

# Inferential Statistics

## L4 - Point estimation

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

# Contents

- 1 Statistics
- 2 Methods for computing estimators
- 3 Methods for evaluating estimators
- 4 Further properties: Asymptotics

# Overview

Suppose  $Y_1, \dots, Y_n$  is a random sample with  $Y_i \sim F_\theta$  and,

Nature picks  $\theta = \theta_0$  (secretly) and uses it generate the observed sample  $y_1, \dots, y_n$  from the above random sample.

With this observed sample at hand, one of the aims of statistics is to guess  $\theta_0$ .

Such a guess is called an estimate of the unknown parameter  $\theta_0$ . In this lecture we'll study methods estimating a parameter.

In the first part of this lecture we will see what properties we wish our estimates should satisfy. In the second part we will see methods for building such estimates.

# Statistics (dejavu')

Let  $Y_1, \dots, Y_n$  be a random sample with  $Y_i \sim F_\theta$ , with pdf  $f_\theta$  and unknown parameter  $\theta$ .

If  $T_n = T(Y_1, \dots, Y_n)$ , with  $T_n : \mathbb{R}^n \rightarrow \mathbb{R}^d$  doesn't depend on any unknown quantity, then it is called a statistic.

All summary statistics we saw in L2 are all examples of statistics.

In L2 we didn't pay much attention, but  $T_n$  is a rve, thus it has a df.

The point is that, when  $T_n$  is chosen with care, it reveals us something useful about  $\theta$ .

## Example 1

For the iid random sample  $Y_1, \dots, Y_n$ , assume that  $E(X_i) = \mu$ , and  $\text{var}(X_i) = \sigma^2$ . Then, the sample average  $\bar{Y} = (Y_1 + \dots + Y_n)/n$  is a statistic and

$$E(\bar{Y}) = E((Y_1 + \dots + Y_n)/n) = n^{-1}E(Y_1 + \dots + Y_n) = \mu,$$

and

$$\text{var}(\bar{Y}) \stackrel{\text{expand sum and squares, covariance is 0}}{=} \sigma^2/n$$

Thus if we are interested in learning the expected value of a population, i.e.  $\theta = \mu$ , then the sample average is a good candidate.

Furthermore, let  $Y_i \sim N(\mu, \sigma^2)$ , then we can show that

because it's a  
linear  
transformation

$$\bar{Y} \sim N(\mu, \sigma^2/n) \quad \text{or} \quad \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\sigma^2}} \sim N(0, 1).$$

Note that, because  $\mu, \sigma^2$  are unknown,  $\frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\sigma^2}}$  is not a statistic  
standard deviation

If this doesn't make much sense to you, let's make it more concrete. Assume that  $F_\theta$  is discrete and  $Y$  can assume values in  $\{1, 2, 3\}$  with equal probability; so  $\mu = 2$ .

Let  $n = 2$ , and consider  $Y_1, Y_2$  iid sample from  $F_\theta$ . The possible observed samples are

1, 1; 1, 2; 1, 3; 2, 1; 2, 2; 2, 3; 3, 1; 3, 2; 3, 3.

Using the distribution of the sample averages (below) we find that the average of the sample averages is

$$\underline{E(\bar{Y}) = 1 \cdot \frac{1}{9} + 1.5 \cdot \frac{2}{9} + 2 \cdot \frac{3}{9} + 2.5 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} = 2 = \mu.}$$

$\bar{Y}$	$P(\bar{Y} = k)$
1	1/9
1.5	2/9
2	3/9
2.5	2/9
3	1/9

# Average of sample variance = population variance

## Example 2

slightly different from  $S^2$  because we divide by  $n$  instead of  $n-1$

Under the assumptions of Example 1, let  $\hat{\sigma}^2 = n^{-1} \sum_i (Y_i - \bar{Y})^2$  be (a version of) the sample variance. Then

*linearity of expectation and i.i.d. sum and subtract  $\mu$*

$$E(\hat{\sigma}^2) = E[(Y_1 - \bar{Y})^2] = \frac{n-1}{n} \sigma^2.$$

*Controlly!*

*$E[\frac{1}{n} \sum_i (Y_i - \bar{Y})^2]$*

On the other hand for the sample variance  $S^2 = (n-1)^{-1} \sum_i (Y_i - \bar{Y})^2$ , we have

$$E(S^2) = E[n\hat{\sigma}^2/(n-1)] = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2.$$

This is the reason why we defined  $S^2$  dividing by  $n-1$ . It can be show that

$$\text{var}(S^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3},$$

where  $\mu_k = E(Y_1^k)$ , is the  $k$ th moment of  $Y_1$ .

