Example 8

Let $Y_1, \dots, Y_m$ be a random vector with distribution $\text{Mn}(n, \theta_1, \dots, \theta_m)$ with $0 < \theta_i < 1$ for all i, $\sum_i \theta_i = 1$ and $n = \sum_i Y_i$.

For example, in a sample of Chinese population of Hong Kong in 1937, blood types occur with the following frequencies, where M and N are red cell antigens

	Blood Type			
	M	MN	N	Total
Frequency	342	500	187	1029

Intuitively, the estimated probability of each blood type is the ratio of the observed frequency divided by n, i.e. $\widehat{\theta}_i = y_i/n$ for all i. This is the MLE of θ.

Indeed, the log-likelihood is

$$\ell(\theta) = \log n! - \sum_{i=1}^{3} \log y_i! + \sum_{i=1}^{3} y_i \log \theta_i .$$

Example 8 (cont'd)

To find the MLE, this time we have to be more careful, due to the constraint $\sum_i \theta_i = 1$.

For this we use the method of Lagrange multiplier, and get the augmented log-likelihood

$$\ell_a(\theta, \lambda) = \ell(\theta) + \lambda \left(\sum_i \theta_i - 1 \right)$$

Taking partial derivatives w.r.t θ_i and solving the equations leads to

$$\theta_i = -y_i / \lambda.$$

Summing both sides of the equations we get $1 = -\sum_i y_i / \lambda$, thus $\lambda = n$.

Replacing back to the equation we get the solution $\widehat{\theta} = y_i / n$ as conjectured. The estimated cell probabilities are thus (0.332, 0.486, 0.182)

Methods for evaluating estimators: Sufficiency

A sufficient statistic for the parameter θ is a statistic that, intuitively, captures all the information about θ in the sample.

Formally, T_n is a sufficient statistic for θ if the conditional distribution of the sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ given the value of $T(\mathbf{Y})$ does not depend on θ.

To use this definition we must check that for any $\mathbf{y} = (y_1, \ldots, y_n)$ and t, the conditional probability $P_\theta(\mathbf{Y} = \mathbf{y} | T(\mathbf{Y}) = t)$ is the same for all θ. But

$$\begin{aligned} P_\theta(\mathbf{Y} = \mathbf{y} | T(\mathbf{Y}) = t(\mathbf{y})) &= \frac{P_\theta(\mathrm{Y}{=}\mathrm{y} \text{ and } T(\mathrm{Y}){=}t(\mathrm{y}))}{P_\theta(T(\mathrm{Y}){=}t(\mathrm{y}))} \\ &= \frac{P_\theta(\mathrm{Y}{=}\mathrm{y})}{P_\theta(T(\mathrm{Y}){=}t(\mathrm{y}))} \\ &= \frac{f(\mathrm{y};\theta)}{q(t(\mathrm{y});\theta)}, \end{aligned}$$

where q is the pdf of $T(\mathbf{Y})$.

Example 9

Let $Y_1, \dots, Y_n$ be an iid random sample with $Y_i \sim \text{Ber}(\theta)$. We show that $T = Y_1 + \cdots + Y_n$ is sufficient for θ.

For, let $t = y_1 + \cdots + y_n$ and note that $T(Y) \sim \text{Bin}(n, \theta)$ and thus

$$P_\theta(\mathbf{Y} = \mathbf{y} | T(\mathbf{Y}) = t(\mathbf{y})) = \frac{\prod_i \theta_i^y (1-\theta)^{1-y_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Since this ratio doesn't depend on θ, $T(\mathbf{Y})$ is sufficient for θ.

The definition of sufficiency may be difficult to apply because:

- the computation of the conditional probability is tedious
- we may have no candidate statistic T in mind.

The Likelihood factorisation criterion is much easier:
A statistic $T(\mathbf{Y})$ is a sufficient statistic for θ, iff there is a function $g(t;\theta)$ and $h(\mathbf{y})$ such that, for all sample points $\mathbf{y}$ and all parameter points θ,

$$f(\mathbf{y};\theta) = g(T(\mathbf{y});\theta)h(\mathbf{y}).$$

Example 10

Let $\mathbf{Y}$ be a random sample with $Y_i \sim \text{Poi}(\lambda)$. The joint distribution of the sample is

$$f(\mathbf{Y}) = \prod_i \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = e^{-n\lambda}\lambda^{\sum_i y_i}\left(\prod_i y_i!\right)^{-1}.$$

From this we see that $g(t(\mathbf{y});\lambda) = e^{-n\lambda}\lambda^{t(\mathrm{y})}$, where $t(\mathbf{y}) = y_1 + \cdots + y_n$.

