GM – Probabilité et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Lecture 6

- Sampling
- Methods of *point* estimation
  - method of maximum likelihood (MLE)
- Properties of MLEs
- (Statistical / Fisher) Information
- (Asymptotic) confidence intervals

Probability vs. statistics

- For a *known value* of p, we can calculate the probability of any possible result
- This is *probability*
- In many practical situations though, we do not know the value of p, but instead have some data that will be used to *estimate* the value of p
- This is *statistics*

Sampling

- The goal of a statistical study is to obtain some knowledge about a *population*, that is, **estimation of a parameter**
- Since a complete enumeration of the population is rarely practical, we need other, more practical means
- $\Rightarrow$ **Sampling** consists of choosing among the population a certain number of individuals ('sampling units') for which we will obtain observations (data)
- Our data can be considered as arising from a *random process* : if we were to repeat the collection of the data, the results would be different, and this could influence any conclusions drawn based on the data
- That is, our conclusions are subject to *random variation*

Utility of sampling

- A gardener has two million practically identical seeds, some of which produce white flower and some of which produce red flowers
- He would like to know in advance *the percentage* of white flowers des fleurs blanches (so that he can sell them without deceiving his clients)
- If he wants to be *absolutely certain* which color flowers are produced, he would need to plant *every seed*
- **And thus he would have no more seeds to sell ! !**
- $\Rightarrow$ A **sample** is necessary
- (Even when the process is not destructive, it is most often *impossible or infeasible* (time, cost) to measure every individual from the population)

Representativeness

- Based on his observations, the gardener could make an *estimate* of the number of white/red flowers among the two million seeds
- $\Rightarrow$ Then *generalize* to the *population* the knowledge acquired on the basis of *some observations*
- We cannot be *absolutely certain* of our prediction, since we only consider a fraction of the total population : $\Rightarrow$ Imprecision due to sampling
- Generally, there will be a *deviation* between the observations obtained from the sample and those from the *totality* of the population
- But : if the sample is chosen in a *scientific manner*, it is possible to make a *probabilistic* evaluation
- $\Rightarrow$ Possible to *evaluate the error*, and determine the *precision of the estimation*

Sampling methods

Arbitrary sampling

- Impossible to quantify the associated probabilities, thus difficult to estimate parameters and standard deviation of the estimation (standard error, SE)
- For example, the first ten people to enter the room
- $\Rightarrow$ **NOT recommended ! !**

Random sampling

- Corresponds to methods of sampling draws where each unit in the population has *a positive, known probability* of being chosen
- These methods permit population parameter estimation, and also allow us to obtain a measure of the SE
- For us, the most important methods correspond to sampling WITH replacement (independent), and sampling WITHOUT replacement (simple random sampling (SRS) / échantillonnage aléatoire simple (EAS))

Estimation

- The procedure of using sampling information that permits us to make deductions regarding the entire population is called **estimation**
- The *unknown* population value (to estimate based on the sample) is called a **parameter**
- For example : the mean (μ) ; the proportion (or percentage) (p)
- The population parameter is estimated by a **statistic** calculated based on the sample
  $\Rightarrow$ a statistic is *a function of the data*
- An **estimator** is a statistic used to estimate (guess the value of) a parameter θ ; that is, it is a rule that lets us calculate an approximation of θ based on the sample values $X_1, \ldots, X_n$
- An **estimate** is a value observed (calculated) of the estimator for a sample

Quality of an estimator

- To answer the question : 'how to choose between candidate estimators', we should examine what makes a 'good' estimator
- Thus, we consider the (statistical) qualities of the estimators
- Some important qualities :
  - bias
  - variance
  - mean square error (MSE) / erreur quadratique moyenne (EQM)
  - consistent (tends toward the correct result)
  - robustness

Bias

- The **bias** of an estimator T of the parameter θ is defined by :

$$b(T) = E[T] - \theta,$$

(that is, the difference between the *expected value* of the sampling distribution of the estimator T and the *true value* of the parameter θ)

- An estimator is **unbiased**) if the bias equals 0

Example7.1 What is the bias of the estimator $\overline{X}$ for the population mean μ **...**

Variance

- Another quality that we can consider is the *variance* of the estimator :

