Some Asymptotic Bayesian Inference
(background to Chapter 2 of Tanner's book)

Principal Topics

- The approach to *certainty* with increasing evidence.
- The approach to *consensus* for several agents, with increasing shared evidence.
- A role for *statistical models* in these asymptotic results.
  - symmetry/independence assumptions in these results.
  - data reduction
  - asymptotic Normal inference for these results.

Generalizing the ***coin-tossing*** **example** from last lecture:

Sample space of (observable) outcomes:

A 2-sided coin is repeatedly tossed, indefinitely,

$$X = <X_1, X_2, \ldots, X_n, \ldots>$$

$X_j = 0$, or $X_j = 1$ as the coin lands *tails up* or *heads up* on the j^{th} flip.

So that, $\boldsymbol{x} = < x_1, x_2, \ldots, x_n, \ldots>$ is a point of the space $\boldsymbol{\Omega} = \{0,1\}^{\aleph_0}$

Of course, at any one time we observe only a finite, initial segment.

The ***events*** that make up the σ-algebra, A, are the (smallest) σ-field of sets including all the (historical) observable events, of the form,

$$\boldsymbol{H}_n = < x_1, x_2, \ldots, x_n, \{0,1\}, \{0,1\}, \ldots.>$$

The Statistical Model:

Introduce a statistical quantity, a parameter θ, such that the events in A have a determinate conditional probability, given the parameter.

Bernoulli (i.i.d.) Coin flipping (continued):

$\mathbf{P}(X_j = 1 | \theta) = \theta$ ($j = 1, \ldots$), for $0 \leq \theta \leq 1$

$\mathbf{P}(\boldsymbol{H}_n \mid \theta) = \theta^k(1-\theta)^{n-k}$, where k of the first n coordinates of $\boldsymbol{H}_n$ are 1

and n-k of the first n coordinates $\boldsymbol{H}_n$ of are 0.

Now, if we are willing to make θ into a random variable (by expanding the σ-algebra accordingly), we can write Bayes theorem for the parameter:

$$\mathbf{P}(\theta \mid H_n) = \mathbf{P}(H_n \mid \theta)\, \mathbf{P}(\theta) / \mathbf{P}(H_n)$$

$$\propto \mathbf{P}(H_n \mid \theta)\, \mathbf{P}(\theta)$$

OR

The ***posterior probability for*** θ is proportional to

the product of the *likelihood for* θ and its *prior probability*.

With the conjugate ***Beta***(α , β) prior for θ

$$\mathbf{P}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\alpha, \beta > 0, 0 \leq \theta \leq 1,$$

the posterior distribution $\mathbf{P}(\theta \mid \boldsymbol{H}_n)$ is given by the distribution ***Beta***($\alpha+k$, $\beta+n-k$)

having *mean* $(\alpha+k) / (\alpha+k+\beta+n-k) = (\alpha+k) / (\alpha+\beta+n)$

and *variance* $(\alpha+k)(\beta+n-k) / (\alpha+\beta+n)^2(\alpha+\beta+n+1)$

Note: Here we may reduce the historical data of n-bits to two quantities (k, $n-k$). That is, the two likelihood functions: $\mathbf{P}(H_n \mid \theta)$ and $\mathbf{P}(k, n-k \mid \theta)$ are the same.

$$\mathbf{P}(H_n \mid \theta) = \mathbf{P}(k, n-k \mid \theta)$$

Now, by the *Strong Law of Large Numbers*: for each $\varepsilon > 0$ and given θ

$$\mathbf{P}(\lim_{n\to\infty} \left| k/n - \theta \right| < \varepsilon \mid \theta) = 1.$$

Hence, with probability 1, the sequence of posterior probabilities for θ

$$\lim_{n\to\infty} \mathbf{P}(\theta \mid \boldsymbol{H}_n) = \lim_{n\to\infty} \boldsymbol{Beta}(\alpha+n\theta, \beta+n(1-\theta))$$

have a limit distribution with *mean* θ and *variance* 0, independent of α and β.

