Backgroud for Machine Learning

Using python.

Iván Andrés Trujilllo Abella

ivantrujilo1229@gmail.com

Preamble

For this lesson you need remember some concepts as **Probability Distribution Function(PDF)**, **joint distribution function**, **expectation** and **variance** of random varaibles please check the following material:

-

This note we are constructed with several references that are listed in references.

Problems

Quality control

Suppose that you need estimate what is the probability of see two in a slot of 5 computers.

Problems

Quality control

Suppose that you need estimate what is the probability of see two in a slot of 5 computers.

Estimating population

A study uses to count the number of mobile sex workers in bangladesh (to be continued...)

Related problems!

Both problems require are without replacement!

What is the expected value?

What is the expected value?

...

Is the average value outcome of the realization the experiment a large number of times....

Simulations

Therefore we can construct a program that predefined conditions replicated a number of times the experiment...

Dice game

...

Rolling two dice if the sum of each outcome is equal to seven or five, the you will gain 5, in otherwise you pay 1. Which will be your profit playing a lot of times (long run)? **(Click here)**

Dice game

...

Rolling two dice if the sum of each outcome is equal to seven or five, the you will gain 5, in otherwise you pay 1. Which will be your profit playing a lot of times (long run)? **(Click here)**

Conclusion

Statistics is prediction! a powerful tool!!!

Simulation

Quality control simulation

We are going to simulate the probability of get i failed computers on a sample of j where there are m damaged computers in a batch of N. **(Click here)**.

Combinatorial identities

Properties

- $\binom{N}{x} = \frac{N}{N-x}\binom{N-1}{x}$

Empirical results

Vandermonde's indentity

$\mathbf{A_1, A_2, ..., A_N}$	$\mathbf{A_N, A_{N+1}, ..., A_M}$
$\binom{N}{x}$	$\binom{M}{k-x}$

Insight

For a sample of k elements you could select from *left box* $\binom{M}{1}$ and in*right box* $\binom{N}{k-1}$ and by the multiplication rule could get a total of combination $\binom{M}{1}\binom{N}{k-1}$, thus for $x = 0, 1, ..., k$ and finally

$$\sum_{x=0}^{k} \binom{N}{x}\binom{M}{k-x} = \binom{N+M}{k} \tag{1}$$

Generalization of the problem

The problem: *Find the probability of get* (x) *in a sample of* (k) *where* (m) *fill a property in a total of* (N).

deriving

there are $\binom{m}{x}$ possible ways of get x among m and by each one of the possible combination there are $\binom{N-m}{k-x}$ in a total of $\binom{N}{k}$ combinations, therefore the probability will be:

$$\frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}} \tag{2}$$

(2) is also known as **Hypergeometric distribution**.

Ω

hyper

$$\sum_{x=0}^{k} \frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}} = \frac{1}{\binom{N}{k}} \sum_{x=0}^{k} \binom{m}{x}\binom{N-m}{k-x} \tag{3}$$

