Jensen-Shannon Measures for the Parsing of Crosstabulation Tables

Patrick Habecker, MA - UNL

Bilal Khan, Ph.D. - CUNY

Kirk Dombrowski, Ph.D. - UNL

Imagine for a few minutes that you work for this man, Police Commissioner Gordon of the Gotham Police.

- **Your job is to analyze the meth markets in Gotham**
- **To do this you conducted a survey of meth users asking about attributes, relationships, and sources of meth**
- **Now you have to start making policy recommendations**

Crosstabulations

You decided to see if people of different sexual ID obtain meth differently.

Table 1: Frequencies

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	19	22	24	65
MSM/W	7	14	10	31
Other	14	8	14	36
Total	40	44	48	

Real NYC data

Do different groups source meth differently?

What tools do you have? χ^2, theory, and intuition.

Jensen-Shannon Divergence

- Measure the amount of information lost if categories are combined
- Information loss by variable for all possible collapse combinations
- Empirical method
- Has been used in the fields of cell biology, physics, and data mining

How does this work?

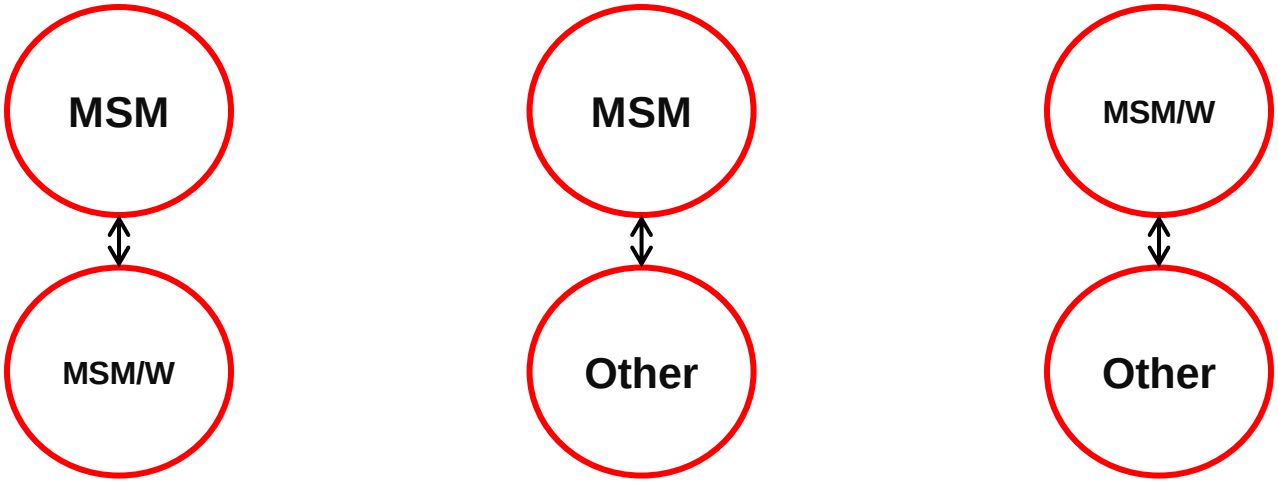
1. **Examine the different possible collapse combinations for our categories.**
2. **Then calculate the Jensen-Shannon Divergence for each possible grouping.**

Table 1: Frequencies

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	19	22	24	65
MSM/W	7	14	10	31
Other	14	8	14	36
Total	40	44	48	

