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 <u>all three</u> 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* $a \ldots h$. (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 this model and the reference used in (b1) [originally: "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

Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
MIDTERM EXAM Fall 2019 (Oct 30, 2019)

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

I. BASICS [32 points]

1. [14 points]
(a) A deck of cards has 13 ranks (Ace, 2-10, Jack, Queen, King) and 4 suits (hearts, diamonds, clubs, spades), 52 different cards in all. I pick a card. Telling you (honestly) that I picked a diamond reduces your uncertainty by how much? (Circle one; 2 pts)

i. $\log_2(52/13)$	ii. $\log_2(52/4)$	iii. $\log_2(13/4)$
iv. $\log_2(13/52)$	v. $\log_2(4/52)$	vi. $\log_2(4/13)$

(b) 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)

(c) For the following contingency table, with probabilities a+b+c+d=1, if X and Y are independent, then (circle all that are true; 2 pts)

	Y_1	Y_2
X_1	a	b
X_2	c	d

i. a/b = c/d ii. a/c = b/d iii. a/d = b/c iv. none of the above

(d) Let the level of knowledge of a student about information theory be a dichotomous variable, K, whose value is high (h) or low (l). Assume that an exam testing students on their knowledge makes only a binary discrimination, i.e., the grade, G, is also either h or l. The probabilities of possible situations give the following table, where a+b+c+d = 1:

		G: l	G: h
K	l	a	b
	h	c	d

(d1) Write an expression in terms of probabilities a, b, c, d for H(K|G), the uncertainty which remains about a student's knowledge when his/her grade is known. (2 pts)

(d2) It would clearly be desirable that H(K|G) be zero, i.e., that there were no uncertainty about a student's actual knowledge once his/her grade is known. If this were true, what must be true about the numerical values of the probabilities a, b, c, d? (2 pts)

(d3) Write an (equality and/or inequality) equation involving *only* the univariate uncertainties, H(K), H(G), and the transmission, T(K:G), which would hold if the exam not only resolved all uncertainty about K but was actually *more detailed and complex* than needed to resolve this uncertainty. (2 pts)

(e) Circle all choices which give a correct expression for transmission, T(X:Y) in terms of the following uncertainties and conditional uncertainties of x and y (2 pts):

i. H(x|y) + H(y|x) – H(x,y)
ii. H(x,y) – H(x|y) – H(y|x)
iii. H(x,y) + H(x|y) + H(y|x)
iv. H(x) – H(x|y)
v. H(x) – H(y|x)
vi. H(y) – H(y|x)
vii. H(y) – H(x|y)
viii. H(x|y) + H(y|x)
ix. H(x) + H(y) – H(x,y)
x. H(x,y) – H(x) – H(y)

2. [8 points] Let the data be the table below, with known probabilities, a…h.

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

(a) Give an expression for T(A:Z) in terms of parameters, a…h, by using the definition of T(A:Z) in terms of model and data entropy values (2 pts).

(b) Give an expression for T(A:Z) in terms of parameters, a…h, using the definition of T(A:Z) as a sum of p log (p/q) terms. Give only the *first two terms* of this sum (2 pts).

(c) Give an expression for T_B(A:Z) in terms of parameters, a…h (2 pts).

(d) Let the DV = Z and the IVs = A, B, C. If one interprets "the predictive effect of A on Z, *controlling for* B and C" to mean how much knowing A tells us about Z beyond what B and C tells us about Z, which transmission value gives the magnitude of the predictive effect of A, controlling for B and C? (circle one; 2 pts)

i. T(BCZ:AZ) ii. T(ABC:AZ) iii. T(ABC:BCZ) iv. None of these

3. [4 points] Consider relations defined over time.
(a) For an AB table where A = x(t) and B = x(t+1), which of the following, if true, implies that AB system *must* be stochastic (not deterministic)? (circle one; 2 pts)

i. H(A) = H(B) ii. H(A) > H(B) iii. H(B) > H(A)
iv. H(A|B) = 0 v. H(B|A) = 0 vi. None of these

(b) Let time be a variable, as follows. Consider ABC_1 and ABC_2 distributions below, where C_1 = t and C_2 = t+1, with probabilities, a+b+c+d = 1 and e+f+g+h = 1. Treating the AB distribution at time t as the reference, I wish to test the hypothesis that the AB distribution at t+1 is the same as it was earlier at t. Using the p log (p/q) expression for transmission, write the T in terms of a…h that will let me test this hypothesis. (2 pts)

C_1

	B_1	B_2
A_1	a	b
A_2	c	d

C_2

	B_1	B_2
A_1	e	f
A_2	g	h

T =

4. [6 points]
(a) For neutral system ABCD, I consider a disjoint model AB:CD. It is always the case that (circle one; 2 pts):

i. T(A:B) + T(C:D) = T(AB:CD)
ii. T(A:B) + T(C:D) + T(AB:CD) = 0
iii. T(A:B) + T(C:D) + T(AB:CD) = T(ABCD)
iv. T(A:B) + T(C:D) + T(AB:CD) = T(A:B:C:D)
v. None of these

(b) 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., $H(Z) = T(A{:}Z) + T_A(B{:}Z) + H_{AB}(Z)$. (A proof of this assertion or its negation is *not* required for this question.)

(c) The following table is for a 3-variable Borromean rings, where there is non-zero constraint in ABC, but no constraint in any of the bivariate projections.