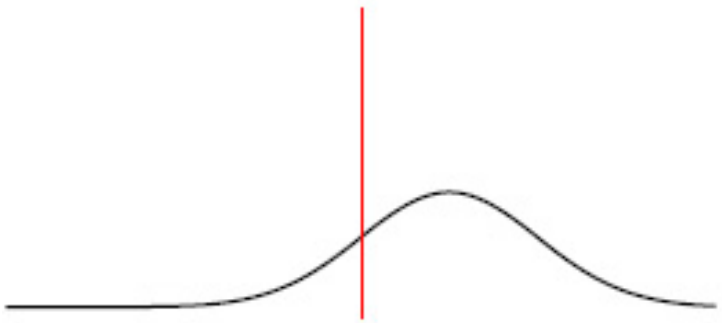
$$Var(T) = E[(T - E[T])^2]$$

- Among two unbiased estimators of θ, one is *more efficient* than the other if *its variance is smaller*

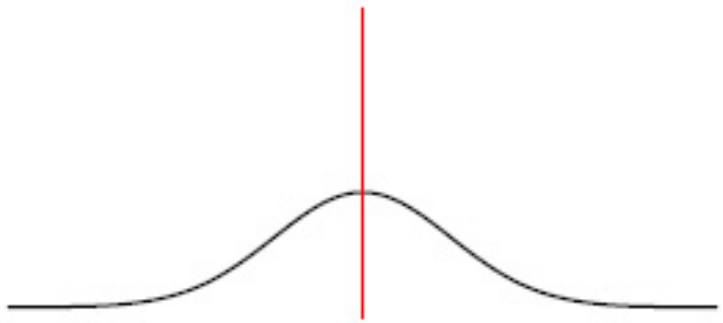
Example 7.2 Now let's consider the variance of some candidate estimators of the population mean μ **...**

Bias and variance of an estimator T

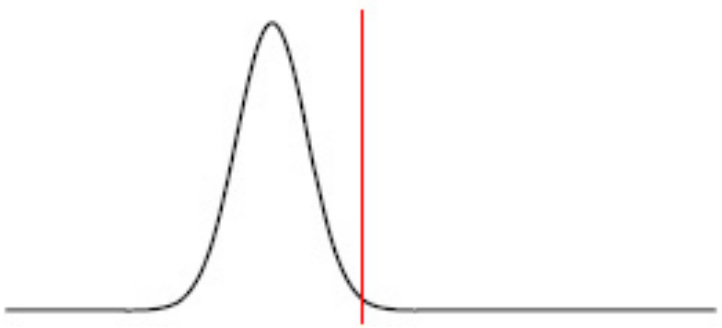
big bias, big variance



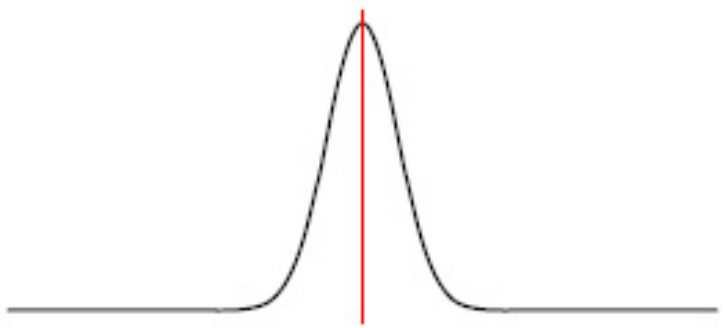
no bias, big variance



big bias, low variance



no bias, low variance



Mean Square Error (MSE)

- Another quality that we can consider is the **mean square error (MSE) / erreur quadratique moyenne (EQM)** of an estimator

$$EQM(T) = E[(T - \theta)^2]$$

- This is different from the variance when the estimator T is biased
- Sometimes we would like to use a slightly biased estimator if its variance is much smaller than thet best unbiased estimator (*bias-variance tradeoff*)
- It is simple to demonstrate that the MSE can be expressed as a combination of the bias and variance :