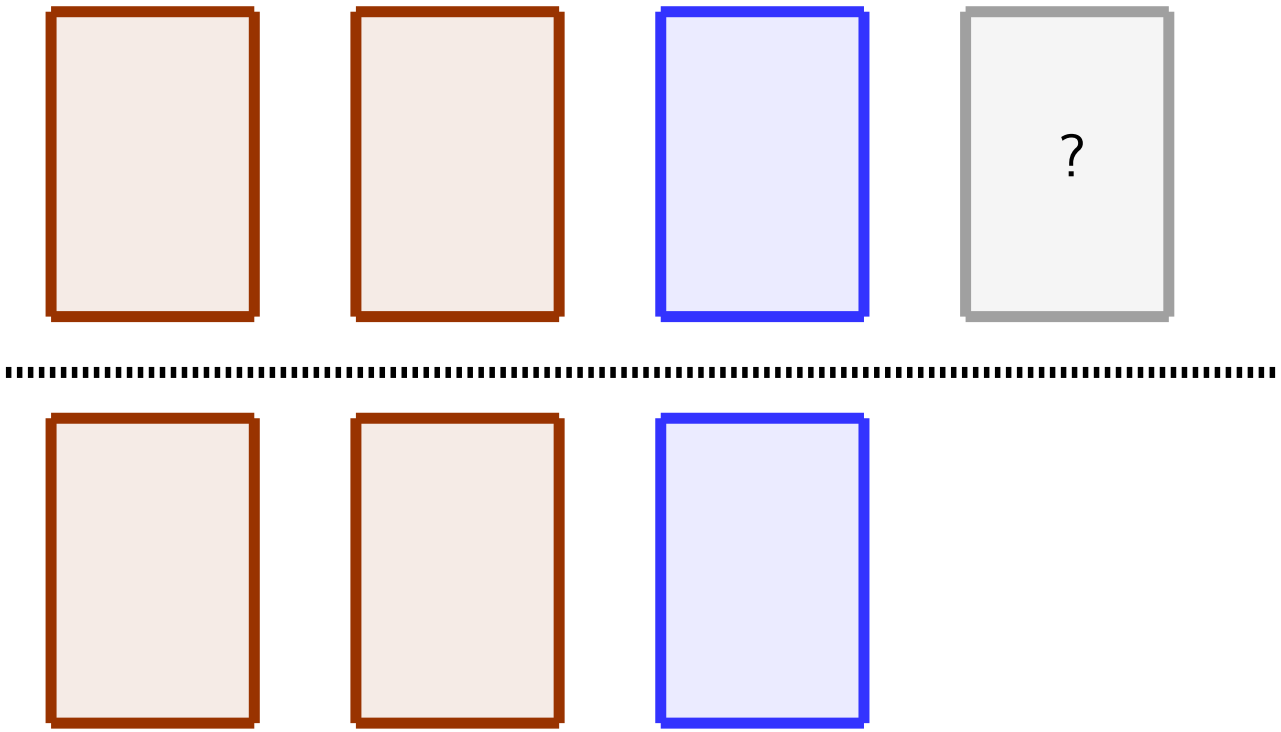
Assuming that $N = m$ and $M = N - m$ and using (1) we have:

$$\sum_{x=0}^{k} \binom{m}{x}\binom{N-m}{k-x} = \binom{N}{k}$$

Therefore:

$$\sum_{x=0}^{k} \frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}} = \frac{\binom{N}{k}}{\binom{N}{k}} = 1 \tag{4}$$

Freund ilustration



Estimate Θ by maximun likelihood

The **Method of maximun likelihood** consist in estimate the parameter Θ the number of *red cards* that maximize the probability of see the **data** (three red cards and a blue car), in this case Θ could be 2 or 3.

$$\frac{\binom{3}{2}\binom{1}{1}}{\binom{4}{3}} > \frac{\binom{2}{2}\binom{2}{1}}{\binom{4}{3}} \tag{5}$$

Scipy

in Scipy the notation is:

$$p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \tag{6}$$

```
from scipy.stats import hypergeom
k,M,n,N = 2,4,3,3
print(hypergeom.pmf(k,M,n,N))
k, M,n,N = 2,4,2,3
print(hypergeom.pmf(k,M,n,N))
```

Maximun Likelihood Estimation (MLE)

Suppose that you data it is generated by a theoretical distribution, the inverse problem is determine the most probable parameter that generate the data.

Insights about MLE

we are going to say in a general term that $f(x_i)$ is PDF or PMF of a random variable.

Considerations

Consider the following:

Useful identity

$$\binom{N}{x} = \binom{N-1}{x} \frac{N}{N-x} \tag{7}$$

Now consider a hill then you could know how is the maximun if

Maximize

For a function $f(x)$ (how is enough)

$$\frac{f(x)}{f(x-1)} > 0 \tag{8}$$

Lincoln-Petersen

Using the properties:

$$H_N = \frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}} = \frac{\binom{m}{x}\frac{N-m}{(N-m)-(k-x)}\binom{(N-1)-m}{k-x}}{\frac{N}{N-k}\binom{N-1}{k}} \tag{9}$$

Thus:

Ratio to maximize

$$\frac{H_N}{H_{N-1}} = \frac{(N-m)(N-k)}{N(N-m-k+x)} \tag{10}$$

Lincoln-Petersen

Max

we are searching that $\frac{H_N}{H_{N-1}} > 1$, therefore:

$$(N - m)(N - k) > N(N - m - k + x) \tag{11}$$

Let to one side N

$$\frac{mk}{x} > N \tag{12}$$

The greater int that not exceded $\frac{mk}{x}$.

Properties

It is a estimator of N is **unbiased** for large samples however for smalls samples not, Chapman estimator produce better results.

