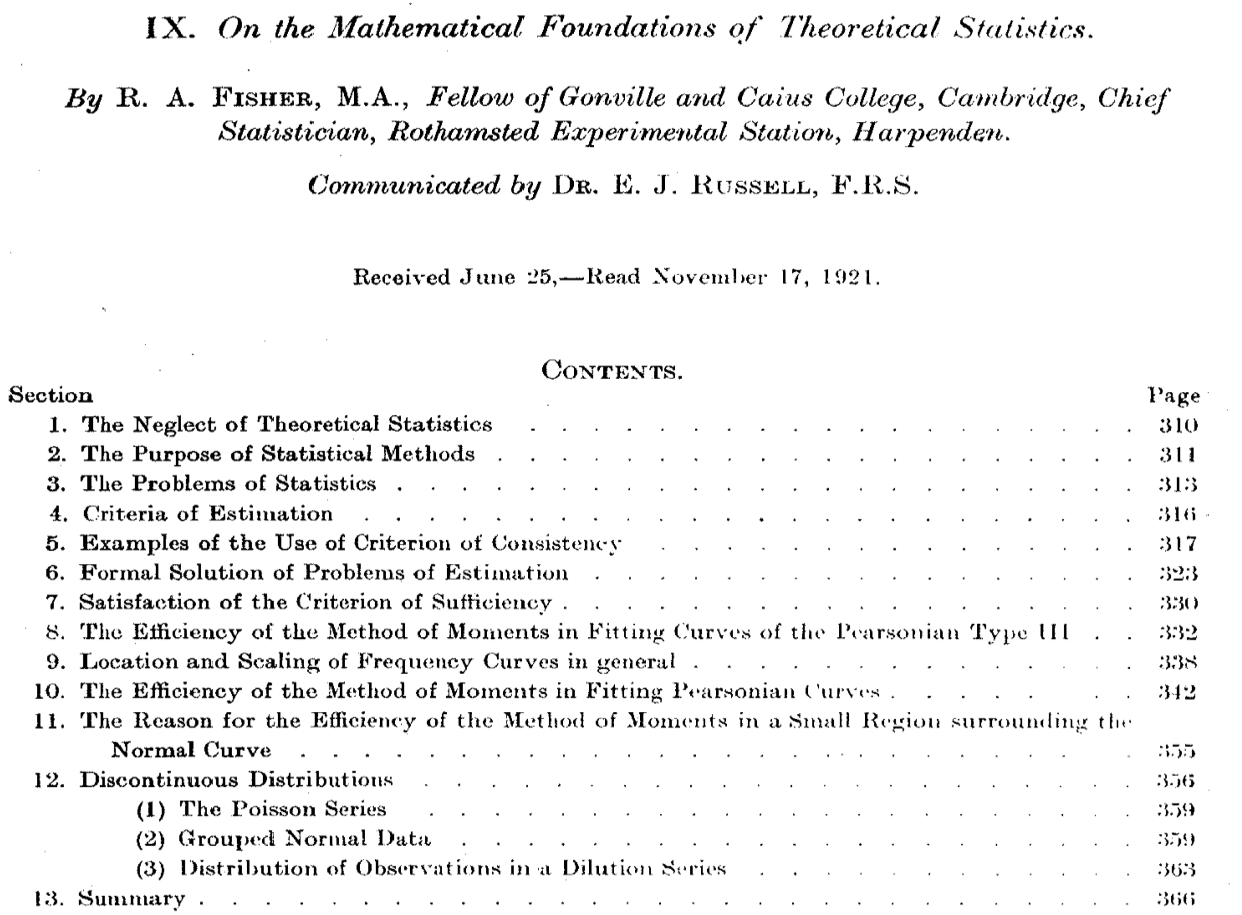
Main results 1912-1919

- Proved the result of Gosset in 1912
- Introduce degrees of freedom
- Present a general procedure for finding sampling distributions.
- Propose transformations to get symmetry in distributions.

3. From 30 to 52 (1920-42)

- *On the mathematical foundations of theoretical statistics, (1922)*
- *Statistical Methods for Research workers (1925)*
- Fellow of Royal Society (1929)
- Honorary Fellow of American Stat. Ass. (1930)
- Galton Professor in Eugenics (1933), University College, London.
- *The Design of Experiments* (1935)
- Honorary PH.D. degree, Harvard University, (1936)
- Viaje a la India invitado por Mahalanobis (1938)