In general  $\bar{Y}$  and  $S^2$  are not independent, except if  $Y_i \sim N(\mu, \sigma^2) \dots$

## Example 2 (cont'd)

Indeed, let  $Y_i \sim N(\mu, \sigma^2)$ . Then  $\bar{Y}$  and  $S^2$  are independent and

$$\text{var}(\chi^2_{n-1}) = 2(n-1) \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

It follows that  $\text{var}(S^2) = 2(\sigma^2)^2/(n-1)$ . Furthermore, combining this with the results of Example 1, we have that **def of student t**

$$\frac{\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \sim t_{n-1}.$$

Quantities s.t.  $(n-1)S^2/\sigma^2$  or  $\sqrt{n}(\bar{Y} - \mu)/S$ , which are not statistics but have a known distribution, are called pivotal quantities.



They play a key role in the development of confidence intervals and hypothesis testing as we will see in L5 and L6.



### Example 3

Let  $Y_1, \dots, Y_n$  be an iid random sample with  $Y_i \sim \text{Unif}(0, \theta)$ , with  $\theta > 0$ . For the statistics  $\bar{Y}$  and  $Y_{(n)}$ , let's compute some of their features.

For the sample average we have

 **example 1** 

$$E(\bar{Y}) = E(Y_1) = \theta/2, \quad \text{var}(\bar{Y}) = E((\bar{Y})^2) - E(\bar{Y})^2 = \theta^2/(12n).$$

For the maximum we have

$$\begin{aligned} E(X_{(n)}) &= \int_{-\infty}^{+\infty} t f_{X_{(n)}}(t) dt = \int_0^{\theta} t n F(t)^{n-1} f(t) dt \\ &= \int_0^{\theta} t n (t/\theta)^{n-1} (1/\theta) dt = \theta n / (n + 1), \end{aligned}$$

and

$$\text{var}(Y_{(n)}) = \theta^2 n / (n + 1)^2.$$

Note that in the case of  $Y_{(n)}$  we had to use it's pdf in order to compute the two moments.

Some statistics target  $\theta$  the parameter of a distribution (or a component of it).

In that case, we call them estimators and denote them by a Greek letter with a hat, e.g.  $\hat{\theta}$ .

An estimator thus is a function of the random sample, and can tell us something useful about the parameter  $\theta$ .

For example, the sample average is useful when we want to learn about the population average  $\mu$ , the sample median  $Q_2$  is useful for learning about the population median and so on.

We now look at methods for computing estimators and then we will see methods for comparing estimators.

# Method of Moments

Useful when we want to estimate  $\theta$  that can be expressed as a function of moments of  $Y$ .

Is one of the oldest statistical estimation methods, dating back to 1936.

The method consists in equating sample moments, e.g.  $\overline{Y}$ ,  $\overline{Y^2}$ ,  $\overline{Y^3}$ , ... with the corresponding population moments, e.g.  $E(X)$ ,  $E(X^2)$ ,  $E(X^3)$  and solving these equations in terms of the parameter  $\theta$ .

### Example 4

Suppose  $Y_1, \dots, Y_n$  is an iid sample from some distribution  $F_\theta$  and let  $E(Y_1) = \mu$  be the unknown parameter.

Equating the sample moment with the corresponding population moment leads to

$$\overline{Y} = E(Y) = \mu.$$

So  $\overline{Y}$  is the method of moment estimator for  $\mu$ , i.e.  $\hat{\mu}_{MM} = \overline{Y}$  (reads: the method of moments estimator for  $\mu$  is  $\overline{Y}$ ).

## Example 5

Let  $Y_1, \dots, Y_n$  be an iid random sample with  $E(Y_1) = \mu$  and  $\text{var}(Y_1) = \sigma^2$ , with  $\mu, \sigma^2$  unknown.

