AMS 132, Winter 2017: Classical and Bayesian Inference

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ucsc.edu
www.ams.ucsc.edu/~draper

CLASS WEB PAGE:
ams132-winter17-01.courses.soe.ucsc.edu

Lecture Notes, Part 1 (Foundations) [Version 1]
Example (Krnjajić, Kottas, Draper [KKD] 2008): In-home geriatric assessment (IHGA). In an experiment conducted in the 1980s (Hendriksen et al. 1984), 572 elderly people, representative of $\mathcal{P} = \{\text{all non-institutionalized elderly people in Denmark}\}$, were randomized, 287 to a control ($C$) group (who received standard health care) and 285 to a treatment ($T$) group (who received standard care plus IHGA: a kind of preventive medicine in which each person’s medical and social needs were assessed and acted upon individually).

One important outcome was the number of hospitalizations during the two-year life of the study:

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Hospitalizations</th>
<th>$n$</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>$n_C^0$ $n_C^1$ $\ldots$ $n_C^k$</td>
<td>$n_C = 287$</td>
<td>$\tilde{y}_C$</td>
<td>$s_C$</td>
</tr>
<tr>
<td>Treatment</td>
<td>$n_T^0$ $n_T^1$ $\ldots$ $n_T^k$</td>
<td>$n_T = 285$</td>
<td>$\tilde{y}_T$</td>
<td>$s_T$</td>
</tr>
</tbody>
</table>

Let $\mu_C$ and $\mu_T$ be the mean hospitalization rates (per two years) in $\mathcal{P}$ under the $C$ and $T$ conditions, respectively.

Here are four statistical questions that arose from this study:
Q1: Was the mean number of hospitalizations per two years in the IHGA group different from that in control by an amount that was large in practical terms? \[ \text{description involving } \left( \overline{y}_T - \overline{y}_C \right) \]

Q2: Did IHGA (causally) change the mean number of hospitalizations per two years by an amount that was large in statistical terms? \[ \text{inference about } \left( \frac{\mu_T - \mu_C}{\mu_C} \right) \]

Q3: On the basis of this study, how accurately can you predict the total decrease in hospitalizations over a period of \( N \) years if IHGA were implemented throughout Denmark? \[ \text{prediction} \]

Q4: On the basis of this study, is the decision to implement IHGA throughout Denmark optimal from a cost-benefit point of view? \[ \text{decision-making} \]

These questions encompass almost all of the discipline of statistics: describing a data set \( D \), generalizing outward inferentially from \( D \), predicting new data \( D^* \), and helping people make decisions in the presence of uncertainty (I include sampling/experimental design under decision-making; omitted: data quality assurance (QA), ...).
1. (definition) **Statistics** is the study of **uncertainty**: how to measure it well, and how to **make good choices** in the face of it.

2. (definition) **Uncertainty** is a state of incomplete **information** about something of interest to **You** (Good, 1950: a **generic person** wishing to **reason sensibly** in the presence of **uncertainty**).

3. (axiom) (Your uncertainty about) "Something of interest to You" can always be **expressed** in terms of **propositions**: true/false statements $A, B, \ldots$

   **Examples**: You may be **uncertain** about the **truth status** of

   - $A = \text{(Hillary Clinton will be elected U.S. President in 2016)}$, or
   - $B = \text{(the in-hospital mortality rate for patients at hospital } H \text{ admitted in calendar 2010 with a principal diagnosis of heart attack was between 5\% and 25\%)}$.

4. (implication) It follows from 1–3 that **statistics** concerns **Your information** (NOT Your beliefs) about $A, B, \ldots$
5 (axiom) But Your information cannot be assessed in a vacuum: all such assessments must be made relative to (conditional on) Your background assumptions and judgments about how the world works vis à vis A, B, ... .

6 (axiom) These assumptions and judgments, which are themselves a form of information, can always be expressed in a set $B$ of background propositions, all of which You believe to be true.

Examples of $B$:

- In the IHGA study, based on the experimental design, $B$ would include the propositions
  
  (Subjects were representative of [like a random sample from] $P$),

  (Subjects were randomized into one of two groups, treatment (standard care + IHGA) or control (standard care)).

