What we need to know about likelihood – a summary

1. Maximum Likelihood Estimation

1.1 What is estimation

Statistical model: Data random variable $\mathbf{X}$, whose value $\mathbf{x}$ is observed, has one of a family of probability distributions $\{P_\theta; \theta \in \Theta\}$, indexed by a **parameter** θ in **parameter space** Θ.

Goals of estimation:
(i) make inferences about which P_θ gave rise to the observed $\mathbf{x}$.
(ii) Assess the uncertainty associated with this inference.

A **statistic** is a function of the data random variables $\mathbf{X}$.

An **estimator** $T \ = \ T(\mathbf{X})$ is a statistic taking values in Θ.

The **estimate** is $T(\mathbf{x})$, the value taken by the estimator, which is used to estimate θ.

The **likelihood** is $L_{\mathbf{x}}(\theta) \ = \ P_\theta(\mathbf{x})$, a function of θ. The likelihood provides the connection between the data $\mathbf{x}$ and the probability model P_θ.

Example: Suppose X_i, $i = 1, ..., n$ i.i.d. $B(1, \theta)$, the indicators of success in n independent trials, each with success probability θ.. So $P_\theta(x) = \theta^x(1-\theta)^{1-x}$, for each trial.
So $L_n(\theta) = \prod_1^n (\theta^{x_i}(1-\theta)^{1-x_i})$ and

$$l_n(\theta) = (\sum_1^n x_i)\log(\theta) + (n - \sum_1^n x_i)\log(1-\theta)$$

Note that the (log)-likelihood depends only on $t = \sum_1^n x_i$, the total number of successes.

1.2 Maximum likelihood estimation.

The MLE maximises the likelihood as a function of θ, to give the value of θ that “best explains” the data $\mathbf{x}$. To obtain the estimator, maximise $P_\theta(\mathbf{X})$, or $\log P_\theta(\mathbf{X})$ w.r.t. θ.

Example: If T is $B(n, \theta)$

$$L(\theta) = P_\theta(T = t) = \left(\begin{array}{c} n \\ t \end{array}\right)\theta^t(1-\theta)^{n-t}$$
$$l(\theta) = const + t\log(\theta) + (n-t)\log(1-\theta)$$

maximising this w.r.t. θ we get $\hat{\theta} = t/n$.

1.3 Desirable properties of estimators

1.3.1 Lack of bias:

$T(\mathbf{X})$ is **unbiased** if, for any $\theta \in \Theta$, $\mathbf{X} \sim P_\theta \implies \mathrm{E}(T(\mathbf{X})) \ = \ \theta$.
(We write this as $\mathrm{E}_\theta(T(\mathbf{X})) \ = \ \theta$, the subscript indicating the ”true” θ-value.)
That is, an unbiased estimator is ”right, on average”, over repetitions of the experiment.

The **bias** of estimator $T(\mathbf{X})$ is $b_T(\theta) \ = \ \mathrm{E}_\theta(T(\mathbf{X})) - \theta$.

Example: if T is $B(n, \theta)$, then $\mathrm{E}_\theta(T) = n\theta$, so the MLE is unbiased. Unbiasedness alone is a very weak criterion. Some unbiased estimators may be very bad, and many “good” estimators are biased. In particular many MLE’s are biased.

1.3.2 Small mean square error

The **mean square error** of estimator $T(\mathbf{X})$ is $\mathrm{MSE}_\theta(T) \;=\; \mathrm{E}_\theta((T(\mathbf{X})-\theta)^2)$.

If $T(\mathbf{X})$ is unbiased, $\mathrm{MSE}_\theta(T) \;=\; \mathrm{var}_\theta(T)$.

In general, $\mathrm{MSE}_\theta(T) \;=\; \mathrm{var}_\theta(T) \;+\; (b_T(\theta))^2$.

Example: For the estimator T/n of the binomial parameter θ,

$$\mathrm{var}(T/n) = \mathrm{var}(T)/n^2 = n\theta(1-\theta)/n^2 = \theta(1-\theta)/n$$

1.3.3 Consistency:

Consider an n-sample $\mathbf{X}^{(n)} = (X_1, ..., X_n)$, X_i i.i.d. and a sequence of estimators $T_n = T(\mathbf{X}^{(n)})$. Then the sequence of estimators (T_n) is **consistent for** θ if, for every $\theta \in \Theta$, and every $\epsilon > 0$, $P_\theta(|T_n - \theta| > \epsilon) \to 0$ as $n \to \infty$.
That is, for every θ, if $\mathbf{X}^{(n)} \sim P_\theta$, $T_n \xrightarrow{\mathrm{p}} \theta$.

