Getting to know your probabilities:

Three ways to frame personal probabilities for decision making.

Teddy Seidenfeld – CMU

An old, wise, and widely held attitude in Statistics is that modest intervention in the design of an experiment followed by simple statistical analysis may yield much more of value than using very sophisticated statistical analysis on a poorly designed existing data set.

In this sense, good inductive learning is active and forward looking, not passive and focused exclusively on analyzing what is already given.

In this talk I review three different approaches for how a decision maker might actively frame her/his *probability space* rather than being passive in that phase of decision making.

***Method 1*: Assess precise/determinate probabilities only for the set of random variables that define the decision problem at hand. Do not include other "nuisance" variables in the space of possibilities. In this sense, over-refining the space of possibilities may make assessing probabilities infeasible for good decision making.**

***Example 1.1*:**

Random sampling: the "nuisance" of individual tags and designing *an experiment to prove*. (K-S, 1990).

***Example 1.2*:**

Juhl's (1993) *incompleteness* for formal learning with computable Bayesian methods.

Example 1.1

- **Simple Random Sampling – informal version.**

Design *an experiment to prove* to a general readership what is the percentage k_Z in a large population (> 10^6) that bear property *Z*.

- **A familiar approach is to use overt randomization to select a sample (using random-numbers) and to perform routine statistical inference on the observed *z*-values in the sample.**

For instance, with a sample of 100 randomly selected individuals from the population, the probability is at least .95 that the percentage of *Z* in the sample, $\bar{z}$, differs from k_Z by no more than 10%.

$$P(\,|k_Z - \bar{z}| \leq .10\,) \geq .95 \qquad \text{(approximately)}$$

However, in order to apply overt randomization, in order to use random numbers to sample the population, the individuals require tags

$$t_i \ (i = 1, \ldots, 10^6).$$

Then a straightforward formalization of the probability space for the inference about the percent of Z in the population, k_Z, has as the sample space for the data the 100 pairs

$$\{(z_j, t_j): j = j_1, \ldots, j_{100}\}$$

where the j's are the 100 randomly selected numbers.

However, unless the tags are irrelevant about Z,

$$P(|k_Z - \bar{z}| \leq .10) \neq P(|k_Z - \bar{z}| \leq .10 \mid \{t_{j1}, \ldots, t_{j100}\}).$$

For example, let the tags be individual Social Security numbers, which reveal considerable information about, e.g., age and gender. Then the tags introduce "nuisance" parameters into the statistical reasoning.

If, e.g., Latanya Sweeney (2006) is among the readership of your publication, the familiar statistical inference based on overt randomization will no longer be compelling for her once the tags for the sampled individuals are revealed.
BUT – the clever statistician can be careful to include the *z*-values but NOT to include the tags in the sample space for probabilistic analysis.

I.J.Good (1971, #679) notes that sometimes a Bayesian can make sense of a Classical Statistical procedure by avoiding parts of the data, employing what he calls a *Statistician's Stooge*.

I.Levi (1980, chapter 17) makes a similar distinction between

***data as evidence* and *data as input*!**

***Example 1.2*: Juhl's (1993) *incompleteness* for formal learning with computable Bayesian methods.**

Let *T* be a recursively enumerable but not recursive set of integers, e.g., the Godel-numbers of theorems of a particular first order theory. The formal learning problem is to decide whether an integer *k* belongs to *T* or not relative to a "data stream" $\{d_i\}$ of the elements of T.

The challenge Juhl sets for Bayesian theory is to construct a straightforward probability analysis where, e.g., the (posterior) probability for the event E_k: $k \in T$, given the growing data stream $\{d_i\}$, converges to the truth value of E_k.

$$\lim_{m\to\infty} \mathrm{Prob}(E_k \mid d_1, \ldots, d_m\} = \text{indicator for } E_k.$$

There are two familiar but significant impediments that block a straightforward Bayesian solution of the kind Juhl requests.