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

It is possible to construct a 4-variable analog to the above table. Give this analog by filling in the following table with numerical values. (2 pts)

	D_1				D_2			
	C_1		C_2		C_1		C_2	
	B_1	B_2	B_1	B_2	B_1	B_2	B_1	B_2
A_1								
A_2								

II. STRUCTURES [28 points]

5. [7 points]
(a) For some ABZ data, $T_A(B:Z) = 0$. Circle all models where T(model) is also 0. (2 pts)

i. ABZ	ii. AB:AZ:BZ	iii. AB:AZ	iv. AB:BZ	v. AZ:BZ
vi. AB:Z	vii. AZ:B	viii. BZ:A	ix. A:B:Z	x. none of these

(b) Consider a directed system in which A, B, and C are IVs, and Z is a DV. List all the specific structures (not general structures merely illustrated by specific structures), i.e., list all the structures that one would need to consider in modeling actual data. *Circle* all structures which do *not* have loops (5 pts).

6. [8 points]
(a) Let CBWAITO mean "cannot be written algebraically in terms of." Circle all choices that are true for models with loops. (4 pts)

i. H(model) CBWAITO the uncertainties of model components
ii. T(model) CBWAITO uncertainties of the model & the data
iii. *calculated* model probabilities CBWAITO *observed* data probabilities
iv. df(model) CBWAITO df values of model relations & their projections

(b) True or false? (circle; 2 pts): Given the convention that directed systems must have an IV component which includes all the IVs, a directed system model that has two or more ("predicting") relations that involve the same DV must necessarily have a loop.

(c) Suppose a problem has few variables, so it is practical to evaluate all models in the Lattice of Structures, and I use the data as my reference in a top-down analysis. I know that model complexity (df) varies monotonically as one goes down the LOS and model error (T) also varies monotonically as one goes down the LOS. True or False? (circle one; 2 pts): The "best model," which would have minimum complexity and minimum error, is easy to choose, since no tradeoff is needed between complexity and error.

7. [8 points] One can sometimes come up with *multiple* good models. It would be handy to have a measure for the similarity or difference between any two models, m_1 and m_2 where this measure would be 0 if the models have nothing in common and 1 if the models were identical. Consider a measure constructed from two quantities, $f(m_1 \cup m_2)$ and $f(m_1 \cap m_2)$, where f is some suitable property of these models gotten by taking the union and intersect.

(a) Which of the following ratios of these quantities varies *directly*, not inversely, with the degree of *similarity* of the two models? (2 pts, circle one)

i. $f(m_1 \cup m_2) / f(m_1 \cap m_2)$ ii. $f(m_1 \cap m_2) / f(m_1 \cup m_2)$
iii. both iv. neither

(b) Actually, one wants *two* different types of measures: (i) a measure of similarity between two structures *without regard to data*, and (ii) a measure of similarity between models, where 'model' means a structure *applied to some actual data*.

(b1) What is a good choice for f for a data-*<u>in</u>dependent* measure of the similarity or dissimilarity of the two structures? (circle one; 2 pts)

i. H ii. df iii. Neither of these

(b2) What is a good choice for f for a data-*dependent* measure of the similarity or dissimilarity of the two models? (circle one; 2 pts)

i. H ii. df iii. Neither of these

(c) For NCA = nearest common ancestor & NCD = nearest common descendent, circle the statement below that is true (2 pts):

i. $m_1 \cup m_2 =$ NCA; $m_1 \cap m_2 =$ NCD ii. $m_1 \cup m_2 =$ NCD; $m_1 \cap m_2 =$ NCA iii. Neither

8. [5 points]
(a) p(ABZ) can be written as p(AB)$*$p(Z|AB). We might use for this the notation $AB{:}_{AB}Z$ where the subscript to the left means 'conditioned on.' Since p(AB) and p(Z|AB) can be independently specified, df(ABZ) = df(AB) + df(Z|AB). Given that |A|=|B|=|Z|=2,

(a1) How many parameters does it take to specify AB? ___ Give a number. (1 pt)

(a2) How many parameters does it take to specify p(Z|AB)? ___ Give a number. (2 pts)

(b) Klir's equivalence classes, called ρ structures, each represented by an ordinary graph which indicates which variables are directly linked by only one relation, where each class has a most complex C-structure and a least complex P structure, is (circle one; 2 pts)

i. a way to search the lattice for loopless models
iii. a way to search the lattice for disjoint searches
iii. a way to search the lattice hierarchically

Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
MIDTERM EXAM Spring 2021 (April 28, 2021)

60 points, open book; 1 hr. + 50 min. + 30 min. extra. Γ(a, b,...) = – a log a – b log b - ...

1. [18 points]
(a) 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 squared distance by the probability of the value of x. Entropy is a nominal data measure that *resembles* variance in that it also measures the *spread* of a distribution. The statement that "The entropy of distribution I = the entropy of distribution II" is (choose one; 2 pts)

i. True ii. False iii. Insufficient information to decide.

