

---

# MAXIMUM ENTROPY HANDBOOK

## Formulations, Theorems, M-files

---

### 1 MOST UNCERTAIN FINITE DISTRIBUTIONS

*According to Theorem ET1-1.5 the uniform probability mass function, which is an example of the most uncertain state, maximizes Shannon's entropy. It corresponds with the famous Laplace 'Principle of Insufficient Reason'. The principle postulates that probability mass functions should be considered uniformly if we know nothing about the appropriate random variables.*

*In the following we shall look for the most uncertain probability mass functions of random variables where their mathematical expectations are available. In other words we shall find values of the respective probabilities  $p = (p_1, p_2, \dots, p_m)$  so as to maximize  $H(p)$  (or other entropies) subject to known mean value of the associated random variable. It is a current task in practice because what we generally know is often expressed by mathematical expectations of random variables and what we need is a probability mass function which ignores no possibility. Prescribed mean values may have form of  $E(X), E(X^2), E(\ln(X))$  etc. We shall prefer classical form of the mean values.*

**Notation 1:** We shall denote (arithmetic) mean value by  $E$ . If  $X$  is a random variable with finite alphabet  $\Sigma_X$  and  $X \sim p(x)$  then the mean value of  $X$  is written

$$E_p(X) = \sum_{x \in \Sigma_X} x p(x)$$

or more simply as  $E(X)$  when the probability mass function is understood from the context.

**Notation 2:** We shall denote geometric mean value by  $G$ . If  $X$  is a random variable with finite alphabet  $\Sigma_X$  and  $X \sim p(x)$  then the geometric mean value of  $X$  is written

$$G_p(X) = \prod_{x \in \Sigma_X} x^{p(x)}$$

or more simply as  $G(X)$  when the probability mass function is understood from the context.

**Formulation 1:** *Principle of Maximum Entropy.* Consider an entropy  $H^{(?)}(p)$ , where  $X \sim p(x)$  and a mean value of  $X$  is prescribed. The Principle of Maximum Entropy is formulated as choosing of probability mass function that maximizes entropy  $H^{(?)}(p)$ . We choose from the set of all probability mass functions compatible with given mean value(s).

**Remark 1:** In Formulation 1 we may have arithmetic or geometric mean value independently or both together. The ? symbolizes the possibility of use of different entropies such as Shannon's entropy (most frequently), Relative entropy etc.

**Theorem 1:** *Maximum-Shannon's Entropy Distribution.* Applying the Principle of Maximum Entropy we choose the probability mass function that maximizes Shannon's entropy

$$H(p) = H(p_1, \dots, p_m) = - \sum_{j=1}^m p_j \ln p_j$$

subject to the constraints

$$p_j \geq 0, \quad (j = 1, \dots, m); \quad \sum_{j=1}^m p_j = 1$$

and

$$E = \sum_{j=1}^m p_j x_j.$$

As the solution we obtain the so called *Maximum-Shannon's Entropy Distribution* with respective probabilities

$$p_j = \frac{e^{-\beta_0 x_j}}{\sum_{r=1}^m e^{-\beta_0 x_r}}, \quad (j = 1, \dots, m),$$

where  $\beta_0$  is the solution of the exponential equation

$$\sum_{j=1}^m [x_j - E] e^{-\beta(x_j - E)} = 0.$$

If the random variable  $X$  is nondegenerate (i.e., if  $X$  takes on at least two different values), such a solution exists and is unique.

**Corollary 1:** Due to Theorem ET1-1.5 such a probability distribution is the 'largest one'; it will ignore no possibility, being the most uniform one, subject to the given constraints.

**Corollary 2:** If the geometric mean value is prescribed, we use a logarithm in the constraints

$$\ln G = \sum_{j=1}^m p_j \ln x_j$$

rather than its original definition.

