Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
MIDTERM EXAM Fall 2018 (Oct 24, 2018)

60 points; 2 hours; closed book, no notes. *Answer questions on exam sheets, and put **your name** on them.* EXCEPT for question 1(h), you can use Γ(a, b...) = -a log a - b log b -....

1. [24 points]
(a) Shannon entropy can be interpreted as diversity when probabilities are fractional proportions. Consider four counties of Oregon with fractions of employment in three economic sectors (A, B, C), as given below. D(1), D(2), D(3), D(4) are the employment diversities of the four counties. The ordinal ranking of these four diversity magnitudes is: D( ) > D( ) > D( ) > D( ): Put the right numbers in the parentheses on the left. (2 pts)

	Sector			
County	A	B	C	Diversity
1	.50	.50	.00	D(1)
2	.25	.25	.50	D(2)
3	.33	.33	.33	D(3)
4	.00	.00	1.00	D(4)

(b) Probability distribution I has values p(x) = .4, .3, .2, .1, for x = 1, 2, 3, 4 respectively, and probability distribution II has values p(x) = .4, .1, .2, .3 for x = 1, 2, 3, 4 respectively. One can calculate variances for these two distributions, where variance is the sum of square distances of values of x from the average x, weighting each square distance by the probability of that value. Entropy is a nominal data measure that *resembles* variance in that it also measures the *spread* of a distribution. Question: True or false? (circle one; 2 pts): The entropy of distribution I = the entropy of distribution II.

(c) (2 pts) For the contingency table whose X and Y margins are shown below, write an expression in terms of the constants, a, b, c, d for T_{max}, the maximum value that T(X:Y) could possibly have: T_{max} =

	Y_1	Y_2	
X_1			a
X_2			b
	c	d	

(d) Data ABZ is fit with the AZ:BZ model. Is the following statement True or False? (circle one; 2 pts): The AB distribution projected from the calculated ABZ for this AZ:BZ model *can* exhibit some non-zero-strength constraint, i.e., it is possible that, for this projected AB distribution, T(A:B) > 0.

(e) Given a directed system with variables A, B, C, and Z, the difference of transmissions, T(ABC:AZ) – T(ABC:ABZ) = (circle all that are true, 4 pts)

i. H(Z|A) – H(Z|AB)	ii. H(Z|A) + H(Z|AB)		
iii. H(Z|B) – H(Z|AB)	iv. H(Z|B) + H(Z|AB)		
v. H(Z|AB) – H(Z|A)	vi. H(Z|AB) + H(Z|A)		
vii. H(Z|AB) – H(Z|B)	viii. H(Z|AB) + H(Z|B)		
ix. T(A:Z)	x. T(B:Z)	xi. T_B(A:Z)	xii. T_A(B:Z)
xiii. T(AB:Z)	xiv. T(ABZ)	xv. none of the above	

(f) Let the data be the contingency table below, with known probabilities, a…h. Give an expression for T(A:Z | B_1) in terms of parameters, a…h (2 pts).

	Z:	0		1	
	B:	0	1	0	1
A:	0	a	b	c	d
	1	e	f	g	h

(g) True or false? (circle; 2 pts): For a directed ABZ system, the entropy of Z equals the transmission between Z and one predictor, plus the transmission given this predictor between Z and the other predictor, plus an unexplained (not reduced) entropy, i.e., $H(Z) = T(A:Z) + T_A(B:Z) + H_{AB}(Z)$. (You can prove this assertion or its negation, but a proof is *not* required here.)

(h) Consider the following table. Give NUMERICAL answers. IN *THIS* QUESTION, THE GAMMA FUNCTION IS NOT ALLOWED.

	C_1		C_2	
	B_1	B_2	B_1	B_2
A_1	.25	0	0	.25
A_2	0	.25	.25	0

Calculate the entropies of the relations in the *Lattice of Relations*:

(h1) H(ABC) =_____(1 pt)

(h2) H(AB) =_____; H(AC) =_____; H(BC) =_____(1 pt)

(h3) H(A) =_____; H(B) =_____; H(C) =_____(1 pt)

Calculate now the entropies of the models in the *Lattice of Structures*:

(h4) H(AB:AC:BC) =_____(2 pts) In general, one can't *directly* calculate entropy for models with loops, but *in this particular case*, one can know its value by reasoning.

(h5) H(AB:BC)=_____; H(AC:AB)=_____; H(BC:AC)=_____(1 pt)

(h6) H(AB:C) =_____; H(AC:B) =_____; H(BC:A)_____(1 pt)

(h7) H(A:B:C) =_____(1 pt)

2. [6 points]
(a) True or false? (circle one; 2 pts): For data m_0 and models m_j and m_k that are fit to the data, if $df(m_j) = df(m_k)$, then $T(m_j) = T(m_k)$.

(b) True or false? (circle one; 2 pts): H(m), the Shannon entropy of model m, does *not* depend on whether the reference is the top or the bottom.