(b) Consider the contingency (probability) table, where a+b+c+d = 1

	y_1	y_2
x_1	a	b
x_2	c	d

Let x be one word in a message; let y be the next word that follows x. Using the Γ function write an expression, in terms of parameters, a, b, c, & d, for the amount of information which receiving y provides that is beyond what one already knows having previously received x. The answer should be some sum of positive and/or negative Γ terms. To illustrate the format of the answer being asked for: a possible answer (one that is totally wrong) for this question might look like "Γ(a, d) + Γ(b, c) – Γ(a, b, c)". (2 pts)

(c) List all choices which give a correct expression for transmission, T(X:Y), as a function of entropies and conditional entropies of x and y (2 pts):

i. H(x|y) + H(y|x) – H(x,y)
ii. H(x,y) – H(x|y) – H(y|x)
iii. H(x,y) + H(x|y) + H(y|x)
iv. H(x) – H(x|y)
v. H(x) – H(y|x)
vi. H(y) – H(y|x)
vii. H(y) – H(x|y)
viii. H(x|y) + H(y|x)
ix. H(x) + H(y) – H(x,y)
x. H(x,y) – H(x) – H(y)

(d) A genomics researcher wants to use information theory to quantify the strength of the *interaction effect* between two causal variables, A and B, representing two genes (or DNA bases, called SNPs), and some effect variable, Z, which is the presence or absence of a disease. The researcher is interested in the joint effect of A and B on Z that is *not* due to the occurrence of the *separate* effects of A on Z and B on Z, and proposes to use the following two measures to quantify this *3-way interaction effect*:

(1) $-H(A) - H(B) - H(Z) + H(AB) + H(AZ) + H(BZ) - H(ABZ)$, and

(2) $H(A) + H(B) + H(Z) - H(ABZ)$

A critic of this proposal argues that neither (1) nor (2) is a valid measure of the strength of the 3-way interaction effect.

(d1) Is the critic right about measure (1)? (Choose one; 2 pts)

i .Yes ii. No

(d2) Is the critic right about measure (2)? (Choose one; 2 pts)

i .Yes ii. No

(e) (6 pts) The frequency tables below describe patients who are either given medical treatment (M=1) or not (M=0) and who either recover from their disease (R=1) or not (R=0). Assume that recovery from disease brings major positive benefits to patients, and that the treatment itself has very minor negative effects.

(e1.1) The table below is data on *male* patients. The value of T(M:R), the strength of constraint between medical treatment (M) and recovery (R) is: (Choose one; 2 pts)

	R	0	1
M	0	100	0
	1	300	100

i. $\Gamma(.1, .4) + \Gamma(.4, .1) - \Gamma(.1, .1, .3)$
ii. $-\Gamma(.1, .4) - \Gamma(.4, .1) + \Gamma(.1, .1, .3)$
iii. $\Gamma(.2, .8) + \Gamma(.8, .2) - \Gamma(.2, .2, .6)$
iv. $-\Gamma(.2, .8) - \Gamma(.8, .2) + \Gamma(.2, .2, .6)$
v. None of the above

(e1.2) Given this (e1.1) table, should one treat male patients? (Choose one; 1 pt)

i. Yes ii. No

(e2) The table below is data on *female* patients. Based on this data below, should one treat female patients? (Choose one; 1 pt)

i. Yes ii. No

	R	0	1
M	0	200	300
	1	0	200

(e3.1) Suppose that *instead* of the above tables, you received the table below, where male and female patients are *aggregated*. The value of T(M:R), the strength of constraint between medical treatment (M) and recovery (R) for patients whose gender is unspecified is: (Choose one; 1 pt)

	R	0	1
M	0	300	300
	1	300	300

i. 9 ii. 6 iii. 3 iv. 2
v. 1 vi. 0 vii. None of these

(e3.2) If one were given only this (e3.1) table, which has no gender information, should one treat patients? (Choose one; 1 pt)

i. Yes ii. No

(f) You want to predict variable Z; that is, you want to reduce its entropy. Assume you can purchase knowledge of p(AZ) or p(BZ) but not both and also not p(ABZ). Knowing either A or B will reduce the entropy of Z. To help you decide whether to purchase p(AZ) or p(BZ), you are told the values of four numbers: H(A), H(B), H(AZ), H(BZ). You should purchase p(AZ) and not p(BZ), i.e., A will be better than B as a predictor of Z, if (choose one; 2 pts)

i. H(A) < H(B)
ii. H(A) > H(B)
iii. H(AZ) < H(BZ)
iv. H(AZ) > H(BZ)
v. H(AZ) + H(B) < H(BZ) + H(A)
vi. H(AZ) + H(B) > H(BZ) + H(A)
viii. Insufficient information to decide

2. [12 points]
(a) For *any* ABC data (choose one; 2 pts):

i. $T(A:B) \geq T_C(A:B)$ ii. $T(A:B) \leq T_C(A:B)$ iii. Either i or ii could be true

(b) (choose one; 2 pts) $T(ABC:BZ) - T(ABC:ABZ) =$

i. $H(Z) - H(Z|A)$ ii. $H(Z|A) - H(Z)$
iii. $H(Z) - H(Z|B)$ iv. $H(Z|B) - H(Z)$
v. $H(Z) - H(Z|AB)$ vi. $H(Z|AB) - H(Z)$
vii. $H(Z|AB) - H(Z|A)$ viii. $H(Z|A) - H(Z|AB)$
ix. $H(Z|AB) - H(Z|B)$ x. $H(Z|B) - H(Z|AB)$
xi. None of these

(c) Consider the following frequency table.

