

Entropy, which is not to be confused with entropy from physics, comes from **information theory**. [Information theory](#) is based on probability and statistics, and deals with the transmission, processing, utilization, and extraction of information. A key concept in information theory is the notion of a **bit of information**. One bit of information is one [unit of information](#).

We can represent a bit of information as a binary number because it either has the value 1 or 0. Suppose there's an equal probability of tomorrow being sunny (1) or not sunny (0). If I tell you that it will be sunny, I've given you one bit of information.

We can also think of entropy in terms of information. If we flip a coin where both sides are heads, we know upfront that the result will be heads. We gain no new information by flipping the coin, so entropy is 0. On the other hand, if the coin has a heads side and a tails side, there's a 50% probability that it will land on either. Thus, flipping the coin gives us one bit of information -- which side the coin landed on.

Entropy can be much more complex, especially when we get to cases with more than two possible outcomes, or differential probabilities. A deep understanding of entropy isn't necessary for constructing decision trees, however. If you'd like, you can read more about entropy [at Wikipedia](#).

The formula for entropy looks like this:

$$- \sum_{i=1}^c P(x_i) \log_b P(x_i)$$

→ Vu que l'on calcule l'entropie de la variable **prédite** (colonne « cible »), on pourrait noter plutôt :

$$H(Y) \text{ (ou } H(T) \text{ dans la suite)} = - \sum_{i=1}^c P(y_i) \log_b P(y_i)$$

We iterate through each unique value in a single column (in this case, `high_income`), and assign it to x_i . We then compute the probability of that value occurring in the data ($P(x_i)$). Next we do some multiplication, and sum all of the values together. b is the base of the logarithm. We commonly use the value 2 for this, but we can also set it to 10 or another value.

Let's say we have this data:

age	high_income
25	1
50	1
30	0
50	0
80	1

We could compute its entropy like this:

$$- \sum_{i=1}^c P(x_i) \log_b P(x_i) = - \left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \approx 0.97$$

We get less than one "bit" of information -- only .97 -- because there are slightly more 1s in the sample data than 0s. This means that if we were predicting a new value, we could guess that the answer is 1 and be right more often than wrong (because there's a .6 probability of the answer being 1). Due to this prior knowledge, we gain less than a full "bit" of information when we observe a new value.

→

Hi,

On page 8 of the "Introduction to Decision Trees" mission, you write: "If you'd like, you can read more about entropy [at Wikipedia](#)."

The URL link references the "Entropy" article (in *physics*!), I think it should rather be:

[https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))

Oops! Good catch — I've logged this to be fixed by our content team.

Best,

Dustin

→ [https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))

Definition

Named after [Boltzmann's H-theorem](#), Shannon defined the entropy H (Greek capital letter [eta](#)) of a [discrete random variable](#) X with possible values $\{x_1, \dots, x_n\}$ and [probability mass function](#) $P(X)$ as:

$$H(X) = E[I(X)] = E[-\ln(P(X))].$$

Here E is the [expected value operator](#), and I is the [information content](#) of X .^{[4][5]} $I(X)$ is itself a random variable.

The entropy can explicitly be written as

$$H(X) = - \sum_{i=1}^n P(x_i) I(x_i) = - \sum_{i=1}^n P(x_i) \log_b P(x_i),$$

where b is the [base](#) of the [logarithm](#) used. Common values of b are 2, [Euler's number](#) e , and 10, and the corresponding units of entropy are the [bits](#) for $b=2$, [nats](#) for $b=e$, and [bans](#) for $b=10$.^[6]