**M-function:** Computes the Maximum-Shannon's Entropy Distribution from Theorem 1.

```
function [q,b]=et2_smx(x,e,met)
```

```
if (nargin<=2) met='RID'; end; % Initialization stage
if ((nargin<2)|(nargin>3))
    error('Function must have 2 or 3 input arguments'); end;
if ((e>=max(x))|(e<=min(x)))
    error('Expectation must lie in (max(X),min(X)) open interval'); end;
```

```

MAX_LOOPS=20; % Iterations limit
[m,n]=size(x); h=1; tol=10+log(12); etol=exp(-tol);
b=2*e*sum(x)-n*e^2-sum(x.^2);
if b~=0
    b=(n*e-sum(x))/b; % First order beta approximation
    x1=b-log(abs(b)+2); x2=b+log(abs(b)+2);
    fl=sum((x-e).*exp(-x1*(x-e))); fh=sum((x-e).*exp(-x2*(x-e))); rts=b;

    if (strcmp(upper(met),'SEC')) % Secant method
        if (abs(fl)<abs(fh))
            rts=x1; x1=x2; h1=f1; f1=fh; fh=h1; else x1=x1; rts=x2; end;
        for m=1:MAX_LOOPS
            if (fh==f1) h=0; break, end;
            dx=(x1-rts)*fh/(fh-f1); x1=rts; f1=fh; rts=rts+dx;
            fh=sum((x-e).*exp(-rts*(x-e)));
            if ((abs(dx)<=etol*etol)|(abs(fh)<=etol)) h=0; break, end;
        end;

    elseif (strcmp(upper(met),'NEW')) % Newton-Raphson method
        for m=1:MAX_LOOPS
            f=sum((x-e).*exp(-rts*(x-e))); f1=sum((x-e).*exp(-(rts+sqrt(eps))*(x-e)));
            df=(f1-f)/sqrt(eps);
            if (df==0) h=0; break, end; dx=f/df; rts=rts-dx;
            if ((abs(dx)<=etol*etol)|(abs(f)<=etol)) h=0; break, end;
        end;

    elseif (strcmp(upper(met),'APX')) % First order approximation
        rts=b; h=0;

    else % Ridders' method (default)
        for m=1:MAX_LOOPS
            xm=.5*(x1+x2); fm=sum((x-e).*exp(-xm*(x-e)));
            if ((fm*fm-fl*fh)>=0) s=sqrt(fm*fm-fl*fh); else, s=0; end;
            if (s==0) h=0; break, end;
            xnew=xm+(xm-x1)*((fl>=fh)-(fl<fh))*fm/s;
            if (abs(xnew-rts)<=etol*etol) h=0; break, end;
            rts=xnew; fnew=sum((x-e).*exp(-rts*(x-e)));
            if (abs(fnew)<=etol) h=0; break, end;
            if (sign(fnew)*abs(fm)~=fm)
                x1=xm; fl=fm; x2=rts; fh=fnew;
            elseif (sign(fnew)*abs(fl)~=fl) x2=rts; fh=fnew;
            elseif (sign(fnew)*abs(fh)~=fh) x1=rts; fl=fnew; end;
            if (abs(x2-x1)<=etol*etol) h=0; break, end;
        end;
    end; if (h~=0) b=sign(b)*tol; else, b=rts; end;
    if (abs(b)>tol) b=sign(b)*tol; end;
else b=0; end;
q=exp(-b*x)/sum(exp(-b*x)); % Resulting probability mass function

```

**Synopsis:** `[p,beta0]=et2_smx(X,E, met)`

**Directions for Use:**  $X$  is a vector of random values,  $E$  is known expectation. Since an exponential equation must be solved to obtain resulting Maximum-Shannon's Entropy Probability Distribution  $p$ , we have used several methods to solve it. Using parameter  $met$  one select a method by the following table:

'sec'	Secant Method
'rid'	Ridder's Method
'new'	Newton's Method
'apx'	Approximation Method
No $met$	Ridder's Method

In case of the approximation we use Taylor series of the first order. For  $\beta_0$  we may have the following equation

$$\beta_0 = \frac{mE(X) - \sum_{j=1}^m x_j}{2E(X) \sum_{j=1}^m x_j - mE^2(X) - \sum_{j=1}^m x_j^2}.$$

This quick approximation is suitable for low distances between  $\min(X)$  and  $\max(X)$  only.

