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):

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

Claim: $\mathbf{g}(\boldsymbol{X}) = \boldsymbol{Y} = <\sum_i X_i, n-\sum_i X_i>$ 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: $\mathbf{P}(\boldsymbol{x},\boldsymbol{y} \mid \theta) = \mathbf{P}(\boldsymbol{x} \mid \theta) = \theta^k (1-\theta)^{n-k}$

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

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 $<\overline{X}, n>$ 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 $\mathbf{p}(X \mid \mu) =$

$$(2\pi)^{-n/2}\, exp\Big(-\Sigma_i (X_i - \bar{X})^2 / 2\Big)\; exp\Big(-n(\mu - \bar{X})^2 / 2\Big),$$

where

$$\underbrace{(2\pi)^{-n/2}\, exp\Big(-\Sigma_i (X_i - \bar{X})^2 / 2\Big)}_{\Uparrow \atop \boldsymbol{h}(X)}\; \underbrace{exp\Big(-n(\mu - \bar{X})^2 / 2\Big)}_{\Uparrow \atop \boldsymbol{j}(Y,\theta)}$$

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

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

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.

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

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 ($\mathfrak{A}$) 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 $\mathfrak{A}$ 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

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

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)

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

- Fisher Information is additive for independent data.
- $I_X(\theta) = I_Y(\theta)$ whenever Y is sufficient for θ (with respect to X).
- Fisher Information is a differential form of Kullback-Leiber information.

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

the *maximum likelihood estimate* (*MLE*) of the parameter.

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

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

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

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