	C_0		C_1	
	B_0	B_1	B_0	B_1
A_0	12	18	15	15
A_1	28	42	35	35

What is the numerical value of $T_C(A:B)$? Choose one (2 pts) (To answer this, you do not actually need to do any logarithmic calculations. Hint: break table up into C_0 & C_1 parts.)

i. 0 ii. > 0 but ≤1 iii. > 1 but ≤ 2
iv. > 2 but ≤ 3 v. None of these

(d) Two variables are 'directly linked' if they are involved in the same relation. Suppose $T_B(A:Z) = 0$. This means (choose one; 2 pts)

i. A is not directly linked to B ii. A is not directly linked to Z
iii. B is not directly linked to Z iv. all of the above
v. None of these

(e) The statement, "$T(A:B:C) = T(A:B) + T(AB:C)$" is (choose one; 2 pts)

i. True ii. False

(f) The statement, "$H(Z) – H(Z|AB) = T(B:Z) + T_B(A:Z)$" is (choose one; 2 pts):

i. True ii. False

3. [12 points]
(a1) The statement, "If a structure has a loop, then a child of this structure (a descendant, a structure lower in the lattice) will necessarily have a loop" is (choose one; 1 pt).

i. True ii. False

(a2) The statement, "If a structure has a loop, then a parent of this structure (an ancestor, a structure higher in the lattice) will necessarily have a loop" is (choose one; 1 pt).

i. True ii. False

(b) Calculating $\Delta df = df(m_i) – df(m_j)$ without actually calculating either $df(m_i)$ or $df(m_j)$ individually (choose one; 2 pts)

i. can be done by the Krippendorff method of df calculation
ii. can be done by the log-linear method of df calculation
iii. can be done by both methods
iv. cannot be done by either method

(c) Consider ABC, where |A|=2, |B|=3, |C|=4, where | | means cardinality (# of states).
(c1) State the numerical value of df(AB:AC) by the *Krippendorff* method by writing an equation where the right hand side of the equation is the df value and the left hand side indicates all the added and/or subtracted numbers that total to the right hand side value. To illustrate the format of the answer here being asked for: an example of such an equation is "2+3+4–1–2=6"; this is not the correct answer for this question. (2 pts)

(c2) State the numerical value of df(AB:BC) by the *log-linear* method; that is, write an equation where the right hand side of the equation is the df value and the left hand side indicates all the added and/or subtracted numbers that total to the right hand side value. To illustrate the format of the answer here being asked for: an example of such an equation is "2+3+4–1–2=6"; this is not the correct answer for this question. (2 pts)

(d) The statement, “Since models that have loops are topologically more complicated than models that don’t have loops, models that have loops will always have higher degrees of freedom than models without loops involving the same set of variables” is (choose one; 2 pts)

i. True ii. False

(e) The statement, “For models with loops, df must be calculated iteratively” is (choose one; 2 pts).

i. True ii. False

4. [18 points]
(a) ABZ is data on a directed system. In each of questions (a1) through (a3) below, your answer should either be *one* of the following transmissions or a *difference between two* of the following transmissions:

T(AB:Z) T(AB:AZ) T(AB:BZ) T(AB:AZ:BZ) T(ABZ)

(a1) What is the strength of the purely triadic interaction of A, B, and Z? That is, what is the error in a model with no triadic interaction but only dyadic ones? (2 pts)

(a2) What is the magnitude of the *error* in model AB:AZ:BZ? (2 pts)

(a3) What is the magnitude of the information *captured* in model AB:AZ:BZ? (2 pts)

(b) Suppose we allow AZ:BZ to be an acceptable directed system model, even though it doesn’t have an component consisting of all the IVs. Suppose also that T(AZ:BZ) = 0. The statement that “T(A:B) must therefore also be 0” is (choose one; 2 pts)

i. True ii. False

(c) Suppose I have a neutral system, ABCD, with disjoint subsystems AB and CD. The total internal constraint in the subsystems is (choose one; 2 pts)

i. T(ABCD) ii. T(AB:CD) iii. T(AC:BD)+T(AD:BC)
iv. T(AB) + T(CD) v. T(A:B) + T(C:D) vi. T(A:B:C:D)
vii. T(A:C) + T(B:D)+T(A:D)+T(B:C) viii. None of these.

(d) H(ABZ) – H(AB:AZ:BZ) = (choose one; 2 pts)

i. ≥ 0 ii. ≤ 0 iii. Either ≥ 0 *or* ≤ 0

(e) Given an XY data distribution as follows

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

(e1) I want to consider the state-based model X_0Y_0 that specifies that $q(x_0,y_0) = .1$. Write an expression for transmission, T, the error in this model, in terms of the Γ function and numerical constants, i.e., as a sum of Γ terms, with appropriate numerical arguments, with these Γ terms having appropriate positive or negative signs. (2 pts)

(e2) What is the value of Δdf = df(XY) – df(X_0Y_0)? (State a number; 1 pt)

(e3) Suppose instead of the above state-based model, I propose a model in which the probability distribution of the table is uniform.

(e3.1) In terms of the Γ function and numerical constants of the table or other numbers, write an expression (in the same format as e1) for T, the error in this model. (2 pts)

(e3.2) What is the df of this model that asserts uniformity? (State a number; 1 pt)

Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
FINAL EXAM Fall 2018 (12/3/2018)

60 points; 2 hours; closed book, no notes. *Answer question on exam sheets, and put* ***your name*** *on them.* You can use Γ(a, b,...) = –a log a – b log b – ... Do not use red ink.

---

1. [16 points]

(a) True or false? (circle; 2 pts): For the data, H(Z) – H(Z|AB) = T(A:Z) + T_A(B:Z)

(b) 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

(c) For directed system ABCDZ, I(ABCD:ABCZ→ABCD:BCZ) = H(Z|X) – H(Z|Y), where X and Y are subsets of the IVs. Fill in (2 pts): X = _______ ; Y = _______

(d) Let the probabilities of *micro*states a, b, c, d be .1, .2, .3, .4, respectively, and the probabilities of *macro*states I, II be .3, .7, respectively, where I includes a and b and II includes c and d. Write an expression involving only numbers (probability values) for the entropy of the microstate given the macrostate. (2 pts)

H(microstate | macrostate) =

(e) On the left is an observed probability table (p) for a *directed* system, with sample size N. None of parameters (a...h) is 0. On the right is the calculated table (q) for AB:AZ.