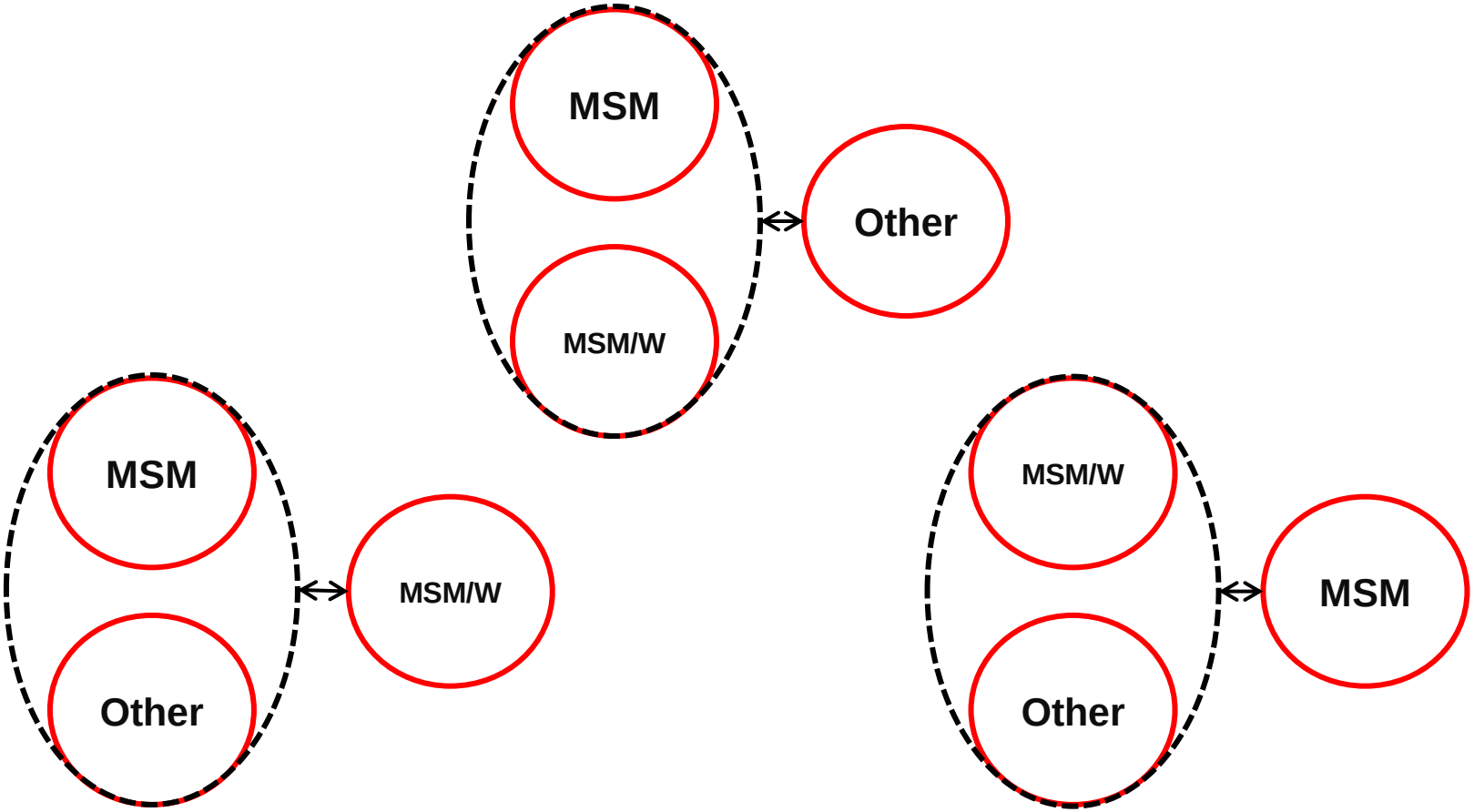
The Groupings

3 groupings that ignore the third category



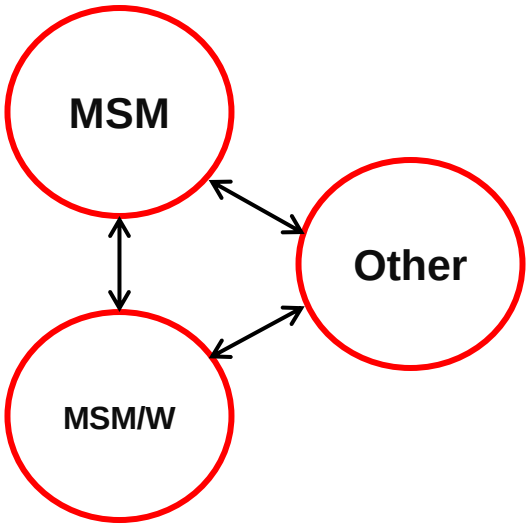
The Groupings

3 groupings that combine two and compare with the third



The Groupings

1 final grouping that compares all three individually, the default



Calculating the JSD

- **Difference between**
  - The *entropy of the average* of a set of *n* probability distributions
  - The *average of the entropy* of each taken separately

- **Step 1: Pick a grouping and calculate the probabilities**

Table 1: Frequencies

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	19	22	24	65
MSM/W	7	14	10	31
Total	26	36	34	

Table 2: Probabilities

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	0.292	0.338	0.369	1.0
MSM/W	0.226	0.452	0.323	1.0
Average	0.259	0.395	0.346	1.0

Average of the Entropy

- **Step 2: Calculate the average of the entropy of each group separately**
  - $x * \log_2\left(\frac{1}{x}\right)$ for each cell and sum across categories (rows)
  - $\left[0.292 * \log_2\left(\frac{1}{0.292}\right)\right] + \left[0.338 * \log_2\left(\frac{1}{0.338}\right)\right] + \left[0.369 * \log_2\left(\frac{1}{0.369}\right)\right] = 1.578$
  - $\left[0.226 * \log_2\left(\frac{1}{0.226}\right)\right] + \left[0.452 * \log_2\left(\frac{1}{0.452}\right)\right] + \left[0.323 * \log_2\left(\frac{1}{0.323}\right)\right] = 1.529$
- **Step 3: Calculate the average of the group entropies.**
  - $Average\ of\ the\ Entropy = \frac{1.578+1.529}{2} = 1.554$

Table 2: Probabilities

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	0.292	0.338	0.369	1.0
MSM/W	0.226	0.452	0.323	1.0
Average	0.259	0.395	0.346	1.0

Entropy of the Average

- **Step 4: Calculate the entropy of the average of the set of probabilities**
  - $0.259 * \log_2\left(\frac{1}{0.259}\right) = 0.505$
  - $0.395 * \log_2\left(\frac{1}{0.395}\right) = 0.529$
  - $0.346 * \log_2\left(\frac{1}{0.346}\right) = 0.530$
  - $Entropy\ of\ the\ Average = 0.505 + 0.529 + 0.530 = 1.564$

