Special Topics in Computational Biology

November 2, 2005

Stanley Sawyer, Department of Mathematics, WashU

Likelihoods and Likelihood Ratio Tests for Nested Hypotheses:

General framework:
H_1: Model with n_1 parameters
H_0: Model with n_0 parameters

Assume $H_0 \subseteq H_1$ and $n_0 < n_1$
Are the extra parameters necessary?
Does H_1 fit the data significantly better than H_0?

Easy example:
Toss a coin 100 times, get 61 heads, 39 tails.
H_1: Coin has $\Pr(H) = p$ for unknown p
H_0: Coin is fair: That is, $p = 0.50$

(Toss a coin 100 times, get 61 heads, 39 tails.)

This is not a very interesting example of a nested hypothesis framework since we can easily test $H_0 : p = 0.50$ using the Central Limit Theorem:

$$Z \;=\; \frac{\text{Binom}(n,p) - np}{\sqrt{np(1-p)}} \;\approx\; \text{Norm}(0,1)$$

For $p = 0.50$ and $n = 100$:

$$Z_{\text{obs}} \;=\; \frac{61 - 50}{\sqrt{100/4}} \;=\; 2.20$$

Therefore we can reject $H_0 : p_0 = 0.50$ with the two-sided P-value:

$$P \;=\; P(|Z| \geq 2.20) \;=\; 0.0278 \;<\; 0.05$$

Harder example:

Suppose that we observe for 1000 nucleotides from one strand of DNA:

(A)212 (T)219 (C)253 (G)316

In particular, this implies:

(AT)431 (CG)569

so that the strand appears CG-rich, but does the data show (within) AT or CG strand bias? Can we test

$H_0 : p_A = p_T$ and $p_C = p_G$?

This is a nested hypothesis test with

H_1: 3 free parameters (p_A, p_T, p_C)

H_0: 1 free parameter p_A

since $p_G = 1 - p_A - p_T - p_C$, and for H_0:
$p_A = p_T$, $p_C = p_G$, $p_A + p_A + p_C + p_C = 1$.

Can we develop a general theory for testing nested hypotheses?

We go through four steps:

I. *Likelihoods of H_0 and H_1:* Let $L_{H_1}(p, X)$ be the probability of observing counts X, assuming for simplicity *in a particular order*, so that

$$L_{H_1}(p, X) = {p_A}^{n_A}\, {p_T}^{n_T}\, {p_C}^{n_C}\, {p_G}^{n_G}$$

without any combinatorial coefficients.

II. Define the *maximum likelihood estimator* (MLE) $\widehat{p} = \widehat{p}(X)$ of $p = (p_A, p_T, p_C, p_G)$ (for H_1) as the solution of

$$\max_p L_{H_1}(p, X) \;=\; L_{H_1}\big(\widehat{p}(X),\, X\big)$$

Since

$$\log L_{H_1}(p, X) = n_A \log p_A + \cdots + n_G \log p_G$$

This is the same as solving

$$\frac{\partial}{\partial p_A} \log L_1(p, X) \;=\; \frac{p_A}{n_A} - \frac{p_G}{n_G} \;=\; 0$$

since $p_G = 1 - p_A - p_T - p_C$ and $\frac{\partial}{\partial p_A} p_G = -1$. Similarly

$$\frac{\partial}{\partial p_T} \log L_1(p, X) \;=\; \frac{p_T}{n_T} - \frac{p_G}{n_G} \;=\; 0$$
$$\frac{\partial}{\partial p_C} \log L_1(p, X) \;=\; \frac{p_T}{n_C} - \frac{p_G}{n_G} \;=\; 0$$

From this it follows that the MLEs are the sample proportions

$$\widehat{p}_A = \frac{n_A}{n}, \quad \widehat{p}_T = \frac{n_T}{n}, \quad \widehat{p}_C = \frac{n_C}{n}, \quad \widehat{p}_G = \frac{n_G}{n}$$

III. We find the *estimated likelihood of the data* X for models H_1 and H_0: Our best guess for the likelihood of X for H_1 is then

$$\begin{aligned}\widehat{L}(H_1, X) &= L_{H_1}(\widehat{p}_1(X), X) \\ &= \left(\frac{n_A}{n}\right)^{n_A} \left(\frac{n_T}{n}\right)^{n_T} \left(\frac{n_C}{n}\right)^{n_C} \left(\frac{n_G}{n}\right)^{n_G}\end{aligned}$$