Note: The posterior variances for θ are $O(1/n)$. That is, in advance, we can bound from below the precision (that is, bound from above the variance) of the posterior distribution for the parameter by choosing the sample size to observe.

Asymptotic Certainty

THUS, under each conjugate (Beta) prior:

With probability 1, the posterior probability for θ converges to the

(0-1) ***Delta*** distribution, concentrated on the *true* parameter value.

Aside: (Doob, 1949) if the parameter space is *finite-dimensional*, this almost sure convergence occurs for **each** value of the parameter in the *support* of the prior.

(The *support* of the prior in the parameter set is the smallest closed set with prior prob. 1)

From the perspective of the posterior probability for θ,

through the likelihood function, the data "swamp" the prior.

Asymptotic Consensus
(Merging of Posterior Probability Distributions)

As a metric (distance) between two distributions **P** and **Q** over the algebra A, consider a strict standard, *uniform distance*,

$$\rho(\mathbf{P},\mathbf{Q}) = sup_{E\in \mathsf{A}} \left| \mathbf{P}(E) - \mathbf{Q}(E) \right|$$

Let $\mathbf{P}^n = \mathbf{P}(\theta \mid \boldsymbol{H}_n)$ and $\mathbf{Q}^n = \mathbf{Q}(\theta \mid \boldsymbol{H}_n)$ (n = 1, 2, …) be two sequences of posterior probability distributions for the parameter θ based on two (conjugate) Beta priors.

Then, it is not hard to show that

$$\lim_{n\to\infty} \rho(\mathbf{P}^n, \mathbf{Q}^n) = \lim_{n\to\infty} sup_{\Theta} \left| \mathbf{P}(\theta \mid \boldsymbol{H}_n) - \mathbf{Q}(\theta \mid \boldsymbol{H}_n) \right| = 0.$$

In other words, the two systems of posterior probabilities *for the parameter*, based on shared evidence, merge together.

Question: What about posterior probability distributions over the algebra generated by the observable events, A?

- Recall that the *i.i.d.* Bernoulli statistical model for the data is shared between these two investigators: $(\forall E \in \mathsf{A})\ \mathbf{P}(E \mid \theta) = \mathbf{Q}(E \mid \theta)$.
- Also, with conjugate priors from the Beta family, the prior probability is positive for each "historical" event H_n. That is, $(\forall H_n, 0 < \theta < 1)\ \mathbf{P}(H_n \mid \theta) > 0$.

  Moreover, $\mathbf{P}(H_n) = \int_\Theta \mathbf{P}(H_n \mid \theta) d\mathbf{P}(\theta)$. Therefore, $\mathbf{P}(H_n) > 0$, and likewise $\mathbf{Q}(H_n) > 0$.

Answer: $\mathbf{P}(E \mid H_n) = \int_\Theta \mathbf{P}(E \mid \theta, H_n)\, d\mathbf{P}^n(\theta).$

$$= \int_\Theta [\mathbf{P}(E, H_n \mid \theta) / \mathbf{P}(H_n \mid \theta)]\, d\mathbf{P}^n(\theta)$$

and as $\mathbf{P}^n$ merges with $\mathbf{Q}^n$ for large n,

$$\approx \int_\Theta [\mathbf{P}(E, H_n \mid \theta) / \mathbf{P}(H_n \mid \theta)]\, d\mathbf{Q}^n(\theta)$$

and as the two investigators agree on the statistical model

$$= \int_\Theta [\mathbf{Q}(E, H_n \mid \theta) / \mathbf{Q}(H_n \mid \theta)]\, d\mathbf{Q}^n(\theta)$$

$$= \int_\Theta \mathbf{Q}(E \mid \theta, H_n)\, \mathbf{Q}(\theta \mid H_n) d\mathbf{Q}^n(\theta)$$

$$= \mathbf{Q}(E \mid H_n).$$

Thus, the two posterior *predictive* distributions (over A) also merge.