First note that  $\sigma^2 = E(Y^2) - E(Y)^2 = E(Y^2) - \mu^2$ . Equating the first two moments leads to

$$\begin{aligned}\bar{Y} &= E(Y) = \mu, \\ E(Y^2) &= \sigma^2 + \mu^2 = \overline{Y^2}.\end{aligned}$$

We conclude that

$$\hat{\sigma}_{MM}^2 = \overline{Y^2} - \bar{Y}^2 = \hat{\sigma}^2,$$

(we have encountered this estimator previously.) and  $\hat{\mu}_{MM} = \bar{Y}$ .

So  $(\bar{Y}, \hat{\sigma}^2)$  is the method of moments estimator for  $(\mu, \sigma^2)$ .

# A signal+noise problem

In many situations, we have measurements  $Y_i$ , which can be thought of as the sum of a physical signal  $g_i(\beta)$  and a noise  $\epsilon_i$ , i.e.

$$Y_i = g_i(\beta) + \epsilon_i,$$

where  $\epsilon_i$  is some unknown component whereas  $g_i(\beta)$  is the signal, which may depend on some unknown parameter  $\beta$ .

For instance, when measuring the resistance of an electronic circuit with a multimeter, we observe a realisation of  $Y_i$ , say  $y_i = 12 \Omega$ .

This value depends, in part, on the real resistance of the equipment  $g_i(\beta)$  (e.g. the sum of the resistance of all its component) and for the other part, on the accuracy of the instrument; see also L2 for another example.

# Linear regression

The signal  $g_i(\beta)$  may depend on some known features  $x_{i1}, \dots, x_{ip}$ , through the linear function

$$g_i(\beta) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \quad i = 1, \dots, n,$$

with unknown parameters  $\beta = (\beta_0, \dots, \beta_p)$ .

The aim is to understand the impact of  $x_{ij}$ 's on  $Y_i$ , i.e.  $\beta_i$ .

We have the pair  $y_i, g_i(\beta)$ , where  $y_i$  is what we measure and  $g_i(\beta)$  is how the system should behave according to modeller's view.

This is commonly known as a linear regression problem, and one of the points is how to estimate  $\beta$  using  $y_i, g_i(\beta)$  for all  $i$ .

# Method of Least Squares

For any fixed  $\beta$ , the deviances  $y_i - g_i(\beta)$ , tell's us by how much our model  $g_i(\beta)$  misses the observed value. It seems intuitive then to look for a  $\beta$  that leads to smallest deviances.

The method of least squares consists in estimating  $\beta$  by the value that leads to smallest sum of squared deviances.

That is, the least squares estimator is defined by

$$\hat{\beta}_{LS} = \arg \inf_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - g_i(\beta))^2.$$



## Example 6

Let  $y_1, \dots, Y_n$  be counts of bacteria in a culture of cells, measured at time points  $t_1, \dots, t_n$ . Aim: study bacteria growth rate.

A possible model for this problem is

$$\begin{aligned} Y_i &= g(t_i; \beta) + \epsilon_i, \\ g(t_i; \theta) &= \beta_0 + \theta_1 t_i, \quad i = 1, \dots, n, \end{aligned}$$

where  $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$  is unknown.

In words: (we assume) the counts follow a linear equation with time, and we wish to learn about the parameters of this line.

To find the LS estimator we solve in  $\beta_0, \beta_1$  the system

$$\frac{d}{d\beta} \sum_{i=1}^n (y_i - g_i(\theta))^2 = 0,$$

## Example 6 (cont'd)

This system is

$$\begin{aligned}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1}) &= 0 \\ \sum_{i=1}^n x_{i1} (y_i - \beta_0 - \beta_1 x_{i1}) &= 0\end{aligned}$$

and the solution is

$$\hat{\beta}_0 = \bar{y} - \frac{s_{y,x}}{s_x^2} \bar{x},$$

and

$$\hat{\beta}_1 = \frac{s_{y,x}}{s_x^2}.$$

The vector  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$  as above is the least squares estimator for  $\beta$ .

# Method of Maximum Likelihood

Let  $Y_1, \dots, Y_n$  be an iid random sample from with  $Y_i \sim F_\theta$  and pdf  $f$  and unknown parameter  $\theta$ .

