

# CSYS 300 PoCS Assignment 5

Thomas Waters

October 5th 2018

**Code is located at** [https://github.com/Evelios/PoCS\\_Assignment\\_05](https://github.com/Evelios/PoCS_Assignment_05)

## 1 Problem 1

This problem is about evaluating a diversity index  $D$ , which is easier to intuit about than the entropy number. This diversity index is supposed to represent a different hypothetical text  $T'$  which shares the same entropy  $H$  as the original text  $T$ . The text  $T'$  has all the words appearing in equal frequency  $1/D$ , so there is  $p_i = 1/D$  for  $i = 1, \dots, D$ . This means that we set  $H = H'$ , substituting the  $p_i$  from  $T$  with  $p'_i = 1/D$  in  $H'$  from  $T'$ . Also remembering that the sum of the text is from  $i = 1$  to  $D$ .

### Simpson Concentration

$$S = \sum_{i=1}^n p_i^2 \quad (1)$$

$$S = S' \quad (2)$$

$$\sum_{i=1}^n p_i^2 = \sum_{i=1}^D (1/D)^2 \quad (3)$$

$$\sum_{i=1}^n p_i^2 = \sum_{i=1}^D (1/D)^2 \quad (4)$$

$$\sum_{i=1}^n p_i^2 = D(1/D)^2 \quad (5)$$

$$\sum_{i=1}^n p_i^2 = (1/D) \quad (6)$$

$$D = \left( \sum_{i=1}^n p_i^2 \right)^{-1} = S^{-1} \quad (7)$$

### Gini Index

$$G \equiv 1 - S = 1 - \sum_{i=1}^n p_i^2 \quad (8)$$

$$G = G' \quad (9)$$

$$1 - \sum_{i=1}^n p_i^2 = 1 - \sum_{i=1}^n (1/D)^2 \quad (10)$$

$$\sum_{i=1}^n p_i^2 = \sum_{i=1}^D (1/D)^2 \quad (11)$$

$$\sum_{i=1}^n p_i^2 = D(1/D)^2 \quad (12)$$

$$\sum_{i=1}^n p_i^2 = (1/D) \quad (13)$$

$$D = \left( \sum_{i=1}^n p_i^2 \right)^{-1} = S^{-1} \quad (14)$$

The diversity index for the Simpson Concentration and the Gini Index come out to be the same! The diversity index for the Gini Index can even be written in terms of the Simpson Concentration.

### Shannon's Entropy

$$H = - \sum_{i=1}^n p_i \ln p_i \quad (15)$$

$$H = H' \quad (16)$$

$$H = - \sum_{i=1}^D (1/D) \ln(1/D) \quad (17)$$

$$H = -D(1/D) \ln(1/D) \quad (18)$$

$$H = -\ln(1/D) = \ln D \quad (19)$$

$$D = e^H = e^{-\sum_{i=1}^n p_i \ln p_i} \quad (20)$$

**Renya Entropy,  $q \neq 1$**

$$H_q^{(R)} = \frac{1}{q-1} \left( -\ln \sum_{i=1}^n p_i^q \right) \quad (21)$$

$$H_q^{(R)} = (H_q^{(R)})'$$

$$\frac{1}{q-1} \left( -\ln \sum_{i=1}^n p_i^q \right) = \frac{1}{q-1} \left( -\ln \sum_{i=1}^D (1/D)^q \right) \quad (22)$$

$$-\ln \sum_{i=1}^n p_i^q = -\ln \sum_{i=1}^D (1/D)^q \quad (23)$$

$$\sum_{i=1}^n p_i^q = \sum_{i=1}^D (1/D)^q \quad (24)$$

$$\sum_{i=1}^n p_i^q = D(D)^{-q} \quad (25)$$

$$\sum_{i=1}^n p_i^q = D^{q+1} \quad (26)$$

$$D = \left( \sum_{i=1}^n p_i^q \right)^{\frac{1}{q+1}} \quad (27)$$

**Generalized Tsallis Entropy,  $q \neq 1$**

$$H_q^{(T)} = \frac{1}{q-1} \left( 1 - \sum_{i=1}^n p_i^q \right) \quad (28)$$

$$H_q^{(T)} = (H_q^{(T)})'$$

$$\frac{1}{q-1} \left( 1 - \sum_{i=1}^n p_i^q \right) = \frac{1}{q-1} \left( 1 - \sum_{i=1}^D (1/D)^q \right) \quad (29)$$

$$1 - \sum_{i=1}^n p_i^q = 1 - \sum_{i=1}^D (1/D)^q \quad (30)$$

$$\sum_{i=1}^n p_i^q = \sum_{i=1}^D (1/D)^q \quad (31)$$

$$\sum_{i=1}^n p_i^q = \sum_{i=1}^D (1/D)^q \quad (32)$$

$$\sum_{i=1}^n p_i^q = D(D)^{-q} \quad (33)$$

$$\sum_{i=1}^n p_i^q = D^{q+1} \quad (34)$$

$$D = \left( \sum_{i=1}^n p_i^q \right)^{\frac{1}{q+1}} \quad (35)$$

We see that the Generalized Tallis Entropy and the Renya Entropy come out to be the same!