$$EQM(T) = Var(T) + [b(T)^2]$$

Methods of point estimation

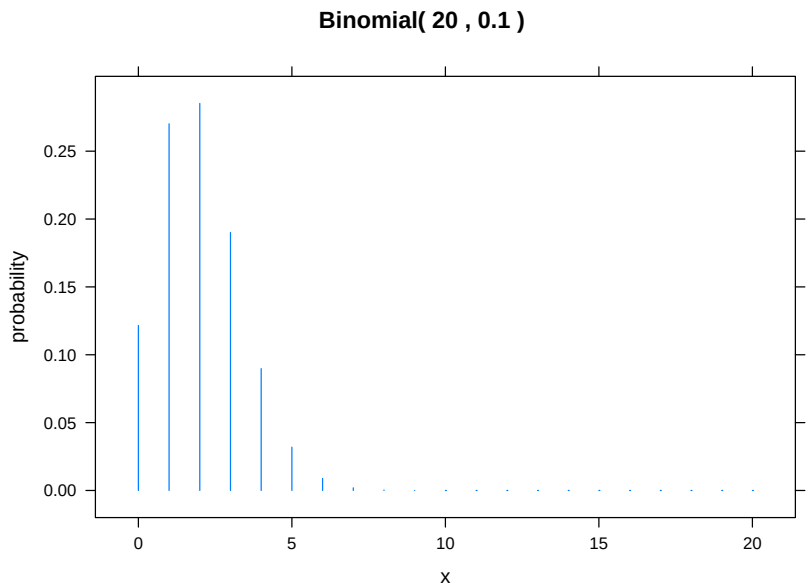
- *Method of maximum likelihood* (which often produces 'intuitive' estimators)
- *Method of moments* – will be illustrated on (VIDEO ONLY), but
  ⇒ **willt NOT be part of the exam**
- *Method of least squares* (later, with 'regression')
- Method of minimum absolute deviations
- Bayesian estimation

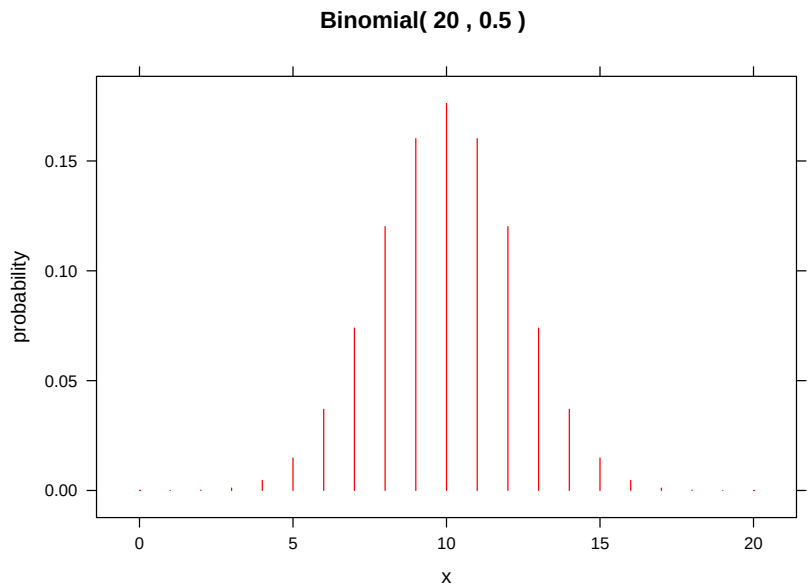
Likelihood

- For a known value p, we can express the probability of any set of possible data
- On the other hand, we can *consider the observations as known* and consider the probability as a function of the unknown parameter p
- The probability function viewed in this way is called the **likelihood**

Likelihood illustrated

20 tosses of a coin ; we observe *??* Heads





Of these two distributions, from which is it more likely that the sample came ?

Definition of the likelihood

- **Definition** : Let $X \sim f(x;\theta)$. The **likelihood** and **log likelihood** are :

$$L(\theta) \propto f(x;\theta), \quad \ell(\theta) = \log L(\theta),$$

considered as functions of θ for given x.

- Let $x = (x_1, \ldots, x_n)$ be a realization of *iid* RVs $X_1, \ldots, X_n$. Then

$$L(\theta) = f(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta),$$

where $f(x_i;\theta)$ is the law/density of x_i.

- **NOTE** : log = log base e = **natural log**.

