# I. BASIC CONCEPTS



Exercises
Midterm 2018: 1
Midterm 2019: 1-4
Midterm 2021: 1-2
Final 2021: 1ab

# 1. UNIVARIATE UNCERTAINTY, H; DIVERSITY, INFORMATION

### *1. Univariate focus initially*

We're considering only <u>one</u> variable, not a system (defined as involving >1 variable.)

### *2. Probabilistic uncertainty = Shannon entropy*

$H(x) = -\sum p(x_j) \log_2 p(x_j)$ ; for simplicity, $H = -\sum p_j \log_2 p_j = \Gamma(p_1, p_2, ...)$

$\sum$ is 1 to **n**, #number states (cardinality) of x, not sample size, **N**. Kripp reverses n &N.

H goes up with (i) n, the <u>number</u> of possible x states, & (ii) <u>uniformity</u> of $p_j$ distribution

H is average ***surprise*** (assume repeated x measurements). Surprise is log (1/p); then weight by p, then sum.

Units of H is "bits" which are non-physical, using base-2 logarithm.

### *3. Assumptions in derivation of uncertainty (Shannon & Weaver, p. 45)*

$H = \log_2 n$ for equal probabilities; univariate decomposability (discussed later)

### *4. No necessary relation to 2nd Law*

$H(x(t+1))$ need not be $\geq H(x(t))$; comment on Markovian doubly stochastic systems.

### *5. Uncertainty & diversity*

Uncertainty is a measure of **diversity**: economic diversity or ecological (species) diversity or population diversity (e.g., within a species as evolution progresses).

Related to **Ginni** coefficient (often quantifies income inequality). Entropy nice since it's decomposable.

Attaran's 1984 dissertation: relation of economic diversity & per capita income or unemployment in Oregon counties. Found statistically significant relation, but weak. Note difference between statistical **significance** and **size** of effect. Articles widely cited.

### *6. Uncertainty & information*

**1. Information is a change (decrease) in uncertainty**

Now we need **2** probability distributions, one initial & other final, both for one variable: *x*

$\text{information}(x) = -\Delta H = -(H_{final} - H_{initial}) = H_{initial} - H_{final}$

**2. Dangers of sign confusion**

Information and uncertainty have connotation of being opposites,

But note that if $H_{final} = 0$, **information = $\mathbf{H_{initial}}$** : information & $H_{initial}$ are opposites.

## 2. MEASURES & MODELS

Now consider systems. Initially assume 2 variables: x & y.

Notation: **XY** is model; **x, y** are variables. For 2 variables only **two models** need to be considered: **XY and X:Y**; these are the **only models** in the **Lattice of Structures**. XY is the data, the **top** of the lattice, "saturated model," and X:Y is the model that says that variables are independent, the **bottom** of the lattice, "independence model."

A model is a hypothesis, and actually there are possible hypotheses that are not models in the Lattice of Structures, so even with 2 variables, there are other possible models, but we won't consider them here. (We will later.)

For 3 or more variables there are more than two models, i.e., models **in-between top and bottom**; we will not consider them now.

You can think about information theory methodology as
(a) giving you **measures**, like uncertainty (H), Transmission (T), or
(b) letting you assess **models**.

When you talk about **measures without specifying any model, the model=data** is assumed. But one could calculate **measures for particular models** as well.

Notation: probability, p, for a particular model will be called $\mathbf{q_{model}}$ (so q does *not* mean not 1 – p), e.g., probabilities for independence model are written as $\mathbf{q_{X:Y}(x,y)}$.
p(x,y) will be probabilities from the data; it would be same as $\mathbf{q_{XY}(x,y)}$ .

Shannon entropy, H, can be calculated for the data (p) or for a particular model (q).

Notation: In the past, I've used u instead of H, because u goes with uncertainty, and H meant heap, i.e., independence model, but I've decided now to use the same notation that Krippendorff uses (also Shannon and nearly everyone else!).

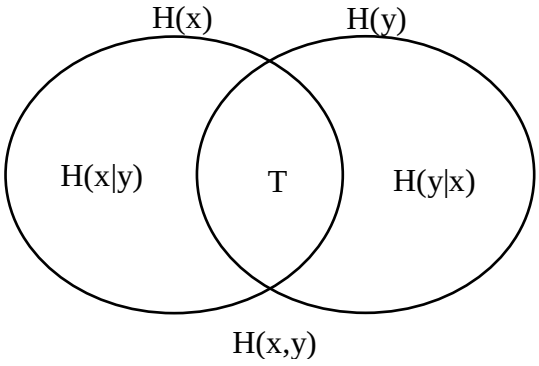
Full notation: $H_{XY}(x, y)$, but sometimes will write H(XY) and sometimes H(x, y); in latter case XY assumed; $H_{X:Y}(x, y)$ is the uncertainty (about x, y) for the independence model.

Can also **condition entropy on one or more variables**. Can be written in two ways: H(y|x) or $H_x(y)$. Again, conditional uncertainties are for **data** or for some **model**.
H( z | x,y ) would be for data, (not bothering to subscript the XYZ data)
$H_{XY:XZ:YZ}$( z | x,y ) is conditional distribution for z, given x & y, for model XY:XZ:YZ.

# 3. BIVARIATE & CONDITIONAL UNCERTAINTIES

### *1. Visualization with uncertainty circles*

Not a Venn diagram (since in diagrams of this sort, areas can be negative!)




