“On the Theoretical Foundations of Mathematical Statistics”

RA Fisher

presented by Blair Christian
10 February 2003

Ronald Aylmer Fisher

1890-1962 (Born in England, Died in Australia)

1911- helped found Eugenics Society (Cam. Univ)

1912 BA in astronomy (Cam. Univ)

1913-1919- mathematics teacher (vision too bad to fight, subsistence farming)

1919-1933?-Rothemstad biologist/statistician (ANOVA, exp design, “Statistical Methods for Research Workers”

1931,1936- visitor at Iowa State

1933- Professor of Eugenics, conflict with Neyman, E Pearson

1943- Professor of Genetics

1957- retired, moved to Australia (1959)

Outline:

1. Theoretical Statistics before 1922
    - Pearson's curves and the method of moments
    - Neglected for several reason
2. Fisher's Contributions
    - Brings rigour to mathematical statistics
    - Bridges key conepts, resulting in...

Fisher’s Contributions in 1922:

- The problems of statistics: specification, estimation and sampling distribution
- Sufficiency, ancillarity and factorization
- Maximum Likelihood Estimation , invariance
- (Fisher) consistency
- Efficiency (partially)
- Contrasts to methods of Bayes and Pearson

Definitions for Clarity:

specification-the distribution of our sample ($x_i \sim f(x, \theta)$)

estimation-estimating parameters of our distribution ($T(x)$)

sampling distribution-the distribution of our estimator ($T(x) \sim G$, Fisher's est's usually based on MLE, Pearson's on MOM)

efficiency-comparison of variance of estimator with the variance of the most efficient (lowest variance) estimator ($\frac{Var(T_1(x))}{Var(T_2(x))}$, ie Fisher information)

Some order in Pearson's Approach

- advantages:
  - 8 types of curves fit using first four moments
  - worked well in most practical situations
- disadvantages:
  - existence problems (cauchy dist)
  - inefficiency increases with distance from normality (see figure in paper)

Statistical tools before 1922 (keyword is *ad hoc*):

- Pearson's curves (1895), goodness of fit (1900) (specification)

- MOM, Bayes' 1763 paper still influential (cited), other methods (estimation)

- Many scattered ideas (regression, correlation, ANOVA, normality/probability,...)

- Student's 1908 work on the t-distribution, understanding of the chi-square distribution (samp dist)

Statistical Tools Missing or Incomplete Before 1922:

- data reduction, relationship to estimation in a formal framework
- criterion for efficiency (Fisher, 1920)
- basic concepts (Bayes methods okay? Distinction between parameters and statistics? Relationship between statistics and probability?)
- Clear framework for specification, estimation and sampling distribution
- Distribution of regression coefficients

Fisher Begins with 'The Neglect of Theoretical Statistics'

> "...in spite of the immense amount of fruitful labour... the basic principles [of TS] are still in a state of obscurity,... fundamental problems have been ignored and fundamental paradoxes left unresolved."

Hopes to work on basic principles, which has been hindered due to 1) the idea that it is impossible to quantify error and 2) verbal confusion between parameters and statistics.

Sec. 2, “The Purpose of Mathematical Statistics”

Statisticians should:

1. reduce data as much as possible without losing the relevant information (sufficient and ancillary statistics)

2. construct a hypothetical infinite population of which the data are a random sample (ie as your sample tends to infinity, the histogram/FP converges to the true distribution)

3. relate this to probability theory via a distribution specifed by few parameters

Sec. 3, “The Problems of Statstics”

The reduction of data leads to 3 types of problems:

1. Specification (“a matter entirely for the practical statistician”)

2. Estimation (focus of article, MLE)

3. Sampling Distribution (“very little progress has been made”, chi-square and t-distributions)

After specifying the distribution (“We must confine ourselves to those forms which we know how to handle”), estimation and sampling distribution are related-

- want easy to compute estimates whose sampling distributions are known

Sec. 4, “Criteria of Estimation”

Def. A statistic is *consistent (Fisher consistent)* if, when calculated from the whole population, it is equal to the parameter describing the probability law.

Ex. Consider estimating the mean of a normal distribution, $T_n = \bar{x}_n$ is Fisher consistent, but $T_n = \bar{x}_n + \frac{1}{n}$ is not.

Ex. When estimating σ from a normal sample (μ unknown), here are two consistent estimators-which is better?

$$\begin{aligned} \hat{\sigma}_1 &= \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum_{i=1}^{n}|x_i - \bar{x}| \\ \hat{\sigma}_2 &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Which is better? Both are Fisher consistent. Need another criterion.

- $\widehat{\sigma}_1$ is easier to compute

- $\widehat{\sigma}_2$ is more efficient ($\frac{Var(\widehat{\sigma}_1)}{Var(\widehat{\sigma}_2)} = 1.14$), ie we need 14% more data when using $\widehat{\sigma}_1$ in order to get the same variance as $\widehat{\sigma}_2$

- small annoyance- lack of uniqueness for efficient statistics?