(1) Given ordinary mathematical background knowledge, in each measure space the random variable E_k is a constant – either it is 1 (if $k \in T$) or it is 0 (if $k \notin T$). So, a coherent $P(\bullet)$, has $P(E_k) = 1$, or $P(E_k) = 0$, respectively.

(2) But as set T is *re* and not recursive – theoremhood is undecidable the coherent probability from (1) is not computable.

This leads Juhl (1993) to conclude:

COROLLARY 1. There exist problems solvable by a recursive method but that no computable coherent Bayesian can possibly solve.

Aside: **The problem is *solvable* by positing "$k \notin T$" and changing to "$k \in T$" if and only if *k* appears among the data stream $\{d_1, \ldots, d_m, \ldots\}$.**

However, the computable Bayesian decision maker faced with this formal learning problem can solve the problem by taking charge of the measure space over which probability is defined.

(Counter) Example 1.2^{+}.

Let *X* be an integer random variable. Partially define the probability distribution for *X* as follows:

- **$P(\ X = d_m \mid X \in T\ ) = 2^{-m}$ Given that $X \in T$, let $P(X = d_m) = 2^{-m}$.**
- **$P(X \in T) = .4$. Unconditionally, $P(X \in T) < P(X \notin T)$.**

The *Statistician's Stooge* knows that *X* = k, but that is <u>*not*</u> part of the Statistician's evidence. The *Stooge* checks whether $X = d_m$ or not and reports just that fact to the Statistician as the evidence d_m.

Then $\lim_{m\to\infty} \mathrm{Prob}(\ X \in T \mid d_1, \ldots, d_m)$

is a coherent, computable Bayesian solution to the learning problem.

***Method 1* for getting to know your probabilities is to avoid including more in the sample space than is required for *robust* inference – inference free of *nuisance parameters*: about which there may be conflicted personal opinions or infeasible computations, and about which the experiment may be silent.**

- **In example 1.1, overt random sampling, the key to constructing the measure space is to avoid including the tags in the sample space.**
- **In example 1.2/1.2^{+}, Juhl's formal learning problem for an *re* set, the key to constructing the measure space is to avoid including the (name of) the number tested in the sample space.**

In both examples, the statistician restricts the measure space to a proper subset of the "input space" used to solve the problem!

***Method 2*: With respect to a particular decision problem, choose wisely the set of events $\mathcal{E}$ that you can assess with probabilities.**

Coherence (as in de Finetti's theory) requires that you extend these probabilities to the *linear span* generated by $\mathcal{E}$, which may be a smaller and simpler set than the Boolean algebra generated by $\mathcal{E}$.

If $\mathcal{E}$ is wisely chosen, the decision problem at hand may be solved by the assessments over the smaller space.

Let us review de Finetti's (1974) two related theorems.

***Coherence* – de Finetti's notion of (2-sided) coherence begins with**

a partition of states, $\Omega = \{\omega_i : i \in I\}$, and

a collection of real-valued variables, $\chi = \{X_j : j \in J\}$, defined on Ω.

***Note*: In some settings de Finetti begins instead with logical variables and forms a partition of "constituents," i.e., the smallest common refinement under which all the variables are measureable.**

For each random variable $X \in \chi$, the rational agent, <u>the bookie</u>, has a *prevision* P(X) which is to be interpreted as a *fair* price both for *buying* and *selling* units of X.

Hence, the prevision is *2-sided* price.

For real $\beta > 0$ small enough the bookie enters the market with a fair buy/sell price P(X) for X.

The bookie

- **is willing to pay βP(X) in order *to buy* βX in return.**

and,

- **is willing to accept βP(X) in order *to sell* βX in return.**

In symbols, the bookie will accept the gamble

$$\beta[X - P(X)]$$

as a change in fortune, for all sufficiently small, positive or negative β.