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

(e1) Write an expression in terms of a…h for the amount of constraint that data ABZ *adds* to the constraint that is already captured in the AB:AZ model. (2 pts)

(e2) Write an expression for q_5 in terms of the parameters a through h. (2 pts)

(f) Suppose A is a quantitative variable that needs to be binned. We want to maximize its power to predict Z, but not use degrees of freedom *inefficiently*, which might prevent, for small sample sizes, the use also of B to predict Z. One reasonable *heuristic* is to choose the number of bins for A which maximizes: (circle one; 2 pts)

i. H(Z|A)	iv. [H(Z) – H(Z|A)] * |A|
ii. H(Z) – H(Z|A)	v. H(Z|A) / |A|
iii. H(Z|A) * |A|	vi. [H(Z) – H(Z|A)] / |A|

(g) Suppose the conditional probabilities for model m = AB:AZ:BZ are

A	B	p(AB)	p(Z=1|AB)	p(Z=2|AB)	q_m(Z=1|AB)	q_m(Z=2|AB)
1	1	a	b	c	d	e
1	2	f	g	h	i	j
2	1	k	l	m	n	o
2	2	r	s	t	u	v
		marginals	w	x	y	z

Assume that you know the sample size. In terms of (a…z) in the above table (and any necessary constants), write an expression for transmission T that you would use to test the null hypothesis that the differences between the calculated conditional Z probabilities for (A,B)=(1,1) and the Z marginal probabilities are not statistically significant. (2 pts)

2. [14 points]

(a) For the table given in question 1(e) and for model AB:AZ,

(a1) write the full set of linearly independent constraints in the form M **q** = M **p** by filling in rows in matrix M (starting from the top of M) for the correct number of constraints in this model. In these rows, fill in only the 1's; leave the 0's blank. (The bottom row is the constraint that is true of all models, namely that Σq = Σp.) (2 pts)

Matrix, M

1	1	1	1	1	1	1	1

$$\begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \\ q_7 \\ q_8 \end{bmatrix} = M \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{bmatrix}$$

(a2) Add one more constraint that is true, that follows from the fact that q(AZ) = p(AZ), as an *additional* row in M. *Circle* this additional row to distinguish it from the rows that answer (a1). Since this constraint is *not* linearly independent of the rows listed for (a1), this added row must be a sum and/or difference of some rows above it. Write the equation for this linear dependence in the form "Row X = Row Y + … – Row Z – … where X is the added row and Y, Z, etc., are rows above it. (2 pts):

To answer the IPF question below, fill in the AB and AZ tables for p and q.

Projected from p

A \ B:	0	1
0		
1		

A \ Z:	0	1
0		
1		

Projected from q

A \ B:	0	1
0		
1		

A \ Z:	0	1
0		
1		

Use IPF to obtain q_5 by writing the sequence of values of q_5 in terms of the parameters, a through h, after first imposing AB and then also AZ. The initial value is $q_5^{\text{initial}} = 1/8$.

(a3) By imposing AB, we get $q_5^{AB} = q_5^{\text{initial}} *$ ____________________________(2 pts)

(a4) By imposing also AZ, we get (2 pts)

$q_5^{AB:AZ} = q_5^{AB} *$ __________________________=

(b) For model ABC:BCD:CDE, write an algebraic expression for q(ABCDE) in terms of p(ABCDE) and its various projections (2 pts).

(c) Suppose I have a state-based model A_1B_2 for an AB neutral system, where A and B have cardinality 2. Let q_1 be the calculated distribution for A_1B_2. I want to pick a second state to add to this state based model; then the state-based model will be $A_1B_2{:}A_iB_j$, for some particular i and j. I have three possible choices for this second state, namely A_1B_1, A_2B_1, and A_2B_2. (Ignore the possibility of choosing the second state from the margins.) Call the distribution I get after choosing the second state q_2. I want to pick that state among the three possible choices that maximizes (circle one; 2 pts)

i. Σ p log $[q_1/q_2]$ ii. Σ p log $[q_2/q_1]$
iii. Σ q_1 log $[q_1/q_2]$ iv. Σ q_1 log $[q_2/q_1]$
v. Σ q_2 log $[q_1/q_2]$ vi. Σ q_2 log $[q_2/q_1]$

(d) True or false (circle; 2 pts): A state that is chosen for a state-based model must have a probability (in the data or a projection) *greater* than the probability it would have if all states (in the data or projection) were equally probable, i.e., only *more* probable, not *less* probable, single states are used in RA state-based models.

3. [18 points]

(a) Suppose I pick a best BIC model, then notice that the 2^{nd} best BIC model has a ΔBIC that is only slightly less than the ΔBIC of the best model. I wonder if I should look at the predictions of this 2^{nd} best BIC model and compare them to the predictions of the best BIC model to see if the two models predict similarly or differently. I should definitely make this comparison if which of the following measures is large compared to 1 (2 pts).

i. $I(m_1 \cup m_2 \rightarrow m_{ind}) / I(m_1 \cap m_2 \rightarrow m_{ind})$ ii. $I(m_1 \cap m_2 \rightarrow m_{ind}) / I(m_1 \cup m_2 \rightarrow m_{ind})$

(b) The Transmission of which model in the Lattice of Structures for ABZ will specify the amount of information that A tells me about Z if I already know B? (2 pts).