In the Binomial example, above, $T(\mathbf{X}) = \frac{1}{n}\sum_1^n X_i$ is unbiased for θ, has

$$\mathrm{MSE}_\theta(T) = \mathrm{var}_\theta(T) = \theta(1-\theta)/n \longrightarrow 0 \;\; \text{as } n \to \infty$$

and hence is consistent.

1.4 Properties of maximum likelihood estimators

1.4.1 Basic properties of MLE

(i) If $\theta = \widehat{\theta}$ maximises $L_\mathbf{x}(\theta)$ then $\theta = \widehat{\theta}$ maximises $l_\mathbf{x}(\theta) = \log L_\mathbf{x}(\theta)$ since log is an increasing function. The equation $\frac{\partial l_\mathbf{x}}{\partial \theta} = 0$ is known as the **likelihood equation**. It is usually the easiest way to find the MLE if $l(\theta)$ is differentiable w.r.t. θ.

(ii) Let $\alpha(\theta)$ be a 1-1 differentiable function of θ then

$$\frac{\partial l_\mathbf{x}}{\partial \alpha} \;=\; \frac{\partial l_\mathbf{x}}{\partial \theta}\frac{d\theta}{d\alpha} \quad \text{and} \quad \frac{\partial^2 l_\mathbf{x}}{\partial \alpha^2} \;=\; \frac{\partial^2 l_\mathbf{x}}{\partial \theta^2}\left(\frac{d\theta}{d\alpha}\right)^2 \;+\; \frac{\partial l_\mathbf{x}}{\partial \theta}\frac{d^2\theta}{d\alpha^2}$$

So the two likelihood equations (w.r.t. θ and w.r.t. α) have the same roots, and the same sign of the second deriative at any local stationary point of $l_\mathbf{x}$. That is, $\widehat{\alpha} \;=\; \alpha(\widehat{\theta})$.

(iii) MLE's can often be biased, but under very broad conditions they are asympotatically unbiased.

1.4.2 Asymptotic properties

(i) There is a neat result which we won't prove here, which says that if we consider $\mathrm{E}_{\theta_0}(\log(P_\theta(\mathbf{X})))$ this is maximized w.r.t θ by $\theta = \theta_0$. Informally, this says *the expected log-likelihood is maximized at the true value of the parameter.* The difference

$$\mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{X}))) \;-\; \mathrm{E}_{\theta_0}(\log(P_\theta(\mathbf{X})))$$

is known as the Kullback-Leibler information.

(ii) One of the fairly immediate consequences of (i) is that under very broad conditions MLEs are consistent.

(iii) There is another neat result, which we also won't prove here, which says that (subject to a few conditions) no unbiased estimator can have a variance smaller than

$$\left[\mathrm{E}_{\theta_0}\left(-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{X}))\right)\right]^{-1}$$

This quantity is known as the Fisher information. The larger the information, the smaller the variance can be.

(iv) Another result says that, subject to a few more conditions, MLE's are asymptotically approximately Normal, with mean θ_0, the true parameter value, and variance the inverse of the Fisher Information. This says that, *in large samples,* MLEs, are the *best estimators.*

(v) All these conditions etc. are satisfied in all the binomial and other examples we shall talk about.

(vi) Of course, we don't know θ_0, but at least for large samples, we know the MLE $\hat{\theta}$ is close to the true value θ_0, so we compute the Fisher information, and plug in $\hat{\theta}(\mathbf{X})$ for θ_0 to get an estimate of the variance of the MLE.

(vii) Sometimes the expectation in the Fisher Information can be hard to compute. Then, at least for large samples, we just compute

$$\left[-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{X}))\right]^{-1}$$

and evaluate it by plugging in the observed $\mathbf{x}$ for $\mathbf{X}$ and $\hat{\theta}(\mathbf{x})$ for θ_0.
There is lots of theory that says all this is ok.

Example:

T is $B(n, \theta)$. As before

$$l(\theta) = const + t\log(\theta) + (n-t)\log(1-\theta)$$

and we know the MLE is T/n which has expectation θ and variance $\theta(1-\theta)/n$.

Now,

$$l''(\theta) = -\frac{t}{\theta^2} - \frac{(n-t)}{(1-\theta)^2}$$

so the Fisher information comes out as $n/\theta(1-\theta)$. Thus in this example, the MLE has the smallest possible variance.

