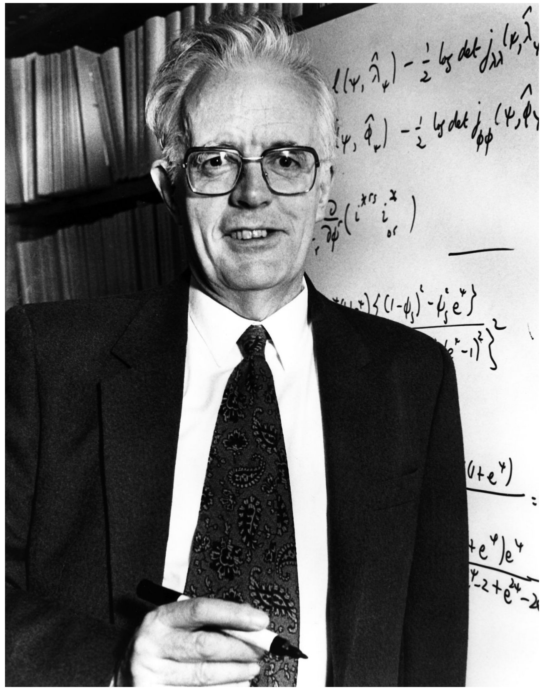
Introduction

Statistical Inference - PhD EcoStatData

Tommaso Rigon
Università degli Studi di Milano-Bicocca

Homepage



"I would like to think of myself as a scientist, who happens largely to specialise in the use of statistics."

Sir David Cox (1924-2022)

- This course will cover the following **topics**:
  - Point estimation
  - Exponential families
  - Generalized linear models
  - ...and more advanced topics
- This is a **Ph.D.-level** course, so it is assumed that you have already been exposed to all these topics to some extent.
- We aim to (briefly!) touch upon many **key concepts** of **classical statistical inference** from the **20th century**.
- Fundamental topics such as **hypothesis testing** are not covered here, as they are addressed in another module.
- To introduce the main ideas, I will borrow the words of Davison (2001) — a source you are **encouraged to read!**

Statistics of the 20th century

Biometrika (2001), **88,** 1, *pp.* 13–52

Printed in Great Britain

Biometrika Centenary: Theory and general methodology

By A. C. DAVISON

Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland

anthony.davison@epfl.ch

Summary

Contributions to statistical theory and general methodology published in *Biometrika*, 1901–2000, are telegraphically reviewed.

Some key words: Bayesian inference; Estimating function; Foundations of statistics; Generalised regression model; Graphical method; Graphical model: Laplace approximation; Likelihood; Missing data; Model selection; Multivariate statistics; Non-regular model; Quasilikelihood; Saddlepoint: Simulation; Spatial statistics.

- **Biometrika** is among the most prestigious journals in Statistics. Past editors include Karl Pearson, Sir David Cox, and Anthony Davison.

Foundations and Bayesian statistics

2. Foundations

2·1. *Objective theories*

The foundations of statistical inference are discussed sporadically in early issues of *Biometrika*. Karl Pearson's views were apparently close to what was then called inverse probability and is now called a Bayesian approach, though he seems to have believed that the prior should have some frequency interpretation, and certainly his son Egon Sharpe Pearson believed that it should be subject to empirical assessment. Inverse probability had dominated discussions of inference for much of the 19th century, but for the first half of the 20th century it was largely eclipsed by the ideas of Fisher, Neyman and Egon Pearson. An important exception to this was the attempt to put objective Bayesian inference on a secure footing summarised by Jeffreys (1939). Wilk's (1941) verdict now seems ironic:

> From a scientific point of view it is doubtful that there will be many scholars thoroughly familiar with the system of statistical inference initiated by R. A. Fisher and extended by J. Neyman, E. S. Pearson, A. Wald and others who will abandon this system in favour of the one proposed by Jeffreys in which inverse probability plays the central role.

The rats have subsequently abandoned ship in numbers unthinkable in the 1940s. Bayesian contributions in *Biometrika* are briefly surveyed in § 7.