Statistician at Rothamsted 1919-1933



IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

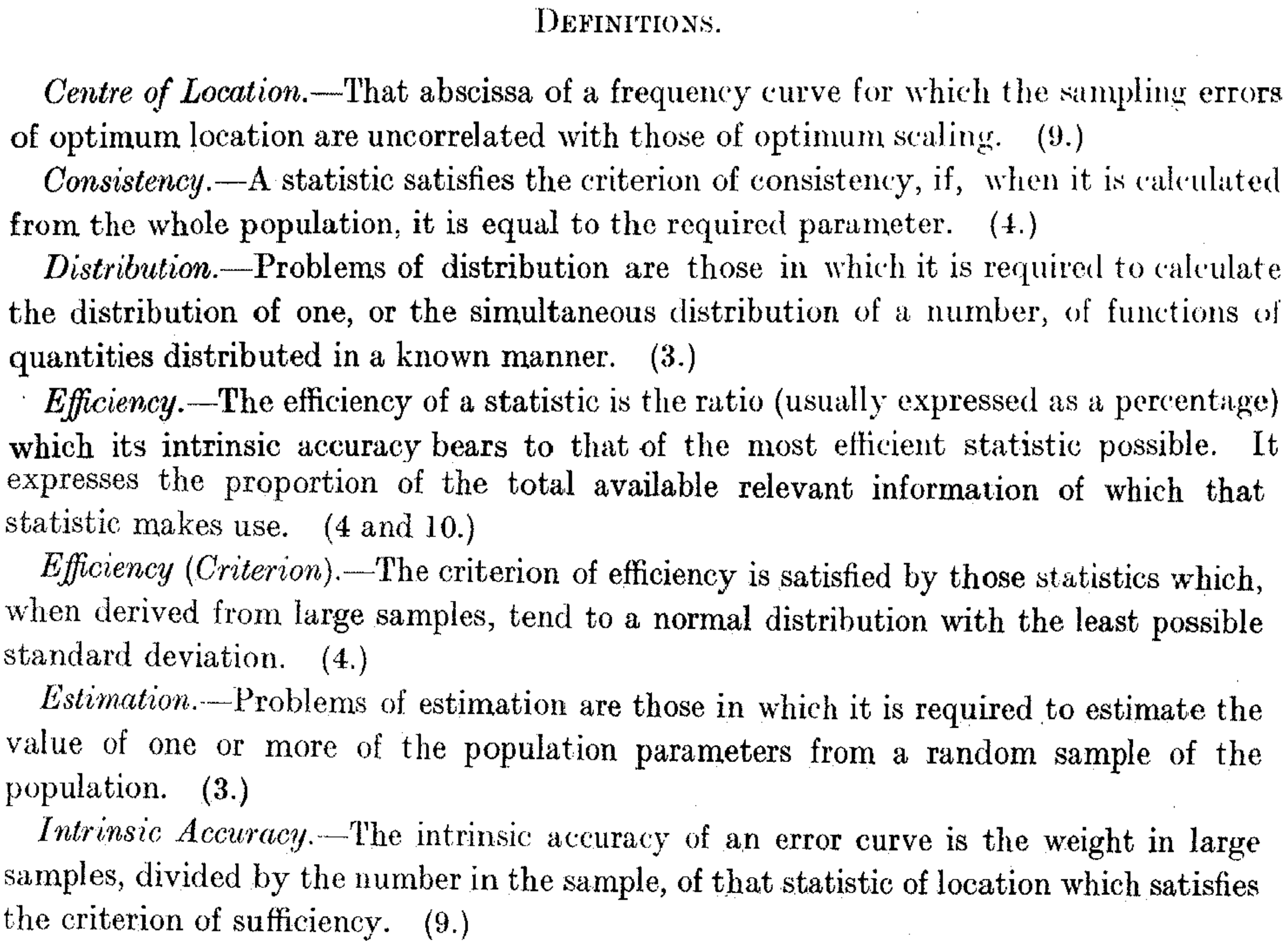
Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Section — Page

1. The Neglect of Theoretical Statistics . . . 310
2. The Purpose of Statistical Methods . . . 311
3. The Problems of Statistics . . . 313
4. Criteria of Estimation . . . 316
5. Examples of the Use of Criterion of Consistency . . . 317
6. Formal Solution of Problems of Estimation . . . 323
7. Satisfaction of the Criterion of Sufficiency . . . 330
8. The Efficiency of the Method of Moments in Fitting Curves of the Pearsonian Type III . . . 332
9. Location and Scaling of Frequency Curves in general . . . 338
10. The Efficiency of the Method of Moments in Fitting Pearsonian Curves . . . 342
11. The Reason for the Efficiency of the Method of Moments in a Small Region surrounding the Normal Curve . . . 355
12. Discontinuous Distributions . . . 356
    (1) The Poisson Series . . . 359
    (2) Grouped Normal Data . . . 359
    (3) Distribution of Observations in a Dilution Series . . . 363
13. Summary . . . 366



Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

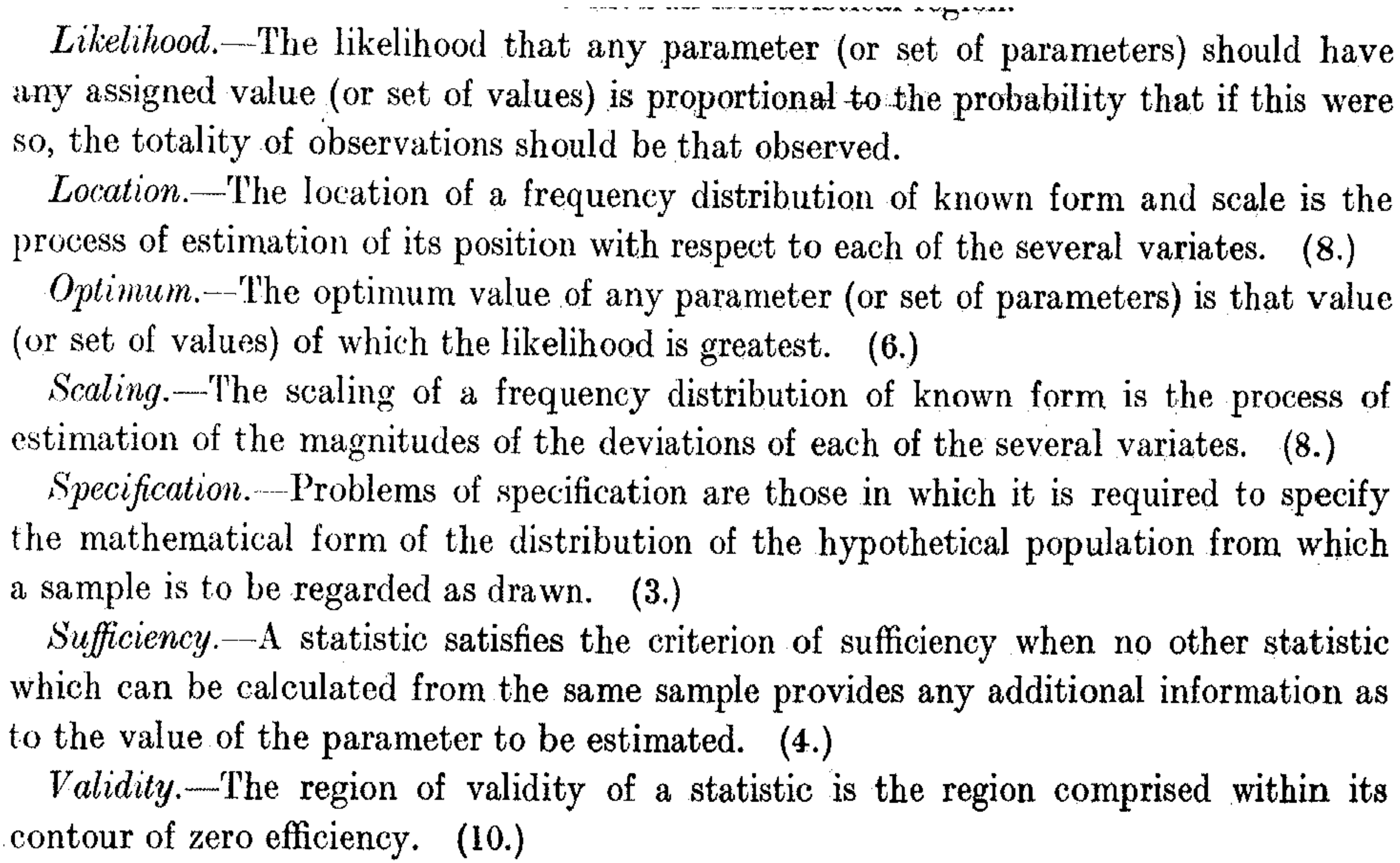
Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It expresses the proportion of the total available relevant information of which that statistic makes use. (4 and 10.)

Efficiency (Criterion).—The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation. (4.)

Estimation.—Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population. (3.)

Intrinsic Accuracy.—The intrinsic accuracy of an error curve is the weight in large samples, divided by the number in the sample, of that statistic of location which satisfies * the criterion of sufficiency. (9.)



Likelihood.—The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

Location.—The location of a frequency distribution of known form and scale is the process of estimation of its position with respect to each of the several variates. (8.)

Optimum.—The optimum value of any parameter (or set of parameters) is that value (or set of values) of which the likelihood is greatest. (6.)

Scaling.—The scaling of a frequency distribution of known form is the process of estimation of the magnitudes of the deviations of each of the several variates. (8.)

Specification.—Problems of specification are those in which it is required to specify the mathematical form of the distribution of the hypothetical population from which a sample is to be regarded as drawn. (3.)

Sufficiency.—A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated. (4.)

Validity.—The region of validity of a statistic is the region comprised within its contour of zero efficiency. (10.)

In this Landmark paper, Fisher:

Make, for the first time, a clear distinction between parameter and statistic

Propose three problems:

(1) specification (choosing the model);

(2) estimation of the parameters;

(3) sampling distribution of the estimates.

Propose estimation by Maximum Likelihood, (instead of the moment method proposed by Pearson) introduce sufficiency, consistency and efficiency and obtain the large sampling variance equation for ML estimates.