The Maximum Likelihood Estimator (MLE) is defined by

$$\hat{\theta} = \arg \sup_{\theta \in \Theta} L(\theta).$$

Under standard regularity conditions, the MLE is also defined as the solution to the likelihood equation

$$\frac{d\ell(\theta)}{d\theta} = 0.$$

Note that  $\theta$  may be  $d$ -dimensional vector, in which case the likelihood equation consists in  $d$  simultaneous equations.

## Example 7

Let  $Y_1, \dots, Y_n$  be an iid random sample with  $Y_i \sim \text{Ber}(\theta)$ , with  $\theta$  unknown. Let's compute the maximum likelihood estimator for  $\theta$ .

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \\ &= \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i}, \end{aligned}$$

(regularity conditions holds)

Since  $L(\theta) > 0$ , for all  $\theta$ , we can apply the log to get the log-likelihood function. So we solve the likelihood equation  $d\ell(\theta)/d\theta = 0$ , i.e.

$$\frac{\sum_i y_i}{\theta} - \frac{n - \sum_i y_i}{1 - \theta} = 0,$$

to get the solution  $\hat{\theta} = \bar{y}/n$ .  MLE of  $\theta$  is  $\bar{Y}$

Furthermore,  $d^2\ell(\theta)/d\theta^2$  at  $\theta = \hat{\theta}$  is negative, so  $\hat{\theta}$  is a local maximum, thus it's the MLE of  $\theta$ . In this case, the MLE coincides with the MME.

we should also check the second derivative

## Example 8

we won't need the product of densities

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.  $\hat{\theta}_i = y_i/n$  for all  $i$ . This is the MLE of  $\theta$ .

in this case pdf( $y_1, y_2, y_3$ ) is the same as  $L(\theta)$  depending on which parameters are fixed

Indeed, the log-likelihood is

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

(doesn't matter for argmax)

## Example 8 (cont'd)

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

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  $\theta_i$  and solving the equations leads to

$$\theta_i = -y_i / \lambda. \quad \text{estimator}$$

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  $\hat{\theta} = y_i / n$  as conjectured. The estimated cell probabilities are thus

(0.332, 0.486, 0.182) **maximum likelihood estimate**

**we should also check the hessian/-information matrix**

# Methods for evaluating estimators: Sufficiency

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

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

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

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

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  $T = Y_1 + \dots + Y_n$  is sufficient for  $\theta$ .

For, let  $t = y_1 + \dots + y_n$  and note that  $T(Y) \sim \text{Bin}(n, \theta)$  and thus (sum of n iid bernoulli is binomial)

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

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



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  $\theta$ , iff there is a function  $g(t; \theta)$  and  $h(\mathbf{y})$  such that, for all sample points  $\mathbf{y}$  and all parameter points  $\theta$ ,

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

## Example 10

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

$$f(\mathbf{Y}) \stackrel{\text{prod of marginals}}{=} \prod_i \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} = \overbrace{e^{-n\lambda} \lambda^{\sum_i y_i}}^g \overbrace{\left( \prod_i y_i! \right)^{-1}}^h.$$

From this we see that  $g(t(\mathbf{y}); \lambda) = e^{-n\lambda} \lambda^{t(\mathbf{y})}$ , 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  $\lambda$

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

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

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  $\theta$ .

# Unbiasedness

Given  $\hat{\theta}$  an estimator of  $\theta$ , based on a random sample  $Y_1, \dots, Y_n$  from some distribution  $F_\theta$ , the bias is defined as

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

Here  $E_\theta$  is the expectation with respect to the distribution  $F_\theta$ .

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

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  $Y_1, \dots, Y_n$ , the sample average  $\bar{Y}$  is an unbiased estimator for  $\mu$ . Indeed,  $b(\mu; \bar{Y}) = E(\bar{Y}) - \mu = 0$ , for any  $\mu$ . On the

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

### Example 13

For a random iid sample  $Y_1, \dots, Y_n$ , with  $Y_i \sim \text{Unif}(0, \theta)$  both, the sample average  $\bar{Y}$  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  $\bar{Y}$ .

# MSE

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

latter

The bias thus is limited to the location of the distribution of  $\hat{\theta}$ .