7 (definition) Call the “something of interest to You” $\theta$; in applications $\theta$ is often a vector (or matrix, or array) of real numbers, but in principle it could be almost anything (a function,
Axiomatization (continued)

an image of the surface of Mars, a phylogenetic tree, ...).

IHGA example: \( \theta = \text{mean relative decrease}\left(\frac{\mu_T - \mu_C}{\mu_C}\right) \) in hospitalization rate in \( P \).

8 (axiom) There will typically be an information source (data set) \( D \) that You judge to be relevant to decreasing Your uncertainty about \( \theta \); in applications \( D \) is often again a vector (or matrix, or array) of real numbers, but in principle it too could be almost anything (a movie, the words in a book, ...).

9 (implication) The presence of \( D \) creates a dichotomy:

- Your information about \( \theta \) \{internal, external\} to \( D \).

(People often talk about a different dichotomy: Your information about \( \theta \) \{before, after\} \( D \) arrives (prior, posterior), but temporal considerations are actually irrelevant.)

10 (implication) It follows from 1–9 that statistics concerns itself principally with five things (omitted: description, data QA, ...):

1) Quantifying Your information about \( \theta \) internal to \( D \) (given \( B \)), and doing so well (this term is not yet defined);
(2) Quantifying Your information about $\theta$ external to $D$ (given $B$), and doing so well;

(3) Combining these two information sources (and doing so well) to create a summary of Your uncertainty about $\theta$ (given $B$) that includes all available information You judge to be relevant (this is inference);

and using all Your information about $\theta$ (given $B$) to make

(4) Predictions about future data values $D^*$ and

(5) Decisions about how to act sensibly, even though Your information about $\theta$ may be incomplete.

**Foundational question:** How should these tasks be accomplished?

This question has **two parts**: probability and statistics.

The probability foundations (addressed here first), have an interesting and unfortunate history, in which much of the 20th century will (in my view) be seen in the 21st century to have been a series of missed scientific opportunities.
From the 1650s (Fermat, Pascal) through the 18th century (Bayes, Laplace) to the period 1860–1930 (Venn, Boole, von Mises), three different approaches for how to think about uncertainty quantification — classical, Bayesian, and frequentist probability — were put forward in an intuitive way, but no one ever tried to prove a theorem of the form \{given these premises, there’s only one sensible way to quantify uncertainty\} until Kolmogorov, de Finetti, and RT Cox.

— Kolmogorov (1933): following (and rigorizing) Venn, Boole and von Mises, probability is a function on (possibly some of) the subsets of a sample space $\Omega$ of uncertain possibilities, constrained to obey some reasonable axioms; this is excellent, as far as it goes, but many types of uncertainty cannot (uniquely, comfortably) be fit into this framework (examples follow).

Kolmogorov was trying to make precise the intuitive notion of repeatedly choosing a point at random in a Venn diagram and asking how frequently the point falls inside a specified set, i.e., his concept of probability had a repeated-sampling, frequentist character:
Frequentist Probability: Kolmogorov

“The basis for the applicability of the results of the mathematical theory of probability to real ‘random phenomena’ must depend on some form of the frequency concept of probability, the unavoidable nature of which has been established by von Mises in a spirited manner.”

∗ Example: You’re about to roll a pair of dice and You regard this dice-rolling as fair, by which You mean that (in Your judgment) all \(6^2 = 36\) elemental outcomes in \(\Omega = \{(1,1), (1,2), \ldots, (6,6)\}\) are equally probable; then the Kolmogorov probability of snake eyes \((1,1)\) exists and is unique (from Your fairness judgment), namely \(\frac{1}{36}\); but

∗ Example: You’re a doctor; a new patient presents saying that he may be HIV positive; what’s the Kolmogorov probability that he is?

What’s \(\Omega\)? This patient is not the result of a uniquely-specifiable repeatable “random” process, he’s just a guy who walked into Your doctor’s office, and — throughout the repetitions of whatever repeatable phenomenon anyone might imagine — his HIV status is not fluctuating “randomly”: he’s either HIV positive or he’s not.
The closest you can come to making Kolmogorov’s approach work here is to imagine the set $\Omega$ of all people \{similar to this patient in all relevant ways\} and ask how often you’d get an HIV-positive person if you repeatedly chose one person at random from $\Omega$, but to make this operational you have to specify what you mean by “similar to, in all relevant ways,” and if you try to do this you’ll notice that it’s not possible to do so uniquely (in such a way that all other reasonable people would unanimously agree with you).