That is, for all finite n and all small, real valued $\beta_1, \ldots, \beta_n$ and all $X_1, \ldots, X_n \in \chi$, the bookie will accept the combination of gambles

$$\sum_{i=1}^{n} \beta_i[X_i - P(X_i)].$$

The $|\beta_i|$ must be small enough so that, also, the bookie is prepared to accept all *__finite__* sums of gambles of the preceding form.

For positive β_i, the bookie buys β_i-units of X_i for a price of $\beta_i P(X_i)$.

For negative β_i,, the bookie sells β_i-units of X_i for a price of $\beta_i P(X_i)$.

Thus, the combinations of the bookie's "fair" previsions cover the *span* (= all finite linear combinations) of the variables in χ.

Consider an opponent to the bookie, <u>the gambler</u>, who may select which of the bookie's fair contracts to accept. The bookie's previsions are *incoherent* if there is a *uniformly negative <u>finite</u>* combination of gambles that are acceptable to the bookie.

That is, the bookie's previsions are *incoherent* if the gambler can choose finitely many non-zero β_i where $\varepsilon > 0$ and for each $\omega \in \Omega$,

$$\sum_{i=1}^{n} \beta_i[X_i(\omega) - P(X_i)] < -\varepsilon.$$

Otherwise the bookie's previsions are *<u>coherent</u>*.

- **Where previsions are incoherent, the book that indicates this constitutes a combination of gambles uniformly, strictly dominated by *not-betting* (= 0).**

Notes:

- **Checking for coherence/incoherence of a finite set of previsions is a linear programming problem.**

- **When a set of previsions are incoherent, many different "books" may be constructed. These may indexed in different ways in order to quantify a degree of incoherence. -- See, e.g., SSK (2003) *Measures of Incoherence*.**

- **We return to this theme with Method 3.**

1st de Finetti *Theorem for Coherent Previsions.*

- **A set of previsions are coherent if and only if they are the expected values for the respective random variables under a (finitely additive) probability distribution over Ω.**

- **When the variables are indicator functions for events (subsets of Ω), coherent previsions are exactly those in agreement with a (finitely additive) probability. And then the $|\beta_i|$ are the stakes in winner-take-all bets, where the previsions fix betting rates,**

$$P(X_i) : 1\text{-}P(X_i).$$

Method 2 for getting to know your probabilities

2^{nd} de Finetti theorem: The *Fundamental Theorem of Previsions*

Suppose *coherent* previsions are given for all variables in a set χ defined with respect to Ω.

Let Y be a real-valued function defined on Ω but possibly not in χ.

Define: $\underline{A} = \{X: X(\omega) \leq Y(\omega)$ **and X is in the span of** $\chi\}$

$\overline{A} = \{X: X(\omega) \geq Y(\omega)$ **and X is in the span of** $\chi\}$

Let

$$\underline{P}(Y) = \sup_{X \in \underline{A}} P(X) \quad \text{and} \quad \overline{P}(Y) = \inf_{X \in \overline{A}} P(X)$$

***Aside*: Think of inner/outer measure fixed by the set χ.**

<u>Then</u> the prevision, P(Y), may be any finite number from $\underline{P}$ (Y) to $\overline{P}$ (Y) and the resulting enlarged set of previsions is coherent.

Outside this interval, the enlarged set of previsions are incoherent.

That is,

The interval for a new prevision [$\underline{P}$ (Y) $\overline{P}$ (Y)] given by the *Fundamental Theorem* constrains a new prevision for a variable, Y$\notin\chi$, while preserving coherence of the previsions already assigned to X$\in\chi$.

***Aside*: This is an instance of Imprecise Probabilities – IP theory**

***Advertisement*: See www.sipta.org !**

Thus, coherence for previsions does <u>not</u> require the rational bookie to identify precise previsions beyond those in the span of the variables in the set χ.

Specifically, the rational agent is not required by *coherence* to have determinate probabilities defined on an algebra of events, let alone on a power-set of events.

It is sufficient to have probabilities defined *<u>as-needed</u>* for the arbitrary set χ, as might arise in a particular decision problem.