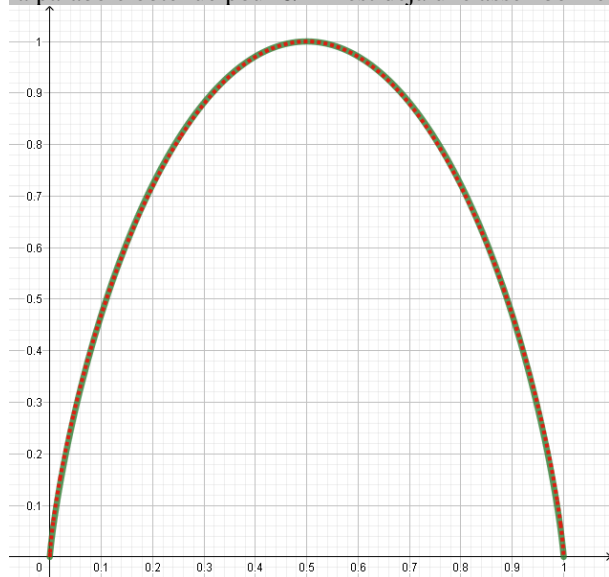
One may also define the [conditional entropy](#) of two events X and Y taking values x_i and y_j respectively, as

$$H(X|Y) = - \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(y_j)}$$

where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_j$. This quantity should be understood as **the amount of randomness in the random variable X given the event Y** .

→ Ci-dessous, la courbe de l'entropie binaire (ou entropie de Bernoulli) $x \mapsto -(x \log_2 x + (1-x) \log_2(1-x))$ dont le maximum, sans surprise, est atteint pour $x = 1-x = \frac{1}{2}$.

Pour la petite curiosité, j'ai ajouté en pointillés le remarquable ajustement $x \mapsto (4x(1-x))^\alpha$ pour $\alpha = 0,74$ (après avoir remarqué que la parabole obtenue pour $\alpha = 1$ est déjà une assez bonne approximation).



→ Notons le *très vague* « air de famille » entre la formule de l'entropie binaire et celle de la fonction « coût » pour un réseau de neurones à deux couches :

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^m (y_i \log(h_\Theta(x_i)) + (1-y_i) \log(1-h_\Theta(x_i))) \text{ où } h_\Theta(x_i) = \frac{1}{1 + \exp(-\Theta^T x_i)}.$$

→ À ce propos, dans la leçon « Machine Learning in Python - Intermediate » (p. 40), je m'étais posé la question incongrue de la presque aussi vague ressemblance entre chaque terme $y_i \log(h_\Theta(x_i)) + (1-y_i) \log(1-h_\Theta(x_i))$ et le **logarithme de la « probabilité binomiale »** $p_i = h_\Theta(x_i)^{y_i} (1-h_\Theta(x_i))^{1-y_i}$.

L'ironie est que la **loi binomiale** possède effectivement un **lien**... mais avec l'**entropie binaire**, pas avec la fonction « coût » !

https://en.wikipedia.org/wiki/Bernoulli_process#Law_of_large_numbers,_binomial_distribution_and_central_limit_theorem

Let us assume the canonical process with H ("heads") represented by 1 and T ("tails") represented by 0.

If the probability of flipping heads is given by p , then the total probability of seeing a string of length n with k heads is

$$P([S_n = k]) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $S_n = \sum_{i=1}^n X_i$ and $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. The probability measure thus defined is known as the **Binomial distribution**.

Of particular interest is the question of the value of S_n for a sufficiently long sequences of coin flips, that is, for the limit $n \rightarrow \infty$. In this case, one may make use of **Stirling's approximation** to the factorial, and write

$$n! = \sqrt{2\pi n} n^n e^{-n} (1 + O(\frac{1}{n}))$$

Inserting this into the expression for $P(k, n)$, one obtains the **Normal distribution**; this is the content of the **central limit theorem**, and this is the simplest example thereof.

The combination of the law of large numbers, together with the central limit theorem, leads to an interesting and perhaps surprising result: the **asymptotic equipartition property**. Put informally, one notes that, yes, over many coin flips, one will observe H exactly p fraction of the time, and that this corresponds exactly with the peak of the Gaussian. **The asymptotic equipartition property** essentially states that this peak is infinitely sharp, with infinite fall-off on either side. That is, **given the set of all possible infinitely long strings of H and T occurring in the Bernoulli process, this set is partitioned into two: those strings that occur with probability 1, and those that occur with probability 0**. This partitioning is known as the **Kolmogorov 0-1 law**.