In practice, we estimate the variance as

$$\hat{\theta}(1-\hat{\theta})/n \;=\; t(n-t)/n^3$$

We can note that in fact we get the same result if we plug in t/n for θ in $l''(\theta)$, without going through the expectation step. That is not true in general, but it is for this particular example.

2. Testing multinomial probabilities

2.1 Generalized likelihood ratio tests

(i) Just as the maximum likelihood estimate is the value of the parameter that best explains the observed data, the maximized value of the likelihood is a measure of how well this value is supported by the data, relative to how well other values are supported by the observation of these data.

Accordingly we define

$$L(\Theta_0) \quad = \quad \max_{\theta \in \Theta_0}(L(\theta))$$

as a measure of support for the hypothesis $\theta \in \Theta_0$, and

$$\Lambda(\Theta_1 : \Theta_0) \quad = \quad L(\Theta_1)/L(\Theta_0)$$

as a measure of the relative support for the two hypotheses $\theta \in \Theta_1$ and $\theta \in \Theta_0$.

(ii) In the case when $\Theta_0 \subseteq \Theta_1$, $\Lambda \geq 1$, $2\log_e \Lambda \geq 0$, and asymptotically, if $\theta \in \Theta_0$ is true, then $2\log_e \Lambda$ is approximately distributed as a χ^2 random variable, with degrees of freedom equal to $\dim(\Theta_1) - \dim(\Theta_1)$. If $\theta \in \Theta_0$ is not true, then $2\log_e \Lambda \to \infty$ at a rate which depends on the Kullback-Leibler information.

There are, of course, various regularity conditions in order that these results hold, but these are essentially the same as the ones needed for the asymptotic results about MLEs. They hold for all the sorts of examples we shall talk about.

2.2 Multinomial likelihoods

(i) Suppose there are k possible outcomes, having probabilities $p_i, i = 1, \ldots, m$, and a vector of parameters θ, so p_i is $p_i(\theta)$. The log-likelihood is

$$\begin{aligned}
\ell &= \mathit{const} + \sum_{i=1}^{m} n_i \log p_i \\
\text{so } \frac{\partial \ell}{\partial \theta_j} &= \sum_{i=1}^{m} \frac{n_i}{p_i} \frac{\partial p_i}{\partial \theta_j} \\
\text{and } \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} &= -\sum_{i=1}^{m} \frac{n_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} - \sum_{i=1}^{m} \frac{n_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k} \\
\text{so } \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right) &= \sum_{i=1}^{m} \frac{np_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} + \sum_{i=1}^{m} \frac{np_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k} \\
&= n \sum_{i=1}^{m} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} + n \sum_{i=1}^{m} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k} \\
&= n \sum_{i=1}^{m} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} + n \frac{\partial^2 \sum_{i=1}^{m} p_i}{\partial \theta_j \partial \theta_k} \\
&= n \sum_{i=1}^{m} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k}
\end{aligned}$$

since $\sum_{i=1}^{m} p_i \equiv 1$. This is sometimes known as Fisher's formula, since he was one of the first people to write it down in this way.

(ii) **Example:** $m = 2$, $p_1(\theta) = \theta^2$, $p_2 = 1 - \theta^2$.

$$\begin{aligned} \frac{\partial p_1}{\partial \theta} &= 2\theta \text{ and } \frac{\partial p_2}{\partial \theta} \;=\; -2\theta \\ \text{so } \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \theta^2}\right) &= n\left(\frac{1}{\theta^2}(2\theta)^2 + \frac{1}{1-\theta^2}(-2\theta)^2\right) \\ &= \frac{4n}{(1-\theta^2)} \end{aligned}$$

(iii) Case of the general model, $\sum_{i=1}^{m} p_i = 1$ with no other constraints.

$$\begin{aligned} \ell &= \sum_{i=1}^{m} n_i \log p_i = \sum_{i=1}^{m-1} n_i \log p_i + n_m \log(1 - \sum_{i=1}^{m-1} p_i) \\ \frac{\partial \ell}{\partial p_i} &= \frac{n_i}{p_i} - \frac{n_m}{p_m} \end{aligned}$$

for $i = 1, \ldots, m-1$, giving the MLE $\widehat{p_i} = n_i/n$, and the maximized value of the log-likelihood is

$$\hat{\ell} = \sum_{i=1}^{m} n_i \log \widehat{p_i} = \sum_{i=1}^{m} n_i \log n_i - n \log n$$

3. Gene counting and the EM algorithm

(i) We have seen that where genotypes are observable, estimating allele frequencies is just a matter of *counting the genes*. In a slightly more general sense, the same is true when we cannot observe the genotypes fully. The formal derivation of this is the theory of the EM algorithm, which dates to 1977, but *gene-counting* was well established for estimating ABO allele frequencies for example, at least 20 years earlier.