For example, the probability that the next flip lands heads given H_n is:

$$\mathbf{P}(X_{n+1} \mid H_n) \quad = \mathrm{E}_{\mathbf{P}n}[\theta] = (\alpha+k) / (\alpha+\beta+n),$$

which for large n, $\quad \approx \; k/n$

and by parallel reasoning

$$\approx \; \mathbf{Q}(X_{n+1} \mid H_n).$$

Note, that the agreement between $\mathbf{P}(E \mid H_n)$ and $\mathbf{Q}(E \mid H_n)$ takes a stronger form for cases when the historical observation H_n precludes E, when $(E \cap H_n) = \varnothing$.

Then, $\quad \mathbf{P}(E \mid H_n\text{'}) = \mathbf{Q}(E \mid H_n\text{'}) = 0$ for all $n\text{'} \geq n$.

Question: What parts of these asymptotic results for the algebra of events A depends upon the (shared) statistical model?

Answers:

(1) *Asymptotic Certainty* is automatic with the Bayesian framework!

$(\forall \boldsymbol{E} \in \mathsf{A})$ with **P**-probability 1,

$$\lim_{n\to\infty} \mathbf{P}^{\mathbf{n}}(E) = \chi(E), \text{ i.e.} \qquad \text{(Halmos, 1948)}$$

(2) *Asymptotic Consensus* requires only agreement on "null" events.

Assume that $(\forall \boldsymbol{E} \in \mathsf{A})$ $\mathbf{P}(E) = 0$ *if and only if* $\mathbf{Q}(E) = 0$.

With **P**-(or **Q**-) probability 1, with respect to A

$$\lim_{n\to\infty} \rho(\mathbf{P}^{\mathbf{n}}, \mathbf{Q}^{\mathbf{n}}) = 0. \qquad \text{(Blackwell \& Dubins, 1962)}$$

However, the statistical model **is** needed for each of the following:

(1) data reduction

(2) rates of convergence to certainty

(3) rates of merging for Bayesian investigators with shared evidence

Next, we explore/review several themes for Bayesian asymptotics:

- A role for *statistical models* in these asymptotic results.
  - symmetry/independence assumptions in these results.
  - data reduction
  - asymptotic Normal inference for these results.

A puzzlement?

We have two investigators (***T*** and ***J***) for our coin-tossing problem.
They share the same statistical (*i.i.d.* Bernoulli) model for coin flips,
and they have the ***same*** (conjugate) ***Beta*** prior for θ.

They collect (shared) evidence by flipping the coin until one says, "Stop."
In fact, they observe the sequence

(*H*,*H*,*T*,*H*,*T*,*T*,*H*,*H*,*T*,*H*)

at which point they both (simultaneously) say "Stop!"

However:

T's plan was to flip the coin exactly 10 times and stop
and ***J***'s plan was to flip until there were 6 "Heads" and stop.

Exercise: Give the Bayes analysis for ***T*** and for ***J*** of these data.

Roles for Statistical Models

- Data Reduction and factorization of the likelihood function.
    - *Sufficient* Statistics
    - *Ancillary* Statistics

- Symmetry and Independence assumptions
    - deFinetti's theorem on *exchangeable sequences*

- Properties of *Maximum Likelihood*

Data Reduction Concepts for Statistical Models

Defn: The (dimensional) random variable $Y = \mathbf{g}(X)$ is *sufficient* for the parameter θ (with respect to X) *iff*

$$\mathbf{P}(X \mid Y, \theta) = \mathbf{P}(X \mid Y), \text{ independent of } \theta.$$

Theorem: The likelihood for θ given a sufficient (set of) statistic(s) Y is the same as the likelihood for θ given the (dimensional) variable X for which Y is sufficient.