Principles: sufficiency, conditionality and likelihoods

2·2. *Principles of inference*

Most attempts to put statistical inference on a consistent footing rest on some subset of the sufficiency, conditionality, likelihood and invariance principles. The sufficiency principle simply says that two datasets from the same model that yield identical minimal sufficient statistics should give the same inferences on the model parameter θ. The conditionality principle rests on the notion of an ancillary statistic, namely a function A of the minimal sufficient statistic whose distribution does not depend on the parameter, and says that inference should be conducted using the relevant subset of the sample space, i.e. that portion of it in which A equals its observed value a. [redacted]

The strong likelihood principle states that, if the likelihoods under two possibly different models but with the same parameter are proportional, then inferences about θ should be the same in each case; its weak form is equivalent to the sufficiency principle. In particular, this implies that inference should not be influenced by elements of the sample space that were not in fact observed, appearing to rule out use of procedures such as significance tests and confidence intervals and paving a path towards some form of Bayesian inference.

Both the sufficiency and conditionality principles are accepted more widely than the likelihood principle, so Birnbaum's (1962) article in the *Journal of the American Statistical Association* caused consternation when he showed that acceptance of the first two entails acceptance of the third. Later work somewhat reducing the force of this result includes Kalbfleisch (1975), who distinguishes between experimental and mathematical ancillaries.

Likelihood

3. Likelihood

3·1. *Primary notions*

Likelihood is central to much statistical theory and practice. The ideas of likelihood, sufficiency, conditioning, information and efficiency are due to R. A. Fisher, whose work established the excellent asymptotic properties of the maximum likelihood estimator and showed that likelihood gives a basis for exact conditional inference in location and scale models. The last 25 years have seen a second flowering of this theory, stimulated by influential work in the late 1970s and early 1980s.

- The study of the **likelihood** has gone **far beyond** the classical **textbook** description. Specialized topics that have attracted considerable attention include:
  - Likelihood ratio tests and their large-sample properties
  - Conditional and marginal likelihoods
  - Modified profile likelihoods
  - Restricted maximum likelihood

Estimating functions

5. Estimating functions

Elementary discussion of maximum likelihood estimators usually stresses asymptotic efficiency rather than finite-sample properties, which are difficult to obtain because of the implicit nature of the score function. Starting in the 1960s finite-sample theory was developed from a different point of view, taking as basis score-like estimating functions that determine estimators rather than estimators themselves. If ψ is a scalar parameter to be estimated using data Y, the function $g(Y; \psi)$ is said to be an unbiased estimating function if $E\{g(Y; \psi)\} = 0$ for all ψ. With $g(y; \psi) = \partial \log f(y; \psi)/\partial\psi$ this determines the maximum likelihood estimator, but it also encompasses least squares, minimum chi-squared and robust estimators. In a paper in the *Annals of Mathematical Statistics* and under regularity conditions, Godambe (1960) defined g^* to be optimal in the class of all unbiased estimating functions if it minimised the ratio $E\{g^*(Y; \psi)^2\}/E\{\partial g^*(Y; \psi)/\partial\psi\}^2$ for all ψ. An asymptotic basis for this choice is that this ratio is the large-sample variance of $\hat{\psi}$ determined as the root of the equation $g(Y; \psi) = 0$. A Cramér–Rao argument then establishes that the estimating function $g(y; \psi) = \partial \log f(y; \psi)/\partial\psi$, the score function, is optimal in this finite-sample sense; this result extends to vector ψ. It was extended to give a finite-sample non-

Generalized linear models

10. GENERALISED REGRESSION

10·1. Generalised linear models

One of the most important developments of the 1970s and 1980s was the unification of regression provided by the notion of a generalised linear model (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) and its associated software, though the concept had appeared earlier (Cox, 1968). In such models the response Y is taken to have an exponential family distribution, most often normal, gamma, Poisson or binomial, with its mean μ related to a vector of regressor variables through a linear predictor $\eta = x^{\mathrm{T}}\beta$ and a link function g, where $g(\mu) = \eta$. The variance of Y depends on μ through the variance function $V(\mu)$, giving $\mathrm{var}(Y) = \phi V(\mu)$, where ϕ is a dispersion parameter. Special cases are:

for fitting regression models. The estimating equations for a generalised linear model for independent responses $Y_1, \ldots, Y_n$ and corresponding covariate vectors $x_1, \ldots, x_n$ may be expressed as

$$\sum_{j=1}^{n} x_j \frac{\partial \mu_j}{\partial \eta_j} \frac{Y_j - \mu_j}{V(\mu_j)} = 0, \qquad (10)$$

or in matrix form

$$D^{\mathrm{T}} V^{-1}(Y - \mu) = 0, \qquad (11)$$

where D is the $n \times p$ matrix of derivatives $\partial \mu_j / \partial \beta_r$, and the $n \times n$ covariance matrix V is diagonal if the responses are independent but not in general. Taylor expansion of (11)

