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 $\Omega = \{0,1\}^{\aleph_0}$

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

The ***events*** that make up the σ-algebra, $\mathfrak{A}$, 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 $\mathfrak{A}$ 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

$\mathbf{P}(\theta \mid \boldsymbol{H}_n)$ 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.

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 $\mathfrak{A}$, consider a strict standard, *uniform distance*,

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

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, $\mathfrak{A}$?

- Recall that the *i.i.d.* Bernoulli statistical model for the data is shared between these two investigators: $(\forall E \in \mathfrak{A})\ \mathbf{P}(E \mid \theta) = \mathbf{Q}(E \mid \theta)$.
- 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: $\mathbf{P}(E \mid H_n) = \int_\Theta \mathbf{P}(E \mid \theta, H_n)\, \mathbf{P}(\theta \mid H_n) d\mathbf{P}^n(\theta).$

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

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

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

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

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

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

Thus, the two posterior *predictive* distributions (over $\mathfrak{A}$) 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, $\quad \mathbf{P}(E \mid H_{n'}) = \mathbf{Q}(E \mid H_{n'}) = 0$ for all $n' \geq n$.

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

Answers:

(1) *Asymptotic Certainty* is automatic with the Bayesian framework!
$(\forall \boldsymbol{E} \in \mathfrak{A})$ with **P**-probability 1,

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

(2) *Asymptotic Consensus* requires only agreement on "null" events.
Assume that $(\forall \boldsymbol{E} \in \mathfrak{A})$ $\mathbf{P}(E) = 0$ *if and only if* $\mathbf{Q}(E) = 0$.
With **P**-(or **Q**-) probability 1, with respect to $\mathfrak{A}$

$$\lim_{n\to\infty} \rho(\mathbf{P}^{\mathbf{n}}, \mathbf{Q}^{\mathbf{n}}) = 0.$$

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

In the next lectures we will explore 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 ***Z***) 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 Z's plan was to flip until there were 6 "Heads" and stop.

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