Thus, by the factorisation criterion, $T(\mathbf{Y}) = Y_1 + \cdots + Y_n$ is a sufficient statistic for λ

Note: sufficient statistics need not be unique. Indeed, if $T(\mathbf{Y})$ is sufficient and g is a bijective function (with suitable domain and codomain), then $g(T(\mathbf{Y}))$ is also sufficient.

Example 11

Suppose $Y_1, \dots, Y_n$ are iid uniform random variables on the interval $(\theta, \theta + 1)$, i.e. $Y_i \sim \text{Unif}(\theta, \theta + 1)$, $\theta \in \mathbb{R}$.
The joint pdf of $\mathbf{Y}$ is

$$f(\mathbf{y}) = \begin{cases} 1 & \text{if } \theta < y_i < \theta + 1, \quad i = 1, \dots, n, \\ 0 & \text{otherwise.} \end{cases}$$

For this joint to take value 1 it's sufficient that the minimum and the maximum are in the interval $(\theta, \theta + 1)$, so the joint pdf can be written as

$$f(\mathbf{y}) = \begin{cases} 1 & \text{if } \max_i y_i - 1 < \theta < \min_i y_i \\ 0 & \text{otherwise.} \end{cases}$$

or as

$$f(\mathbf{y}) = 1_{t_2(\mathrm{y}) - 1 < \theta < t_1(\mathrm{y})} = g(t; \theta),$$

where $t = (t_1, t_2) = (\min_i y_i, \max_i y_i)$ so by the factorisation criterion, $Y_{(1)}, Y_{(n)}$ is a sufficient statistic for θ.

Unbiasedness

Given $\widehat{\theta}$ an estimator of θ, based on a random sample $Y_1, \ldots, Y_n$ from some distribution F_θ, the bias is defined as

$$b(\theta; \widehat{\theta}) = E_\theta(\widehat{\theta}) - \theta.$$

Here E_θ is the expectation with respect to the distribution F_θ.

Furthermore, $\widehat{\theta}$ is an unbiased estimator of θ if $b(\theta; \widehat{\theta}) = 0$ for all θ.

Unbiased estimators are to be preferred, but many useful estimators have non-zero bias, which tend to 0 as n grows. These are called asymptotically unbiased estimators.

Example 12

For random iid sample $Y_1, \ldots, Y_n$, the sample average $\overline{Y}$ is an unbiased estimator for μ. Indeed, $b(\mu; \overline{Y}) = E(\overline{Y}) - \mu = 0$, for any μ. On the

other hand, the sample variance $\widehat{\sigma}^2$ is only an asymptotically unbiased estimator for σ^2, i.e. $E(\widehat{\sigma}^2) = (n-1)\sigma^2/n$.

Example 13

For a random iid sample $Y_1, \ldots, Y_n$, with $Y_i \sim \text{Unif}(0, \theta)$ both, the sample average $\overline{Y}$ and the maximum $Y_{(n)}$ are biased. However, $Y_{(n)}$ is

asymptotically unbiased; thus the maximum is to be preferred to the sample average.

Given the sufficiency, it is not surprising that $Y_{(n)}$ beats $\overline{Y}$.

MSE

The bias tells us on average by how much we would miss θ when the later is estimated by $\widehat{\theta}$; the lower the bias the better the estimator.

The bias thus is limited to the location of the distribution of $\widehat{\theta}$. An overall measure of performance of an estimator that takes its variability into account is the Mean Squared Error

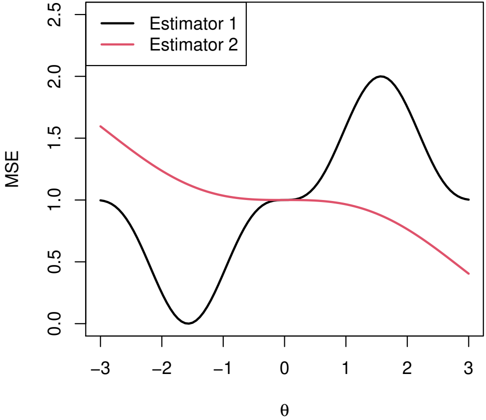
$$\text{mse}(\theta; \widehat{\theta}) = E_\theta(\widehat{\theta} - \theta)^2 = \text{var}(\widehat{\theta}) + (\text{bias}(\theta; \widehat{\theta}))^2.$$

There is nothing special about MSE, we could define our own measure if we wished to.