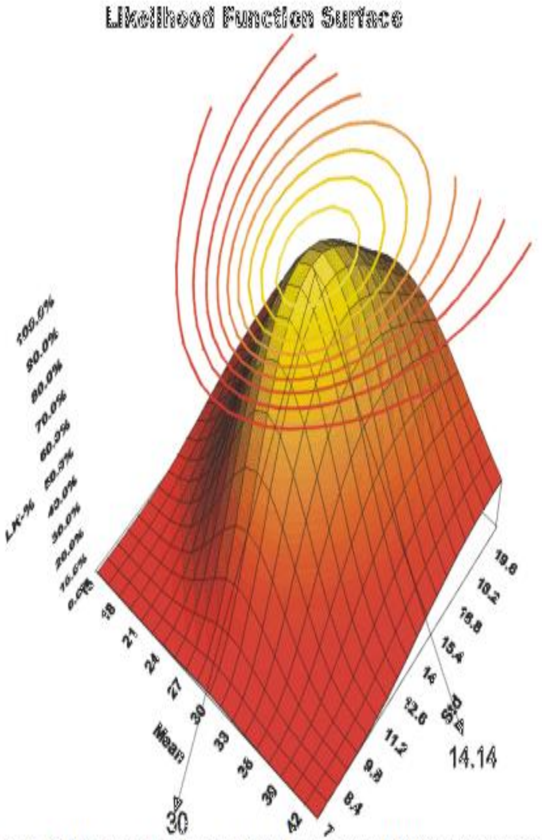
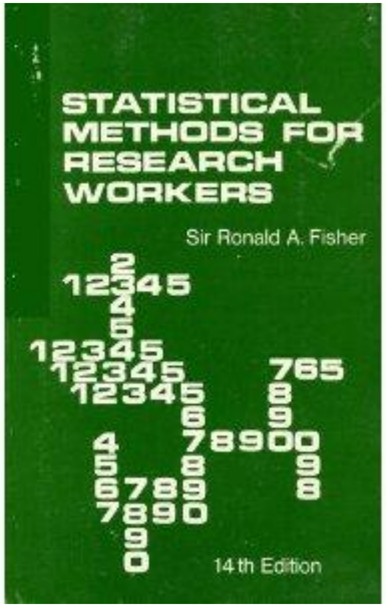


Figure A-1: Likelihood surface plot for MLE normal distribution example.

Other key work at Rothamsted 1919-1933

- Statistical Methods for Research Workers, 1925



STATISTICAL METHODS FOR RESEARCH WORKERS

Sir Ronald A. Fisher

14th Edition

- The Design of Experiments, 1935

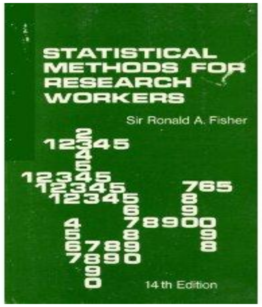
The
Design of Experiments

By

R. A. Fisher, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
and American Academy of Arts and Sciences
Galton Professor, University of London

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 33 Paternoster Row, E.C.
1935



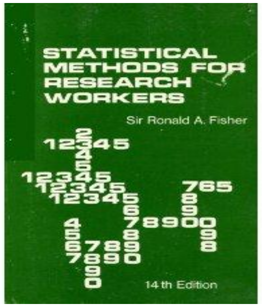
STATISTICAL METHODS FOR RESEARCH WORKERS

Sir Ronald A. Fisher

14th Edition

The book includes

- Diagrams and plotting data
- Normal and Binomial Distributions
- Contingency tables, chi-squared tests
- Test of significance for mean, variance, difference of means, simple and multiple regression and the linear model, correlations, partial and multiple correlations, and ANOVA models.
- Maximum Likelihood estimation
- Many real data set and their analysis is the largest part of the book



Many critics from Academia:

- Too much of Fisher´s own work…
- The procedures are not yet well acepted…
- Lack of rigor in the proofs…

14th editions of the book and one of the most influential in Statistics ever. Test of significance are used in all branches of science.

The
Design of Experiments

By
R. A. Fisher, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
and American Academy of Arts and Sciences
Galton Professor, University of London

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 33 Paternoster Row, E.C.
1935

- “A Lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea was first added to the cup.”

(The Lady was real, the experiment was made, and she only made one error in eight cups)

Randomization

Replication

Blocking

Factorial Designs

Also, in Rothamsted

- Transfer function in dynamic models (studies in crop variation: yield and rain)
- Sampling distribution of partial and multiple correlation coefficients
- Nonparametric tests
- Distribution of extreme values
- Fiducial probability

University College (1933-43)

- The Department of K. Pearson is split into Eugenics, (Fisher) and Statistics (Egon Pearson), 1933.
- Neyman join Pearson, 1934.
- Fisher: estimation, properties of MLE,

information, sufficiency, likelihood.