### *2. Algebraic derivation*

$H(XY) = H(x, y) = H(x) + H(y|x) = H(y) + H(x|y)$

DERIVATION from joint probabilities into conditional probabilities

$$H(x, y) = -\sum\sum p(x_j , y_k ) \log_2 p(x_j, y_k )$$

$$= -\sum\sum p(x_j)\, p(y_k \mid x_j ) \log_2 [\, p(x_j)\, p(y_k \mid x_j ) \,]$$

$$= -\sum\sum p(x_j)\, p(y_k \mid x_j ) [\log_2 p(x_j) + \log [\, p(y_k \mid x_j ) \,]$$

$$= -\sum\sum p(x_j)\, p(y_k \mid x_j ) [\log_2 p(x_j) \,] - \sum\sum p(x_j)\, p(y_k \mid x_j ) \log [\, p(y_k \mid x_j ) \,]$$

$$= -\sum_x p(x_j) \log_2 p(x_j) \sum_y p(y_k \mid x_j ) - \sum\sum p(x_j)\, p(y_k \mid x_j ) \log [\, p(y_k \mid x_j ) \,]$$

$$= -\sum_x p(x_j) \log_2 p(x_j) - \sum\sum p(x_j)\, p(y_k \mid x_j ) \log [\, p(y_k \mid x_j ) \,]$$

$$= H(x) - \sum\sum p(x_j)\, p(y_k \mid x_j ) \log [\, p(y_k \mid x_j ) \,]$$

$$= H(x) - \sum_x p(x_j) \sum_y p(y_k \mid x_j ) \log [\, p(y_k \mid x_j ) \,]$$

$$= H(x) + \sum_x p(x_j)\, H(y \mid x_j )$$

$$= H(x) + H(y \mid x )$$

For X:Y, i.e., x & y independent, $H(y \mid x) = H(y)$, so $H(x,y) = H(x) + H(y)$

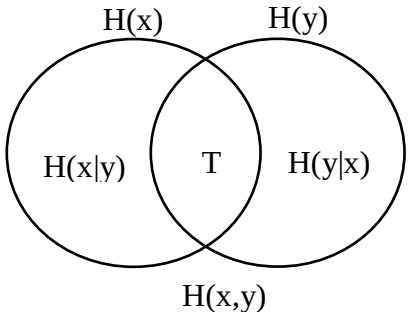
## 4. TRANSMISSION, T (MUTUAL INFORMATION, CONSTRAINT)

H(X:Y) = H(x) + H(y)

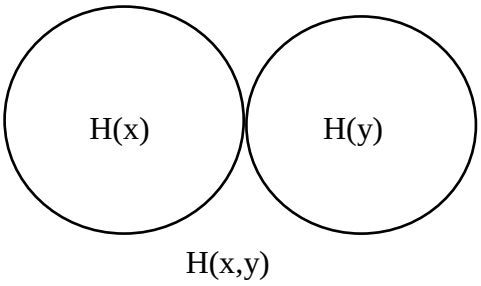
DERIVATION from joint probabilities (above), since p(x, y) = p(x) p(y).

Constraint in the data = T = H(X:Y) – H(XY) Also called “mutual information”

= H(x) + H(y) – H(x, y) = H(x) – H(x|y) = H(y) – H(y|x)




**XY (data)**




**X:Y (model)**

If T = 0, variables independent, heap, H(XY) = H(X:Y), **circles** separate, so H increases.

Transmission is a measure of **association**. It is ***like* correlation**, but T ≥ 0, i.e., T is only positive or zero, not negative. T is for nominal (categorical, qualitative, symbolic) variables, while correlation is for (quantitative) interval or ratio variables.

If you had two variables that are correlated *either* positively or negatively, and you binned them appropriately, T would be positive. But you could **have an association** between quantitative variables **but zero correlation**. **T could detect** this association.

**Two meanings of T:**

1. constraint in XY

2. error in X:Y

In lattice, **constraint** goes **up** from X:Y to XY; **error** goes **down** from XY to X:Y.

Which is **more general**, if go to **more than two** variables? How should we define T if we have XYZ at top, X:Y:Z at bottom, and some model in between, like XY:YZ?

T will be defined as **error**, so write T(X:Y) which may/may not be zero, but T(XY) = 0.

# 5. COMPUTATIONS on CONTINGENCY TABLES

## *5.1 Two variables*

To illustrate how these uncertainty measures are actually computed, consider the following table of data (these are referred to as "contingency tables"):

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | a | b | a+b |
| $x_2$ | c | d | c+d |
| | a+c | b+d | N=a+b+c+d |

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | .1 | .2 | .3 |
| $x_2$ | .3 | .4 | .7 |
| | .4 | .6 | 1 |

These tables are normalized to obtain probabilities:

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | $p(x_1,y_1)$ | $p(x_1,y_2)$ | $p(x_1)$ |
| $x_2$ | $p(x_2,y_1)$ | $p(x_2,y_2)$ | $p(x_2)$ |
| | $p(y_1)$ | $p(y_2)$ | 1 |

Uncertainties are given by

$H(x) = -\sum p(x_i) \log p(x_i)$
$H(y) = -\sum p(y_j) \log p(y_j)$
$H(x,y) = -\sum\sum p(x_i, y_j) \log p(x_i, y_j)$

Conditional uncertainties can be calculated **two ways**:

1. $H(x|y) = p(y_1) H(x|y_1) + p(y_2) H(x|y_2)$, where

$H(x|y_1) = -p(x_1|y_1) \log p(x_1|y_1) - p(x_2|y_1) \log p(x_2|y_1)$ where

$p(x_1|y_1) = p(x_1,y_1)/p(y_1) = a/(a+c)$
$p(x_2|y_1) = p(x_2,y_1)/p(y_1) = c/(a+c)$

2. $H(x|y) = H(x,y) - H(y) = \Gamma(a/N, b/N, c/N, d/N) - \Gamma((a+c)/N, (b+d)/N)$

Do this for the above **numerical** distribution.