Lincoln-Petersen, Works?

Erdos doubt?

Erdos

Brillant mathematician, doubt until see simulation!

Simulating

Assume the following, you need probe to a committee that an estimator return acceptable results. **(Click here)**

Parameter estimation!

All the problems consist in estimate a Unknown parameter with the information of a sample!

Parameter estimation!

All the problems consist in estimate a Unknown parameter with the information of a sample!

How describe a random variable?

- Mean
- Standard Deviation

The fundamental question is if we can get the population values from a sample?, Therefore we could try uses MLE to find the parameters.

$\hat{\sigma}^2_{MLE}$

Assume a sample of $x_1, x_2, ..., x_n \sim N(\mu, \sigma)$ therefore applying the MLE principle of maximizing the probability of get the data we need maximize

$$\mathbb{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right) \tag{13}$$

Maximizing the expression

First Order Condition (FOC)

$$\frac{\partial \mathbb{L}}{\partial \sigma^2} = 0 \tag{14}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{15}$$

Simulation

Benchmark of estimators

...

$$S_n^2 = \frac{\sum(x_i - \bar{x})^2}{n} \tag{16}$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{17}$$

(Click here)

Precision and bias

Ilustrate the concepts..

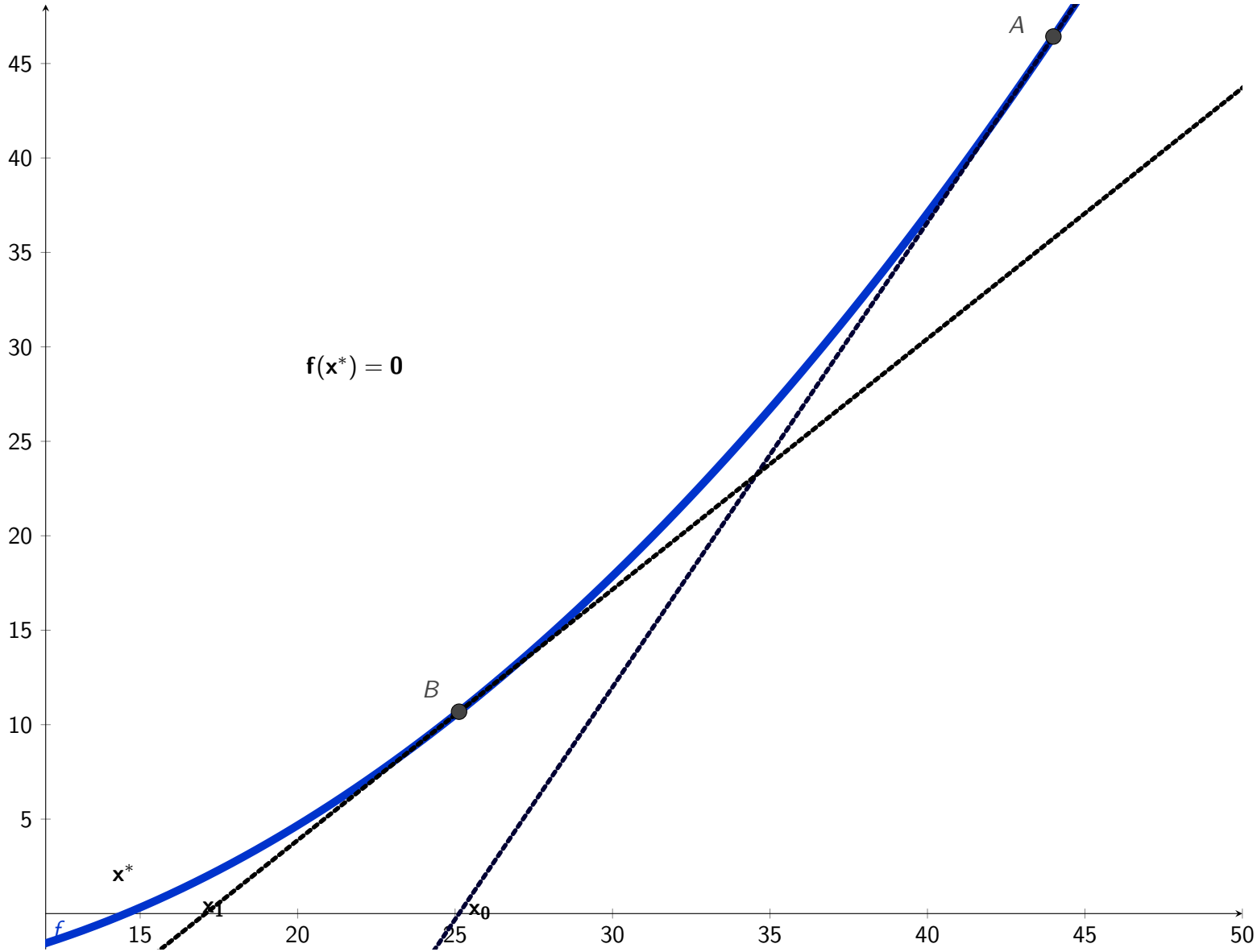
Introduction

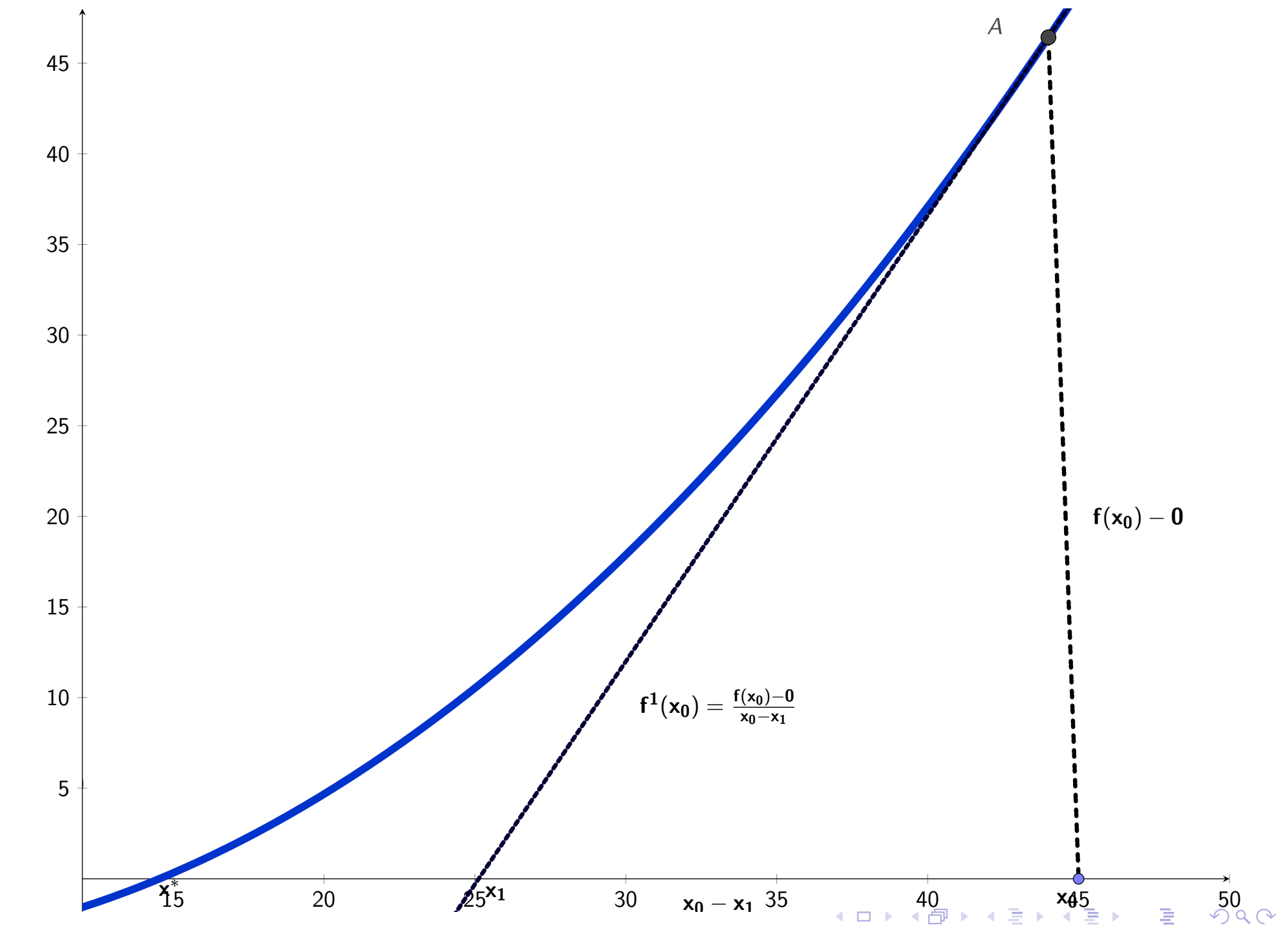
Logistic regression is used broadly in empirical works, it used in economics, engineering, epidemiology and clinical research. In a simplified way the logistic regression it is used to binary problems.