→ Je ne suis pas sûr de ce que signifie précisément l'avant-dernière phrase : doit-on comprendre que,

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(\frac{S_n}{n} (= \bar{X}) \in [p - \varepsilon, p + \varepsilon]) = 1 \text{ (et donc } \lim_{n \rightarrow \infty} P(|\frac{S_n}{n} - p| > \varepsilon) = 0) ?$$

Sauf que cette interprétation n'est qu'*asymptotique* (ne concernant que des suites *finies* de $\{H, T\}^n$) tandis que la phrase semble parler d'une **partition** « fixe » de l'ensemble des suites *infinies* de $\{H, T\}^\mathbb{N}$...

La partition en question serait-elle alors justement le « **pic de Dirac** » $\{(X_i)_{i \in \mathbb{N}}, \text{ tq } \lim_{n \rightarrow \infty} \frac{S_n}{n} = p\}$, de probabilité 1, et son **complémentaire** (c'est-à-dire toutes les suites de « **moyenne asymptotique** » — limite des moyennes de Cesàro — autre que p !), de probabilité 0 ?

The size of this set is interesting, also, and can be explicitly determined: the logarithm of it is exactly the **entropy** of the Bernoulli process. Once again, consider the set of all strings of length n . The size of this set is 2^n . Of these, **only a certain subset are likely**; the size of this set is 2^{nH} for $H \leq 1$. By using Stirling's approximation, putting it into the expression for $P(k, n)$, solving for the **location** and **width** of the peak, and finally taking $n \rightarrow \infty$ one finds that

$$H = -p \log_2 p - (1-p) \log_2 (1-p)$$

This value is the **Bernoulli entropy** of a Bernoulli process. Here, H stands for entropy; do not confuse it with the same symbol H standing for *heads*.

→ Ce paragraphe elliptique pour ne pas dire lapidaire mérite d'amples explications !

« Only a certain subset are likely » : si mon interprétation précédente est correcte, **les suites « likely » (disons : « presque certaines ») sont celles pour lesquelles $\frac{S_n}{n} = \frac{k}{n}$ est proche de p et même tend vers p asymptotiquement.** Il s'agirait donc des suites

telles que $S_n = k =$ l'entier le plus proche de $p n$. Ces suites sont au nombre de $\binom{n}{k}$.

« The size of this set is 2^{nH} »: ceci définit H comme une « **proportion logarithmique** » :

$$H = \frac{nH}{n} = \frac{\log_2(\text{nombre de suites « quasi-certaines »})}{\log_2(\text{nombre total de suites})}$$

Pour alléger, on notera dans la suite $\log = \log_2$.

Ainsi $H = \frac{1}{n} \log \binom{n}{k}$, c'est déjà une jolie définition synthétique de l'entropie binaire. → Je crains que cette prémisse soit erronée, voir à la fin pourquoi...

Mais ce n'est qu'un début car c'est là que Stirling entre en scène et que ça va devenir remarquable.

$$n! = \sqrt{2\pi n} n^n e^{-n} (1 + O(\frac{1}{n})) \text{ donne } \log(n!) = n \log n - \frac{1}{\ln 2} n + O(\log n)$$

donc, à $O(\log n)$ près :

$$\log \binom{n}{k} \approx n \log n - \frac{1}{\ln 2} n - (n-k) \log(n-k) + \frac{1}{\ln 2} (n-k) - k \log k + \frac{1}{\ln 2} k = n \log n - n \log(n-k) - k \log k + k \log(n-k)$$

Pour faire apparaître des $\log \frac{k}{n}$ et $\log \frac{n-k}{n}$ (objectif de symétrie...), on introduit deux termes en télescopage :

$$\begin{aligned} \log \binom{n}{k} &\approx k \log n - k \log k + n \log n - k \log n - n \log(n-k) + k \log(n-k) \\ &= -k \log \frac{k}{n} + (n-k) \log n - (n-k) \log(n-k) = -k \log \frac{k}{n} - (n-k) \log \frac{n-k}{n} (+ O(\log n)) \end{aligned}$$

$$\text{Ainsi, } H = \frac{1}{n} \log \binom{n}{k} \approx -\frac{k}{n} \log \frac{k}{n} - \left(1 - \frac{k}{n}\right) \log \left(1 - \frac{k}{n}\right) (+ O(\frac{\log n}{n}))$$

Or $\frac{k}{n} \rightarrow p$ donc $H(p) \approx -p \log p - (1-p) \log(1-p)$: **formule de l'entropie binaire** (à $O(\frac{\log n}{n}) \subset o(1)$ près) !!