Quasi likelihoods

10·2. *Quasilikelihood*

Data are often overdispersed relative to a textbook model. For example, although the variance of count data is often proportional to their mean, the constant of proportionality ϕ may exceed the value anticipated under a Poisson model, so $\text{var}(Y) = \phi\mu$ for $\phi > 1$. One way to deal with this is to model explicitly the source of overdispersion by the incorporation of random effects; see § 3·5. The resulting integrals can considerably complicate computation of the likelihood, however, and a simpler approach is through quasilikelihood (Wedderburn, 1974).

Quasilikelihood is perhaps best seen as an extension of generalised least squares. To see why, note that (11) is equivalent to $U(\beta) = 0$, where $U(\beta) = \phi^{-1}DV^{-1}(Y - \mu)$. Asymptotic properties of $\hat{\beta}$ stem from the relations $E(U) = 0$ and $\text{cov}(U) = -E(\partial U/\partial\beta)$, corresponding to standard results for a loglikelihood derivative. However, these properties do not depend on a particular probability model, requiring merely that $E(Y) = \mu$ and $\text{cov}(Y) = \phi V(\mu)$, subject also to some regularity conditions. Hence $\hat{\beta}$ has the key properties of a maximum likelihood estimator, namely consistency and asymptotic normality, despite not being based on a fully-specified probability model. Moreover, it may be computed simply by solving (11), that is, behaving as if the exponential family model with variance function $V(\mu)$ were correct. The scale parameter ϕ is estimated by $\hat{\phi} = (n-p)^{-1}(Y - \hat{\mu})^{\mathrm{T}}V(\hat{\mu})^{-1}(Y - \hat{\mu})$, and the asymptotic covariance matrix of $\hat{\beta}$ is $\hat{\phi}(D^{\mathrm{T}}VD)^{-1}$ evaluated at $\hat{\beta}$. A unified asymptotic treatment of such estimators from overdispersed

Nonparametric (local) models

10·4. *Local models*

One major change during the last two decades has been the development, implementation and now widespread use of smoothing procedures. A wide range of methods for local density and curve estimation, each with its advantages and disadvantages, is now available to the data analyst. Contributions in *Biometrika* to this area are reviewed in Hall (2001), and here we simply note some connections with the regression models discussed above. One approach to local estimation of the mean $\mu(x)$ of a response Y as a function of the scalar covariate x is through weighting the contribution to a system of estimating equations according to their distance from the point at which local estimation is required. Then (10) becomes

$$\sum_{j=1}^{n} h^{-1} w\left(\frac{x_j - t}{h}\right) \frac{\partial \mu_j}{\partial \beta} \frac{Y_j - \mu_j}{V(\mu_j)} = 0, \qquad (12)$$

where t is the value of x at which it is required to estimate μ, $w(.)$ is a weighting function such as the normal density, and h is a bandwidth. As $h \to \infty$ the system reduces to (10), while as $h \to 0$ the estimation is based entirely on the observations closest to t. One

Bayesian methods

4. Predictive inference

Prediction has long been a weak point of likelihood theory. Bayes's theorem gives a basis for inference not only on parameters but also on an unobserved random variable Z. Given the observed data $Y = y$, the posterior predictive density of Z may be written

$$f(z \mid y) = \int f(z \mid y; \theta)\pi(\theta \mid y)\, d\theta = \frac{\int f(z \mid y; \theta) f(y; \theta)\pi(\theta)\, d\theta}{\int f(y; \theta)\pi(\theta)\, d\theta}, \qquad (4)$$

where $\pi(\theta)$ is a prior density for θ and $\pi(\theta \mid y)$ is the corresponding posterior density.

7. Bayesian statistics

The Bayesian revival that began in the 1950s soon led to *Biometrika* publishing investigations of particular models important in applications, such as the linear model (Tiao & Zellner, 1964) and random effects models (Tiao & Tan, 1965, 1966), but also to broader methodological discussions, concerning particularly the robustness of Bayesian inferences. Examples are Box & Tiao (1962, 1964), who assess the sensitivity of posterior densities to distributional assumptions, replacing the normal with a heavier-tailed density in work prefiguring current practice, and the investigation of modelling of outliers by mixtures in Box & Tiao (1968).