By the same arguments, the estimated likelihood of X for H_0 is

$$\begin{aligned}\widehat{L}(H_0, X) &= L_{H_1}(\widehat{p}_0(X, H_0), X) \\ &= \left(\frac{n_A + n_T}{2n}\right)^{n_A+n_T} \left(\frac{n_C + n_G}{2n}\right)^{n_C+n_G}\end{aligned}$$

IV. Finally, let d be the *difference* between the *numbers of parameters* in H_1 and H_0. In our case, $d = 3 - 1 = 2$.

Then, the *likelihood ratio test* (LRT) of H_0 is to compare

$$Q = 2 \log \left(\frac{\widehat{L}(H_1, X)}{\widehat{L}(H_0, X)} \right) \approx \chi^2_d$$

where here $d = 2$. That is, the P-value is $P = \Pr\left(\chi^2_d \geq Q_{\text{obs}}\right)$. In our case,

$$\begin{aligned}
\log \widehat{L}(H_1, X) &= 212 \log \left(\frac{212}{1000} \right) + 219 \log \left(\frac{219}{1000} \right) \\
&\quad + 253 \log \left(\frac{253}{1000} \right) + 316 \log \left(\frac{316}{1000} \right) \\
&= -1373.190
\end{aligned}$$

$$\begin{aligned}
\log \widehat{L}(H_0, X) &= 431 \log \left(\frac{431}{2000} \right) + 569 \log \left(\frac{569}{2000} \right) \\
&= -1376.742
\end{aligned}$$

Thus

$$\begin{aligned} Q &= 2\log\left(\frac{\widehat{L}(H_1, X)}{\widehat{L}(H_0, X)}\right) \\ &= 2\Big(\log \widehat{L}(H_1, X) - \log \widehat{L}(H_0, X)\Big) \\ &= 2(1376.742 - 1373.190) = 7.1034 \end{aligned}$$

The P-value is

$$P = P(\chi^2_2 \geq 7.1034) = 0.0287 < 0.05$$

and we reject H_0: The data does show significant evidence for strand asymmetry between A and T and/or between C and G, for data

(A)212 (T)219 (C)253 (G)316

and

(AT)431 (CG)569

Bayesian Analysis and Conjugate Priors:

Example: Toss a coin $n = 10$ times. Let X be the number of heads ($0 \leq X \leq 10$). How should we estimate $p = \Pr(\text{Head})$? Here, the likelihood and MLE are:

$$L(p, X) = p^X (1-p)^{10-X}, \qquad \widehat{p}(X) = \frac{X}{10}$$

What if $X = 0$? What should $\widehat{p}(X)$ be in that case?

The problem may be that we are treating p and X differently (as a parameter and a random variable, respectively):

We might have a better idea of how to handle odd questions such as this if we could put p and X somehow on the same footing. (Or, at least, that was Bayes' original idea.)

The first step is to force p to be a random variable by saying that it has a probability distribution $\pi_0(p)$ for $0 \leq p \leq 1$. This is called a *prior distribution* for p.

The prior distribution $\pi_0(p)$ makes p into a random variable. Then (p, X) together have the joint probability distribution

$$\pi_1(p, X) = \pi_0(p)L(p, X) = \pi_0(p)\, p^X (1-p)^{10-X}$$

Finally, we notice that X is constant (because we have just observed it), so that we can form the *posterior (conditional) distribution*

$$\pi_1(p \mid X) = \frac{\pi_0(p)L(p, X)}{\int_0^1 \pi_0(x)L(x, X)\, dx}$$

We can use the posterior distribution $\pi_1(p \mid X)$ to define the *Bayes estimator*

$$\widehat{p}_B(X) = \int_0^1 p\pi_1(p, X)\, dp = \frac{\int_0^1 p\pi_0(p)L(p, X)\, dp}{\int_0^1 \pi_0(p)L(p, X)\, dp}$$

Note that $\widehat{p}_B(X)$ makes sense even if $X = 0$, but depends on $\pi_0(p)$. Also, we should be careful when we use this method, since we may end up having to evaluation not only one, but two nasty integrals. In contrast, finding MLEs only requires derivatives but not integrals.

Example: $\pi_0(p) = 1$. Then the full likelihood is

$$\pi_1(p, X) = \pi_0(p)L(p, X) = p^X(1-p)^{10-X}$$

so that the posterior distribution is

$$\pi_1(p \mid X) = C(X)p^X(1-p)^{10-X}$$

Note that, as a function of p, the posterior density

$$\begin{aligned}\pi_1(p \mid X) &= C(X)p^X(1-p)^{10-X} \\ &\approx \text{Beta}(X+1, 11-X)\end{aligned}$$

where $\text{Beta}(\alpha, \beta)$ is the beta distribution with density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

Since

$$E\big(\text{Beta}(\alpha,\beta)\big) = \frac{\alpha}{\alpha+\beta}$$

it follows that

$$\begin{aligned}\widehat{p}_B(X) &= \int_0^1 p\pi_1(p \mid X)\, dp \\ &= \frac{X+1}{12}\end{aligned}$$

In particular, if $X = 0$, $\widehat{p}_B(0) = 1/12$.