- Neyman-Pearson: Optimal Hypothesis

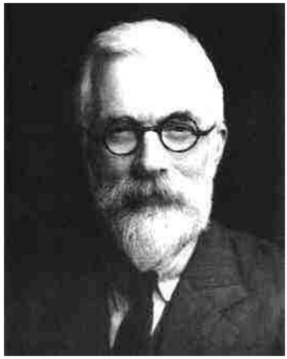
testing and decisions.

- Neyman moved to Berkeley in 1938.
- When the war starts Fisher moved to Rothamsted as University College was closed.

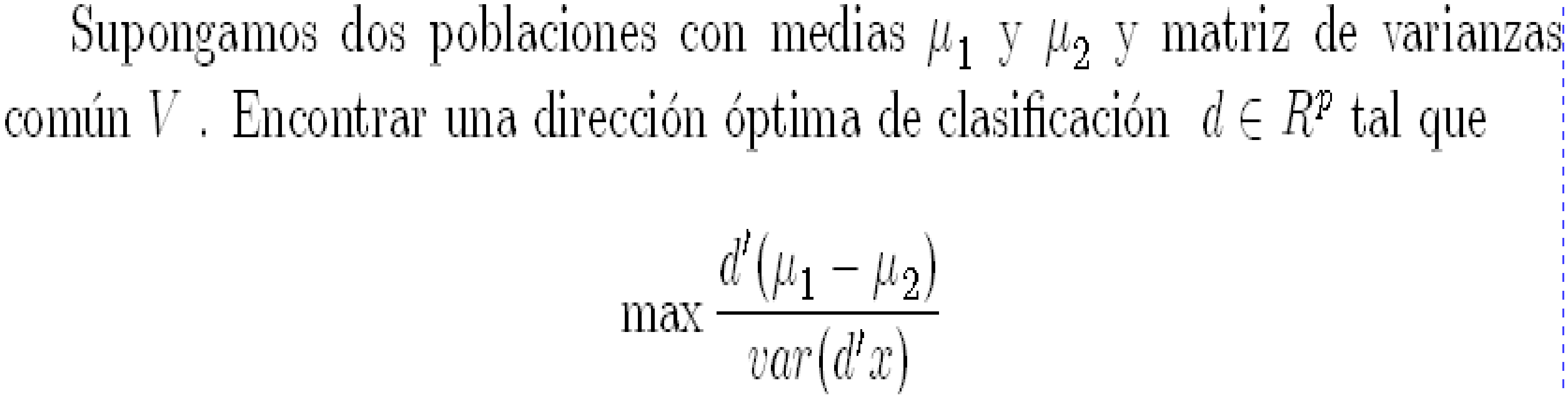
Multivariate Statistics

- Linear Discriminant Function (1936)
- Multivariate ANOVA (1939)
- Optimal Scoring and singular value decomposition of cell frequencies (multiple correspondence analysis) (1938, 1941)
- Relation between Hotelling’s T, Mahalanobis distance and Discriminant functions