### *5.2 More than 2 variables*

For > 2 variables & for non-dichotomous variables, the procedure is essentially the same:

| | $z_1$ | | $z_2$ | |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_1$ | $y_2$ |
| $x_1$ | a | b | c | d |
| $x_2$ | e | f | g | h |

N=a+b+...+h. So a, b, ...h are frequencies

The entropy (uncertainty) of the overall **system**, H(x,y,z) is just

$H(x,y,z) = - a/N \log a/N - b/N \log b/N \ldots - h/N \log h/N$

$= \Gamma (a/N, b/N, \ldots, h/N)$

**Redefine a, b, …h as probabilities**, not frequencies, so don't need to divide by N.

The entropy of any **single** variable is gotten by an appropriate aggregation:

$H(x) = - \sum p(x_1) \log p(x_1) - p(x_2) \log p(x_2)$

where $p(x_1) = (a + b + c + d)$ and $p(x_2) = (e + f + g + h)$

$H(x) = \Gamma (a + b + c + d, \quad e + f + g + h)$

$H(y) = \Gamma (a + e + c + g, \quad b + f + d + h)$

$H(z) = \Gamma (a + b + e + f, \quad c + d + g + h)$

Show the entropy of all three **pairs** of variables H(x,y), H(x,z), H(y,z).

$H(x,y) = \Gamma (a + c, \quad b + d, \quad e + g, \quad f + h)$

$H(x,z) = \Gamma (a + b, \quad c + d, \quad e + f, \quad g + h)$

$H(y,z) = \Gamma (a + e, \quad b + f, \quad c + g, \quad d + h)$

**Conditional uncertainties** for more than one variable are easily calculated in **2 ways**:

1. $H(x|yz) = H(x|y_1,z_1)\ p(y_1,z_1)+H(x|y_1,z_2)\ p(y_1,z_2)+H(x|y_2,z_1)\ p(y_2,z_1)+H(x|y_2,z_2)\ p(y_2,z_2)$

where $H(x|y_1,z_1) = -\ a/(a+e) \log a/(a+e) - e/(a+e) \log e/(a+e)$; $p(y_1,z_1) = (a+e)$

2. also 2nd way, easier: $H(x|yz) = H(xyz) - H(yz)$

### *5.3 Testing association directly from probabilities*

Independence model, X:Y. what are its probabilities? Give example with .1,.2,.3,.4
$q_{X:Y}(x,y) = p(x)\ p(y)$.

In general, for >2 variables, when all multivariate probabilities are simply the products of marginals, i.e., if $p(x_i, y_j, z_k, ...) = p(x_i) * p(y_j) * p(z_k) ...$, then there is no association,

### *5.4 Calculated probabilities maximize entropy subject to constraints*

This model q distribution is the **solution** to **maximizing entropy, $H_{X:Y}(x,y)$, subject to constraint, $q_{X:Y}(x) = p(x)$ & $q_{X:Y}(y) = p(y)$**. (Can be proven using LaGrange multipliers)

Consider the following data, **p(x,y)**

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | .1 | .2 | .3 |
| $x_2$ | .3 | .4 | .7 |
| | .4 | .6 | 1 |

| | $y_1$ | $y_2$ |
|---|---|---|
| $x_1$ | $p_1$ | $p_2$ |
| $x_2$ | $p_3$ | $p_4$ |

1

The independence model distribution $q_{X:Y}(x,y)$ gotten by multiplying p(x) by p(y)

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | .12 | .18 | .3 |
| $x_2$ | .28 | .42 | .7 |
| | .4 | .6 | 1 |

**This distribution is the solution to following optimization subject to constraints: maximize $\Gamma(q_1, q_2, q_3, q_4)$ subject to $q_{X:Y}(x) = p(x)$ & $q_{X:Y}(y) = p(y)$**.

| | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | $q_1$ | $q_2$ | .3 |
| $x_2$ | $q_3$ | $q_4$ | .7 |
| | .4 | .6 | 1 |

**A model tells us <u>constraints</u>, i.e., what we know. Otherwise the <u>calculated distribution</u> is <u>maximally uniform</u>. How many constraints? 2.**

**Linear algebra form ("structure matrix) 2 independent rows (rank = 2):**

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} .1 \\ .2 \\ .3 \\ .4 \end{bmatrix}$$

## 6. A STATE DECOMPOSITION of UNIVARIATE UNCERTAINTY

Information theory is nice because information measures are **decomposable**.