Estimation by maximum likelihood

- An intuitive estimation method is **maximum likehihood estimation**
- For example, the most 'obvious' estimator of p is $\hat{p} = X/n$ turns out to be **the maximum likelihood estimator (MLE)**
- In general, the MLE is the value that makes the probability as large as possible – it's the value that makes *the observed data most probable*
- The usual way of finding the MLE : using calculus – find the derivative of the (log) likelihood function, set equl to zero and solve :

$$\frac{d \log L(\hat{\theta})}{d\theta} = 0, \quad \frac{d^2 \log L(\hat{\theta})}{d\theta^2} < 0$$

- (This method does not always work)
- We are supposing that the first equation has a *unique* solution (not alway true though)

MLE, cont

- The MLE $\hat{\theta}$ satisfies the condition

$$L(\hat{\theta}) \geq L(\theta) \quad \text{for all } \theta,$$

  which is equivalent to $\log L(\hat{\theta}) \geq \log L(\theta)$, $L(\hat{\theta})$ and $\log L(\hat{\theta})$ are obtained at the same (maximizing) value $\hat{\theta}$
- The MLE can :
  - exist and be unique
  - not be unique, or
  - not exist
- In practice, it is typically necessary to use numerical algorithms to obtain $\hat{\theta}$ and $d^2 \log L(\hat{\theta})/d\theta^2$

Advantages/disavantages of the method

- For a 'sufficiently large' sample size, the MLE is :
  - unbiased
  - consistent
  - efficient (minimal MSE ; thus at least as powerful as MOM)
  - *Normally distributed*
  - therefore practical for statistical inference
- On the other hand, the MLE :
  - can be very biased is the sample size is small
  - can be very complicated to evaluate (possibly necessary to evaluate numerically)

PAUSE

Example

Example 7.3 Let $X \sim Bin(n, p)$. Find the MLE of p **...**

Example

Example 7.4 Let $X_1, \ldots, X_n \sim$ iid $Pois(\lambda)$, $\lambda > 0$. Calculate

1. $L(\lambda)$
2. $\log L(\lambda)$
3. $\hat{\lambda}_{MLE}$ (verify that the extremum is a maximum)

Solution

1. $L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \propto e^{-n\lambda} \lambda^{\sum x_i} (= e^{-n\lambda} \lambda^{n\bar{x}})$

2. $\ell(\lambda) = n\bar{x} \log \lambda - n\lambda$

3. $\hat{\lambda}_{MLE} : \frac{d\ell(\lambda)}{d\lambda} = \frac{n\bar{x}}{\lambda} - n = 0 \implies \frac{\bar{x}}{\lambda} = 1 => \hat{\lambda}_{MLE} = \bar{x}$

- *Verify max :*

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = \frac{d\ell(\lambda)}{d\lambda}\left[\frac{n\bar{x}}{\lambda} - n\right] = -\frac{n\bar{x}}{\lambda^2} < 0\,;$$

since $n\bar{x} > 0, \lambda > 0$, so $\frac{n\bar{x}}{\lambda^2} > 0 \Rightarrow -\frac{n\bar{x}}{\lambda^2} < 0$,

thus the extremum $\left(\hat{\lambda}_{MLE}\right)$ is a *maximum*

Example

NOT important for us ! !

Example 7.5 Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. Find the MLEs of μ and σ^2.

Solution : The normal density is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\},$$

so the log likelihood for a random sample (iid) $y_1, \ldots, y_n$ is

$$\ell(\mu, \sigma) = \log L(\mu, \sigma) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

Solution, cont

Taking the derivative, we have

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum (y_i - \mu) = 0 \qquad (*)$$

and $$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (y_i - \mu)^2 = 0 \qquad (**)$$

Solving (*), we have (for any value of σ^2) :

$$\sum (y_i - \mu) = 0 \quad => \quad \sum y_i = n\mu \quad => \quad \hat{\mu} = \sum y_i / n = \overline{y}$$

Solving (**) (using $\hat{\mu}$ in the place of μ), we have :

$$-n\sigma^2 + \sum (y_i - \hat{\mu})^2 = 0 \quad => \quad \sum (y_i - \hat{\mu})^2 = n\sigma^2 \quad => \quad \hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \hat{\mu})^2$$