Proof:

$$\begin{aligned} \mathbf{P}(x \mid \theta) &= \mathbf{P}(x, y \mid \theta) && \text{as } Y = \mathbf{g}(X) \\ &= \mathbf{P}(x \mid y, \theta)\, \mathbf{P}(y \mid \theta) && \text{multiplication axiom} \\ &= \mathbf{P}(x \mid y)\, \mathbf{P}(y \mid \theta) && \text{by sufficiency of } Y \\ &\propto \mathbf{P}(y \mid \theta) \end{aligned}$$

Corollary (Factorization of the likelihood function):

$Y = \mathbf{g}(X)$ is *sufficient* for the parameter θ (with respect to X) *iff*

The likelihood (probability or density) function can be written as the product of two functions of this form:

$$\mathbf{P}(X \mid \theta) = \boldsymbol{h}(X)\, \boldsymbol{j}(Y,\theta).$$

Recall: $Y = \mathbf{g}(X)$

Example 1 (coin-tossing, again):

$\boldsymbol{X} = \langle X_1, \ldots, X_n \rangle$ are *iid* Bernoulli trials given θ, with $\mathbf{P}(X_1=1|\theta) = \theta$, $0 < \theta < 1$.

Claim: $\mathbf{g}(\boldsymbol{X}) = \boldsymbol{Y} = \langle \sum_i X_i, n-\sum_i X_i \rangle$ is a sufficient reduction to the two statistics, #1's $= \sum_i X_i = k$ and #0's $= n-k$ in the sequence $\boldsymbol{X}$.

Proof:

$$\mathbf{P}(\boldsymbol{x},\boldsymbol{y} \mid \theta) = \mathbf{P}(\boldsymbol{x} \mid \theta) = \theta^k (1-\theta)^{n-k}$$

$$\mathbf{P}(\boldsymbol{y} \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Thus $\mathbf{P}(\boldsymbol{x} \mid \boldsymbol{y}, \theta) = \mathbf{P}(\boldsymbol{x},\boldsymbol{y} \mid \theta) / \mathbf{P}(\boldsymbol{y} \mid \theta) = \mathbf{P}(\boldsymbol{x} \mid \boldsymbol{y}) = k!(n-k)!/n!$

That is, $\mathbf{P}(\boldsymbol{X} \mid \boldsymbol{y}, \theta)$ is a discrete, uniform distribution over all sequences H_n that begin with k 1's and $(n-k)$ 0's, independent of θ.

Or, use factorization and note that, alternatively $\langle \overline{X}, n \rangle$ are sufficient for θ as

$$\mathbf{P}(\boldsymbol{X} \mid \theta) = \theta^{n\overline{X}} (1-\theta)^{n(1-\overline{X})} = \boldsymbol{h}(X)\, \boldsymbol{j}(Y,\theta)$$

where $\boldsymbol{h}(X) = 1$ and $Y = \overline{X} = \sum_i X_i/n$.

Example 2 (Normal distribution, known variance):

$\boldsymbol{X} = \langle X_1, \ldots, X_n \rangle$ are *iid* normal $\boldsymbol{N}(\mu,1)$ trials.

Claim: The pair $\langle \bar{X}, n \rangle$ is sufficient for μ.

Proof: Write $\mathbf{p}(X \mid \mu) =$

$$(2\pi)^{-n/2} \exp\left(-\sum_i (X_i - \bar{X})^2 / 2 \right) \exp\left(-n(\mu - \bar{X})^2 / 2\right),$$

where

$$\underline{(2\pi)^{-n/2} \exp\left(-\sum_i (X_i - \bar{X})^2 / 2 \right)} \quad \exp\left(-n(\mu - \bar{X})^2 / 2\right)$$

$$\Uparrow \qquad\qquad \Uparrow$$

$$\boldsymbol{h}(X) \qquad\qquad \boldsymbol{j}(Y,\theta)$$

Defn: The (dimensional) random variable $Y = \mathbf{g}(X)$ is *ancillary* for the parameter θ (with respect to X) *iff*

$$\mathbf{P}(Y \mid \theta) = \mathbf{P}(Y), \text{ independent of } \theta.$$

Theorem: The likelihood for θ based on an ancillary (set of) statistic(s) Y is *constant*.