**Example 1:** Let

$$X = \left\{ \begin{array}{c} 12 \\ 15 \\ 20 \end{array} \right\}$$

and the mean value  $E(X) = 18.12$ . Demonstrating `et2_smx`, we can obtain

```
>>p=et2_smx([12 15 20],18.12,'sec')
>>p=
```

```
0.1035    0.2104    0.6861
```

or similarly

```
>>p=et2_smx([12 15 20],18.12)
>>p=
```

```
0.1035    0.2104    0.6861
```

or similarly

```
>>p=et2_smx([12 15 20],18.12,'apx')
>>p=
```

```
0.1743    0.2693    0.5564
```

etc.

**Theorem 2:** *Special Case of Theorem 1 for random variables with natural finite alphabets.* Let  $\Sigma_X = \{1, 2, 3, \dots, m\}$ . We have to maximize  $-\sum_{j=1}^m p_j \ln p_j$  subject to  $\sum_{j=1}^m p_j = 1$  and  $E = \sum_{j=1}^m j p_j$ ,  $1 < E < m$ . Maximizing this we get the *Maximum-Shannon's Entropy Distribution* with respective probabilities

$$p_j = ab_o^j, \quad (j = 1, \dots, m),$$

where  $b_o$  is the solution of equation

$$\frac{1}{1-b} - \frac{m}{1-b^m} - E + m = 0$$

and in consequence of it

$$a = \frac{1}{b} \frac{1-b}{1-b^m}.$$

So  $b \leq 1$  for  $E \leq \frac{1}{2}(m+1)$  and vice versa  $b \geq 1$  for  $E \geq \frac{1}{2}(m+1)$ .

**Corollary 3:** Thus the Maximum Entropy Probability Distribution is a *geometric distribution* for which  $p_j$  decreases with  $j$  if  $E \leq \frac{1}{2}(m+1)$ , remains constant if  $E = \frac{1}{2}(m+1)$  and increases with  $j$  when  $E \geq \frac{1}{2}(m+1)$ .

**Example 2:** *Analytic Solution.* Let  $m = 3$ . Maximizing  $-\sum_{j=1}^3 p_j \ln p_j$  subject to  $\sum_{j=1}^3 p_j = 1$ ,  $E = \sum_{j=1}^3 j p_j$  we may obtain an analytic solution

$$p_1 = \frac{1}{2}(3 - E - p_2), \quad p_2 = \frac{1}{3}[(4 - 3(E - 2)^2)^{1/2}], \quad p_3 = \frac{1}{2}(E - 1 - p_2).$$

**Example 3:** Let  $\Sigma_X = \{1, 2, 3, 4, 5, 6\}$ . Investigate cases with  $E = 2.5, 3.5, 4.5$ .

```
>>[p,b]=et2_smx([1 2 3 4 5 6],2.5)
```

```
>>p=
```

```
0.3475 0.2398 0.1654 0.1142 0.0788 0.0544
```

```
>>b=
```

```
0.3710
```

```
>>[p,b]=et2_smx([1 2 3 4 5 6],3.5)
```

```
>>p=
```

```
0.1667 0.1667 0.1667 0.1667 0.1667 0.1667
```

```
>>b=
```

```
0
```

```
>>[p,b]=et2_smx([1 2 3 4 5 6],4.5)
```

```
>>p=
```

```
0.0544 0.0788 0.1142 0.1654 0.2398 0.3475
```

```
>>b=
```

```
-0.3710
```