Decomposability here has 2 meanings (now considering only first meaning):

subsets of **states** (macro, micro); subsystems of **variables** (subsystems)

Uncertainties are decomposable: let x = macro-state, y = micro-state

Consider values $y_1$ and $y_2$ to be subset of $x_1$ and values $y_3$ and $y_4$ to be subset of $x_2$. Note here **subsets are of possible values** of **one variable**, as opposed to **subsets of variables**.

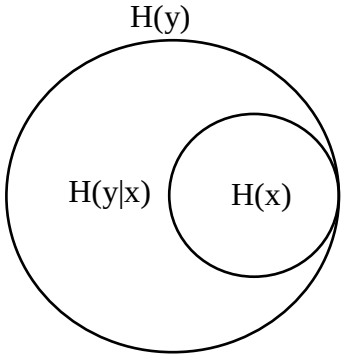
| $x_1$ | | $x_2$ | |
|---|---|---|---|
| $y_1$ | $y_2$ | $y_3$ | $y_4$ |

| .3 | | .7 | |
|---|---|---|---|
| .1 | .2 | .3 | .4 |

Let **within** = within subsets; **between** = between subsets. Like analysis of variance in a population divided into groups; like diversity within groups and between groups. One could have a lot of diversity in every group but there wouldn't be much difference between group averages or very little diversity in every group but great differences between group averages, or anything in between.

$H = H_{within} + H_{between}$

Derivation:
The overall uncertainty in x,y is no bigger and indeed is the same as uncertainty in just y.
CIRCLES: x circle **within** y circle




$H(x, y) = H(y) =$ **$\Gamma$(.1, .2, .3, .4)**

$= H(x) + H(y|x)$

$= H(x) + \sum p(x_i)\, H(y|x_i)$

$= H(x) + \sum p(x_i) \sum p(y_j|x_i) \log p(y_j|x_i)$

where $p(y_j|x_i) = p(x_i, y_j) / p(x_i)$

$=$ **$\Gamma$(.3, .7) + .3 $\Gamma$(.1/.3, .2/.3) + .7 $\Gamma$(.3/.7, .4/.7)**

$=$ H[ between ] + H[ weighted within ]

$=$ H[ between ] + $\sum p(x_i)$ H[ within i ]

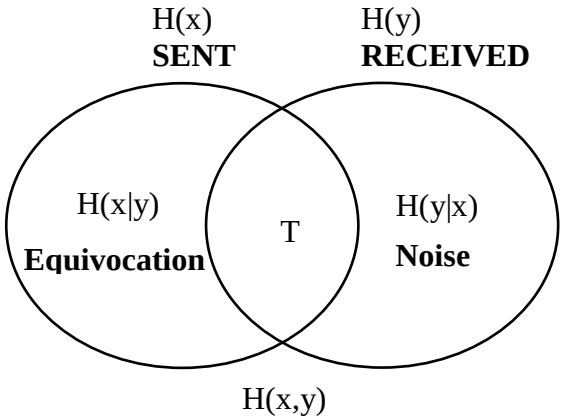
Examples: Attaran's economic decomposition of macro-sectors of economy, Shirazi's analysis of trade between/within geographic regions. ALife paper where decompose **diversity of behavioral actions** of a population into diversity of the population under **identical** environmental conditions plus the diversity of the population under **differing** environmental conditions.

# 7. T in 'TRANSMISSION' & 'SEQUENTIAL' SITUATIONS

## *7.1 Transmission situation*

Sent and received; **T ("transmission") is what is sent and is received**.

Equivocation = H(x|y); Noise = H(y|x)




K, p.21 explains m:1 for equivocation and 1:m for noise. Also see K, p.25, Figure 12.

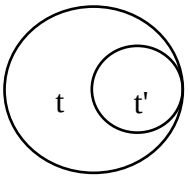
## *7.2 Sequential situation*

(G. A. Miller) x & y are NOT sent & received, but rather **x = z(t), y=z(t+1) or z(t')**

Sequential situation: **temporal linkage**. Redundancy in natural language is illustrated by uncertainty of a second letter in a word, given the first, as in TV show, **Wheel of Fortune**. Statistical approach used in machine translation (calculate, instead of understand word sequences), Google searches, ChatGPT, etc. See my paper on information, constraint, & meaning (**ICM**).

Here dynamic relations are viewed information-theoretically. T measures **temporal order:** T(z(t),z(t')) measures the constraint between values of a variable at two times.

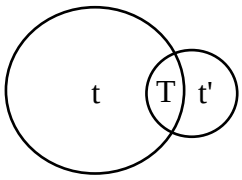
For **deterministic** systems initial time circle includes later time circle, which is smaller. (In 1:1 mappings, circles are identical.) If one knows initial state then one knows absolutely the final state. Even if one doesn't know initial state (initial H is not zero), then final uncertainty must be same or smaller.




Ashby, in *Introduction to Cybernetics*, talks about **decay of variety in deterministic systems**. Ashby speculates that this relates to **Second Law**, but it's the **opposite**! His "Law of experience": in (deterministic) machines with input, changes in input parameters only **preserves or reduces variety**; one always loses information about initial state.

In stochastic systems, T measures **degree of non-stochasticity**.

Can of course have more than one variable changing in time.



# 8. T as LIKELIHOOD RATIO; RELATION to UNCERTAINTY

## *8.1 Likelihood Ratio form of T; T is for a particular model*

$T(X:Y) = H(x) - H(x|y) = H(y) - H(y|x) = H(x) + H(y) - H(x,y)$

**Another way** to represent T is $T(X:Y) = \sum\sum p(x,y) \log [p(x,y)/q_{X:Y}(x,y)]$, where

q(x, y) are **expected probabilities** based on some **hypothesis**. Here, for $q_{X:Y}$, the hypothesis is independence. Krippendorff uses π for q (p.24).