#### Matching the Tallis Diversity as $q \rightarrow 1$ to Shannon Diversity

$$D_H = e^H = e^{-\sum_{i=1}^n p_i \ln p_i} \quad (36)$$

$$D_T = \left( \sum_{i=1}^n p_i^q \right)^{\frac{1}{q+1}} \quad (37)$$

$$\lim_{q \rightarrow 1} D_T = \lim_{q \rightarrow 1} \left( \sum_{i=1}^n p_i^q \right)^{\frac{1}{q+1}} \quad (38)$$

$$\lim_{q \rightarrow 1} D_T = \left( \sum_{i=1}^n p_i \right)^{1/2} \quad (39)$$

Hmmm, doesn't seem quite right

## 2 Problem 2

Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, \dots, p_n) = F(p_1, p_2, \dots, p_n) + \lambda G(p_1, p_2, \dots, p_n) \quad (40)$$

The 'cost over information' is,

$$F(p_1, p_2, \dots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^n p_i \ln(i+a)}{-g \sum_{i=1}^n p_i \ln p_i} \quad (41)$$

The constrain equation is,

$$G(p_1, p_2, \dots, p_n) = \sum_{i=1}^n p_i - 1 \quad (= 0) \quad (42)$$

Minimizing the cost over information...

$$\nabla F(p_1, p_2, \dots, p_n) = \nabla \frac{\sum_{i=1}^n p_i \ln(i+a)}{-g \sum_{i=1}^n p_i \ln p_i} \quad (43)$$

We can look at the gradient component wise, because when taking the partial derivative of the summation, all other terms other than the  $i = j$  terms cancel out for that particular vector entry.

$$\frac{\partial F}{\partial p_j} = \frac{\partial}{\partial p_j} \frac{p_j \ln(i+a)}{-g(p_j \ln p_j)} \quad (44)$$

$$= \frac{\partial}{\partial p_j} \frac{\ln(i+a)}{-g} \frac{p_j}{(p_j \ln p_j)} \quad (45)$$

$$= \frac{\partial}{\partial p_j} \left( \frac{\ln(i+a)}{-g} \right) \left( \frac{1}{\ln p_j} \right) \quad (46)$$

$$= \left( \frac{\ln(i+a)}{-g} \right) \left( \frac{1}{p_j (\ln p_j)^2} \right) \quad (47)$$

Solving for the gradient of the constrain function...

$$\nabla G(p_1, p_2, \dots, p_n) = \nabla \sum_{i=1}^n p_i - 1 \quad (48)$$

$$\frac{\partial G}{\partial p_j} = \frac{\partial}{\partial p_j} (p_i - 1) \quad (49)$$

$$\frac{\partial G}{\partial p_j} = 1 \quad (50)$$

Solving for the critical point,

$$\frac{\partial \Psi}{\partial p_j} = \frac{\partial F}{\partial p_j} + \lambda \frac{\partial G}{\partial p_j} = 0 \quad (51)$$

$$0 = \left( \frac{\ln(i+a)}{-g} \right) \left( \frac{1}{p_j(\ln p_j)^2} \right) + \lambda \quad (52)$$

$$\left( \frac{\ln(i+a)}{-g} \right) \left( \frac{1}{p_j(\ln p_j)^2} \right) = \lambda \quad (53)$$

$$\frac{\ln(i+a)}{-g\lambda} = p_j(\ln p_j)^2 \quad (54)$$

I should have reached the following equation, so working on from here.

$$p_j = e^{-1-\lambda H^2/gC} (j+a)^{-H/gC} \quad (55)$$

Using the constrain equation,

$$\sum_{j=1}^n p_j = \sum_{j=1}^n e^{-1-\lambda H^2/gC} (j+a)^{-H/gC} = 1 \quad (56)$$

$$\left( e^{-1-\lambda H^2/gC} \right) \sum_{j=1}^n (j+a)^{-H/gC} = 1 \quad (57)$$

$$\sum_{j=1}^n (j+a)^{-H/gC} = e^{1+\lambda H^2/gC} \quad (58)$$

Expected Output

$$\alpha = H/gC \quad (59)$$

$$p_j = (j+a)^\alpha \quad (60)$$

### 3 Problem 3

### 4 Problem 4

$$N_{\leq 200} \approx 3.46 \times 10^8 K^{-0.661} \quad (61)$$