**Theorem 3:** Let us consider  $X$  with alphabet  $\{x_1, x_2, \dots, x_m\}$ . Suppose that respective probabilities can be well-approximated by

$$p_i = \frac{1}{m} + \epsilon_i$$

where the  $\epsilon_i$  must all be small. Therefore the Maximum Entropy Principle goes over into a principle of least squares. In other words

$$\max_p - \sum_{i=1}^m p_i \ln p_i \quad \text{leads to} \quad \min_p \sum_{i=1}^m \left(p_i - \frac{1}{m}\right)^2.$$

Any information constraints, such as normalization of the  $p$  and a known  $E$  may be added into this new principle.

**Theorem 4:** *Approximate form of the Principle of Maximum Entropy.* We have to minimize

$$\sum_{i=1}^m \left(p_i - \frac{1}{m}\right)^2$$

subject to  $\sum_{j=1}^m p_j = 1$  and  $E = \sum_{j=1}^m p_j x_j$ . Minimizing this we get the *Approximate Maximum-Shannon's Entropy Distribution* with respective probabilities

$$p_j = \frac{1}{m} - \frac{1}{m} \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} \sum_{j=1}^m x_j + \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} x_j$$

for all  $j = 1, 2, \dots, m$ . For  $E = (m + 1)/2$  case we also verify the uniform distribution.

**Remark 1:** Note that a non-negativity of respective probabilities must be verified in case of bigger distance from the uniform distribution.

**M-function:** Computes the Approximate Maximum-Shannon's Entropy Distribution due to Theorem 4.

```
function p=et2_amx(x,e)

if (nargin~=2)
    error('Must be two input arguments.');
```

$$p_j = \frac{1}{m} - \frac{1}{m} \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} \sum_{j=1}^m x_j + \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} x_j$$

```
end;
[n,m]=size(x);
if ((m <= 1) | (n ~= 1))
    error('1st input argument must be a vector.');
```

$$p_j = \frac{1}{m} - \frac{1}{m} \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} \sum_{j=1}^m x_j + \frac{E - \frac{1}{m} \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2} x_j$$

```
end;

sum_x=sum(x);
sum_xx=sum(x.^2);
p=1/m-1/m*(e-1/m*sum_x)*sum_x/(sum_xx-1/m*sum_x^2);
p=p.+(e-1/m*sum_x).*x/(sum_xx-1/m*sum_x^2);
```

**Synopsis:** `p=et2_amx(X,E)`

**Example 4:** Let  $\Sigma_X = \{1, 2, 3, 4, 5, 6\}$ . Let us investigate cases with  $E = 3, 3.5, 4$ .

```
>>p=et2_amx([1 2 3 4 5 6],3)
>>p=
```

```

0.2381 0.2095 0.1810 0.1524 0.1238 0.0952
>>p=et2_amx([1 2 3 4 5 6],3.5)
>>p=

```

```

0.1667 0.1667 0.1667 0.1667 0.1667 0.1667
>>p=et2_amx([1 2 3 4 5 6],4)
>>p=

```

```

0.0952 0.1238 0.1524 0.1810 0.2095 0.2381

```

Note that the correct probability mass function for  $E = 3$  obtained by `et2_smx` is  $p = (0.2468, 0.2072, 0.1740, 0.1461, 0.1227, 0.1031)$ .