This argument generalizes: Suppose we chose, instead, $\pi_0(p) = Cp(1-p)$ or $Cp^5(1-p)^5$ or, in general

$$\pi_0(p) = Cp^{\alpha-1}(1-p)^{\beta-1}$$

Then the full likelihood is

$$\begin{aligned}
\pi_1(p, X) &= \pi_0(p)L(p, X) \\
&= Cp^{\alpha-1}(1-p)^{\beta-1}p^X(1-p)^{10-X} \\
&= Cp^{\alpha+X-1}(1-p)^{\beta+10-X-1} \\
&\approx \text{Beta}(\alpha + X, \beta + 10 - X)
\end{aligned}$$

and

$$\begin{aligned}
\widehat{p}_B(X) &= E\big(\text{Beta}(\alpha + X, \beta + 10 - X)\big) \\
&= \frac{X + \alpha}{\alpha + \beta + 10}
\end{aligned}$$

This is a special case of a general situation:

Consider a general family of distributions, for example Beta(α, β) or Gam(α, β) or Norm(μ, σ^2).
Let $\pi_0(p, \alpha, \beta)$ be a general prior distribution from that family.

Then, we say that this family is a *conjugate prior* (family) for the likelihood $L(p, X)$ if we always have that

$$\begin{aligned}\pi_1(p, \alpha, \beta \mid X) &= \pi_0(p, \alpha, \beta) L(p, X) \\ &= \pi_0(p, \alpha_1, \beta_1)\end{aligned}$$

is a member of the same family, where $\alpha_1 = \alpha(X)$ and $\beta_1 = \beta(X)$ are called *updating formulas* for α and β. (These can also depend on other constants in $L(p, X)$.)

Note that here $\pi_0(p, \alpha, \beta)$ is a density in p, while $L(p, X)$ is a density in X.

In our case, the posterior density $\pi_1(p \mid X)$ is

$$\begin{aligned} & Cp^{\alpha-1}(1-p)^{\beta-1}\, p^X(1-p)^{10-X} \\ = \; & Cp^{\alpha+X-1}(1-p)^{\beta+10-X-1} \end{aligned}$$

so that the beta distribution family is a *conjugate prior* for binomial sampling. The updating formulas are $\alpha_1(X) = \alpha + X$ and $\beta_1(X) = \beta + 10 - X$.

This generalizes to multinomial sampling:

Recall that a distribution on the simplex (p_1, p_2, p_3, p_4) (that is, $p_i \geq 0$ and $p_1 + p_2 + p_3 + p_4 = 1$) is a *Dirichlet distribution* $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ if

$$\pi_0(p) = Cp_1^{\alpha_1-1}p_2^{\alpha_2-1}p_3^{\alpha_3-1}p_4^{\alpha_4-1}$$

where

$$C = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)}$$

Suppose that we do n independent multinomial trials and obtain, in some order,

$$n_i \text{ outcomes of Type } i \text{ (prob. } p_i \text{ each)}$$

for $i = 1, 2, 3, 4$. Then the likelihood is

$$L(p, X) = p_1^{n_1}\, p_2^{n_2}\, p_3^{n_3}\, p_4^{n_4}$$

for $X = (n_1, n_2, n_3, n_4)$. If we multiply $L(p, X)$ by the Dirichlet prior $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, we obtain the posterior distribution

$$\pi_0(p) = C p_1^{\alpha_1+n_1-1} p_2^{\alpha_2+n_2-1} p_3^{\alpha_3+n_3-1} p_4^{\alpha_4+n_4-1}$$

which is $\mathcal{D}(\alpha_1 + n_1, \alpha_2 + n_2, \alpha_3 + n_3, \alpha_4 + n_4)$.

This means that the family of Dirichlet distributions are a conjugate prior for *multinomial* sampling.

In this case, since

$$E(p_i) = \frac{\alpha_i}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}$$

for a Dirichlet $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, we have

$$E(p_i \mid X) = \frac{\alpha_i + n_i}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + n}$$

for the posterior distribution. If the α_i are small, this is close to the MLE $\widehat{p}_i(X) = n_i/n$.