MSE is ≥ 0, is unbounded and the lower the MSE the better is the estimator.

In general, since MSE is a function of θ, there won't be a "best" estimator since the MSE will cross each other.
In the figure below Estimator1 is better only when for $\theta < 0$.

Two crossing MSE curves



Example 14

Consider the two estimators S^2 and S_b^2 for σ^2, with an iid random sample $Y_1, \ldots, Y_n$ with $Y_i \sim N(\mu, \sigma^2)$.

We have $\mathrm{var}(S^2) = \frac{2(n-1)}{n}$ so

$$\mathrm{mse}(\sigma^2; S^2) = \frac{2\sigma^4}{n-1}.$$

On the other hand, $\mathrm{var}(\widehat{\sigma}^2) = \frac{2(n-1)\sigma^2}{n^2}$, thus

$$\mathrm{mse}(\sigma^2; \widehat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{(n-1)\sigma^2}{n} - \sigma^2\right)^2 = \frac{(2n-1)}{n^2}\sigma^4$$

So $\mathrm{mse}(\sigma^2; \widehat{\sigma}^2) < \mathrm{mse}(\sigma^2; S^2)$.

Use it judiciously

This doesn't mean that we should abandon S^2, after all it is unbiased and MSE is only one way to measure the overall performance of an estimator.

Moreover, MSE penalises equally negative and positive biases. This may be fine for location parameters s.t. μ, but seems unfair for scale parameters s.a. σ^2 which are strictly positive.

Consistency

Intuitively, an estimator is consistent if it's distribution collapses to the true parameter value θ as n diverges.
Formally, an estimator $\widehat{\theta}$ based on a random sample $Y_1, \ldots, Y_n$ is consistent if it converges in probability to θ, the true parameter value, i.e. if for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\theta} - \theta| > \epsilon) = 0$$

To check for consistency one can also directly appeal to the following result. If

(1) $\lim_{n\to\infty} \text{bias}(\theta; \widehat{\theta}) = 0$

(2) $\lim_{n\to\infty} \text{mse}(\theta; \widehat{\theta}) = 0.$

then $\widehat{\theta}$ is consistent.

Example 15

If $Y_1, \ldots, Y_n$ is an iid random sample from $\text{Unif}(0, \theta)$. Then $\overline{Y}$ is not a consistent estimator for θ; though it is a consisten estimator for $\theta/2$.

If $Y_1, \ldots, Y_n$ is an iid random sample from $N(\mu, \sigma^2)$, then $\overline{Y}$ is a consistent estimator for μ.

Indeed, in this case we have

$$\text{bias}(\mu; \overline{Y}) = 0$$

and

$$\lim_{n \to \infty} \text{var}(\overline{Y}) = \lim_{n \to \infty} \sigma^2/n = 0.$$

Distribution of an estimator

So far we were concerned about only some fo the features of $\widehat{\theta}$, e.g. $E(\widehat{\theta})$and $\mathrm{var}(\widehat{\theta})$, but there's much more.

Indeed, the distribution of $\widehat{\theta}$, encapsulates all possible features of $\widehat{\theta}$ that we may ever need.

The distribution of $\widehat{\theta}$ is easy to derive only in simple problems but

in realistic scenarios, the computation of this distribution is tedious or even impossible and approximation methods must be used.

When exact derivation fails, there are two main lines of attack:

(a) Asymptotic approximations s.t. CLT, delta method, saddlepoint approximation, Edgeworth expansions, etc.

(b) Bootstrap or approximations via Monte Carlo simulations.