(ii) The general set-up is as follows. Let $\mathbf{Y}$ be the actual data random variables, while $\mathbf{X}$ are additional related random variables which are not observed, so the likelihood is

$$L(\theta) = f_{\mathbf{Y}}(\mathbf{y}; \theta) = \int_{\mathbf{x}} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \theta) d\mathbf{x}$$

The *complete-data log-likelihood*, which is what the log-likelihood would be *if we could observe* $\mathbf{X}$ *in addition to* $\mathbf{Y}$, is $\ell^*(\mathbf{X}, \mathbf{Y}; \theta) = \log f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}; \theta)$.

The EM algorithm is as follows:
At current parameter values θ_0
E-step: compute $\ell^{**}(\theta_0; \theta) = \mathrm{E}_{\theta_0}(\ell^*(\mathbf{X}, \mathbf{Y}; \theta) \mid \mathbf{Y} = \mathbf{y})$
M-step: maximize $\ell^{**}(\theta_0; \theta)$ w.r.t θ, say at $\theta = \theta_1$
Result: $L(\theta_1) \geq L(\theta_0)$.

(iii) Multinomial case

Since $\ell = \sum n_i \log p_i$, determining ℓ^{**} is just a question of computing the expected value of the complete-data count n_i given the counts actually observed (E-step). Given these complete-data counts, the maximization step (M-step) counts the proportions. Two examples will make this clearer!

(iv) Case of the recessive allele.

In fact, we don't need the EM algorithm here, since we can find the MLE directly, but it's good to see it works. Suppose in a sample size 100 there are $n_1 = 36$ of the recessive type AA. The three genotypes are AA, AB and BB, but we do not see the counts of AB and BB since B is dominant to A. If we did see these counts, n_2 and n_3, we would estimate p, the frequency of A by $(2n_1 + n_2)/2n$. Further,

$$P(AB \mid AB \text{ or } BB) \quad = \quad \frac{2pq}{2pq + q^2} \quad = \quad Q \quad \text{so} \quad \mathrm{E}_p(n_2 \mid n_2 + n_3 = 64) = 64Q$$

	current	current	AA	AB	BB	new
	p	$Q = 2pq/(2pq + q^2)$	$n_1 = 36$	$n_2 + n_3 = 64$		$p = (2n_1 + n_2)/2n$
	0.5	0.667	36	42.67	21.33	0.573
So now	0.573	0.729	36	46.64	17.36	0.593
	0.593	0.745	36	47.66	16.34	0.598
	0.598	0.749	36	47.91	16.09	0.600
	0.600	0.750	36	48.00	16.00	0.600

(v) Case of ABO blood types

The case of ABO blood types is directly analogous, but here we must partition both the A phenotype into AA and AO, and the B phenotype into BB and BO. Once the counts are partitioned, according to current estimates of allele frequencies, the new estimate of the A allele frequency p is $Q_{AA} + (Q_{AO} + Q_{AB})/2$, and the new estimate of the B allele frequency q is $Q_{BB} + (Q_{BO} + Q_{AB})/2$. Here the EM algorithm is in fact one of the easiest ways to get the MLE. Recall, Bernstein sampled 502 individuals, and reported frequencies of the four types 0.422, 0.206, 0.078, and 0.294.

current values				AA	AO	BB	BO	AB	OO	new values	
p	q	$\frac{2r}{p+2r}$	$\frac{2r}{q+2r}$	$Q_A = 0.422$		$Q_B = 0.206$		Q_{AB}	Q_{OO}	p	q
				$= Q_{AA} + Q_{AO}$		$= Q_{BB} + Q_{BO}$		$= 0.078$	$= 0.294$		
0.3	0.3	0.73	0.73	0.115	0.307	0.056	0.150	0.078	0.294	0.308	0.170
0.308	0.170	0.77	0.86	0.096	0.326	0.029	0.177	0.078	0.294	0.298	0.156
0.298	0.156	0.79	0.87	0.091	0.331	0.026	0.180	0.078	0.294	0.295	0.155
0.295	0.155	0.79	0.88	0.089	0.333	0.025	0.181	0.078	0.294	0.295	0.155

Over the iterations, the log-likelihoods is −687.1242, −628.9991, −627.5693, −627.5262, −627.5246.