Note that if q = p, then log p/q = log 1 = 0. then T=0. T measures the **gap between p & q distributions**, the difference log p – log q weighted by observed probabilities, p.

T is directly related to **likelihood-ratio Chi-square**.

Note in conventional statistics, there are (at least) two Chi-square ways of indicating the gap between 2 distributions: see K, p.87 for **ordinary** and **likelihood-ratio** forms.

$T(X:Y) = \mathbf{\sum\sum p(x,y) \log [p(x,y)/q_{X:Y}(x,y)]}$

$= \sum\sum p(x,y) \log p(x,y) - \sum\sum p(x,y) \log q_{X:Y}(x,y)$

Under hypothesis of independence, $q_{X:Y}(x,y) = p(x)\,p(y)$. Hence

$= \sum\sum p(x,y) \log p(x,y) - \sum\sum p(x,y) \log p(x) - \sum\sum p(x,y) \log p(y)$

$= \sum\sum p(x,y) \log p(x,y) - \sum\sum p(x)\,p(y|x) \log p(x) - \sum\sum p(y)\,p(x|y) \log p(y)$

$= -H(x, y) - \sum_x p(x) \log p(x) \sum_y p(y|x) - \sum_x p(y) \log p(y) \sum_y p(x|y)$

$= -H(x, y) - \sum p(x) \log p(x) - \sum p(y) \log p(y)$

$= -H(x, y) + H(x) + H(y)$

Hypothesis of independence is hypothesis that model X:Y has no error. For simplicity, T can be written as T(X:Y) without reference to variables, x & y. When arguments are given, it should be clear whether they're model arguments or variable arguments. The most complete convention would be to write $\mathbf{T_{X:Y}(x, y)}$, i.e., the transmission for model X:Y, which depends on variables x and y.

For 2 variables the saturated model, XY, and independence model, X:Y, are only models possible, but with more than 2 variables, the lattice of structures has other models possible, e.g., we could have $T_{XY:YZ}(x, y, z)$.

$T = \sum p \log [p/q]$ is actually a **more general expression** than the earliest expression introduced just in terms of H's, which **applies only to the independence model**.

T same as I in K, p.87, where I is **information distance** from top to some model, i.e., the information **loss (error)** in the model. I is **NOT** the information **captured** in the model.

$LR = L^2 = 2n \sum p \log_e [p/\pi]$ where **n** is sample size (my **N**) & $\pi$ is my q.

But $\log_e x \equiv \ln x = \ln 2 * \log_2 x$ derived below

So shifting to N for sample size,

$L^2 = 2N \sum p \ln 2 \log_2[p/\pi] = 2 N \ln 2 T = (2 \ln 2) N T =$ **1.3863 N T**

**T** is a **sample size independent** measure of information loss

$\mathbf{L^2}$ is a **sample size dependent** measure of information loss, so can get statistical significance (a **p-value**) for it.

**Proof that $\ln x = \log_2 x * \ln 2$**

**$y = \ln x$** so $x = e^y$ **$z = \log_2 x$** so $x = 2^z$

$e^y = 2^z$ so $e^{y/z} = 2$ so $y/z = \ln 2$ **$y = z \ln 2 = \log_2 x \ln 2$**

**$\ln x = \log_2 x * \ln 2$**

### *8.2 Application to Univariate Uncertainty*

For single variable, q(x) can't be the expected probability distribution for some different structural model, since there is no other **model**, [for 1 variable, no lattice!] but under some **hypothesis**. For example, hypothesis might be that prob. distribution is **uniform**.

So **hypothesis is more general than model**. A hypothesis could be that a particular model (topological connectedness of variables) holds, but one could have some other type (**non-structural**) of hypothesis.

Consider the **hypothesis** of **uniform distribution**, a null hypothesis, what we expect if we know nothing about a distribution, i.e., the least biased distribution, the maximum uncertainty distribution. (the Laplace criterion, Laplace's principle of insufficient reason says that we should assume uniform distribution, maximum ignorance or uncertainty, in absence of information about the distribution.)

$T = \sum p \log p/q$, where $q(x) = 1/n$

**T** $= \sum p \log ( p n ) = \sum p \log p + \sum p \log n = - H(x) + \log n =$ **$H_{max}(x) - H(x)$**

So T measures **difference from uniform** distribution. All we have to do is multiply T by the constant above to get a $L^2$ that we can use, with appropriate degrees of freedom (just n-1 because p's sum to 1), to **test the null hypothesis that distribution is uniform**. What is $H_{max}(x)$? It is entropy of uniform distribution model. What is H(x)? It is data.