Example 16

Let $Y_1, \dots, Y_n$ be an iid random sample with $Y_i \sim \text{Poi}(\lambda)$, with $\lambda > 0$. The log-likelihood function is

$$\ell(\lambda) = -n\lambda + \sum_i y_i \log \lambda - \sum_i \log(y_i!).$$

The MLE is $\widehat{\lambda} = \overline{Y}$. Now, $n\widehat{\theta} = \sum_i Y_i$, thus

$$n\widehat{\theta} \sim \text{Poi}(n\lambda).$$

Thus although $\widehat{\theta}$ doesn't have a known distribution, its scaled version $n\widehat{\theta}$ has an (exact) Poisson distribution

Properties of the MLE

We focus now on the MLE, since it's the most widespread and study some of it's most important properties. For an iid random sample

$Y_1, \ldots, Y_n$ with $Y_i \sim F_\theta$ and pdf f satisfying certain regularity conditions:

(1) If there is a sufficient statistic for θ, then the MLE is a function of the sufficient statistic.

(2) The MLE is equivariant, i.e. if $\tau = g(\theta)$ for any function g, then $\widehat{\tau} = g(\widehat{\theta})$.

(3) The MLE is consistent, i.e. $\widehat{\theta} \xrightarrow{P} \theta$

(4) MLE is asymptotically efficient, roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large n.

(5) MLE is asymptotically normal, i.e. $(\widehat{\theta} - \theta)/\sqrt{\text{var}(\widehat{\theta})} \xrightarrow{d} N(0,1)$ for $n \to \infty$.

Regularity conditions

The regularity conditions need to prove the above properties are too technical for our purpose and not always easy to check. The most intuitive of them are the following two

(i) The parameter is identifiable, which means that if $\theta \neq \theta'$ then $f(y;\theta) \neq f(y;\theta')$.

(ii) The densities $f(y;\theta)$ have common support, and $f(y;\theta)$ is differentiable in θ.

Properties explained

Back to the properties of MLE, (1) tells essentially that if there is a sufficient statistic, then MLE will also be sufficient.

Indeed, if $T(\mathbf{Y})$ is a sufficient statistic, then

$$f(\mathbf{y};\theta) = g(T(\mathbf{y};\theta)h(\mathbf{y}),$$

Thus $\ell(\theta) = \log g(T(\mathbf{y};\theta)) + \text{const}$, so the likelihood depends on the data through t and so does its maximum, i.e. the MLE. For property (2), let g

be invertible, thus $\theta = g^{-1}(\tau)$, and so

$$L(\theta) = L(g^{-1}(\tau)).$$

This means that the likelihood as function of θ is identical to that of $g^{-1}(\tau))$.

Thus, certainly $L(\widehat{\theta}) = L(g^{-1}(\widehat{\tau}))$ and so

$$\widehat{\theta} = g^{-1}(\widehat{\tau}),$$

or $g(\widehat{\theta}) = \widehat{\tau}$

Example 17

Let $Y_1, \ldots, Y_n$ be an iid sample with $Y_i \sim \text{Poi}(\lambda)$. We want to estimate e^{λ}, the probability of observing zero counts.

First, let $\theta = e^{\lambda}$ and note that the MLE of λ is $\widehat{\lambda} = \overline{Y}$. By the equivariance principle (EP), then $\widehat{\theta} = e^{\overline{Y}}$.

Without using the EP, note that $\log\theta = \lambda$, thus the log-likelihood for θ is

$$\ell(\log\theta) = \ell(\lambda) = -n\log\theta + \log\log\theta \sum_i y_i - \sum_i y_i!.$$

Solving the likelihood equation $d\ell(\log\theta)/d\theta = 0$ gives the MLE of θ, i.e. $\widehat{\theta} = e^{\overline{Y}}$.

Property (5) tells us that the MLE has a central limit theorem kind of behaviour. In particular,

we have that

$$(\widehat{\theta} - \theta)/\sqrt{I_n(\theta)} \xrightarrow{d} N(0, 1),\ n \to \infty.$$

Other three equivalent results are

$$(\widehat{\theta} - \theta)/\sqrt{I_n(\widehat{\theta})} \xrightarrow{d} N(0, 1),\ n \to \infty,$$

$$(\widehat{\theta} - \theta)/\sqrt{J_n(\theta)} \xrightarrow{d} N(0, 1),\ n \to \infty,$$

and

$$(\widehat{\theta} - \theta)/\sqrt{J_n(\widehat{\theta})} \xrightarrow{d} N(0, 1),\ n \to \infty,$$

where J_n is the observed information and I_n is the Fisher information for the full sample. Typically, $\sqrt{I_n(\theta)}$ is called standard error or $\mathtt{se}$ for short; thus $\mathrm{se} = \sqrt{I_n(\theta)}$ and $\widehat{\mathrm{se}} = \sqrt{I_n(\widehat{\theta})}$ or the equivalent version based on J_n.