A very important recent development has been the emergence of Markov chain Monte Carlo methods for use in Bayesian applications. This is discussed in § 8.

Prerequisite of this course

- As mentioned, it is assumed that you have already been exposed to courses on statistical inference before.
- **Propedeutical** topics that I will **not** discuss here are:
  - Asymptotic probability theory, $O_p(\cdot)$ and $o_p(\cdot)$ notations
  - Likelihood function: definition and basic properties
  - Sufficiency, ancillarity, Fisher factorization theorem, minimality
  - Tests based on the likelihood (likelihood ratio, score test, Wald test), asymptotically equivalent forms, confidence intervals
  - Linear models, ordinary least squares, exact normal theory
- If you are unfamiliar with any of these, please have a look at Chap. 2 and Chap. 3 of Pace and Salvan (1997), and Davison (2003).

Statistical Inference

- The key assumption is that **observations** $y_1, \ldots, y_n$, seen as realizations of the **random variables** $(Y_1, \ldots, Y_n) \sim P_\theta$, provide information about the **generating process** $P_\theta(\cdot)$.
- We assume that P_θ is only partially known; that is, it belongs to a **model class** specified by the tuple

$$(\mathcal{Y}, P_\theta, \Theta),$$

  where $\mathcal{Y}$ is the **sample space**, P_θ is a **probability measure** over $\mathcal{Y}$ indexed by $\theta \in \Theta$, and Θ is the **parameter space**.
- In this course, we focus on the **parametric case**, where $\Theta \subseteq \mathbb{R}^p$. Hence, $\theta \in \Theta$ is a vector-valued **parameter** that we aim to infer from the data.
- If instead Θ is not a subset of $\mathbb{R}^p$, then we are in the domain of **nonparametric statistics**.
- A basic requirement is **identifiability**, meaning that

$$\text{if} \quad \theta_1 \neq \theta_2, \qquad P_{\theta_1} \neq P_{\theta_2},$$

  that is, there exists a measurable set $A \in \mathcal{B}(\mathcal{Y})$ such that $P_{\theta_1}(A) \neq P_{\theta_2}(A)$.

Dominated statistical models

- We will focus on **dominated families** of distributions, namely we assume there exist a measure $\nu(\mathrm{d}\boldsymbol{y})$ over $\mathcal{B}(\mathcal{Y})$ such that P_θ is absolutely continuous w.r.t. ν for all $\theta \in \Theta$, that is

$$\forall A \in \mathcal{B}(\mathcal{Y}) \quad \text{such that} \quad \nu(A) = 0 \quad \Longrightarrow \quad P_\theta(A) = 0.$$

- **Radon-Nikodym** theorem then ensures there exists a **probability density** $f(\boldsymbol{y};\theta)$ such that

$$P_\theta(A) = \int_A f(\boldsymbol{y};\theta)\nu(\mathrm{d}\boldsymbol{y}).$$

  If $\mathcal{Y} \subseteq \mathbb{R}^d$, then ν is typically the **Lebesgue measure** or the **counting measure**.

- A **dominated statistical model** is therefore identified by the following class of densities:

$$\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\},$$

  or more precisely by the tuple $(\mathcal{Y}, f(\cdot;\theta), \Theta)$, with $\Theta \subseteq \mathbb{R}^p$. We will only consider the dominated case in this course.

Likelihood function

- Let $\mathcal{F}$ be a **dominated (parametric) statistical model** and $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathcal{Y}$ the observed data. Let $c = c(\boldsymbol{y}) > 0$ be a positive **arbitrary constant**, the function $L : \Theta \to \mathbb{R}^+$ defined as

$$L(\theta) = L(\theta; \boldsymbol{y}) = c(\boldsymbol{y}) f(\boldsymbol{y}; \theta), \qquad \theta \in \Theta,$$

  is called **likelihood function**. The log-likelihood function is $\ell(\theta) := \log L(\theta)$.
- Some authors set $c = 1$, but this is **debatable**. Indeed, defining the likelihood **up to a multiplicative factor** can be justified in multiple ways:
- Intuitively, when comparing the coherency of two statistical models with the observed data, we only care about ratios of the form $L(\theta_1; \boldsymbol{y})/L(\theta_2; \boldsymbol{y})$ where the constant simplifies.
- Moreover, this definition does not depend on the choice of the dominating measure ν.
- In particular, the likelihood is **invariant** under **one-to-one transformations** of the data, as the jacobian of the transformation can be incorporated into $c(\boldsymbol{y})$.
- This is also the **original definition** provided by **Fisher** in 1922!