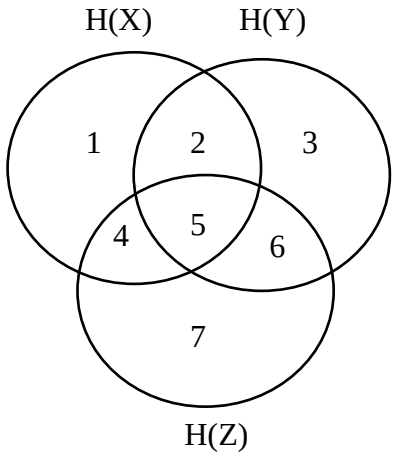
******* In general, **T(model) = H(model) – H(data)** = error in model *******

**Krippendorff incorrectly says that this equation not true for models with loops.**

**What is correct is that cannot write H(model) algebraically for models with loops.**

For 3 variables, that means we **can't write algebraic expression for H(XY:YZ:XZ)**

## 9. T, H for TRIVARIATE (& HIGHER) RELATIONS




Now shift to model notation with capital letters.

Repeat: **T(model) = H(model) – H(data)**

T(X:Y:Z) = H(X:Y:Z) – H(XYZ)

= H(X) + H(Y) + H(Z) – H(XYZ)

= ~~1~~ ~~2~~ **5** ~~4~~ **2** ~~3~~ ~~5~~ ~~6~~ **4 5 6** ~~7~~ – ~~1~~ ~~2~~ ~~3~~ ~~4~~ ~~5~~ ~~6~~ ~~7~~

= 2 4 6 5 5

**Consider a more complex model, XY:Z**

T(XY:Z) = H(XY:Z) – H(XYZ)

= ~~1~~ ~~2~~ ~~3~~ ~~4~~ ~~5~~ ~~6~~ 4 5 6 ~~7~~ – ~~1~~ ~~2~~ ~~3~~ ~~4~~ ~~5~~ ~~6~~ ~~7~~

= 4 5 6

H(XY:Z) = H(XY) + H(Z)

**q(XY:Z) = p(XY) p(Z)**

**Entropies add/substract, probabilities multiply/divide**

Consider a **still more complex** model: one with **<u>overlapping</u> components, XY:YZ**

T(XY:YZ) = H(XY:YZ) – H(XYZ)

H(X**Y**:**Y**Z) = H(XY) + H(YZ) **– H(XY∩ YZ)** third term is XY intersect YZ.

= H(XY) + H(YZ) **– H(Y)**

**q(XY:YZ) = p(XY) p(YZ) / p(Y)**

**A slightly more complex model, XY:YZ:ZA**

T(XY:YZ:ZA) = H(XY:YZ:ZA) – H(XYZA)

H(X**Y**:**Y**Z:ZA) = H(XY) + H(YZ) + H(ZA) **– H(XY∩ YZ) – H(YZ∩ ZA)**

= H(XY) + H(YZ) + H(ZA) **– H(Y) – H(Z)**

**q(XY:YZ:ZA) = p(XY) p(YZ) p(ZA) / [ p(Y) * p(Z) ]**

**Still more complex** model: with **overlapping components <u>AND LOOPS</u>, XY:YZ:XZ**

T(XY:YZ:XZ) = H(XY:YZ:XZ) – H(XYZ)

**But <u>cannot expand H or q expressions since</u>** XY:YZ:XZ has a **<u>loop</u>, so**

**H(XY:YZ:XZ) ≠ H(XY) + H(YZ) + H(XZ) –H(X) –H(Y) –H(Z)**

**q(XY:YZ:XZ) ≠ p(XY) p(YZ) p(XZ) / [p(X) p(Y) p(Z)]**

## *TWO ALGEBRAIC LAWS*

### *Law of Uniform Subscripting: applies to <u>both H & T</u>*

$H(X:Y|Z) = H(X|Z) + H(Y|Z)$ (or, other notation:) $H_Z(X:Y) = H_Z(X) + H_Z(Y)$

$H_Z(XY) = H_Z(X) + H_Z(Y|X) = H_Z(X) + H_{XZ}(Y)$

### *Law of Distribution for <u>conditional</u> measures: applies to <u>T</u> but <u>not H</u>*

**T(A:C)** illustrates **INDEPENDENCE**
**T(A:C|B)** illustrates **<u>CONDITIONAL</u> INDEPENDENCE**

**T(A:C|B) = T(AB:BC)**

**move B into each argument**

$T(A:C|B) = H(A:C|B) - H(AC|B)$ **Uniform Subscripting for T**

$= H(A|B) + H(C|B) - H(AC|B)$ **Uniform Subscripting for H**

$= H(AB) - H(B) + H(BC) - H(B) - [H(ABC) - H(B)]$

$= H(AB) + H(BC) - H(B) - H(ABC)$

Simpler derivation

$T(AB:BC) = H(AB:BC) - H(ABC) = H(AB) + H(BC) - H(B) - H(ABC)$

BUT this **Law of Distribution** does **NOT** apply to **<u>H</u>**.

So **H(A:C|B)** =/= **H(AB:BC)**

<u>Left Hand Side</u>: $H(A:C|B) = H(A|B)+H(C|B) = H(AB) - H(B) + H(BC) - H(B)$

<u>Right Hand Side</u>: $H(AB:BC) = H(AB) + H(BC) - H(B)$

LHS =/= RHS, i.e., H(A:C|B) and H(AB:BC) are **<u>NOT</u>** the same

## 10. A VARIABLE DECOMPOSITION of TRANSMISSION

The transmission measure lets us speak about **organization** of the system. One simple type of organization is division into **disjoint subsystems**.

Consider a system with variables v, w, x, y, divided into two subsystems v, w and x, y.

The transmission of the total system can be broken up into terms giving the transmission **within** each of the subsystems taken separately plus a transmission **between** one subsystem and the other. Shift to **model notation**.