Corollary: The likelihood for θ based on X equals the conditional likelihood for θ based on X, *given* Y.

$$\mathbf{P}(x \mid \theta) = \mathbf{P}(x \mid y, \theta)$$

Proof:

$$\mathbf{P}(x \mid \theta) = \mathbf{P}(x, y \mid \theta) = \mathbf{P}(x \mid y, \theta)\, \mathbf{P}(y \mid \theta)$$

$$\propto \mathbf{P}(x \mid y, \theta).$$

Example 3 (coin-tossing, again):

$\boldsymbol{X} = \langle X_1, \ldots, X_i, \ldots \rangle$ are *iid* Bernoulli trials given θ, with $\mathbf{P}(X_1=1|\theta) = \theta$, $0 < \theta < 1$.

$\mathbf{g}(\boldsymbol{X}) = \boldsymbol{Y} = \langle \sum_i X_i, n - \sum_i X_i \rangle$ is a sufficient reduction for inference about θ.

Version 3a: The *stopping rule* is sample to a fixed sample size n. Then N (sample size) is ancillary ($\mathbf{P}(N=n) = 1$) and, **given** $N = n$, $\sum_i X_i$ is sufficient!
Moreover, $\mathbf{P}(\sum_i X_i \mid n, \theta)$ is given by the *Binomial*(n, θ) distribution.

Version 3b: The *stopping rule* is sample to a fixed number of "heads," say $\sum_i X_i = k$
Then $\sum_i X_i$ (number of heads) is ancillary ($\mathbf{P}(\sum_i X_i = k) = 1$) and, **given** $\sum_i X_i = k$, the number of flips N is sufficient!
Moreover, $\mathbf{P}(N \mid k, \theta)$ is given by the *Neg-Binomial*(k, θ) distribution.

However, regardless of the stopping rule, in either version, the *pair* $\langle \sum_i X_i, N \rangle$ is sufficient!

Recapitulation of data-reduction principles for statistical models

Sufficiency principle: A sufficient statistic preserves all the relevant information about the parameter that is in the full data set

Ancillarity principle: All the relevant information in the data set about the parameter is contained in the conditional model, given the ancillary statistic.

Likelihood principle: All the relevant information in the data set about the parameter is contained in the likelihood function given the data.

Birnbaum's Theorem: The *Likelihood* principle is equivalent to the conjunction of the *Sufficiency* and *Ancillarity* principles.

Identifying statistical models by symmetry & independence involving observables (*deFinetti's Theorem*)

Heuristic Example (coin-tossing yet again!): Let $\boldsymbol{X} = \langle X_1, \ldots, X_i, \ldots \rangle$ be an infinite sequence of binary trials, with the σ-algebra (A) of events generated by the observable "historical" events H_n: $\langle x_1, \ldots, x_n, \{0,1\}, \{0,1\}, \ldots \rangle$.

Defn: Say that a probability **P** over A is:

- *1-exchangeable* if for $\forall$(i,j) $\mathbf{P}(X_i = 1) = \mathbf{P}(X_j = 1)$

- 2-*exchangeable* if $\forall(i_1, i_2,$distinct and j_1, j_2 distinct)

$$\mathbf{P}(Xi_1 = x_1, Xi_2 = x_2) = \mathbf{P}(Xj_1 = x_1, Xj_2 = x_2)$$

- *n-exchangeable* if $\forall(i_1, i_2, \ldots, i_n$ distinct)

$\mathbf{P}(Xi_1 = x_1, Xi_2 = x_2, \ldots, Xi_n = x_n)$ does not depend on the *n* distinct $\langle i_1, i_2, \ldots, i_n \rangle$

- *exchangeable* if **P** is *n*-exchangeable for each *n* (*n* = 1, 2, …).

Theorem (deFinetti): **P** is exchangeable if and only if **P** can be written as

$$\mathbf{P}(E) = \int_{\Theta} \mathbf{P}(E \mid \theta)\, d\mathbf{Q}(\theta)$$

where

- $\mathbf{P}(\mathsf{A} \mid \theta)$ is given by *iid* Bernoulli(θ) trials
- $\mathbf{Q}(\theta)$ is a prior probability distribution over Θ determined uniquely by **P** over A.