On peut aller un peu plus loin en calculant la **probabilité logarithmique asymptotique du pic de Dirac** :

$$\log(P(S_n = k)) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p) \approx n [-p \log p - (1-p) \log(1-p) + O(\frac{\log n}{n}) + p \log p + (1-p) \log(1-p)] = O(\log n)$$

Si l'on pouvait **affiner le D.L. de H** (via le D.A. de Stirling) pour obtenir $\log(P(S_n = k)) = o(1)$, ce serait **cohérent avec l'AEP** (asymptotic equipartition property) : $\lim_{n \rightarrow \infty} P(\frac{S_n}{n} \in [p - \epsilon, p + \epsilon]) = 1$.

... Sauf que $\log(P(S_n = k)) = o(1)$ est **faux** ! Il y a une nuance de taille entre l'AEP et affirmer que $\lim_{n \rightarrow \infty} P(\frac{S_n}{n} = p) = 1$!

En fait, d'après la formule de De Moivre-Laplace (voir plus loin), si $k \sim np$:

$$P(\frac{S_n}{n} = \frac{k}{n}) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi n p(1-p)}} \exp(-\frac{(k - np)^2}{2np(1-p)}) = O(n^{-1/2}) \rightarrow 0 !$$

Il n'est donc pas étonnant que l'on n'obtienne pas mieux que $\log(P(S_n = k)) = O(\log n) !...$

...

→ **Développement asymptotique de Stirling** :

https://fr.wikipedia.org/wiki/Formule_de_Stirling#D%C3%A9veloppement_asymptotique

$$\begin{aligned} n! &= \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left[1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} - \frac{571}{2488320n^4} + \frac{163879}{209018880n^5} + O\left(\frac{1}{n^6}\right)\right] \\ \ln(n!) &= n \ln n - n + \frac{1}{2} \ln(2\pi n) + \sum_{k=1}^K \frac{(-1)^{k+1} B_{k+1}}{k(k+1)n^k} + O\left(\frac{1}{n^{K+1}}\right) \end{aligned}$$

où les B_k sont les **nombre de Bernoulli**.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
B_n	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$	0	$-\frac{691}{2730}$	0	$\frac{7}{6}$

Ici, on a juste besoin d'aller jusqu'à $K = 1$: $\log(n!) = n \log n - \frac{1}{\ln 2} n + \frac{1}{2} \log(2\pi n) + \frac{1}{12(\ln 2)n} + O\left(\frac{1}{n^2}\right)$

→ Calcul des **D.L. de $\log \binom{n}{k}$** puis $\log(P(S_n = k))$ à reprendre...

Prendre garde que, dans $\log(P(S_n = k))$, le bloc $\log(p^k (1-p)^{n-k}) = k \log p + (n-k) \log(1-p)$ ne pourra peut-être plus être remplacé brutalement par l'approximation au premier ordre $n p \log p + n (1-p) \log(1-p)$ si l'on doit tenir compte que $\frac{k}{n} = p + \ll O(\varepsilon) \gg$ (voir plus bas) ?...

Notons que pousser le D.L. de H est également nécessaire pour passer de $H = \frac{1}{n} \log \binom{n}{k} \approx -\frac{k}{n} \log \frac{k}{n} - \left(1 - \frac{k}{n}\right) \log \left(1 - \frac{k}{n}\right)$ à un équivalent de la taille de l'ensemble des suites « presque certaines », qui était initialement 2^{nH} par définition.

Dans « Information Theory, Inference, and Learning Algorithms » (p. 14), on trouve le D.L. obtenu à l'ordre suivant (mais pas encore à $o(1)$ près...) :

$$\log \binom{n}{k} \approx n H\left(\frac{k}{n}\right) - \frac{1}{2} \log [2 \pi n \frac{k}{n} \left(1 - \frac{k}{n}\right)] \text{ où } H \text{ désigne désormais l'entropie binaire.}$$

Le second terme est en $O(\log n)$ et **négatif**, ce qui fournit donc, non un équivalent, mais une **majoration** de la taille 2^{nH} par 2^{nH} . Si l'on conserve la définition initiale de 2^{nH} , cela signifie que la taille de l'ensemble des suites « presque certaines » est **strictement supérieure** à $\binom{n}{k}$, contrairement à la prémisse faite au début, et donc $H > \frac{1}{n} \log \binom{n}{k}$.