As another example, suppose that $X_1, X_2, \ldots, X_n$ are independent *Poisson* random variables. Then the likelihood is

$$L(\mu, X_1, \ldots, X_n) = \prod_{i=1}^{n} \left(e^{-\mu} \frac{\mu^{X_i}}{X_i!} \right)$$

$$= C(X) e^{-n\mu} \mu^{S(X)}, \quad S(X) = \sum_{i=1}^{n} X_i$$

The MLE for μ is

$$\widehat{\mu}(X) = \frac{S(X)}{n} = \overline{X}$$

Again, we may not want to estimate $\widehat{\mu}(X) = 0$ if $X = 0$, so we consider a prior distribution for μ. Since μ satisfies $0 \leq \mu < \infty$ rather than $0 \leq p \leq 1$, we can't use a beta distribution for the prior. However, the gamma density

$$\pi_0(\mu, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} \exp(-\beta\mu)$$

is defined for $0 \leq \mu < \infty$. The posterior distribution of μ is then proportional to

$$\begin{aligned} & \pi_0(\mu, \alpha, \beta) L(\mu, X) \\ &= C(X) \mu^{\alpha-1} \exp(-\beta\mu)\, e^{-n\mu} \mu^{S(X)} \\ &= C(X) \mu^{\alpha+S(X)-1} \exp(-(\beta+n)\mu) \end{aligned}$$

As a function of μ, the posterior distribution $\pi_1(\mu, \alpha, \beta \mid X)$ is then $\mathsf{Gam}(\alpha + S(X), \beta + n)$.

This means that the gamma distributions are a conjugate prior for Poisson sampling.

Since the mean of $\mathsf{Gam}(\alpha, \beta)$ is α/β, the Bayes estimator of μ is

$$\widehat{p}_B(X) = E\big(\mathsf{Gam}(\alpha + S(X), \beta + n)\big) = \frac{\alpha + S(X)}{\beta + n}$$

This is close to $\widehat{p}(X) = \overline{X} = S(X)/n$ if α, β are small.

It is typical to set $\pi_0(\mu) = \mathsf{Gam}(\epsilon, \epsilon)$ for $\epsilon = 0.001$. This distribution has mean one but, due to the $\mu^{\epsilon-1}$ singularity at $\mu = 0$, has the vast majority of its mass very close to 0. For this prior,

$$\widehat{p}_B(0) = \frac{\epsilon}{\epsilon + n}$$

As another example, suppose that $X_1, X_2, \ldots, X_n$ are independent exponentially distributed random variables where r is the *rate* ($E(X_i) = 1/r$). The likelihood is

$$L(r, X_1, \ldots, X_n) = \prod_{i=1}^{n} \big(r \exp(-rX_i)\big)$$

$$= \; r^n \exp(-rS(X)), \qquad S(X) = \sum_{i=1}^{n} X_i$$

If we use a gamma distribution prior for r

$$\pi_0(r, \alpha, \beta) = Cr^{\alpha-1}e^{-\beta r}, \qquad r \geq 0$$

then the posterior density is

$$\begin{aligned} & C(X)\, \pi_0(r, \alpha, \beta) L(r, X) \\ &= \; C(X)\, r^{\alpha-1} \exp(-\beta r)\; r^n e^{-rS(X)} \\ &= \; C(X)\, r^{\alpha+n-1} \exp\big(-\big(\beta + S(X)\big)r\big) \\ &\approx \; \mathrm{Gam}\big(\alpha + n,\; \beta + S(X)\big) \end{aligned}$$

The Bayes estimator of r is then

$$\widehat{r}_B(X) = E\big(\mathsf{Gam}(\alpha + n,\, \beta + S(X))\big) = \frac{\alpha + n}{\beta + S(X)}$$

If α, β are small, this is close to the MLE $\widehat{r}(X) = 1/\overline{X} = n/S(X)$.

Thus the gamma distribution family is a conjugate prior for both Poisson and exponential sampling, but the role of n and $S(X)$ are reversed in the updating formulas:

For Poisson sampling:

$$\pi_1(\mu, \alpha, \beta \mid X) \;\approx\; \mathsf{Gam}\big(\alpha + S(X),\, \beta + n\big)$$

while for exponential sampling:

$$\pi_1(r, \alpha, \beta \mid X) \;\approx\; \mathsf{Gam}\big(\alpha + n,\, \beta + S(X)\big)$$