Fisher information

Under regularity conditions, the Fisher information is defined as

$$\begin{aligned} I_n &= \mathrm{var}(d\ell(\theta)/d\theta) \\ &= \sum_i \mathrm{var}(d \log f(Y_i;\theta) \\ &= nI_1(\theta), \end{aligned}$$

where I_1 is the Fisher information for a single observation.

Alternate formula for I_1 is

$$I_1(\theta) = -E\left(\frac{d^2 \log f(Y;\theta)}{d\theta^2}\right).$$

Example 18

Let $Y_1, \ldots, Y_n$ be an iid random sample from $\text{Poi}(\lambda)$. We saw in Example 16 that a scaled version of the MLE has exact Poisson distribution.

For large n we can use the limiting distribution of the MLE. Note that

$$I_1(\lambda) = E(Y_1/\lambda^2) = 1/\lambda.$$

Thus $\text{se}(\widehat{\lambda}) = \sqrt{n/\lambda}$ and $\widehat{\text{se}} = \sqrt{n/\overline{Y}}$.

So an asymptotic distribution for the MLE is

$$\frac{\sqrt{n}(\overline{Y}-\lambda)}{\sqrt{\overline{Y}}} \stackrel{\cdot}{\sim} N(0,1).$$

The symbol "$\stackrel{\cdot}{\sim} N(0,1)$" reads "approximately distributed as a standard normal for large n". We will see in L5 and L6 how this result is useful for doing useful work.

MLE in the multivariate case

The limiting normal distribution holds even for a vector-valued parameter.

Indeed, when $\theta \in \mathbb{R}^d$ and under the same regularity conditions (suitably adapted to this case), we have

$$I_n(\widehat{\theta})^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N_d(0, I)$$

and

$$J_n(\widehat{\theta})^{-1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N_d(0, I).$$

From these, we can derive similar results for any element of θ. For instance, the standard error for θ_i is $\text{se}(\widehat{\theta}_i) = \sqrt{J_n(\widehat{\theta})^{ii}}$ thus

$$(\widehat{\theta}_i - \theta)/\text{se}(\widehat{\theta}_i) \dot{\sim} N(0, 1).$$

Example 19

Let $0, 4, 5, 1, 1, 0, 1, 3, 0, 0, 2$ be an observed sample from the random sample in Example 16. Then $\widehat{\lambda} = 1.55$. The approximate large sample

distribution for the MLE can be obtained by the result

$$\frac{\sqrt{11}(\widehat{\theta}-\lambda)}{\sqrt{1.55}} \mathrel{\dot\sim} N(0, 1)$$

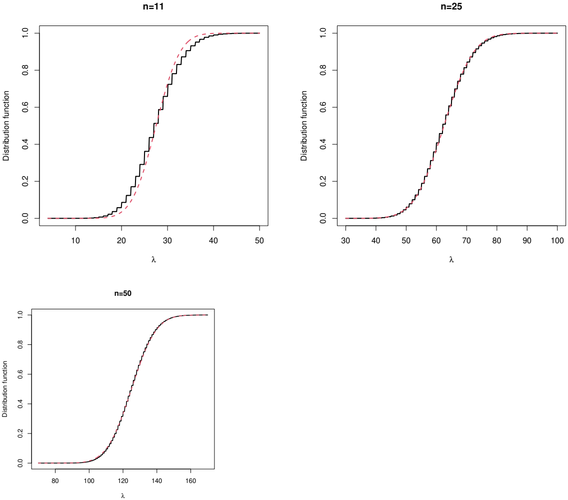
in which we replace $\dot\sim$ by $\sim$. But $\frac{\sqrt{11}(\widehat{\theta}-\lambda)}{\sqrt{2}} \sim N(0, 1)$ implies that

$$\widehat{\theta} \sim N(\lambda, 1.55/11).$$

Thus the MLE has an approximate normal distribution with mean λ and variance $2/11$. Note that this distribution is only an approximation to the true pdf of $\widehat{\theta}$ and the larger n the better it is...

Example 19 (cont'd)

The Figure shows the exact df of $n\widehat{\theta}$ (black) against the asymptotic df (assuming true $\lambda = 2.5$.)


n=11
Distribution function
λ
n=25
Distribution function
λ
n=50
Distribution function
λ