Linear Discriminant function



- Encontrar una dirección de proyección que separe al máximo dos poblaciones

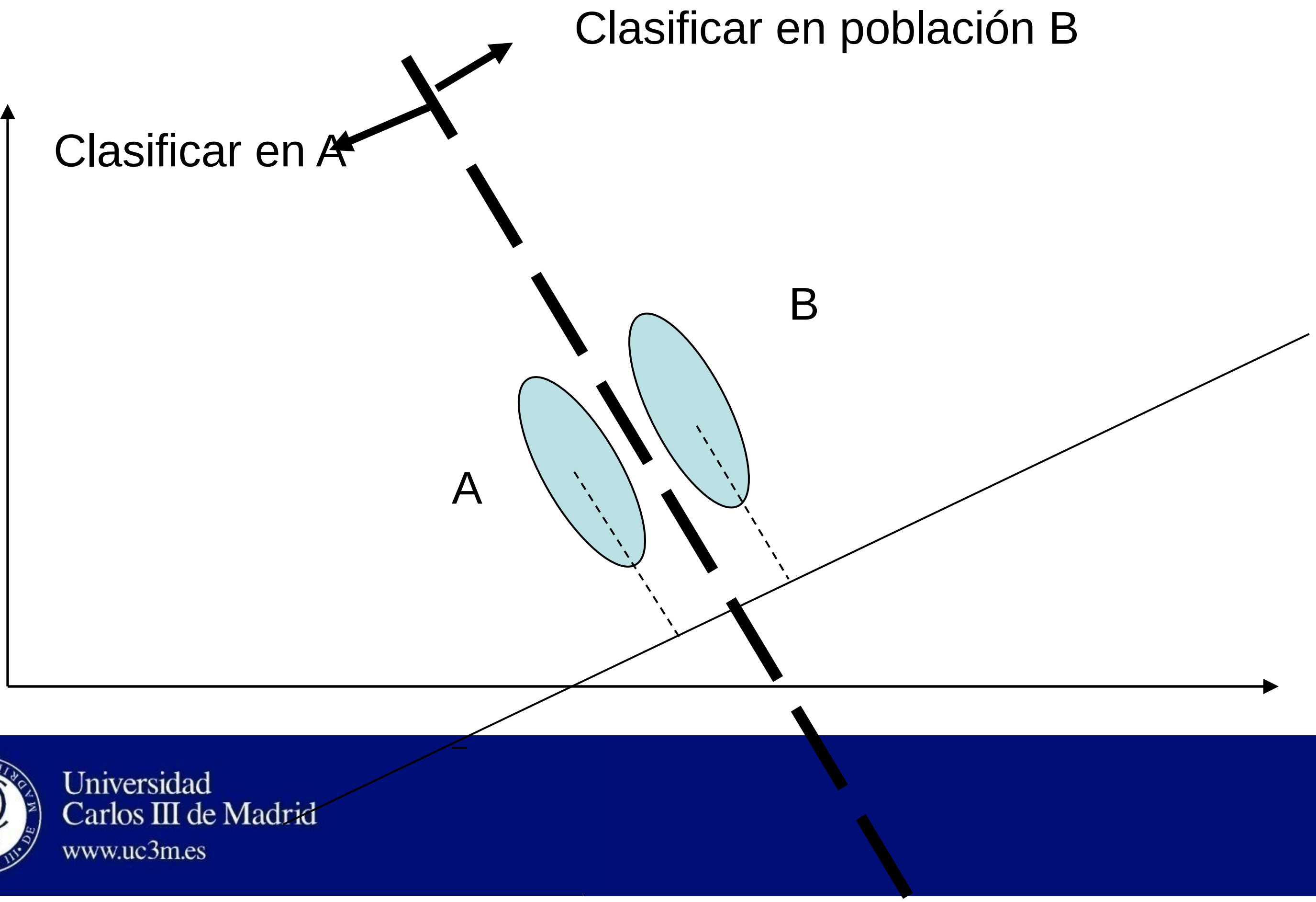


Supongamos dos poblaciones con medias μ_1 y μ_2 y matriz de varianzas común V . Encontrar una dirección óptima de clasificación $d \in R^p$ tal que

$$\max \frac{d'(\mu_1 - \mu_2)}{var(d'x)}$$

La solución es

$$d = V^{-1}(\mu_1 - \mu_2)$$



Clasificar en población B

Clasificar en A

B

A

Mahalanobis´ distance

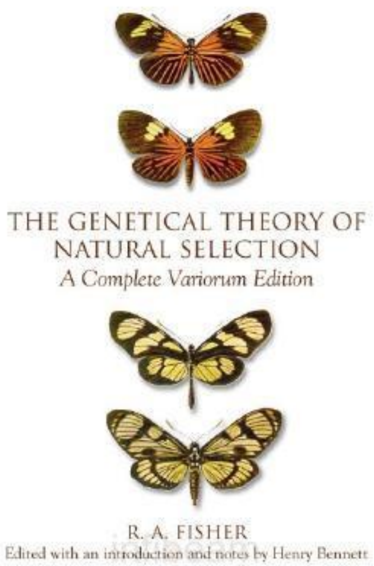
equivale a minimizar las distancias de Mahalanobis

$$\min(x_0 - \mu_i)V^{-1}(x_0 - \mu_i)$$

$$f(\mathbf{x}) = |\mathbf{V}|^{-1/2}(2\pi)^{-p/2}\exp\left\{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

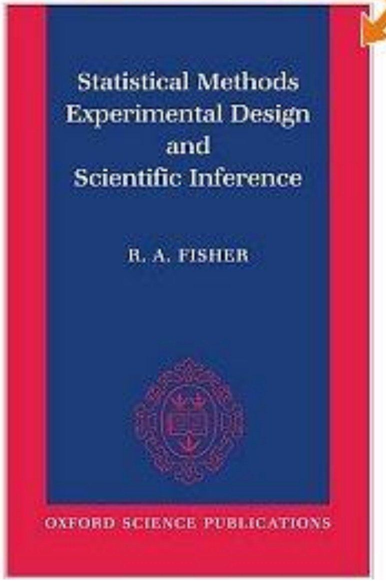
4. From 53 to 72 (1943-62)

- Professor of Genetics, Cambridge University. (1943)
- Many honors: Doctorate (Glasgow, Chicago, Calcutta), Knight, President RSS, ISI,…
- *Statistical Methods and Scientific Inference* (1956)
- Retired from U. Cambridge (1957)
- Move to CSIRO, Adelaide, Australia (1957)
- Died of bowel cancer in Adelaide (1962)



THE GENETICAL THEORY OF NATURAL SELECTION
A Complete Variorum Edition

R. A. FISHER
Edited with an introduction and notes by Henry Bennett



Statistical Methods
Experimental Design
and
Scientific Inference

R. A. FISHER

OXFORD SCIENCE PUBLICATIONS

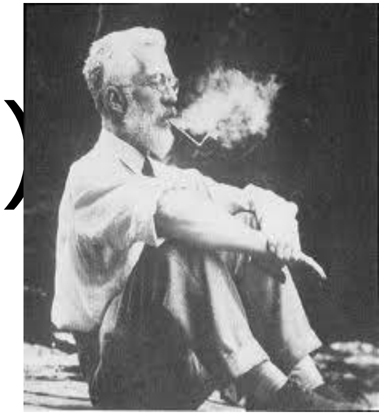
In 1952, when presenting R.A. Fisher for the Honorary degree of Doctor of Science at the University of Chicago, W. Allen Wallis described him in these words.