— de Finetti (1937): rigorizing Bayes, probability is a quantification of betting odds about the truth of a proposition, constrained to obey axioms guaranteeing coherence (absence of internal contradictions); this is more general than Kolmogorov — in fact, it’s as general as you can get: any statement about sets can be expressed in terms of propositions — but betting odds are not fundamental to science.

De Finetti made many important contributions — in particular, his concept of exchangeability is crucial in Bayesian modeling — but (in my view) science is about information, not betting.
RT Cox (1946): following Laplace, probability is a quantification of information about the truth of one or more propositions, constrained to obey axioms guaranteeing internal logical consistency; this is both fundamental to science and as general as You can get.

Cox’s goal was to identify what basic rules $pl(A|B)$ — the plausibility (weight of evidence in favor) of (the truth of) $A$ given $B$ — should follow so that $pl(A|B)$ behaves sensibly, where $A$ and $B$ are propositions with $B$ assumed by You to be true and the truth status of $A$ unknown to You.

He did this by identifying a set of principles making operational the word “sensible” (Jaynes, 2003):

- Suppose You’re willing to represent degrees of plausibility by real numbers (i.e., $pl(A|B)$ is a function from propositions $A$ and $B$ to $\mathbb{R}$);

- You insist that Your reasoning be logically consistent:

  — If a plausibility assessment can be arrived at in more than one way, then every possible way must lead to the same value.
Cox’s Principles and Axioms

— You always take into account all of the evidence You judge to be relevant to the plausibility assessment under consideration (this is the Bayesian version of objectivity).

— You always represent equivalent states of information by equivalent plausibility assignments.

From these principles Cox derived a set of axioms:

• The plausibility of a proposition determines the plausibility of the proposition’s negation; each decreases as the other increases.

• The plausibility of the conjunction $AB = (A$ and $B)$ of two propositions $A, B$ depends only on the plausibility of $B$ and that of $\{A$ given that $B$ is true$\}$ (or equivalently the plausibility of $A$ and that of $\{B$ given that $A$ is true$\}$).

• Suppose $AB$ is equivalent to $CD$; then if You acquire new information $A$ and later acquire further new information $B$, and update all plausibilities each time, the updated plausibilities will be the same as if You had first acquired new information $C$ and then acquired further new information $D$. 
Cox’s Theorem

From these axioms Cox proved a theorem showing that uncertainty quantification about propositions behaves in one and only one way:

**Theorem:** If you accept Cox’s axioms, then to be logically consistent you must quantify uncertainty as follows:

- Your **plausibility operator** $pl(A|B)$ — for propositions $A$ and $B$ — can be referred to as your **probability** $P(A|B)$ that $A$ is true, given that you regard $B$ as true, and $0 \leq P(A|B) \leq 1$, with **certain truth** of $A$ (given $B$) represented by 1 and **certain falsehood** by 0.

- **(normalization)** $P(A|B) + P(\overline{A}|B) = 1$, where $\overline{A} = (\text{not } A)$.

- **(the product rule):**

  $$P(A B|C) = P(A|C) \cdot P(B|A C) = P(B|C) \cdot P(A|B C).$$

The proof (see, e.g., Jaynes (2003)) involves deriving two **functional equations** $F[F(x, y), z] = F[x, F(y, z)]$ and $x S \left[ \frac{S(y)}{x} \right] = y S \left[ \frac{S(x)}{y} \right]$ that $pl(A|B)$ must satisfy and then solving those equations.