- **See, e.g., F. Lad, 1996 for interesting applications of this result.**

Toy Example (Example 2.1)

Ω = {1, 2, 3, 4, 5, 6} the outcome of rolling an 6-sided die.

χ is the set of indicator functions for the following four events

$$\chi = \{ \{1\}, \{3,6\}, \{1,2,3\}, \{1,2,4\} \}$$

Suppose previsions for these four events are given, and agree with the judgment that the die is "fair."

$$P(\{1\}) = 1/6; \; P(\{3,6\}) = 1/3; \; P(\{1,2,3\}) = P(\{1,2,4\}) = 1/2.$$

By the 1st of de Finetti's 2 results, these are coherent previsions.

The set of events for which a determinate prevision is fixed by the previsions for these four events is given by the *Fundamental Theorem*.

- **That set does *<u>not</u>* form an algebra. Only 22 of 64 events (11 pairs of complementary events) have precise previsions.**

For instance, by the Fundamental Theorem,

$$P(\{1,2,3\}) = 1/2$$

likewise $$P(\{1,2,4\}) = 1/2;$$

however, $$\underline{P}(\{1,2\}) = 1/6 < \overline{P}(\{1,2\}) = 1/2.$$

- **Moreover, the smallest algebra containing the 4 events in χ is the power set of all 64 events on Ω.**

***Method 3*: Your probabilistic assessments may be incoherent so that you may be exposed to a sure-loss in your decision making about some specific quantities.**
Nonetheless, you may be able to use familiar algorithms (e.g., Bayes' theorem) to update your views with new data and to improve your incoherent assessments about these quantities.

That is, you may be able to reduce your degree of incoherence about these quantities by active, Bayesian-styled learning. Specifically, by framing your probability space so that incoherence is concentrated in your "prior," you may use Bayesian algorithms to update to a less-incoherent "posterior."

Example 3.1: **How to wager from an incoherent position.**

***Aside*: In this section we restrict ourselves to previsions, rather than working with lower and upper previsions, in order to simplify the analysis of the *Gambler*'s optimal strategy.**

Let $\{E_1, \ldots, E_n\}$ form a partition, and let $0 \leq p(E_i) \leq 1$ be the Bookie's previsions for these n-many events.

- **Assume that no one of these previsions is incoherent, by itself.**

Let $\sum_{i=1}^{n} p(E_i) = q$. It might be that $q \neq 1$, so that the *Bookie*'s previsions are incoherent.

Next, the Bookie is called upon to set a price $p(X)$ for a new random variable X, measurable with respect to this partition, i.e., $$X = \sum_i x_i E_i, .$$

- **What can the Bookie do with the value of $p(X)$ to avoid increasing her/his measure of incoherence?**

Aside: See SSK (2003) for a family of indices of incoherence.

For notational ease, order the events so that $x_1 \leq x_2 \leq \ldots \leq x_n$.

Assume that $x_1 \leq p(X) \leq x_n$, so that by itself $p(X)$ is coherent.

Define $$\mu = \sum_i x_i p_i$$

You may think of μ as the *pseudo-expectation* for X with respect to the *Bookie*'s incoherent *distribution $P(\bullet)$* for the x_i.

Theorem for the ***rate of loss* – at what rate can the Gambler force the Bookie to lose for sure? (Recall: $\sum_{i=1}^{n} p(E_i) = q$.)**

The *Bookie* can avoid increasing the *rate of loss* with a prevision for *X* that satisfies the following conditions:

- **If $q < 1$, choose $p(X)$ to satisfy**

$$\mu + \frac{1-q}{n-1}\sum_{i=1}^{n-1} x_i \le p(X) \le \mu + \frac{1-q}{n-1}\sum_{i=2}^{n} x_i$$