NOTE : this estimator **is different** from the unbiased estimator

$$s^2 = \frac{1}{n-1} \sum (y_i - \hat{\mu})^2$$

Solution, cont

We must verify that the log likelihood is a *maximum* (not min) for the pair of values $(\hat{\mu}, \hat{\sigma}^2)$: Second derivative test :

$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} \cdot \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} - \left(\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu \partial (\sigma^2)} \right)^2 > 0$$

AND $\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} < 0$; $\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} < 0$

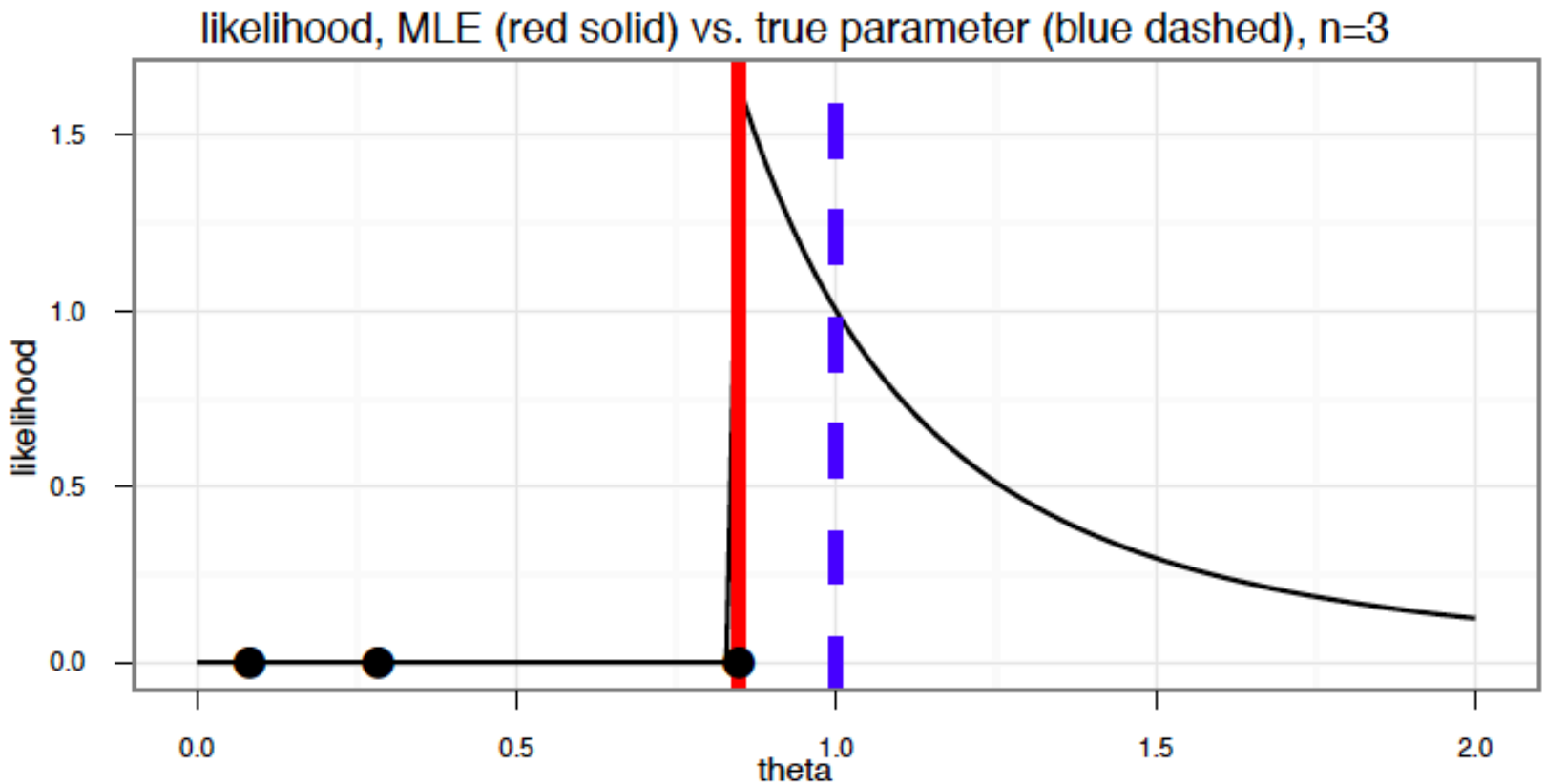
$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu \partial (\sigma^2)} = \frac{-1}{\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0 \; ; \; \frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial \mu^2} = \frac{-n}{\hat{\sigma}^2} < 0$$

$$\frac{\partial^2 \log L(\hat{\mu}, \hat{\sigma}^2)}{\partial (\sigma^2)^2} = \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = \frac{n^3}{2 \sum_{i=1}^{n} (y_i - \hat{\mu})^2} - \frac{n^3}{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}$$

$$= \frac{-n^3}{2} < 0$$

Uniform example – calculus does not work ! !