> He has made contributions to many areas of science; among them are agronomy, anthropology, astronomy, bacteriology, botany, economics, forestry, meteorology, psychology, public health, and—above all—genetics, in which he is recognized as one of the leaders. Out of this varied scientific research and his skill in mathematics, he has evolved systematic principles for the interpretation of empirical data; and he has founded a science of experimental design. On the foundations he has laid down, there has been erected a structure of statistical techniques that are used whenever men attempt to learn about nature from experiment and observation.

Smoking and cancer (I)

- In 1956 a large study in Britain showed a strong association between smoking and lung cancer.
- Fisher argued that correlation is not a prove of causality and fought against this idea.
- He was consultant of the tabacco manufacturers since 1956.

Smoking and cancer (II)



- Fisher defended a hidden genetic cause that produce both cancer and propensity to smoke. He found some small evidence in twin studies.
- He discovered a study showing that inhalers develop less cancer that noninhalers. (More evidence showed few differences).
- He never recognized that whole evidence of data indicates that the causality hypothesis was strongly supported by the data.

5. Personal side

Plate 11. Mrs. Fisher 1938, with daughters, in order of age, Margaret (top right), Joan (bottom right), Phyllis (top left), Elizabeth (bottom left), Rose standing beside her chair, and June in her lap.

Plate 10. R. A. Fisher, 1938, with sons George (aged 19) and Harry (14).

The marriage lasted 25 years. Fisher went alone to Cambridge in 1943
His eldest son George, died in 1943, his plane crushed.
Fisher died away from his family in Adelaide

Politics and Eugenics

- Eugenics, combination of evolutionary theory and genetics.
- Educated men have fewer children than lower class people. This a degradation of the genetic stock. Propose subsidy intelligent people to stimulate them to procreate. The allowance should be proportional to the income.
- Propose legalization of voluntary sterilization.

Personality

- Bad temper and many conflicts with colleagues. He hated to admit he was wrong.
- Generous with young people and scientists.
- Strong minded, hard fighter and political conservative and elitist.

6. The Heritage. Teaching

- The responsability for teaching statistics to mathematians with experience in practical research and data analysis. (India, 1938)
- Teach statistics and a scientific discipline and emphasize project work.
- Avoid mathematics for the sake of it.
- Statistics as the key tool for scientific learning in all stages of the process.

Heritage. Teaching

- Many of the best Department of Statistics in the world have followed his advice

Stanford, Madison-Wisconsin, Chapel Hill NC, Harvard, Chicago, Cambridge, Oxford, Iowa, Toronto, Minesotta…