An overall measure of performance of an estimator that takes its variability into account is the Mean Squared Error

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

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

*if this were  $E[\hat{\theta}]$  this would be  $\text{var}(\hat{\theta})$*

$$\star E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 = E[(\hat{\theta} - E(\hat{\theta}))^2 + b(\theta, \hat{\theta})^2 + 2(\hat{\theta} - E(\hat{\theta}))b(\theta, \hat{\theta})]$$

$$= \underbrace{\text{var}_{\theta}(\hat{\theta})}_{1} + \underbrace{b(\theta, \hat{\theta})^2}_{2} + \underbrace{\sigma^2}_{3} \quad \text{where } b(\theta, \hat{\theta}) = E(\hat{\theta}) - \theta$$

MSE is  $\geq 0$ , is unbounded and the lower the MSE the better is the estimator.

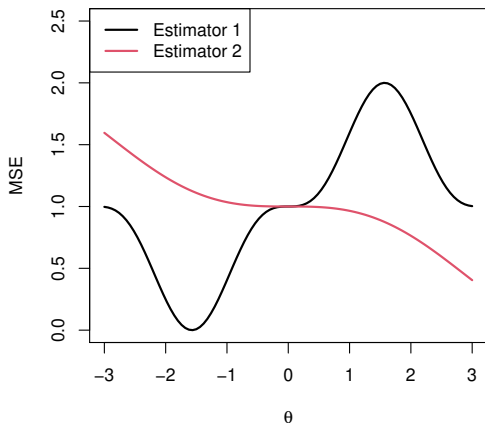
(became  $\text{var}_{\theta}(\hat{\theta})$  depends on  $\theta$ )

In general, since MSE is a function of  $\theta$ , 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



estimator 1 is better for  
 $\theta < 0$ , estimator 2 is  
better for  $\theta > 0$   
(in practice this is rare)

## Example 14

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

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

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

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

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

So  $\text{mse}(\sigma^2; \hat{\sigma}^2) < \text{mse}(\sigma^2; S^2)$ .

$$\sigma^2 = \frac{n-1}{n} S^2$$

$$\text{mse}(\sigma^2, S^2) = \text{var}(S^2) + b(\sigma^2, S^2)^2 = \frac{2\sigma^4}{n-1} + 0$$

$$\text{mse}(\sigma^2, \hat{\sigma}^2) = \text{var}(\hat{\sigma}^2) + b(\sigma^2, \hat{\sigma}^2)^2$$

$$= \text{var}\left(\frac{n-1}{n} S^2\right) + \frac{n-1}{n} \sigma^2 \cdot \sigma^2$$

$$= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} + \dots$$

# 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.  $\mu$ , but seems unfair for scale parameters s.a.  $\sigma^2$  which are strictly positive.

# Consistency

Intuitively, an estimator is consistent if it's distribution collapses to the true parameter value  $\theta$  as  $n$  diverges.

Formally, an estimator  $\hat{\theta}$  based on a random sample  $Y_1, \dots, Y_n$  is consistent if it converges in probability to  $\theta$ , the true parameter value, i.e. if for every fixed  $\epsilon > 0$ ,

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

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

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

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

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

### Example 15

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

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

Indeed, in this case we have

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

and

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

# Distribution of an estimator

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

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

The distribution of  $\hat{\theta}$  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. *→ output is a function*

(b) Bootstrap or approximations via Monte Carlo simulations.  
*→ output are numbers  
(and then maybe hint etc...)*

## 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  $\hat{\lambda} = \bar{Y}$ . Now,  $n\hat{\theta} = \sum_i Y_i$ , thus

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

Thus although  $\hat{\theta}$  doesn't have a known distribution, its scaled version  $n\hat{\theta}$  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

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

- (1) If there is a sufficient statistic for  $\theta$ , then the MLE is a function of the sufficient statistic. *(and MLE will be sufficient) ??? CHIEDI ???*
- (2) The MLE is equivariant, i.e. if  $\tau = g(\theta)$  for any function  $g$ , then  $\hat{\tau} = g(\hat{\theta})$ .
- (3) The MLE is consistent, i.e.  $\hat{\theta} \xrightarrow{P} \theta$
- (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.  $(\hat{\theta} - \theta) / \sqrt{\text{var}(\hat{\theta})} \xrightarrow{d} N(0, 1)$  for  $n \rightarrow \infty$ .