(c) After each measure from (i) to (viii), write whether the measure is MD (monotonically decreasing or staying the same), MI, (monotonically increasing or staying the same), or NM (not monotonically increasing *or* decreasing) for every step going <u>down</u> the Lattice of Structures. (8 pts) **(1 pt for each)**

i. H	ii. T	iii. L^2
iv. α	v. df (not Δdf)	vi. ΔAIC
vii. %correct(train)	viii. %correct(test)	

(d) When the reference is the data, one calculates $\Delta df(m_0 \rightarrow m_j)$ and $L^2(m_j)$ and decides on some α_c. Say that going into the Chi-square table with $\Delta df(m_0 \rightarrow m_j)$ and α_c one gets some L^2_c and suppose that the model $L^2(m_j) > L^2_c$. One would then usually (circle; 2 pts)

i. not reject the null, and consider m_j as possibly a good model
ii. not reject the null, and use the data, m_0, as one's model
iii. reject the null, and consider models higher in the lattice than m_j
iv. reject the null and consider models lower in the lattice than m_j

(e) Occam reports AIC relative to the reference model, but in the literature it is more common to cite 'absolute' AIC which equals $- 2 N \sum p \ln q + 2$ df. Using this absolute AIC measure, a good model is one that has (circle one; 2 pts)

i. minimum AIC ii. maximum AIC iii. AIC = α iv. AIC = $1 - \alpha$

(f) Assume that $q(Z_0|A_iB_j)$ and $q(Z_1|A_iB_j)$ for model m_1 = AB:AZ:BZ are almost equal and that the frequency of A_iB_j is also small, so confidence in predicting Z is low. In this case, it's useful to have a simpler backup model, m_2, and make predictions of Z for A_iB_j from m_2. For a good backup model, $I(m_1 \rightarrow m_2)$ (circle one; 2 pts)

i. should be small ii. should be large iii. is irrelevant

4. [8 points]
Below are results of analyzing some genomic medical data for model IV:AZ:EZ:KZ, where this table has been augmented with three p-value calculations. Z_1 is the diseased state. p and q conditional probabilities are in %. The p-values are as follows:

p_{rule} compares $q(Z|A_jE_kK_l)$ to (50, 50).
$p_{margins}$ compares $q(Z|A_jE_kK_l)$ to the $(q(Z_0), q(Z_1))$ margins, i.e., to (51.8, 48.2).
p_A compares $q(Z|A_jE_kK_l)$ to $q(Z|A_j)$, i.e., to (64.8, 35.2) for j=0 and (28.1, 71.9) for j=1.

	IV			Data obs p(Z | IV)		Model calc q(Z | IV)				correct			
A	E	K	freq	Z_0	Z_1	Z_0	Z_1	rule	p_{rule}	#	%	p_{margin}	p_A
0	0	0	4	0.0	100.0	12.2	87.8	1	0.131	4	100.0	0.113	0.028
0	0	1	8	12.5	87.5	12.4	87.6	1	0.033	7	87.5	0.026	0.002
0	0	2	4	25.0	75.0	29.4	70.6	1	0.409	3	75.0	0.369	0.138
0	1	0	31	64.5	35.5	61.6	38.4	0	0.198	20	64.5	0.277	0.707
0	1	1	37	62.2	37.8	61.9	38.1	0	0.147	23	62.2	0.219	0.714
0	1	2	23	78.3	21.7	82.7	17.3	0	0.002	18	78.3	0.003	0.072
0	2	0	66	63.6	36.4	64.0	36.0	0	0.023	42	63.6	0.047	0.894
0	2	1	61	65.6	34.4	64.4	35.7	0	0.025	40	65.6	0.050	0.942
0	2	2	33	84.8	15.2	84.2	15.8	0	0.000	28	84.8	0.000	0.020
1	0	0	1	0.0	100.0	2.6	97.4	1	0.343	1	100.0	0.325	0.571
1	0	1	7	14.3	85.7	2.6	97.4	1	0.012	6	85.7	0.009	0.134
1	0	2	2	0.0	100.0	7.4	92.6	1	0.228	2	100.0	0.208	0.514
1	1	0	13	30.8	69.2	23.4	76.6	1	0.055	9	69.2	0.041	0.709
1	1	1	24	16.7	83.3	23.7	76.3	1	0.010	20	83.3	0.006	0.633
1	1	2	11	54.5	45.5	47.8	52.2	1	0.884	5	45.5	0.789	0.146
1	2	0	32	21.9	78.1	25.4	74.6	1	0.005	25	78.1	0.003	0.732
1	2	1	39	25.6	74.4	25.6	74.4	1	0.002	29	74.4	0.001	0.735
1	2	2	17	52.9	47.1	50.4	49.6	0	0.973	9	52.9	0.908	0.040
			413	51.8	48.2	51.8	48.2	0		291	70.5		
-	-	-				50.0	50.0						
0	-	-				64.8	35.2						
1	-	-				28.1	71.9						

In (a)-(d), mark a state with X <u>only</u> if its deviation from its reference *is statistically significant*, using the criterion of $p \le 0.05$. In the tables below (2 pts each),
(a) In columns labeled 1, mark with X all IV states whose prediction *rules* are significant.
(b) In columns labeled 2, mark with X all IV states where $q(Z|IV) \ne$ margins.
(c) In columns labeled 3, mark with X all IV states that <u>*increase*</u> the risk of disease *beyond what is predicted by A alone*.
(d) In columns labeled 4 mark with X all IV states that <u>*decrease*</u> the risk of disease *beyond what is predicted by A alone.*