Thus, one can use the computational benefits of sampling from an *iid* statistical model, "as if" it were true, given suitable exchangeability (symmetry) assumptions involving only the algebra of the observable random variables.

Remarks:

- This important theorem generalizes to cover both discrete and continuous random variables.
- Also, there is version dealing with finite sequences (N-exchangeability).
- For a thorough discussion of all this, see chapter 1 of Mark Schervish's book, *Theory of Statistics*, 1995. Springer-Verlag.

Data reduction, Fisher-Information, and Maximum Likelihood

Defn.: *Score* function: $S_{\boldsymbol{X}}(\theta) = \dfrac{\partial (\ln \mathbf{p}(X \mid \theta))}{\partial \theta}$

Fisher Information (under general conditions)

$$\boldsymbol{I}_{\boldsymbol{X}}(\theta) = \mathrm{Var}(S_{\boldsymbol{X}}(\theta)) = \mathrm{E}[-\partial^2 \frac{(\ln \mathbf{p}(X \mid \boldsymbol{\theta}))}{\partial \theta^2}].$$

- Fisher Information is additive for independent data.
- $\boldsymbol{I}_X(\theta) = \boldsymbol{I}_Y(\theta)$ whenever Y is sufficient for θ (with respect to X).
- Fisher Information is a differential form of Kullback-Leiber information.

K-L information inequality

With $\boldsymbol{f}$ and $\boldsymbol{g}$ density functions:

$$\int \mathbf{log}\,[\boldsymbol{f}(\boldsymbol{x})/\boldsymbol{g}(\boldsymbol{x})]\,\boldsymbol{f}(\boldsymbol{x})\,dx = -\int \mathbf{log}\,[\boldsymbol{g}(\boldsymbol{x})/\boldsymbol{f}(\boldsymbol{x})]\,\boldsymbol{f}(\boldsymbol{x})\,dx$$

$$\geq \mathbf{-log(}\int [\boldsymbol{g}(\boldsymbol{x})/\boldsymbol{f}(\boldsymbol{x})]\,\boldsymbol{f}(\boldsymbol{x})\,dx\,\mathbf{)}$$

$$= \mathbf{-log(}\int \boldsymbol{g}(\boldsymbol{x})\,dx\,\mathbf{)}$$

$$\mathbf{= -log\,(}1\mathbf{)}$$

$$\mathbf{=}\ 0$$

where the inequality follows is by an application of Jensen's inequality for the concave **log** function.

The inequality is strict unless almost surely w.r.t. the **F**-distribution, $[\boldsymbol{g}(\boldsymbol{x})/\boldsymbol{f}(\boldsymbol{x})] = 1$.

Corollary: Unless a statistic *t* is sufficient for data *X*, there will be an information (and Fisher Information) gain from adding the full data *X* to *t*.

That is, all and only sufficient statistics preserve all the information in the data.

Defn.: Let $\boldsymbol{\theta^*}$ denote the argmax of the likelihood function $\mathbf{p}(X \mid \theta)$,
the *maximum likelihood estimate* (*MLE*) of the parameter.

Main MLE Theorem (under general regularity conditions on the statistical model):

$$\mathbf{P}(\theta^* \mid \theta_0) \approx \boldsymbol{N}(\theta_0, [\boldsymbol{I}_X(\theta^*)]^{-1}) = \boldsymbol{N}(\theta_0, [n\boldsymbol{I}_{X_i}(\theta^*)]^{-1})$$

So (under "regularity" conditions) the *MLE*:

- Has an asymptotic Normal distribution.
- Is asymptotically consistent (converges to θ_0).
- Is asymptotically sufficient.

Putting these pieces together we have, under the same “regularity” conditions,

convergence of the posterior to the *mle*:

$$\boldsymbol{P}\left[\frac{(\theta^{*}-\theta_{0})}{[\mathrm{I}_{\mathrm{X_n}}(\theta^{*})]^{1/2}} \mid X_n\right] \text{ converges to the standard normal } \boldsymbol{N}(\mathbf{0},1) \text{ distribution.}$$