Sec. 5, “Examples of the Use of the Criterion of Consistency” (criticisms of Pearson’s techniques)

- only efficient in normal case

- only works when first 4 moments are finite, etc. (distributions not satisfying this cannot be determined)

- ie, fails for Cauchy distribution, since $\bar{x}$ has the same distribution as x (Fisher says we must consider median, and should not discard any data)

Def. A statistic is *sufficient* if it summarizes the whole of the relevant information supplied by the data. If θ is to be estimated and T_1 is sufficient, then for any other statistic T_2, we have that T_2 given T_1 is independent of θ.

- Fisher uses a bivariate normal to argue that sufficient statsitics are efficient.

Def. A region is *isostatistical* if each sample from this region gives a statistic with the same value

Sec. 6, "Formal Solution of Problems of Estimation", aka Method of Maximum Likelihood

Since sufficiency is not adequate as a criterion to provide estimators, Fisher proposes the MLE as a method of obtaining a sufficient estimate. Let

$$P\left(x \in A\right) = \int_A f\left(x \mid \theta\right) dx$$

then for n independent observations,

$$\begin{aligned} & P\left(n_1\, x\text{'s} \in A_1, \cdots, n_p\, x\text{'s} \in A_p\right) \\ = & \frac{n!}{\prod_{i=1}^{p} n_i!} \prod_{i=1}^{p} \left\{ \int_{A_i} f\left(x_i \mid \theta\right) dx_i \right\}^{n_i} \end{aligned}$$

Maximizing this quantity (or equivalently its log) with respect to θ gives estimates $\widehat{\theta}$.

In his first attempt, Fisher used a Bayesian approach, but makes clear to differentiate MLE from the Bayesian approach. The main reservation about Bayesian methods was the lack of invariance (later Jeffreys found an invariant prior) see paper for example and discussion.

If we assume that the asymptotic distribution of the MLE is normal, then

- the asymptotic variance of the MLE is the reciprocal of the Fisher information (proof)
- notes that a similar proof use for MOM was incorrect (MOM is actually inefficient)

Sec. 7, “Satisfaction of the Criterion of Sufficinecy” or Sufficiency and Maximum Likelihood

- derived the form of the limiting normal dist for the MLE

- earlier showed that a sufficient estimate has the smallest-variance normal distn in large samples

- if the MLE is sufficient we have a great estimator (proof is incorrect)

Example:

Suppose $x_1, \cdots, x_n \sim Exp(\beta)$. Then the likelihood equation is:

$$
\begin{aligned}
L(\beta \mid \mathbf{x}) &= \prod_{i=1}^{n} \beta^{-1} \exp\left\{-\frac{1}{\beta} x_i\right\} \\
&= \beta^{-n} \exp\left\{-\frac{1}{\beta} \sum_{i=1}^{n} x_i\right\}
\end{aligned}
$$

and $\sum_{i=1}^{n} x_i$ is sufficient. So we can estimate using (equivalently) the log likelihood, taking its derivative and solving for β,

$$
\begin{aligned}
\frac{d}{d\beta}\left(l(\beta \mid \mathbf{x}) = -n \log \beta - \frac{1}{\beta} \sum_{i=1}^{n} x_i\right) = 0 \\
\frac{-n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i = 0 \\
\widehat{\beta} = \bar{x}_n
\end{aligned}
$$

and if we want the sampling distribution, we know that

$$\sum_{i=1}^{n} x_i \sim Gam\left(\frac{1}{\beta}, n\right)$$

so

$$\begin{aligned} E\left(\bar{x}_n\right) &= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n} n\beta \\ &= \beta \end{aligned}$$

and

$$\begin{aligned} V\left(\bar{x}\right) &= \frac{1}{n^2} V\left(\sum x_i\right) \\ &= \frac{1}{n^2} n\beta^2 \\ &= \frac{\beta^2}{n} \end{aligned}$$

Now we can compare our estimate with others.

Sec. 8-13 Are mainly applications showing his methods superior to Pearson's

- focus on efficiency

- MOM efficiency decreases as distributions become less normal

- special case where MOM has zero efficiency

- many examples (mostly using Pearson's curves), also a case where it is impractical to use MLE, but a highly efficient, simple to calculate estimate can be found

What's wrong with the 1922 paper?

- criticisms
- existence of sufficient statistic, MLE not demonstrated
- optimality of MLE not explicit
- MLE when no maximum (uniform, etc)
- bad proof of factorization (sufficiency not fully developed)
- Fisher consistency not adopted

Conclusions:

- a groundbreaking, unifying paper
- not fully appreciated
- some errors, still gaps remaining

References:

Fisher, RA. (1922) “On the Mathematical Foundations of theoretical Statistics”, *Phil. Trans. A.*, 309-368.

Geisser, S. (1989) “Basic Theory of the 1922 Mathematical Statistics Paper”, Ch 7 in “RA Fisher: An Appreciation”, eds Fienberg and Hinkley, 59-66.

Hald, A. (1998) “A History of Mathematical Statistics From 1750 to 1930”, Wiley and Sons:New York.

Box, JF. (1978) “RA Fisher: The Life of a Scientist”, Wiley and Sons:New York.

Savage, LJ (1976) “On Rereading RA Fisher” (with discussion), *Annals of Statistics*, (4)441-500.