A	E	K	1	2	3	4
0	0	0				
0	0	1				
0	0	2				
0	1	0				
0	1	1				
0	1	2				
0	2	0				
0	2	1				
0	2	2				

A	E	K	1	2	3	4
1	0	0				
1	0	1				
1	0	2				
1	1	0				
1	1	1				
1	1	2				
1	2	0				
1	2	1				
1	2	2				

5. [4 points]

(a) For the relation below, what is the *simplest* structure that this relation can be decomposed into with no error? If it can be decomposed into *multiple* structures all equally simple, state all the structures. If it cannot be decomposed at all, say so. (2 pts)

X	Y	Z
0	0	0
1	1	1

(b) The above relation is an example of the smallest number of tuples (2) that a relation between binary variables can have. (If the relation had only one tuple, the variables wouldn't in fact be variables but constants). The *maximum* number of tuples that a relation between binary variables X, Y, and Z can have and still have non-zero constraint is 7. A relation having this number of tuples is (circle one; 2 pts):

i. decomposable ii. non-decomposable iii. either; it depends on the relation

Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
FINAL EXAM Winter 2019 (12/9/2019)

60 points, 2 hours (2 pts/min); closed book. *Answer question on exam, and put **your name** on them.* L^2 = LR. You can use Γ(a, b...) = –a log a – b log b –…Don't use red ink. You can attach pages of your work, but it isn't obligatory.

----

1. [10 points]

(a) Let the level of knowledge of a student about information theory be a dichotomous variable, K, whose value is high (h) or low (l). Assume that an exam testing students on their knowledge makes only a binary discrimination, i.e., the grade, G, is also h or l. The probabilities of possible situations are in the table below, where a+b+c+d = 1. Write an (equality or inequality) equation involving *only* the terms H(K), H(G), and/or T(K:G) (use as many of these terms as necessary) which would hold if the exam resolved all uncertainty about K but was *more detailed* than needed to resolve this uncertainty. (2 pts)

		G: l	G: h
K	l	a	b
	h	c	d

(b) For the data given by the table below, with known probabilities, a…h, give an expression for T(A:Z) in terms of parameters, a…h, using the definition of T as a sum of p log (p/q) terms. Give only the *first two terms* of this sum. (2 pts)

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

(c) Let time be a variable, as follows. Consider ABC_1 and ABC_2 distributions below, where C_1 = t and C_2 = t+1, with probabilities, a+b+c+d = 1 and e+f+g+h = 1. The AB distribution at time t is the reference. I wish to test the hypothesis that AB at t+1 is the same as this reference distribution, i.e., the same as AB at time t. Using the p log (p/q) expression for transmission, write the T in terms of a…h that will let me test this hypothesis. (2 pts)

C_1

	B_1	B_2
A_1	a	b
A_2	c	d

C_2

	B_1	B_2
A_1	e	f
A_2	g	h

T =

(d) For some ABZ data, $T_B(A:Z) = 0$. Circle all models where T(model) is 0. (2 pts)

i. ABZ	ii. AB:AZ:BZ	iii. AB:AZ	iv. AB:BZ	v. AZ:BZ
vi. AB:Z	vii. AZ:B	viii. BZ:A	ix. A:B:Z	x. none of these

(e) Consider the contingency table, where a+b+c+d = 1

	y_1	y_2
x_1	a	b
x_2	c	d

Let x be a word in a message; let y be the word that follows it. Write an expression, in terms of parameters, a, b, c, & d, for the amount of information which receiving y provides beyond what one already knows from having previously received x. (2 pts)

Information(y | x) =

2. [12 points]

(a) Suppose A is a quantitative variable that needs to be binned. We want to maximize its power to predict Z but use degrees of freedom *efficiently*, since this might allow us to also use B to predict Z. One reasonable heuristic is to choose the number of bins for A which maximizes: (circle one; 2 pts)

i. H(Z|A)	iv. [H(Z) – H(Z|A)] * |A|
ii. H(Z) – H(Z|A)	v. H(Z|A) / |A|
iii. H(Z|A) * |A|	vi. [H(Z) – H(Z|A)] / |A|

(b) Suppose our DV is Z and our IVs are A, B, and C. The predictive effect of A on Z, *controlling for* B and C (how much knowing A tells us about Z beyond what B and C tell us about Z) is (circle one; 2 pts)

	i. T(ABCZ)	
	ii. T(ABC:ABZ:ACZ:BCZ	
iii. T(ABC:ABZ:ACZ)	iv. T(ABC:ABZ:BCZ)	v. T(ABC:ACZ:BCZ)
vi. T(ABC:AB:CZ)	vii. T(ABC:ACZ:BZ)	viii. T(ABC:BCZ:AZ)
ix. T(ABC:ABZ)	x. T(ABC:ACZ)	xi. T(ABC:BCZ)
	xii. T(ABC:AZ:BZ:CZ)	
xiii. T(ABC:AZ:BZ)	xiv. T(ABC:AZ:CZ)	xv. T(ABC:BZ:CZ)
xvi. T(ABC:AZ)	xvii. T(ABC:BZ)	xviii. T(ABC:CZ)
	xix. T(ABC:Z)	

(c1) To evaluate the model AB:AZ:BZ relative to the *independence model* as the reference, the appropriate information distance equals (circle all that are true; 2 pts)

i. T(AB:Z) – T(AB:AZ:BZ)	ii. T(AB:AZ:BZ)- T(AB:Z)
iii. T(ABZ) – T(AB:AZ:BZ)	iv. T(AB:AZ:BZ) – T(ABZ)
v. H(AB:Z) – H(AB:AZ:BZ)	vi. H(AB:AZ:BZ) – H(AB:Z)
vii. H(ABZ) – H(AB:AZ:BZ)	viii. H(AB:AZ:BZ) – H(ABZ)