(c) Consider data below for a *neutral* system, with probability values a…h. In the Lattice of Structures presented in class, the bottom model is the independence model, X:Y:Z. An alternative is the uniform distribution which doesn't preserve the X, Y, and Z marginal distributions. There are other models *between* X:Y:Z and the uniform distribution; for example, the X:Y:Φ model, which says that q(XYZ) agrees with the X and Y margins but is uniform (Φ) in Z. Using parameters a through h, what transmission quantity would you use to assess the agreement of this X:Y:Φ model with the data? Express this quantity using $\sum$ p log p/q expression (*not* using entropies) and give only *the first two terms* of the expression. (2 pts)

X	Y	Z	p
0	0	0	a
0	0	1	b
0	1	0	c
0	1	1	d
1	0	0	e
1	0	1	f
1	1	0	g
1	1	1	h

3. [14 points]
(a) Consider a directed system in which A, B, and C are IVs that might affect (or predict) Z, the DV. What *two specific structures* would one compare to find out if there is a tetradic *interaction effect* between all three IVs in their effect on (prediction of) Z? (2 pts)

(b) Consider data, ABC, for three binary variables. What model would you calculate the transmission for to test the null hypothesis that whatever the AB relation is (whether A and B are or are not be mutually constrained), it is independent of C? (2 pts) (The motivation for this question is that if C is time, then this will test whether or not relation AB changes with time.)

(c) What is the nearest common ancestor of ABC:BCD and ABD:CA:CB:CD? (2 pts)

(d) What is the nearest common descendant of ABC:BCD and ABC:DA:DB:DC? (2 pts)

(e) Does ABC:BCD have loops? Yes No (circle one; 1 pt)

(f) Does ABC:DA:DB:DC have loops? Yes No (circle one; 1 pt)

(g) Δdf(ABC:ABD:ACD:BCD $\rightarrow$ AB:AC:AD:BC:BD:CD) = df(ABC:ABD:ACD:BCD) – df(AB:AC:AD:BC:BD:CD). Let A, B, C, & D have cardinalities of 2, 3, 3, & 4, respectively. Use the log-linear method and compute this Δdf. (2 pts)

Δdf =

(h) Suppose I go <u>down</u> the lattice of structures. After each measure, (i) to (iii), write whether the measure is MD (monotonically decreasing or staying the same), MI, (monotonically increasing or staying the same), or NM (not monotonic, i.e., could either increase or decrease) as I go down the lattice (2 pts):

(i) transmission, T
(ii) entropy, H
(iii) degrees of freedom, df

4. [10 points] On the left is an *observed* probability table (p) for a *directed* system, with sample size N. None of parameters (a...h) is 0. Let the *calculated* table (q) for model AB:AZ be the table on the right.

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	a	b	e	f
A_2	c	d	g	h

	Z_1		Z_2	
	B_1	B_2	B_1	B_2
A_1	q_1	q_2	q_3	q_4
A_2	q_5	q_6	q_7	q_8

(a) Fitting the AB:AZ model involves optimizing an expression subject to a set of constraints. In (a1), write the expression that is maximized or minimized. In (a2) and (a3), write the set of linearly independent constraint equations *in terms of $q_1 \ldots q_8$ and a…h*. (The general constraint $\sum q_j = 1$ that holds for all models should not be included.)

(a1) minimize/maximize (*circle* one of these words) *what*? (write the expression that is optimized in terms of $q_1 \ldots q_8$ and/or a…h) (2 pts)

(a2) subject to____ (state *how many* linearly independent) AB constraints. And give the constraint equation(s) here (2 pts):

(a3) and ____ AZ constraints: state here how many, if any, <u>additional</u> AZ constraints there are, beyond the AB constraint(s). And give the constraint equation(s) here. (2 pts)

(b1) Write an expression in terms of a…h for the amount of constraint (information) that is *captured* (*not that is lost*!) in the AB:AZ model. (2 pts)

(b2) What is Δdf, difference in degrees of freedom, between these two models? (2 pts)

Δdf =

5. [6 points] Given an XY data distribution as follows

	y_0	y_1
x_0	.1	.2
x_1	.3	.4

(a1) I want to consider the state-based model X_1Y_1 that specifies that $p(x_1,y_1) = .4$. Write an expression for transmission, T, the error in this model, in terms of the Γ function and numerical constants. (2 pts)

T =

(a2) What is the value of $\Delta df = df(XY) - df(X_1Y_1)$? (1 pt)

Δdf =

(b) Suppose I want to test the hypothesis that the table is "really" uniform, i.e., that the deviations of its probability values from .25 are just due to sampling error.

(b1) In terms of the numerical constants of the table, write an expression for the T that should be used to test this hypothesis (2 pts).

T =

(b2) The model corresponding to this hypothesis has df = (circle one; 1 pt)

0 1 2 3 4 5 6 7 8