Textbooks

- We will use multiple textbooks throughout this course — some more specialized than others. Please treat them as reference materials to consult as needed.
- Roughly speaking, they can be organized as follows:
  - **General references**: Casella and Berger (2002), Davison (2003), and Pace and Salvan (1997)
  - **Point estimation**: Lehmann and Casella (1998) and Keener (2010)
  - **Exponential families**: Pace and Salvan (1997)
  - **Asymptotic statistics**: van der Vaart (1998)
  - **Generalized linear models**: Agresti (2015), McCullagh and Nelder (1989)
- The book by Davison (2003) is perhaps the most accessible among the listed texts. You are encouraged to refer to it if you need to review or catch up on prerequisite material.
- In addition, specialized articles and resources will be discussed throughout the course to complement the textbook material.

The future

12. The future

What would Karl Pearson make of a current issue of *Biometrika*? Statistical theory and methods have developed so much and in so many unexpected ways over the past century that detailed prediction would be foolhardy. One broad trend has been the mathematisation of the subject, which has greatly clarified key notions. It has also enabled ready transfer of ideas from fields such as probability, stochastic processes, algorithmics, optimisation and so forth, despite occasional complaints that journals have become unreadable by non-specialists, comments that were already being made in the early years of the twentieth century!

Perhaps the dominant trend is the effect of the astonishing advances in computation without which much of modern statistics would not have developed. A consequence of this is the increasingly detailed modelling of phenomena in ways unthinkable only 15 years ago, based on data whose form and quantity would then have seemed a dream, or perhaps a nightmare! One result is increasing diversity, as researchers become more immersed in particular areas of application. This brings with it the potential for further fragmentation of the discipline of statistics, so a continuing and increasingly important role for journals such as *Biometrika* is to be a medium of transfer for new theory and methods among sub-fields.

Cynical and questionable advice for a young investigator

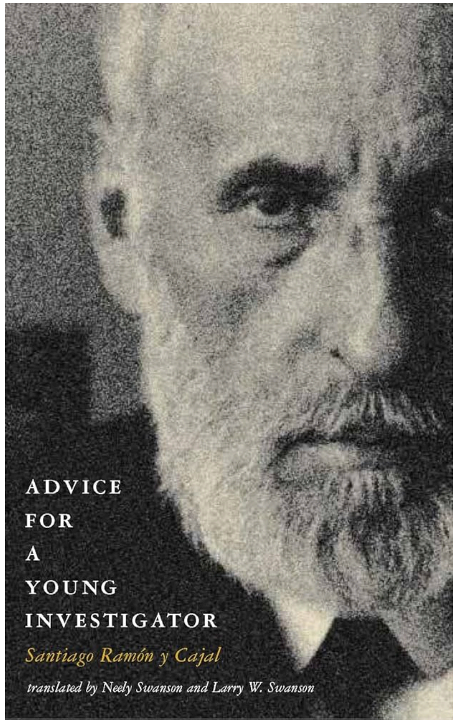
1. Strive to **publish** in **top statistical journals**, such as: *Annals of Statistics, Biometrika, Journal of the American Statistical Association, Journal of the Royal Statistical Society: Series B*.
2. However, keep in mind that both **quality** and **quantity** matter. Aim to have at least 2–4 submitted or published papers by the end of your Ph.D. — the more, the better.
3. Focus on a **niche trending topic**. Make sure you are part of a **large** and **established group** of researchers who actively promote the topic you are working on.
4. Become an **expert** in your niche, and learn how to **write** about it and promote it effectively. In a nutshell, **learn** how to **play the game**.
5. Closely follow the **suggestions** of your **advisor** — they know better than you how to navigate the system and can guide you through many political and scientific challenges.
6. **Do not wast time** on activities that do not produce papers. This include:
   - Teaching to undergraduate students
   - Disseminating your work to the broader community, beyond academia
   - Studying topics unrelated to your niche area

Deconstructing the cynical advice

- The former is a list of concrete recommendations (easier said than done, especially about publishing on top journals) that may help you secure a **permanent position** in **academia**.
- I **do not fully agree** with those rules: there is more to pursuing a Ph.D. than just "getting a job."
- These suggestions may change over time and do not necessarily apply to other fields. Moreover, keep in mind academia, in the **short period**, is also a **game of chance**.
- I recognize their **effectiveness**, but there are, I think, some **uncomfortable** consequences.