Autrement dit, les suites presque certaines ne sont *pas seulement* celles telles $S_n = k =$ l'entier le plus proche de $p n$. Cette « erreur » vient-elle du fait que le pic de Dirac n'est de largeur nulle qu'asymptotiquement et qu'il est nécessaire, pour tout n , de compter également les k tels que $\frac{k}{n} \in [p - \varepsilon, p + \varepsilon]$ — mais alors avec quel « tolérance » ε ?!

Je suis perdu...

Je remarque seulement maintenant la mention du « **width** » dans la phrase :

« Of these, **only a certain subset are likely**; the size of this set is 2^{nH} for $H \leq 1$. By using Stirling's approximation, putting it into the expression for $P(k, n)$, solving for the **location** and **width** of the peak, and finally taking $n \rightarrow \infty$ one finds that

$$H = -p \log p - (1-p) \log(1-p) \gg$$

ce qui rejoint ma problématique de la « tolérance » ε ... Quelle est donc ce « **width** » à laquelle il est fait allusion dans l'article ?!?

Mon interprétation initiale « ponctuelle » du « *likely* » ($S_n = k =$ l'entier le plus proche de $p n$) serait donc incorrecte et par conséquent

le comptage $\binom{n}{k}$ aussi, ce qui expliquerait que l'on obtienne simplement une *minoration* de H ...

...

→ Une piste ?

<https://books.google.fr/books?id=tvQhRUFh7EwC&lpg=PP1&hl=fr&pg=PA21#v=onepage&q&f=false>

(1.4.5) Theorem. Let $0 \leq \lambda \leq \frac{1}{2}$. Then we have

$$(i) \sum_{0 \leq i \leq \lambda n} \binom{n}{i} \leq 2^{nH(\lambda)},$$

$$(ii) \lim_{n \rightarrow \infty} n^{-1} \log \sum_{0 \leq i \leq \lambda n} \binom{n}{i} = H(\lambda).$$

PROOF.

$$\begin{aligned} (i) \quad 1 &= \{\lambda + (1-\lambda)\}^n \geq \sum_{0 \leq i \leq \lambda n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} \\ &\geq \sum_{0 \leq i \leq \lambda n} \binom{n}{i} (1-\lambda)^n \left(\frac{\lambda}{1-\lambda}\right)^{\lambda n} = 2^{-nH(\lambda)} \sum_{0 \leq i \leq \lambda n} \binom{n}{i}. \end{aligned}$$

(ii) Write $m := \lfloor \lambda n \rfloor$. Then $m = \lambda n + O(1)$ for $n \rightarrow \infty$. Therefore we find from Theorem 1.4.2:

$$\begin{aligned} n^{-1} \log \sum_{0 \leq i \leq \lambda n} \binom{n}{i} &\geq n^{-1} \log \binom{n}{m} \\ &= n^{-1} \{n \log n - m \log m - (n-m) \log(n-m) + o(n)\} \\ &= \log n - \lambda \log(\lambda n) - (1-\lambda) \log((1-\lambda)n) + o(1) \\ &= H(\lambda) + o(1) \quad \text{for } n \rightarrow \infty. \end{aligned}$$

The result then follows from part (i). \square

(ii) dit que, pour $\lambda \leq \frac{1}{2}$, $2^{nH(\lambda)}$ est un « **équivalent logarithmique** » de $\sum_{k/n \leq \lambda} \binom{n}{k}$ qui représente l'**effectif cumulé** de la loi binomiale,

c'est-à-dire le **nombre de n -uplets (X_i) comportant une proportion $\leq \lambda$ de « heads ».**