Insights

If we take a point near to the change in concavity the method could produce divergence.

illustration




A
45
40
35
30
25
20
15
10
5
15
20
25
30
35
40
45
50
x^*
x_1
x_0
$x_0 - x_1$
$f(x_0) - 0$
$f^1(x_0) = \frac{f(x_0)-0}{x_0-x_1}$

$$x_1 = x_0 - \frac{f(x_0)}{f^1(x_0)} \tag{18}$$

Thus in i iteration we have:

$$x_{i+1} = x_i - \frac{f(x_i)}{f^1(x_i)} \tag{19}$$

We select the point A the tangent line cross the x-axi in x_0 and again in $f(x_0)$ and again the tangent line of this point (B) cross the x-axis in x_1 each step is also known as a iteration. The algorithm converge to the the value x^* think that this could be used to *optimization problems*.

Newton-raphson program

```
def quadratic(a,b,c, x):
      return a*x**2 + b*x + c
def dfQuadratic(a,b,x):
  return a*x**2 + b*x
a,b,c ,x0 = 1,-3,-4,8
def raphsonQuadratic(a,b,c,x0, error_max=0.0000015,
    iteration_max=100):
  xi = x0
  iter, error = 0, 100
  data = []
  while (iter < iteration_max) and (error > error_max):
    xj = xi - quadratic(a,b,c,xi) / dfQuadratic(a,b,xi)
    error = abs(xj - xi)
    iter += 1
    data.append((xj,xi,error,iter))
    xi = xj
  return data
raphsonQuadratic(a,b,c,4)[-1]
```

$P(\Theta)$ prior distribution, posterior distribution $P(\Theta \mid \mathbf{X})$. likelihood $P(\mathbf{X} \mid \Theta)$. Prior the belief before seen the data.

Max a posteriori

A set of features $\mathbf{X} = \{x_1, ...x_n\}$ assuming a distribution $P(\mathbf{X}, \Theta)$ where *Theta* is a parameter (a random variable).

$$\Theta_{ma} = \max P(\Theta \mid \mathbf{X}) \tag{20}$$

the last equation must be compared regarding the maximun likelihood estimation. that establish the $\max P(\mathbf{X} \mid \Theta)$.

Problem

Problem

We need find the probability that we draw exactly 3 damaged computers of 5 if in the batch there are 12 computers and 5 of them fail.

Problem

you have n balls in a bag where there are j reds and k black thus $n = j + k$. However you do not know the really proportion of each one color. if you draw balls and both are different what is the proportion of black balls Θ.

MLE insight

Then we choose a Θ among all posible values that maximize the probability of seen the data.

Problem

In a formal way we can assume that $x \sim B(n, \Theta)$ and we have seen the data results $\{x_1, x_2, x_3, ..., x_n\}$

$$P(X = x_1 \cap X = x_2, ..., \cap X = x_n) \tag{21}$$

The variables are i.i.d and therefore the joint probability is the result of multiply the marginal probabilities.

$$P(X = x_1 \cap X = x_2, ..., \cap X = x_n) = \prod_{i=1}^{n} P(x_i) \tag{22}$$

Problem

$$
\begin{aligned}
\prod_{i=1}^{n} P(x_i) &= \prod_{i=1}^{n} \binom{m}{x_i} \Theta^{x_i} (1-\Theta)^{m-x_i} \\
&= \prod \binom{m}{x_i} \prod \Theta^{x_i} \prod (1-\Theta)^{m-x_i} \\
&= \prod \binom{m}{x_i} \Theta^{\sum x_i} (1-\Theta)^{\sum (m-x_i)} \\
ln(\prod P(x_i)) &= ln(\binom{m}{x_i} + \sum x_i ln(\Theta) + (nm - \sum x_i) ln(1-\Theta)
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
L(\Theta) =& ln(\prod P(x_i) \\
\frac{dL(\Theta)}{d\Theta} =& \frac{\sum x_i}{\Theta} - \frac{nm - \sum x_i}{1-\Theta} = 0 \\
\frac{1}{\Theta} =& \frac{nm - \sum x_i}{\sum x_i} + 1 \\
\Theta^* =& \frac{\sum x_i}{nm}
\end{aligned}
\tag{24}
$$

Now to proof that $L(\Theta^*)$ is the maximun probability we must show that $\frac{d^2 L(\Theta)}{d\Theta^2} \mid_{\Theta=\Theta^*} < 0$.