A number of important corollaries arise from Cox’s Theorem:
Optimal Reasoning Under Uncertainty

• (the sum rule):
  \[ P(A \text{ or } B|C) \equiv P(A + B|C) = P(A|C) + P(B|C) - P(AB|C). \]

• Extensions of the product and sum rules to an arbitrary finite number of propositions are easy, e.g.,
  \[ P(ABC|D) = P(A|D) \cdot P(B|AD) \cdot P(C|ABD) \]
  \[ P(A + B + C|D) = P(A|D) + P(B|D) + P(C|D) - P(AB|D) - P(AC|D) - P(BC|D) + P(ABC|D). \]

• This framework (obviously) covers optimal reasoning about uncertain quantities \( \theta \) taking on a finite number of possible values; less obviously, it also handles (equally well) situations in which the set \( \Theta \) of possible values of \( \theta \) has infinitely many elements.

— Example: You’re studying quality of care at the 17 Kaiser Permanente (KP) northern California hospitals in 2003–7, before the era of electronic medical records; during that time there was a population \( \mathcal{P} \) of \( N = 8,561 \) patients at these facilities with a primary admission diagnosis of heart attack.
You take a simple random sample of \( n = 112 \) of these admissions and record whether or not each patient had an unplanned transfer to the intensive care unit (ICU), observing \( s = 4 \) who did; \( \theta \) is the proportion of such unplanned transfers in all of \( \mathcal{P} \); here \( \Theta = \left\{ \frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N} \right\} \), which can be conveniently approximated by \( \Theta' = [0, 1] \).

Prior to 2003, the proportion of such unplanned transfers for heart attack patients at KP in the northern California region was about \( q = 0.07 \), so interest focuses on \( P(A|DB) \), where \( A \) is the proposition \( (\theta \leq q) \), \( D \) is the proposition \( (s = 4) \), and \( B \) includes (among other things) details about the sampling experiment (e.g., \( (n = 112) \)).

In this setup \( \theta \) is usually called a (population) parameter, and is not itself the result of any sampling experiment (random or otherwise); for this reason, it’s not possible to (directly) quantify uncertainty about \( \theta \) from the Kolmogorov (set-theoretic) point of view, but it makes perfect sense to do so from the RT Cox (propositional) point of view.
You could now more generally define a function
\[ F(\theta|D\mathcal{B})(q) = P(\theta \leq q|D\mathcal{B}) \]
and call it the cumulative distribution function (CDF) for (not of) \((\theta|D\mathcal{B})\), which is shorthand for the CDF for Your uncertainty about \(\theta\) given \(D\) and \(\mathcal{B}\).

If \(F(\theta|D\mathcal{B})(q)\) turns out to be continuous and differentiable in \(q\) (I haven’t said yet how to calculate \(F\)), it will be convenient to write
\[
F(\theta|D\mathcal{B})(b) - F(\theta|D\mathcal{B})(a) = P(a < \theta \leq b|D\mathcal{B}) = \int_{a}^{b} p(\theta|D\mathcal{B})(q) \, dq, \tag{1}
\]
where the (partial) derivative \(p(\theta|D\mathcal{B})(q)\) of \(F(\theta|D\mathcal{B})\) with respect to \(q\) can be called the density for (not of) (Your uncertainty about) \(\theta\) given \(D\) and \(\mathcal{B}\).

In a small abuse of notation it’s common to write \(F(\theta|D\mathcal{B})\) and \(p(\theta|D\mathcal{B})\) instead of \(F(\theta|D\mathcal{B})(q)\) and \(p(\theta|D\mathcal{B})(q)\) (respectively), letting the argument \(\theta\) of \(F(\cdot|D\mathcal{B})\) and \(p(\cdot|D\mathcal{B})\) serve as a reminder of the uncertain quantity in question.
In the Kolmogorov approach a random variable $X$ is a function from $\Omega$ to some outcome space $O$, and if $O = \mathbb{R}$ you’ll often find it useful to summarize $X$’s behavior through the CDF of $X$: $F_X(x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$, usually written in propositional-style shorthand as $F_X(x) = P(X \leq x)$.

In the RT Cox approach, there are no random variables; there are uncertain things $\theta$ whose uncertainty (when $\Theta = \mathbb{R}^k$, for integer $1 \leq k < \infty$) can usefully be summarized with CDFs and densities.

Jaynes (2003) makes a worthwhile distinction: the statements

There is noise in the room.  

The room is noisy.

seem quite similar but are in fact quite different: the former is ontological (asserting the physical existence of something), whereas the latter is epistemological (expressing the personal perception of the individual making the statement).