$T(V:W:X:Y) = \mathbf{T}_{\text{within subsystems}} + \mathbf{T}_{\text{between subsystems}}$

$= T(V:W) + T(X:Y) + T(VW:XY)$

(This is mentioned in Conant's article, Laws of Information That Govern Systems.)

Derivation:

LHS: $T(V:W:X:Y) = H(V) + H(W) + H(X) + H(Y) - H(VWXY)$

RHS: $T(V:W) = H(V) + H(W) - H(VW)$

RHS: $T(X:Y) = +H(X) + H(Y) - H(XY)$

RHS: $T(VW:XY) = +H(VW) + H(XY) - H(VWXY)$

Add up last three terms in RHS & get overall T of the LHS.

Since transmission is a **measure of organization**, this decomposability is **fundamental to idea of system**. That is, a system has parts, with some internal order, but the parts are organized into a larger whole.

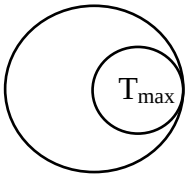
Related to Simon's notion of **nearly decomposable systems:** in most cases

$\mathbf{T}_{\text{between}} << \mathbf{T}_{\text{within}}$

# 11. OTHER INFORMATION THEORETIC FUNCTIONS

### *1. (K, p.24) normalized T*

$T(A:B) / T_{max}(A:B) = T(A:B) / \min\{ H(A), H(B) \}$

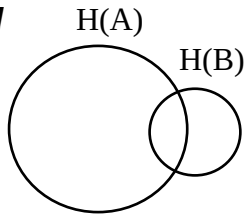



### *2. (McGill & Quastler) fraction of uncertainty explained*

D function of M&Q who call it "coefficient of constraint"

This measure is for predicting B (the DV) from A (the IV)

**D = T(A:B)/H(B)** = [ H(B) – H(B|A) ] / H(B)

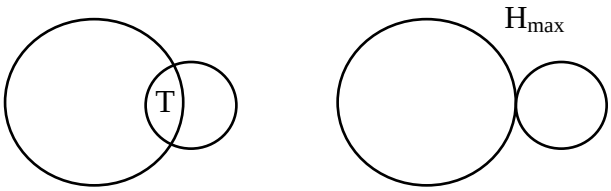



### *3. Predictive power (efficiency of prediction)*

**T(A:B)/H(A)**

### *4. (M&A) Redundancy*

**C** = 1 – H / $H_{max}$ = [ $H_{max}$ – H ] / $H_{max}$ = **T / $H_{max}$**



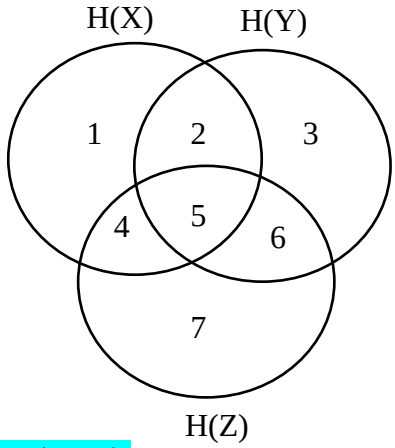

### *5. Interaction, A (sometimes called Q)*

**A** = T(X:Y) – T(X:Y|Z) = 2 + 5 – 2 = 5

= T(X:Z) – T(X:Z|Y) = 4 + 5 – 4 = 5

= T(Y:Z) – T(Y:Z|X) = 6 + 5 – 6 = 5

= 5, the inherently **triadic** interaction




**A** = T(X:Y) – T(X:Y|Z)

= H(X) + H(Y) – H(XY) – T(XZ:YZ)

= H(X) + H(Y) – H(XY) – H(XZ:YZ) + H(XYZ)

= H(X) + H(Y) – H(XY) – [ H(XZ) + H(YZ) – H(Z) ] + H(XYZ)

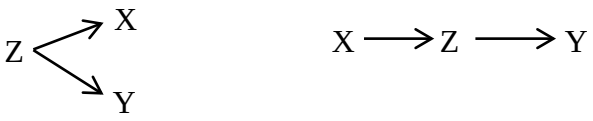
= **H(X) + H(Y) + H(Z) – H(XY) – H(XZ) - H(YZ) + H(XYZ)**

Note the alternating signs. One could also define A as the negation of this.

Natural to assume that region **5 is positive**, i.e., that **T(X:Y) = 2 + 5 ≥ T(X:Y|Z) = 2.** What X tells me about Y is reduced by knowing Z, i.e., expect that **T(X:Y|Z) ≤ T(X:Y)**.

**In the limit, T(X:Y|Z) = 0. What would this mean, if T(X:Y) > 0?**

**It could mean that X-Y association might be due to (explained away by) either a <u>prior</u> effect or an <u>indirect</u> effect.**




Prior effect is like **Factor Analysis**: if X and Y are associated (correlated), knowing (**controlling for**) a common (**prior**) factor, Z, which explains both of them & **explains away** the association (e.g., sibling correlation, given parents). For nominal variables this is **Latent Class Analysis**.

**Indirect effect** comes from **Path Analysis**: if X and Y are associated, then knowing an intermediate (mediating) variable, Z, **explains away** the association.