Example 7.6 Let $y_1, \ldots, y_n$ be a random sample from the uniform distribution $(0, \theta]$, with density $f(y) = 1/\theta$, $0 < y \leq \theta$ $(= 0$ otherwise). Find the MLE $\hat{\theta}$ of θ ...

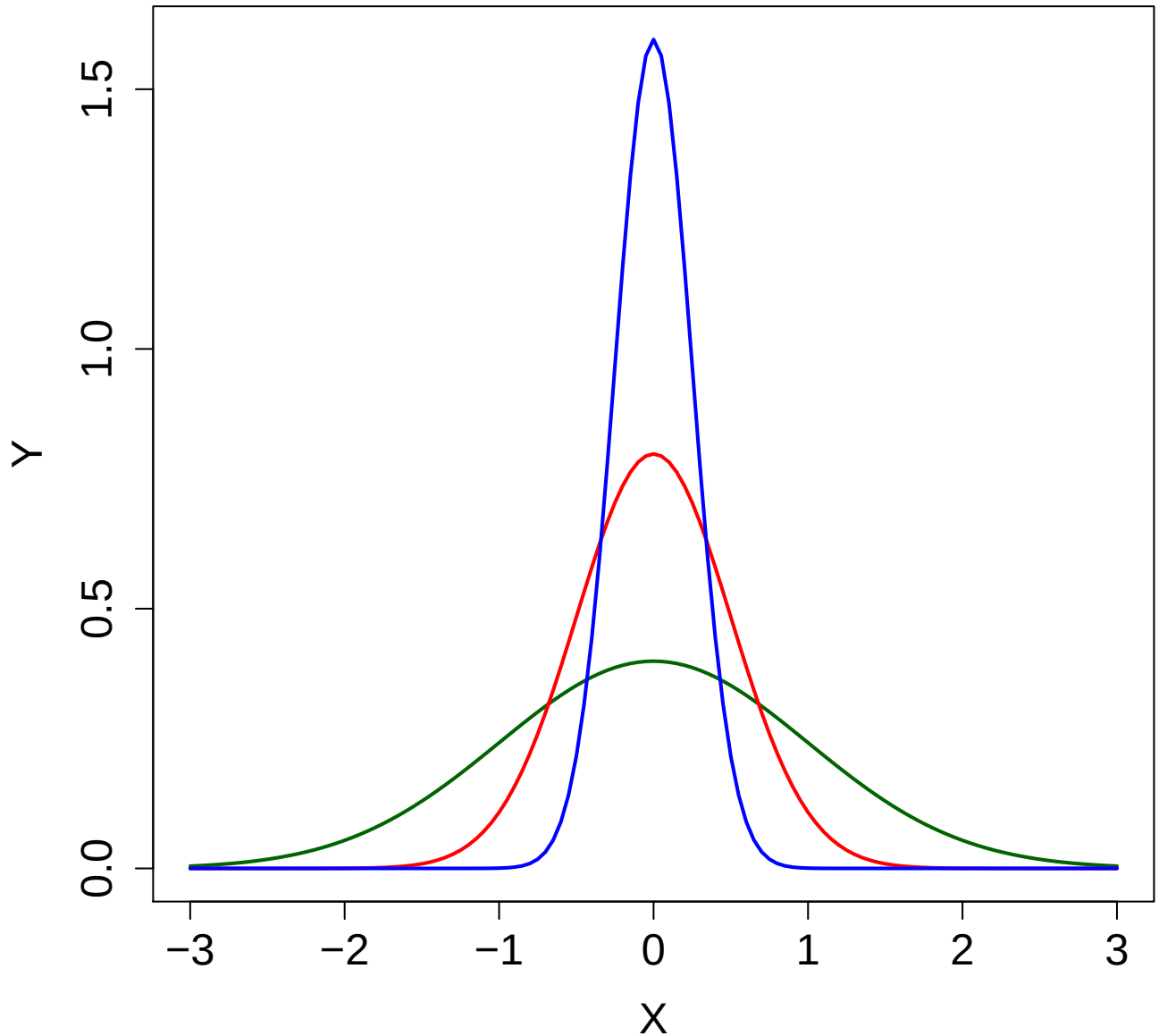


Statistical information

- The *observed information* $J(\theta)$ and the *expected information* (sometimes also called *Fisher information*) $I(\theta)$ are :
  - $J(\theta) = \dfrac{-d^2\ell(\theta)}{d\theta^2}$
  - $I(\theta) = E\{J(\theta)\} = E\left\{\dfrac{-d^2\ell(\theta)}{d\theta^2}\right\}$
- They measure the *curvature* of $-\ell(\theta)$:

  the *larger* the values of $J(\theta)$ and $I(\theta)$, the *more concentrated* are $\ell(\theta)$ and $L(\theta)$

Example : Normal distributions



Properties of MLEs

- **Convergent** : $\lim_{n\to\infty} P(|\ \hat{\theta}_n - \theta\ | < \epsilon) = 1, \forall \epsilon > 0$
- **Invariance** : if $\hat{\theta}$ is the MLE for the parameter θ, then $h(\hat{\theta})$ is the MLE for the parameter $h(\theta)$
- **Asymptotically unbiased** : $b(\theta) \to 0$ as $n \to \infty$
  (for 'small' samples the MLE can be very biased)
- **Asyptotically optimal efficiency** : No other asymptotically unbiased estimator can have a smaller variance than that of the MLE
- **Asymptotic Normality** : the distribution of $\hat{\theta}_n$ as $n \to \infty$ is Normal ; this fact gives us a basis for statistical inference based on the MLE (for example, a CI)
- **Approximate CI** (level $1-\alpha$) for θ : $\hat{\theta} \pm z_{1-\alpha/2} \Big/ \sqrt{J(\hat{\theta})}$

Regularity conditions **(NOT EXAMINED)**