# 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  $\theta$ .

# 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

(2 is also true for non invertible g)

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

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

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

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

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

## Example 17

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

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

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

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

$\Downarrow$   
 $\ell^\lambda(\tau)$

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

$$\frac{d\ell^\lambda(\tau)}{d\tau} = -\frac{n}{\tau} + \sum_i y_i \frac{1}{\log \tau} \cdot \frac{1}{\tau}$$
$$\frac{d\ell^\lambda(\tau)}{d\tau} = 0 \dots \hat{\tau} = e^{\bar{Y}} \text{ much easier with EP! } \triangle!$$

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

we have that

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

*removed the / in all the rows*  
*any  $\theta$  (function)*

Other three equivalent results are

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

*also  $\hat{\theta}$  (specific point)*

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

*easier to compute  $J_n$  than  $I_n$*

and

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

*function for  $\theta$  point in  $\hat{\theta}$*

where  $J_n$  is the observed information and  $I_n$  is the Fisher information for the full sample. Typically,  $1/\sqrt{I_n(\theta)}$  is called standard error or se for short; thus  $se = 1/\sqrt{I_n(\theta)}$  and  $\hat{se} = 1/\sqrt{I_n(\hat{\theta})}$  or the equivalent version based on  $J_n$ .

# Fisher information

Under regularity conditions, the Fisher information is defined as

$$\begin{aligned} I_n &= \text{var}(d\ell(\theta)/d\theta) \\ &= \sum_i \text{var}\left(\frac{d \log f(Y_i; \theta)}{d\theta}\right) \leftarrow \text{converted with derivative} \\ &= nI_1(\theta), \end{aligned}$$

where  $I_1$  is the Fisher information for a single observation.

Alternate formula for  $I_1$  is

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

sometimes its faster  
to compute this way

## Example 18

Let  $Y_1, \dots, Y_n$  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.

HW fallo con la var

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

$$l_1(\lambda) = E(Y_1/\lambda^2) = 1/\lambda.$$

$$\log l(\theta_1, \lambda) = \dots = -\lambda + y_1 \log \lambda$$

$$\frac{\partial \log l(\theta_1, \lambda)}{\partial \lambda} = -1 + \frac{y_1}{\lambda}$$

$$\frac{\partial \log l(y_1, \lambda)}{\partial \lambda^2} = -\frac{y_1}{\lambda^2}$$

Thus  $\text{se}(\hat{\lambda}) = \sqrt{\lambda/n}$  and  $\hat{\text{se}} = \sqrt{\bar{Y}/n}$

So an asymptotic distribution for the MLE is

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

$$-E\left[-\frac{Y_1}{\lambda^2}\right] = \frac{E(Y_1)}{\lambda^2} = \frac{1}{\lambda} = I_1(\lambda)$$

$$\text{se} = \frac{1}{\sqrt{I_1(\lambda)}} = \frac{1}{\sqrt{n I_1(\lambda)}} = \sqrt{\frac{\lambda}{n}}$$

The symbol “ $\sim N(0, 1)$ ” 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.

$$\text{se}(\hat{\lambda}) = \sqrt{\frac{\bar{X}}{n}} = \sqrt{\frac{\bar{Y}}{n}}$$

$n \rightarrow +\infty$

# 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

$d \times d$  matrix  $\longrightarrow J_n(\hat{\theta})^{-1/2}(\hat{\theta} - \theta) \xrightarrow{d} N_d(0, I)$  identity matrix

and

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

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

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

## 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  $\hat{\lambda} = 1.55$ . The approximate large sample distribution for the MLE can be obtained by the result

$$\frac{\sqrt{11}(\hat{\lambda} - \lambda)}{\sqrt{1.55}} \overset{\sim}{\sim} N(0, 1)$$

in which we replace  $\overset{\sim}{\sim}$  by  $\sim$ . But  $\frac{\sqrt{11}(\hat{\lambda} - \lambda)}{\sqrt{1.55}} \sim N(0, 1)$  implies that

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

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



## Example 19 (cont'd)

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