**But is it always the case that T(X:Y) ≥ T(X:Y|Z)**?

**NO! Can get '<u>reverse</u>' latent class analysis**: consider following

| | $z_1$ | | $z_2$ | |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_1$ | $y_2$ |
| $x_1$ | 1/4 | 0 | 0 | 1/4 |
| $x_2$ | 0 | 1/4 | 1/4 | 0 |

| | $y_1$ | $y_2$ |
|---|---|---|
| $x_1$ | 1/4 | 1/4 |
| $x_2$ | 1/4 | 1/4 |

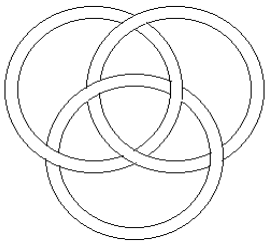
| | $z_1$ | $z_2$ |
|---|---|---|
| $x_1$ | 1/4 | 1/4 |
| $x_2$ | 1/4 | 1/4 |

| | $z_1$ | $z_2$ |
|---|---|---|
| $y_1$ | 1/4 | 1/4 |
| $y_2$ | 1/4 | 1/4 |

Given $z_1$

| | $y_1$ | $y_2$ |
|---|---|---|
| $x_1$ | 1/2 | 0 |
| $x_2$ | 0 | 1/2 |

**Borromean Rings: triadic constraint; dyadic projections uniform (no constraint!!)**



H(XYZ) = 2, H(XY) = 2, H(XZ) = 2. H(YZ) = 2, H(X) = H(Y) = H(Z) = 1

T(X:Y) = H(X) + H(Y) – H(XY) = 1 + 1 – 2 = 0 Not surprising: uniform

T(X:Y|Z) = .5 * T(X:Y|$z_1$) + .5 * T(X:Y|$z_2$)

T(X:Y|$z_1$) = H(X|$z_1$) + H(Y|$z_1$) – H(XY|$z_1$) = 1 + 1 – 1 = 1 = T(X:Y|$z_2$)

T(X:Y|Z) = .5 * 1 + .5 * 1 = 1

So **T(X:Y|Z) = 1 > T(X:Y) = 0**

**A** = T(X:Y) – T(X:Y|Z) = 0 – 1 = –1.

So **T(X:Y|Z)** is **<u>not always</u> less** than or equal to **T(X:Y)**,

But **H(X|Z)** is **always** less than or equal to **H(X)**

Negative interactions. Hence **not** a **Venn** diagram. A can be negative (or positive)!!

In this case, **knowing Z makes association stronger rather than weaker**.

Give **interpretation** of this table for the **couples example**.

**Borromean Rings: triadic constraint, but no dyadic constraints !!**

**T(X:Y:Z)** = H(X:Y:Z) – H(XYZ) = 3 – 2 = **1**,

But **T(X:Y) = T(X:Z) = T(Y:Z) = 0**.

**A *looks like*** the inherently **triadic interaction. We *might think* A = T(XY:XZ:YZ), error in 3 dyadic interactions w/o the triadic interaction. What is this T?**

**T(XY:XZ:YZ) = H(XY:XZ:YZ) – H(XYZ)**

**We might think INCORRECTLY that**

**H(XY:XZ:YZ) = H(XY) + H(XZ) + H(YZ) – H(X) – H(Y) – H(Z)**

**This *looks like it might be* –A, but this H expression is INCORRECT <u>because of loop</u>.**

**If A were positive, entropy would decrease going down the lattice, which can't be.**

**An OPPOSITE (more <u>conventional</u>) example** from Latent Class Analysis book, p.16, where **T(X:Y|Z) = 0 < T(X:Y). This is <u>opposite</u> of Borromean rings.**

| | $z_1$ | | $z_2$ | |
|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_1$ | $y_2$ |
| $x_1$ | 80 | 20 | 15 | 35 |
| $x_2$ | 40 | 10 | 30 | 70 |

| | $y_1$ | $y_2$ |
|---|---|---|
| $x_1$ | 95 | 55 |
| $x_2$ | 70 | 80 |

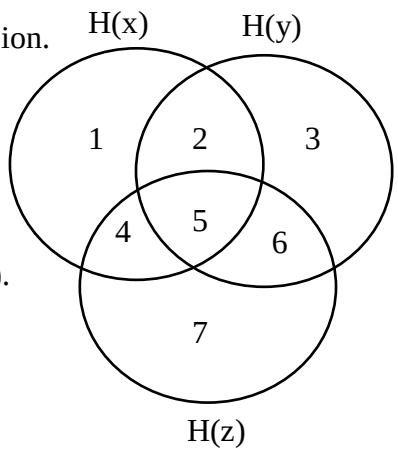
**T(X:Y) is not 0, but T(X:Y|Z) = .5 T(X:Y|$z_1$) + .5 T(X:Y|$z_2$) = 0 + 0**

### *5. Systematic entropy*

Shift back to variable notation.
Useful to define quantity similar to uncertainty called systematic entropy (Krippendorff):

**S(x, y, z) = H(x,y,...z) – H( x | y,z) – H( y | x,z) – H( z | x,y)** = **2 4 6 5**

**Systematic entropy** does **not count region 5 twice**, as does transmission.




Systematic entropy, S, measures **joint variability**, uncertainty of system associated with (that exists despite) **organization** of system:

it **excludes uncertainties of each variable taken singly,** unaffected by knowledge of other variables, which could be considered **monadic noise** (like **unique variance in Factor Analysis**).