In most of the examples before, a likelihood with a single parameter

p for Bernoulli sampling
μ for Poisson sampling
r for exponential sampling

had a conjugate prior with two parameters (Beta(α, β) or Gam(α, β)). There are also conjugate priors for the Gaussian disitribution

$$L(\mu, \sigma^2, X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)(X-\mu)^2}$$

that have *four parameters* (two for μ and two for σ^2). The first step is to rewrite the Gaussian density L in terms of the *precision* $v = 1/\sigma^2$:

$$L(\mu, v, X) = \sqrt{\frac{v}{2\pi}} e^{-(1/2)v(X-\mu)^2}$$

Note that

$$L(\mu, v, X) = \sqrt{\frac{v}{2\pi}} e^{-(1/2)v(X-\mu)^2}$$

is $\text{Gam}\big(3/2,\, (1/2)(X-\mu)^2\big)$ as a function of v. The simplest way to obtain priors for v and μ is to define priors for v and μ separately:

$$\begin{aligned}
\pi_0(v, \alpha_\epsilon, \beta_\epsilon) &\approx \text{Gam}(\alpha_\epsilon, \beta_\epsilon) \quad \text{in } v \\
&= C v^{\alpha_\epsilon - 1} e^{-\beta_\epsilon v} \\
\pi_0(\mu, \mu_\epsilon, v_\epsilon) &\approx \text{Norm}(\mu_\epsilon, v_\epsilon) \quad \text{in } \mu \\
&= C \exp\left(-(1/2) v_\epsilon (\mu - \mu_\epsilon)^2\right)
\end{aligned}$$

for four parameters $\alpha_\epsilon, \beta_\epsilon, \mu_\epsilon, v_\epsilon$, where we ignore factors that don't depend on v or μ. As before, the initial parameters $\alpha_\epsilon, \beta_\epsilon, \mu_\epsilon, v_\epsilon$ will be small, but will become larger after conditioning on data.

For one observations $X \approx \text{Norm}(\mu, v)$ for precision v,

$$\begin{aligned}
&\pi_1(v, \alpha_\epsilon, \beta_\epsilon \mid X, \mu) \\
&\quad = \; C v^{\alpha_\epsilon - 1} e^{-\beta_\epsilon v} \sqrt{\frac{v}{2\pi}} e^{-(1/2)v(X-\mu)^2} \\
&\quad \approx \; \text{Gam}\big(\alpha_\epsilon + 1/2,\, \beta_\epsilon + (1/2)(X-\mu)^2\big)
\end{aligned}$$

$$\begin{aligned}
&\pi_1(\mu, \mu_\epsilon, v_\epsilon \mid X, v) \\
&\quad = \; C e^{-(1/2)v_\epsilon(\mu-\mu_\epsilon)^2} e^{-(1/2)v(X-\mu)^2} \\
&\quad = \; C \exp\left(-(1/2)(v_\epsilon + v)\left(\mu - \frac{v_\epsilon \mu_\epsilon + vX}{v_\epsilon + v}\right)^2 \right) \\
&\quad \approx \; \text{Norm}\left(\frac{v_\epsilon \mu_\epsilon + vX}{v_\epsilon + v},\, v_\epsilon + v \right)
\end{aligned}$$

There are several different ways of setting up conjugate priors for normal sampling. This is the simplest, but not necessarily the best.

These formulas leads to the updating formulas for $X \approx \text{Norm}(\mu, v)$:

$$\begin{aligned}
\text{For } v\text{:} \quad & \alpha_\epsilon \to \alpha_\epsilon + 1/2 \\
& \beta_\epsilon \to \beta_\epsilon + (1/2)(X - \mu)^2 \\
\text{For } \mu\text{:} \quad & \mu_\epsilon \to \frac{v_\epsilon}{v_\epsilon + v}\mu_\epsilon + \frac{v}{v_\epsilon + v}X \\
& v_\epsilon \to v_\epsilon + v
\end{aligned}$$

This generalizes to a formula for updating $(\alpha_\epsilon, \beta_\epsilon, \mu_\epsilon, v_\epsilon)$ for a normal sample $X_1, \ldots, X_n$:

$$\begin{aligned}
\text{For } v\text{:} \quad & \alpha_\epsilon \to \alpha_\epsilon + n/2 \\
& \beta_\epsilon \to \beta_\epsilon + (1/2)\sum_{i=1}^{n}(X_i - \mu)^2 \\
\text{For } \mu\text{:} \quad & \mu_\epsilon \to \frac{v_\epsilon}{v_\epsilon + vn}\mu_\epsilon + \frac{nv}{v_\epsilon + nv}\overline{X} \\
& v_\epsilon \to v_\epsilon + vn
\end{aligned}$$

Thank you for coming.