N'y a-t-il pas une incohérence pour $\lambda = \frac{1}{2}$? $H\left(\frac{1}{2}\right) = 1$ or $\sum_{k/n \leq 1/2} \binom{n}{k} \approx \frac{1}{2} \sum \binom{n}{k} = \frac{1}{2} 2^n = 2^{n-1}$ alors que $2^{nH(1/2)} = 2^n$... mais en

fait en équivalent *logarithmique*, c'est correct : $\frac{1}{n} \log 2^{n-1} = \frac{n-1}{n} \longrightarrow 1 = H\left(\frac{1}{2}\right) !$

Moralité : mettre sur le même plan $2^{nH(\lambda)}$ et $\sum_{k/n \leq \lambda} \binom{n}{k}$ peut induire en erreur quant à l'interprétation correcte du théorème (même si

l'on précise « équivalent logarithmique »). Mieux vaut donc revenir à : $\log \sum_{k/n < \lambda} \binom{n}{k} \sim nH(\lambda)$.

Par corollaire, $\log \sum_{\lambda_1 \leq k/n \leq \lambda_2} \binom{n}{k} \sim n(H(\lambda_2) - H(\lambda_1))$, ce qui donne notamment un outil pour estimer le logarithme du nombre de (X_i) comportant une proportion de « heads » à proximité de p , c'est-à-dire entre $p - \varepsilon$ et $p + \varepsilon$: $n(H(p + \varepsilon) - H(p - \varepsilon))$.

...

→ Corrélé :

<https://link.springer.com/article/10.1007%2FBF02458840>

The binomial distribution is the distribution of the total number of successes S_n , in a collection of n independent trials each with the same individual success probability p .

[...] **the large deviation theory of the binomial distribution**: how to approximate the probability of k or more successes in n independent trials, each with success probability p , when the specified fraction of successes, $a = k/n$, satisfies $0 < p < a < 1$.

For tests of significance, the inequality in Theorem 1:

$$P(S_n \geq a n) < e^{-nH},$$

is often the most useful; it is an upper bound which is valid even before taking the limit $n \rightarrow \infty$. The exponential decay rate in this upper bound, H , is sharp, as shown by the first statement in Theorem 2. The second statement in Theorem 2:

$$P(S_n \geq a n) \sim \frac{1}{1-r} \left(\frac{1}{\sqrt{2\pi a(1-a)n}} \right) e^{-nH},$$

gives two refinements to the estimate in Theorem 1: the factor in brackets which corresponds to the central limit theorem for n independent trials with success probability a , in which the variance is $a(1-a)n$, and a factor $1/(1-r)$, [... ?]

...

→ Corrélé :

https://en.wikipedia.org/wiki/De_Moivre%E2%80%93Laplace_theorem

As n grows large, for k in the neighborhood of np we can approximate^{[2][3]}

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi n p(1-p)}} \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right)$$

...

→ Y a-t-il un rapport direct avec ceci ?

L'entropie de la loi binomiale, c'est-à-dire $-\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} \log\left(\binom{n}{k} p^k (1-p)^{n-k}\right)$ possède l'équivalent suivant :

https://en.wikipedia.org/wiki/Binomial_distribution

Entropy: $\frac{1}{2} \log_2(2\pi e n p(1-p)) + O\left(\frac{1}{n}\right)$ in shannons.

...

→ Et ceci ?

<https://mathoverflow.net/questions/200154/expected-centered-entropy-of-the-binomial-distribution>

$$I_n(p) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} [H(p) - H\left(\frac{k}{n}\right)] < \frac{1}{n} \text{ où } H(x) = -x \log x - (1-x) \log(1-x)$$

...

```
import math
# We'll do the same calculation we did above, but in Python
# Passing in 2 as the second parameter to math.log will take a
base 2 log
entropy = -(2/5 * math.log(2/5, 2) + 3/5 * math.log(3/5, 2))
print(entropy)
p = [ income["high_income"].value_counts()[x] / len(income["high_income"]) for x in [0,
1] ]
print(p)
income_entropy = - ( p[0] * math.log(p[0], 2) + p[1] * math.log(p[1], 2) )
print(income_entropy)
```

Output

```
0.9709505944546686
[0.75919044255397561, 0.24080955744602439]
0.796383955202
```