- **If $q > 1$, choose $p(X)$ to satisfy**

$$max\{x_1,\ \mu - (q-1)x_n\} \le p(X) \le min\{x_n,\ \mu - (q-1)x_1\}$$

- **If $q = 1$, choose $p(X)$ to satisfy the Bayes solution**

$$\mu = p(X).$$

Theorem **for the** ***rate of gain*** **– at what rate can the Gambler force a rate of profit, for sure? (Recall:** $\sum_{i=1}^{n} p(E_i) = q$**.)**

The ***Bookie*** **can avoid increasing the** ***rate of gain*** **by setting a prevision for** ***X*** **as:**

Choose $p(X)$ **to satisfy**

$$\mu + (1\text{-}q)x_1 \leq p(X) \leq \mu + (1\text{-}q)x_n$$

<u>Corollary</u>**:** ***You don't have to be coherent to like Bayes' rule*****!**

Consider a ternary partition

$\{E_1, E_2, E_3\}$ **with previsions** $\{p_1, p_2, p_3\}$**.**

Let X be the called-off *r.v.* for the called-off wager on E_3 *vs* E_1.

$$\underline{E}_1 \qquad\qquad \underline{E}_2 \qquad\qquad \underline{E}_3$$

$$X(E_1) = 0, \quad X(E_2) = p(X), \quad X(E_3) = 1$$

Thus, $\alpha(X - p(X))$ has the respective payoffs:

$$-\alpha p(X) \qquad\qquad 0 \qquad\qquad \alpha(1 - p(X))$$

Then, e.g., with $q \leq 1$, the *Bookie* wants to satisfy the inequalities:

$$p_2 p(X) + p_3 \leq p(X) \leq p_2 p(X) + p_3 + (1-q)$$

If the *Bookie* uses a pseudo-Bayes value, the inequality is

***automatic*, as follows:**

$$p(X) = p(E_3 \,||\, \{E_1, E_3\}) = p_3/(p_1+p_3)$$

"as if" calculating $p(E_3 \mid \{E_1, E_3\})$

Hence, betting like a coherent Bayesian makes sense even if you are incoherent!

Summary – Three ways of getting to know your probabilities.

***Method 1*: Assess precise/determinate probabilities only for the set of random variables that define the decision problem at hand. Do not include other "nuisance" variables in the space of possibilities. In this sense, over-refining the space of possibilities may make assessing probabilities infeasible for good decision making.**

***Method 2*: With respect to a particular decision problem, choose wisely the set of events $\mathcal{E}$ that you can assess with probabilities. Coherence requires assessments over a linear span, which may be a much smaller set than the algebra (i.e., basic logic) of events for the same events.**

***Method 3*: Your probabilistic assessments may be incoherent so that you may be exposed to a sure-loss in your decision making about some specific quantities.**

Nonetheless, you may be able to use familiar algorithms (e.g., Bayes' theorem) to update your views with new data and to improve your incoherent assessments about these quantities.

- **You don’t have to be coherent to like Bayes’ Theorem!**

Selected References

de Finetti, B. (1974) *The Theory of Probability* (2 vols.) New York: Wiley.

Good, I.J. (1971) Twenty Seven Principles of Rationality. In V.P.Godambe and D.A.Sprott (eds.) *Foundations of Statistical Inference*. Holt, Reinhart, and Winston, Toronto: pp. 124-127.

Juhl, C. (1993) Bayesianism and Reliable Scientific Inquiry. *Philosophy of Science* 60: 302-319.

Kadane, J.B. and Seidenfeld, T. (1990) Randomization in a Bayesian Perspective. *J. Stat. Planning and Inference* 25: 329-345.

Lad, F. (1996) *Operational Subjective Statistical Methods*. Wiley: New York.

Levi, I. (1980) *The Enterprise of Knowledge*. MIT Press: Cambridge.

Scherivsh, M.J., Seidenfeld, T. and Kadane, J.B. (2003) Measures of Incoherence. In *Bayesian Statistics 7*. Bernardo, J.M. et al (eds.). Oxford Univ. Press: Oxford.

Sweeney, L. (2006) Protecting Job Seekers from Identity Theft. *IEEE Internet Computing* 10 (2).