## 2 MOST UNCERTAIN INFINITE DISTRIBUTIONS

*We shall take into account cases when the discrete variable takes a countable infinite set of values. In relation to results obtained in the previous section we shall see interesting ‘deformations’ of maximum entropy probability distributions. We call this case as limiting form of the Maximum Entropy Principle.*

**Theorem 1:** *The limiting form of Maximum Entropy Principle.* If  $X$  is a random variable whose range is countable, namely, if

$$\Sigma_X = \{ku \mid u > 0, \quad k = 0, 1, 2, \dots\}$$

and if the mean value  $E(X)$  is given, then the probability distribution

$$p_k > 0, \quad (k = 0, 1, 2, \dots), \quad \sum_{k=1}^{\infty} p_k = 1$$

maximizing the countable entropy

$$H(X) = - \sum_{k=0}^{\infty} p_k \ln p_k$$

is

$$p_k = \frac{u(E(X))^k}{(u + E(X))^{k+1}}, \quad k = 0, 1, 2, \dots$$

The unit  $u$  and the mean value  $E(f)$  completely determine the solution of the Principle of Maximum Entropy.

**Corollary 4:** In case of  $u = 1$  we often express the Maximum Entropy Probability Distribution in the form of a geometrical distribution as (see example 2 of ET1-1)

$$p_k = ab^k = \frac{1}{1 + E} \left( \frac{E}{1 + E} \right)^k, \quad k = 0, 1, 2, \dots$$

### 3 MINIMUM RELATIVE ENTROPY MEASURES

According to properties of the relative entropy presented by section 4 of ET1 part  $D(p||q)$  is a convex function of  $p$  and this measure is non-negative and vanishes if and only if  $p_i = q_i$  for all  $i$ . Minimizing it implies minimizing

$$-\sum_{i=1}^m q_i \ln p_i.$$

Usually we apply this minimization process in cases when an unknown probability mass function  $q$  is observed. In order to know an approximation of  $q$  we select an analytical probability function  $p$  parametrised by some parameters and we minimize a discrepancy between  $p$  and  $q$  which is given by measure of the relative entropy.

**Remark 1:** Let  $q$  be the observed probability mass function, i.e. let  $q_i = x_i/N$  where  $x_i$  is the observed frequency of the  $i$ th random value ( $x_i$  can be zero also) in a random experiment of size  $N$ , so that minimizing  $D(p||q)$  implies the maximizing of

$$\sum_{i=1}^m x_i \ln p_i$$

i.e. of the logarithm of the likelihood function. Thus minimizing the discrepancy measure  $D(p||q)$  is equivalent to maximizing the likelihood function or minimizing the unlikelihood function.

**Theorem 1:** Let  $q$  be an observed probability mass function with  $x_i$  as the observed frequencies for all  $i$ . Consider that estimated probability mass function  $p$  is parametrised by parameters  $\Theta_1, \Theta_2, \dots, \Theta_n$ . The minimum discrepancy estimators  $\Theta_1, \Theta_2, \dots, \Theta_n$  are obtained by minimization of

$$-\sum_{i=1}^m x_i \ln p(x_i, \Theta_1, \Theta_2, \dots, \Theta_n)$$

which leads to solving the equations

$$\sum_{i=1}^m \frac{x_i}{p_i} \frac{\partial p_i}{\partial \Theta_j} = 0, \quad j = 1, 2, \dots, n.$$

### 4 MAXIMIZING WEIGHTED ENTROPY

We now distinguish the values of a random variable by the weighted entropy (see section 5 of ET1). Let us now introduce the expression for the probability mass function, maximizing the weighted entropy.

**Theorem 1:** Consider the probability mass function

$$p_i \geq 0 \quad (i = 1, 2, \dots, m), \quad \sum_{i=1}^m p_i = 1$$

and the weights  $w_i > 0$  ( $i = 1, 2, \dots, m$ ). The weighted entropy  $H_w$  is maximum if and only if

$$p_i = \exp\left(-\frac{\alpha_o}{w_i} - 1\right) \quad (i = 1, 2, \dots, m)$$

where  $\alpha_o$  is the solution of the equation

$$\sum_{i=1}^m \exp\left(-\frac{\alpha}{w_i} - 1\right) = 1.$$

The maximum value of  $H_w$  is given by the quantity

$$\alpha_o + \sum_{i=1}^m w_i \exp\left(-\frac{\alpha}{w_i} - 1\right).$$