Heritage. Research

R. A. FISHER IN THE 21ST CENTURY 111

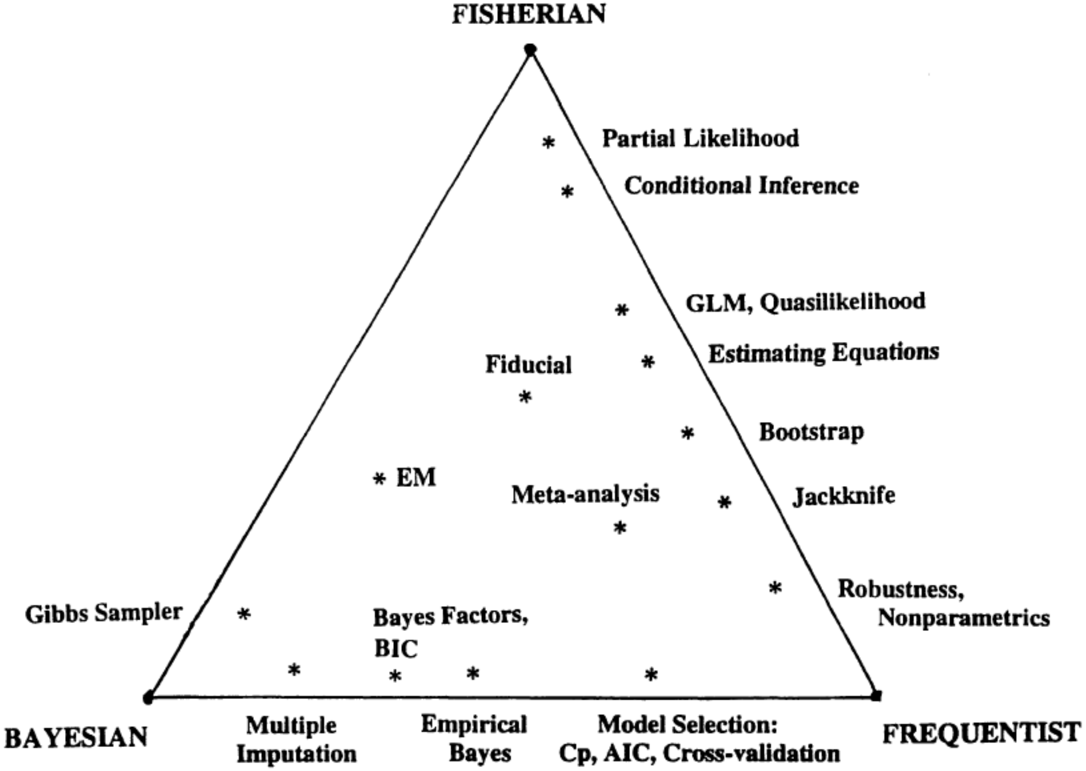


FIG. 8. *A barycentric picture of modern statistical research, showing the relative influence of the Bayesian, frequentist and Fisherian philosophies upon various topics of current interest.*

Disciples

Few people learnt Statistics from Fisher—at least in the conventional way. Rothamsted was a research station and Fisher was professor of *genetics* in London and in Cambridge.

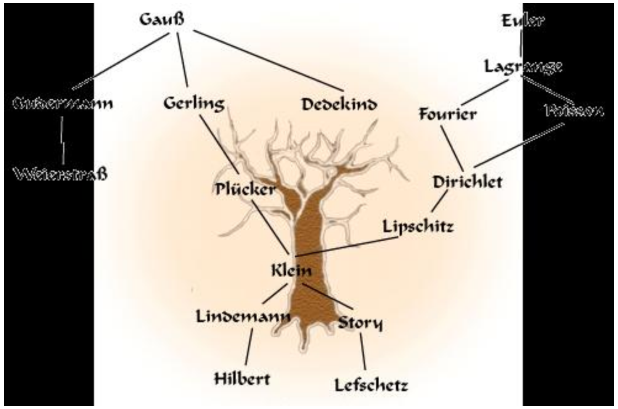
In England:

Harold Hotelling (1895-1973); J. O. Irwin (1898-1982), J. Wishart (1898-1956) and F. Yates (1902-1994),, M. S. Bartlett (1910-2002) and W. G. Cochran (1909-1980), Oscar Kempthorne (1919-2000) George W. Snedecor (1881-1974) . W. J. Youden (1900-1971), E. A. Cornish (1909-1973) M. M. Barnard.

In France: Georges Darmois; In Denmark: Georg Rasch, in India and the US: C. R. Rao (b. 1920) was Fisher's only Cambridge PhD student in Statistics, going to Fisher after he had published the Cramér-Rao and Rao-Blackwell theorems.

Of the people who *never* worked with Fisher, G. A. Barnard (1915-2002) and George Box (1919) were very close to him.

Mathematics Genealogy Project
Ronald A. Fisher



D.Sc. University of Cambridge 1926

- Advisor 1: Sir James Hopwood Jeans
- Advisor 2: F. J. M. Stratton

Students:

- Henry Bennett University of Cambridge 1953
- Walter Bodmer University of Cambridge 8
- Anthony Edwards 56
- J. Irwin University of Cambridge1937 294
- Alan Owen University of Cambridge1948
- C. R. Rao University of Cambridge1948 430

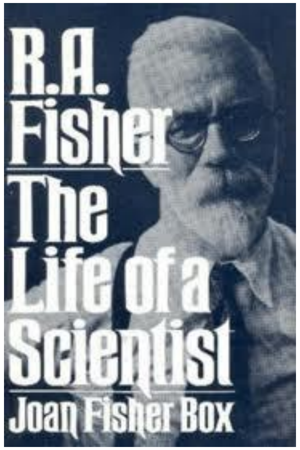
- According to our current on-line database, Ronald Fisher has 6 students and 794 descendants.

Conclusion

- **Harold Hotelling**: To become a statistician practice statistics and mull Fisher *(our Masters)* over with patient, respect and skepticism.

Even Fisher was not always right!

Thanks !



References

[1] Box, G. E. P. (1976) Science and Statistics, *JASA*, 356, 791-799

[2] Efron, B. (1998) R.A. Fisher in the 21st Century, *Statistical Science*, 13, 2, 95-122

[3] Hotelling, H. (1950), The impact of R.A. Fisher in Statistics, *JASA*,

[4] Rao,C.R. (1992), R.A. Fisher: The founder of Modern Statistics, *Statistical Science*, 7, 34-48.

[5] Savage, L.J (1976) On Reading R.A.Fisher, *The Annals of Statistics*, 4, 3, 441-500

http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafframe.htm