Table 2: Probabilities

	Where meth was last obtained			
Sexual ID	Other	Delivery	Supplier's Home	Total
MSM	0.292	0.338	0.369	1.0
MSM/W	0.226	0.452	0.323	1.0
Average	0.259	0.395	0.346	1.0

The Divergence

- **Step 5: Calculate the Jensen-Shannon Divergence**
  - $JSD = Entropy\ of\ the\ Average\ - Average\ of\ the\ Entropy$
  - $JSD = 1.564 - 1.554 = 0.01$
- **We repeat the process for every possible grouping, providing seven JSD figures**

Table 3: JSD Results for LastMeth	
MSM - MSM/W	0.01
MSM - Other	0.01
MSM/W - Other	0.05
MSM+MSM/W - Other	0.02
MSM+Other - MSM/W	0.02
MSM/W+Other - MSM	0.00
MSM - MSM/W - Other	0.03

This is useful, but we can't compare the groupings

Using the JSD

- **By creating a matrix of √JSD ratios we can compare different groupings with each other**

Table 4: √Jensen-Shannon Divergence Ratiod - Sexual Id by Source of Meth

		A	B	C	D	E	F	G
MSM - MSM/W	A		B/A	C/A	D/A	E/A	F/A	G/A
MSM - Other	B	A/B		C/B	D/B	E/B	F/B	G/B
MSM/W - Other	C	A/C	B/C		D/C	E/C	F/C	G/C
MSM+MSM/W - Other	D	A/D	B/D	C/D		E/D	F/D	G/D
MSM+Other - MSM/W	E	A/E	B/E	C/E	D/E		F/E	G/E
MSM/W+Other - MSM	F	A/F	B/F	C/F	D/F	E/F		G/F
MSM - MSM/W - Other	G	A/G	B/G	C/G	D/G	E/G	F/G	

- **Larger numbers indicate that little information is lost by the denominator compared to that lost by the numerator**
- **Looking across rows gives us information about denominator**

Using the JSD

- **What does this tell us?**

Table 4: √Jensen-Shannon Divergence Ratiod - Sexual Id by Source of Meth

		A	B	C	D	E	F	G
MSM - MSM/W	A		1.174	2.153	1.493	1.400	0.194	1.764
MSM - Other	B	0.852		1.835	1.273	1.193	0.166	1.503
MSM/W - Other	C	0.464	0.545		0.694	0.650	0.090	0.819
MSM+MSM/W - Other	D	0.670	0.786	1.442		0.937	0.130	1.181
MSM+Other - MSM/W	E	0.714	0.838	1.538	1.067		0.139	1.260
MSM/W+Other - MSM	F	5.144	6.037	11.075	7.682	7.201		9.072
MSM - MSM/W - Other	G	0.567	0.665	1.221	0.847	0.794	0.110	

- **Large numbers indicate that the denominator is favorable**
- **Numbers near 1 indicate approximate parity**
- **Which grouping is best?**

Using the JSD

- So what do we tell Commissioner Gordon?

- MSM/W and our Other meth users have similar sourcing patterns
- MSM have different sourcing patterns from MSM/W & Other

- Efforts that don't take this difference into account may miss certain meth using groups

One step further

- **By taking the row mean of the √JSD ratio matrix we can look at how these values vary across multiple variables**

Table 5: √SJD Row Means Across A Range of Variables

Variables	*MSM-MSM/W*	*MSM - Other*	*MSM/W - Other*	*MSM+MSM/W - Other*	*MSM+Other - MSM/W*	*MSM/W+Other - MSM*	*MSM - MSM/W - Other*
Number of sex partners last month	1.23	0.81	0.85	0.83	1.18	1.08	1.21
Attitudes about sex with meth	1.01	1.09	0.94	1.10	1.01	1.29	0.72
Where meth was last obtained	1.36	1.14	0.54	0.86	0.93	7.70	0.70
How meth was paid for	0.71	2.08	0.65	1.80	0.70	2.12	0.75
Engaged in sex work	0.58	1.50	1.19	4.95	0.72	0.91	0.75
Number of methsex last month	1.81	0.62	0.96	0.71	4.02	0.98	0.76
Never sex without meth	0.67	0.95	2.99	1.53	0.91	0.81	0.79
Never or seldom sex without meth	0.70	0.73	30.63	1.17	1.16	0.72	0.77
Dependent on meth (self-rating)	0.84	2.44	0.56	1.26	0.71	3.99	0.69

Words of caution

- **Distinguishing the degree to which a √JSD ratio, or √JSD ratio row mean is meaningfully larger than another large value is difficult (7 vs 5)**
- **There is no magic number (1.2 vs 1.0)**
- **Blind use of this could be problematic, theory has a role**
- **Looking at more than 4 categories is problematic and even 4 results in 23 possible groupings**

That said

- **The JSD can empirically highlight differences between groups that may be overlooked with other methods**

Thank you!

reach.unl.edu soc.unl.edu