- These rules may lead to an **unhealthy competition** among peers, who struggle to **publish or perish**, which has **negative psychological** effects and it favors **incremental** contributions.
- Even if they work in the **short period**, in the **long run**, if the niche you decided to focus on is declining, transitioning towards different topics is hard if have not studied anything else.
- If the academic system rewards specialization, why study **classical statistics** or other topics at all? What about the role of Universities in **preserving** and **disseminating knowledge**?
- These suggestions apply to academia and do not consider **working** in **industry** after the Ph.D., which is what many (most?) Ph.D. students will do.

Advice for a young investigator



Santiago Ramon Y Cajal
(1852–1934)

- The former list of practical advice is probably **effective** but **questionable**. For sure, it lacks **perspective**.
- In looking for **principles** defining a **good researcher**, I once again need to borrow the words of somebody else.
- **Santiago Ramón y Cajal** is a fascinating personalities in science. He was one of the most important **neuroanatomist** of his century.
- Cajal was also a **thoughtful** and inspired **teacher**.
- "The advice" became vehicle for Cajal to write down the thoughts and anecdotes he would give to students and colleagues about how to make **important original contributions** in any branch of **science**.
- This book was written in 1898. The world was different, and so was academia. Yet, the book feels remarkably **modern**.

Introduction

On general philosophical principles

It is important to note that the most brilliant discoveries have not relied on a formal knowledge of logic. Instead, their discoverers have had an ***acute inner logic*** *that generates ideas [...]*

Let me assert without further ado that ***there are no rules of logic*** *for* ***making discoveries*** *[...]*

Must we therefore abandon any attempt to instruct and educate about the process of scientific research? Shall we leave the beginner to his own devices, confused and abandoned, struggling without guidance or advice along a path strewn with difficulties and dangers?

Definitely not. In fact, just the opposite — we believe that by ***abandoning*** *the* ***ethereal realm of philosophical principles*** *and abstract methods we can descend to the* ***solid ground*** *of* ***experimental science****, as well as to the sphere of ethical considerations involved in the process of inquiry. In taking this course, simple, genuinely useful advice for the novice can be found.*

Beginner traps

Undue admiration of authority

*I believe that **excessive admiration** for the work of **great minds** is one of the most **unfortunate preoccupations** of intellectual youth — along with a conviction that certain problems cannot be attacked, let alone solved, because of one's relatively limited abilities.*

*Inordinate respect for genius is based on a commendable sense of fairness and modesty that is difficult to censure. However, when foremost in the mind of a novice, it **cripples initiative** and prevents the formulation of **original work**. Defect for defect, arrogance is preferable to diffidence, boldness measures its strengths and conquers or is conquered, and undue modesty flees from battle, condemned to **shameful inactivity**. [...]*

*Far from humbling one's self before the great authorities of science, those beginning research must understand that [...] their destiny is to **grow a little** at the **expense of the great one**'s reputation. [...]*

By way of classic examples, recall Galileo refuting Aristotle's view of gravity, Copernicus tearing down Ptolemy's system of the universe, Lavoisier destroying Stahl's concept of phlogiston, and Virchow refuting the idea of spontaneous generation held by Schwann, Schleiden, and Robin. [...]

*It could be said that in our times, when so many idols have been dethroned and so many illusions destroyed or forgotten, there is little need for resorting to a **critical sense** and **spirit of doubt**. [...] However, old habits die hard — too often one still encounters the pupils of illustrious men wasting their talents on **defending the errors** of their **teachers**, rather than using them to solve new problems.*

Beginner traps

The most important problems are already solved

Here is another ***false concept*** *often heard from the lips of the newly graduated:* "Everything of major importance in the various areas of science has already been clarified. What difference does it make if I add some minor detail or gather up what is left in some field where more diligent observers have already collected the abundant, ripe grain. Science won't change its perspective because of my work, and my name will never emerge from obscurity."