About estimators

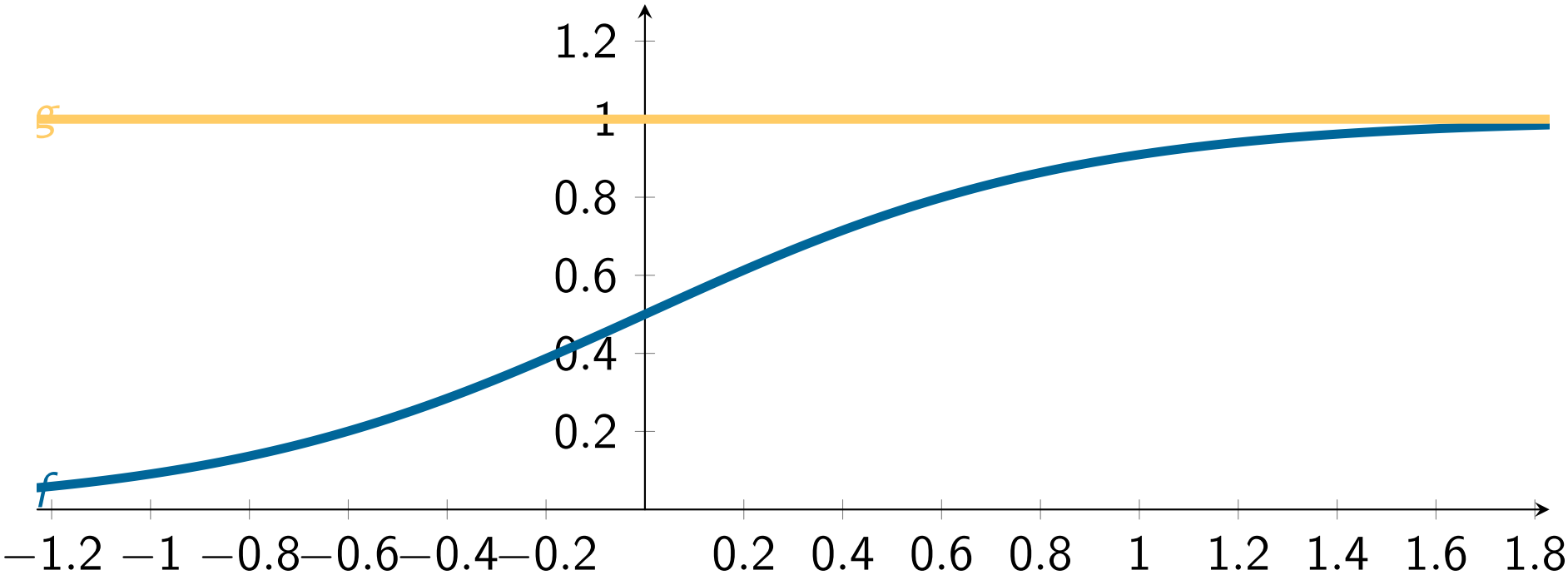
The quality of estimators:

- **Unbiased estimator :** The mean of the estimator is equal to the mean of parameter.
- **Variance:** the dispersion of the estimations regarding the mean value of the same.
- **quadratic value mean:** offer information about another two.

Suppose that do you have a function $f(x, y) = x^2 + y^2$, then we can define the gradient as follow:

$$\nabla f(x, y) = \begin{bmatrix} \frac{df}{dx} \\ \frac{df}{dy} \end{bmatrix} \tag{25}$$

Logistic equation



logistic equation

$$f(x) = \frac{\kappa}{1 + e^{-\alpha(x - x_0)}} \tag{26}$$

Where κ it is the maximun value.

logistic equation

Population growth

$$\frac{dy}{dt} = ry(1 - \frac{y}{k}) \tag{27}$$