Talking about “the density of $\theta$” would be to confuse ontology and epistemology;
The Mind-Projection Fallacy

Jaynes calls this confusion of \{the world\} (ontology) with \{Your uncertainty about the world\} (epistemology) the mind-projection fallacy, and it’s clearly a mistake worth avoiding.

Returning to the corollaries of Cox’s Theorem,

- Given the set $B$, of propositions summarizing Your background assumptions and judgments about how the world works as far as $\theta$, $D$ and future data $D^*$ are concerned:

  (a) It’s natural (and indeed You must be prepared in this approach) to specify two conditional probability distributions:

  — $p(\theta|B)$, to quantify all information about $\theta$ external to $D$ that You judge relevant; and

  — $p(D|\theta B)$, to quantify Your predictive uncertainty, given $\theta$, about the data set $D$ before it’s arrived.

  (b) Given the distributions in (a), the distribution $p(\theta|DB)$ quantifies all relevant information about $\theta$, both internal and external to $D$, and must be computed via Bayes’s Theorem:
Optimal Inference, Prediction and Decision

\[ p(\theta|DB) = c \ p(\theta|B) \ p(D|\theta \ B), \quad \text{(inference)} \quad (2) \]

where \( c > 0 \) is a normalizing constant chosen so that the left-hand side of (2) integrates (or sums) over \( \Theta \) to 1;

(c) Your predictive distribution \( p(D^*|DB) \) for future data \( D^* \) given the observed data set \( D \) must be expressible as follows:

\[ p(D^*|DB) = \int_{\Theta} p(D^*|\theta \ D \ B) \ p(\theta|DB) \ d\theta ; \]

often there’s no information about \( D^* \) contained in \( D \) if \( \theta \) is known, in which case this expression simplifies to

\[ p(D^*|DB) = \int_{\Theta} p(D^*|\theta \ B) \ p(\theta|DB) \ d\theta ; \quad \text{(prediction)} \quad (3) \]

(d) to make a sensible decision about which action a You should take in the face of Your uncertainty about \( \theta \), You must be prepared to specify (i) the set \( (A|B) \) of feasible actions among which You’re choosing, and
(ii) a utility function $U(a, \theta|B)$, taking values on $\mathbb{R}$ and quantifying Your judgments about the rewards (monetary or otherwise) that would ensue if You chose action $a$ and the unknown actually took the value $\theta$; without loss of generality You can take large values of $U(a, \theta|B)$ to be better than small values;

then the optimal decision is to choose the action $a^*$ that maximizes the expectation of $U(a, \theta|B)$ over $p(\theta|DB)$:

$$a^* = \arg\max_{a \in (A|B)} E_{(\theta|DB)} U(a, \theta|B) = \arg\max_{a \in (A|B)} \int_{\Theta} U(a, \theta|B) p(\theta|DB) \, d\theta . \quad (4)$$

The equation solving the inference problem is traditionally attributed to Bayes (1764), although it’s just an application of the product rule (page 13), which was already in use by (James) Bernoulli and de Moivre around 1715, and Laplace made much better use of this equation from 1774 to 1827 than Bayes did in 1764; nevertheless the Laplace/Cox propositional approach is typically referred to as Bayesian reasoning (the subject of AMS 206 or 206B).
Cox’s Theorem is equivalent to the assertion

If You wish to quantify Your uncertainty about an unknown $\theta$ (and make predictions and decisions in the presence of that uncertainty) in a logically internally consistent manner (as specified through Cox’s axioms), on the basis of data $D$ and background assumptions/judgments $B$, then You can achieve this goal with Bayesian reasoning, by specifying $p(\theta|B)$, $p(D|\theta B)$, and $\{(A|B), U(a, \theta|B)\}$ and using equations (2–4).

This assertion has not rendered Bayesian analyses ubiquitous, although the value of Bayesian reasoning has become increasingly clear to an increasingly large number of people in the last 20 years, now that advances in computing have made the routine use of equations (2–4) feasible.