The (not very interesting ! !) technical conditions that the asymptotic normality proof depends on **(just here for completeness, you don't need to study these ! !)** :

- The true value θ_0 of θ is *interior* to the parameter space Θ, which has *finite dimension* and is *compact*
- The densities defined by any two different values of θ are *distinct* ; *i.e.*, θ is *identifiable*
- There is a neighborhood of θ_0 within which the *first 3 derivatives* of ℓ exist almost surely (*i.e.* with probability 1), and for which the expectation of the 3rd derivative is uniformly bounded for θ in the neighborhood
- We can *interchange differentiation and integration* (*i.e.* we can differentiate under the integral sign)

Example

Example 7.7 Suppose $X_1, \ldots, X_n \sim$ iid $Bernoulli(p)$.

Calculate

1. $L(p)$
2. $\ell(p)$
3. $\hat{p}_{EMV}$
4. $J(p)$
5. $I(p)$

Example 7.7, cont.

6. an approximate 95% CI for p using the data :

- $n = 10$ (number of Heads $= 9$)

- $n = 20$ (number of Heads $= 16$)

- $n = 100$ (number of Heads $= 67$)

Example

Example 7.8 Let $X_1, \ldots, X_n \sim$ iid $Pois(\lambda)$, $\lambda > 0$.
Calculate

1. $\hat{\lambda}_{EMV}$ supposing that $\sum X_i > 0$
2. $\hat{\lambda}_{EMV}$ supposing that $\sum X_i = 0$
3. the MLE of $P(X = 0)$
4. $J(\lambda)$
5. $I(\lambda)$
6. an approximate 95% CI for λ ...

Warning

- Estimation by the method of maximum likelihood is *attractive* :
  - conceptually simple
  - intuitive interpretation
- However, some difficulties ; regularity conditions on the likelihood function that cannot be ignored :
  - difficult to establish
  - difficult to interpret
  - difficult to verify in practice
- Thus, even though very often useful, MLE is not a panacea that makes other estimation methods obsolete