This is often ***indolence masquerading as modesty****. [...]*

Instead, bear in mind that even in our own time science is often built on the ***ruins of theories*** *once thought to be* ***indestructible****. It is important to realize that if certain areas of science appear to be quite* ***mature****,* ***others*** *are in the process of* ***development****, and yet others remain to be born. [...]*

It is fair to say that, in general, no problems have been exhausted; instead, men have been exhausted by the problems. [...] ***Fresh talent*** *approaching the analysis of a problem* ***without prejudice*** *will always see new possibilities — some aspect not considered by those who believe that a subject is fully understood. Our knowledge is so fragmentary that unexpected findings appear in even the most fully explored topics.*

Beginner traps

Preoccupation with applied science

Another corruption of thought that is important to battle at all costs is the false distinction between **theoretical** *and* **applied** *science, with accompanying* ***praise of the latter*** *and* ***deprecation*** *of the* ***former****.*

This lack of appreciation is definitely shared by the ***average citizen****, often including lawyers, writers, industrialists, and* ***unfortunately*** *even distinguished* ***statesmen****, whose initiatives can have serious consequences for the cultural development of their nation. [...]*

People with little understanding fail to observe the ***mysterious threads*** *that bind the factory to the laboratory, just as the stream is connected with its source. Like the man in the street, they believe in good faith that scholars may be divided into two groups — those who waste time speculating about unfruitful lines of pure science, and those who know how to find data that can be applied immediately to the advancement and comfort of life.*

Is it really necessary to dwell on such an ***absurd point of view****? Does anyone lack the common sense to understand that* ***applications*** *derive immediately from the* ***discovery of fundamental principles*** *and new data? [...]*

For the present, let us cultivate ***science for its own sake****, without considering its* ***applications****. They will always come, whether in years or perhaps even in centuries. It matters very little whether scientific truth is used by our sons or by our grandsons. [...] Accept the view that* ***nothing*** *in nature is* ***useless****, even from the human point of view. Even in the rare instance where it may not be possible to use particular scientific breakthroughs for our comfort and benefit, there is one positive benefit — the noble satisfaction of our curiosity and the incomparable gratification and feeling of power that accompany the solving of a difficult problem.*

Beginner traps

Perceived lack of ability

*Some people claim a **lack of ability for science** to justify failure and discouragement. [...] but the great majority of those professing incompetence really so? Might they exaggerate how difficult the task will be, and **underestimate their own abilities**? I believe that this is often the case. [...]*

*As many teachers and thinkers have noted, discoveries are **not** the fruit of **outstanding talent**, but rather of **common sense** enhanced and strengthened by **technical education** and a habit of **thinking about scientific problems**. [...]*

*What we refer to as a great and special talent usually implies superiority that is **expeditious** rather than **qualitative**. In other words, it simply means doing quickly and with brilliant success what ordinary intellects carry out slowly **but well**.*

*Instead of distinguishing between mediocre and great minds, it would be preferable and more correct in most instances to classify them as slow and facile. The latter are certainly more brilliant and stimulating — there is no substitute for them in conversation, oratory, and journalism, that is, in all lines of work where time is a decisive factor. However, in scientific undertakings the slow prove to be as useful as the fast because scientists like artists are judged by the **quality of what they produce**, **not** by the **speed of production**.*

References

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, Wiley.

Birnbaum, A. (1962), "On the foundations of statistical inference," *Journal of the American Statistical Association*, 57, 269–306.

Cajal, S. R. Y. (1999), *Advice for a young investigator*, MIT Press.

Casella, G., and Berger, R. L. (2002), *Statistical Inference*, Duxbury.

Davison, A. C. (2001), "Biometrika centenary: theory and general methodology," *Biometrika*, 88, 13–52.

Davison, A. C. (2003), *Statistical Models*, Cambridge University Press.

Keener, R. W. (2010), *Theoretical Statistics*, Springer.

Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation, Second Edition*, Springer.

McCullagh, P., and Nelder, J. A. (1989), *Generalized linear models*, Chapman & Hall/CRC.

Pace, L., and Salvan, A. (1997), *Principles of statistical inference from a Neo-Fisherian perspective*, Advanced series on statistical science and applied probability, World Scientific.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.