Advantages include a unified probabilistic framework: e.g., in my earlier ICU example, Kolmogorov’s non-Bayesian approach does not permit direct probability statements about a population parameter, but Cox’s Theorem permits You to make such statements (summarizing all relevant available information) in a natural way.
It’s **worth noting**, however, that **there really is a theorem here**, of the form \( A \rightarrow B \), from which \( \overline{B} \rightarrow \overline{A} \); this **comes close to the assertion**

If You employ **non-Bayesian reasoning** then You’re **open to the possibility of logical inconsistency**, and indeed there have been some **embarrassing moments** in **non-Bayesian inference** over the past **100 years** (e.g., **negative estimates** for quantities that are **constrained to be non-negative**).

**Challenges:** These **corollaries** to **Cox’s theorem** solve problems (3–5) above (page 7) — they leave **no ambiguity** about how to draw **inferences**, and make **predictions** and **decisions**, in the presence of **uncertainty** — but problems (1) and (2) are still **unaddressed**: to implement this **logically-consistent approach** in a given application, You have to **specify**

- \( p(\theta|B) \), usually called Your **prior information** about \( \theta \) (given \( B \); this is **better understood** as a **summary of all relevant information** about \( \theta \) **external** to \( D \), rather than by appeal to any **temporal** (before-after) considerations);
The Specification Burden (continued)

- $p(D|\theta \mathcal{B})$, often referred to as Your **sampling distribution** for $D$ given $\theta$ (and $\mathcal{B}$; this is **better understood** as Your **conditional predictive distribution** for $D$ given $\theta$, before $D$ has been **observed**, rather than by appeal to other data sets that might have been observed); and

- Your **action space** ($\mathcal{A}|\mathcal{B}$) and the **utility function** $U(a, \theta|\mathcal{B})$ for decision-making purposes.

The results of **implementing** this approach are

- $p(\theta|D \mathcal{B})$, often referred to as Your **posterior** distribution for $\theta$ given $D$ (and $\mathcal{B}$; as above, this is **better understood** as the **totality of Your current information** about $\theta$, again without appeal to temporal considerations);

- Your **posterior predictive distribution** $p(D^*|D \mathcal{B})$ for future data $D^*$ given the **observed data set** $D$; and

- the **optimal decision** $a^*$ given all **available information** (and $\mathcal{B}$).
To summarize: Inference and prediction require You to specify \( p(\theta|B) \) and \( p(D|\theta \ B) \); decision-making requires You to specify the same two ingredients plus \( (A|B) \) and \( U(a, \theta|B) \); how should this be done in a sensible way?

Cox’s Theorem and its corollaries provide no constraints on the specification process, apart from the requirement that all probability distributions be proper (integrate or sum to 1).

In my view, in seeking answers to these specification questions, as a profession we’re approximately where the discipline of statistics was in arriving at an optimal theory of probability before Cox’s work: many people have made ad-hoc suggestions (some of them good), but little formal progress has been made.

Developing (1) principles, (2) axioms and (3) theorems about optimal specification could be regarded as creating a Theory of Applied Statistics, which we need but do not yet have.

This is one of my main areas of research, and I’m looking for Ph.D. students to work with me on it.
Curiously, the Bayesian approach to statistical inference, prediction and decision-making was not the dominant paradigm in the 20th century; the Kolmogorov (frequentist) approach — also known as the classical approach — was employed by most people last century (and many still in this century).

This fact is curious because (as we’ll see) the frequentist approach has many difficulties and drawbacks associated with it, and (in my view) falls dramatically short of providing a comprehensive solution to problems of

— **inference:** to the extent that the frequentist story has a strong point, this is it, but (as we’ll see) even in inference the frequentist approach is subject to paradoxes, contradictions and failures of interpretation;

— **prediction:** as we’ll see, the frequentist approach can often provide single-number predictions of future data values, but it can be quite difficult in this approach to fully assess the uncertainty of these single-number predictions; and
Abraham Wald, a prominent frequentist, proved a theorem in 1950 that informally says that all good decisions are Bayes decisions, and from that moment forward there was no frequentist narrative on decision-making; I won’t talk about this topic much in 206/132.

So why did so many people use the frequentist approach in the 20th century, and why are many people still using it in the 21st century?

Answering this question requires examining the history of probability and statistics, which is our next topic.