(c2) To evaluate the model AB:AZ:BZ relative to the *data* as the reference, the appropriate information distance equals (circle all that are true; 2 pts)

i. T(AB:Z) – T(AB:AZ:BZ)	ii. T(AB:AZ:BZ)- T(AB:Z)
iii. T(ABZ) – T(AB:AZ:BZ)	iv. T(AB:AZ:BZ) – T(ABZ)
v. H(AB:Z) – H(AB:AZ:BZ)	vi. H(AB:AZ:BZ) – H(AB:Z)
vii. H(ABZ) – H(AB:AZ:BZ)	viii. H(AB:AZ:BZ) – H(ABZ)

(d1) True or false? (circle one; 2 pts): T(A:B:C) = T(A:B) + T(AB:C)

(d2) True or false? (circle one; 2 pts) $T(ABC:Z) = T(A:Z) + T_A(B:Z) + T_{AB}(C:Z)$

3. [12 points]
(a) Reference = top.
(a1) A Type I error will probably result in a model that is (circle one; 2 pts)

i. more complex than necessary
ii. too simple to fit the data

(a2) A Type II error will result in a model that is (circle one; 1 pt)

i. more complex than necessary
ii. too simple to fit the data

(b) Reference = bottom.
(b1) A Type I error will result in a model that is (circle one; 2 pts)

i. less complex than is statistically justified
ii. too complex to be statistically justified

(b2) A Type II error will probably result in a model that is (circle one; 1 pt)

i. less complex than is statistically justified
ii. too complex to be statistically justified

(c) A test is given to detect a disease. A ‘negative’ test result means that this condition is not detected, i.e., the patient is judged to be free of the disease; a ‘positive’ result means that the condition is detected, i.e., the patient is judged to have the disease. The actual condition of the patient and the test conclusions are summarized in these frequencies: TN = true negatives, FP = false positives, FN = false negatives, TP = true positives. The frequency marginals of the actual distribution are N = TN + FP, P = FN + TP. Assume that the null hypothesis is ‘negative.’

	Test	(-)	(+)	
Actual	(-)	TN	FP	N
	(+)	FN	TP	P

Which of these is true (circle one; 2 pts)?

i. FP and FN are both Type I errors
ii. FP and FN are both Type II errors
iii. FP are Type I errors; FN are Type II errors
iv. FP are Type II errors; FN are Type I errors
v. none of the above

(d) Suppose the reference is *independence*. For some m_j, one calculates $\Delta df(m_j \rightarrow m_{ind})$ and $\Delta L^2(m_j \rightarrow m_{ind})$. One also decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some L^2_c and that the model $\Delta L^2 > L^2_c$. One would then say that (circle; 2 pts)

i. $m_j = m_0$, so one must use the bottom, m_{ind}, as one’s model
ii. $m_j \neq m_{ind}$, so one is happy with m_j (*or* one *tries* to go higher in the lattice)
iii. $m_j \neq m_{ind}$, so is *un*happy with m_j and *must* go lower in the lattice

(e) When the reference is the *data*, one calculates $\Delta df(m_0 \rightarrow m_j)$ and $L^2(m_j)$ and decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some L^2_c and suppose that the model $L^2(m_j) > L^2_c$. One would then usually (circle; 2 pts)

i. not reject the null, and consider m_j as possibly a good model
ii. not reject the null, and use the data, m_0, as one’s model
iii. reject the null, and consider models higher in the lattice than m_j
iv. reject the null and consider models lower in the lattice than m_j

4. [10 points] Consider on the left an *observed* (data) 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:BZ 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) Solve for q_2 *algebraically* for model AB:BZ in terms of a…h. (2 pts)

q_2 =

(b) Use IPF to obtain q_3^{AB} (not q_2!) and then $q_3^{AB:BZ}$ in terms of parameters a…h. $q_3^{initial} = 1/8$.

(b1) After imposing AB, we get (1 pt; write only the factor that multiplies $q_3^{initial}$)

$q_3^{AB} = q_3^{initial} *$ ________________________________

(b2) After imposing also BZ, we get (3 pts: write only the factor that multiplies q_3^{AB})

$q_3^{AB:BZ} = q_3^{AB} *$ ________________________________

(c1) Given the Occam fit output for model m = AB:AZ:BZ immediately below,

A	B	p(AB)	p(Z=1|AB)	p(Z=2|AB)	q_m(Z=1|AB)	q_m(Z=2|AB)
1	1	a	b	c	d	e
1	2	f	g	h	i	j
2	1	k	l	m	n	o
2	2	r	s	t	u	v
		marginals	w	x	y	z

circle the parameters in the table that are necessary and sufficient to evaluate p_{margin} for (A=2, B=2); circle the <u>smallest</u> number of necessary and sufficient parameters (2 pts).

(c2) The uncertainty reduction that one gets for (A=2, B=2) is (circle one; 2 pts)

i. $H(Z) - H(Z|A_2B_2)$ ii. $H(Z|A_2B_2) - H(Z)$
iii. $H(Z) - [H(Z|A_2) + H(Z|B_2)]$ iv. $H(Z|A_2) + H(Z|B_2) - H(Z)$
v. none of the above

5. [6 points]
(a) For the following table, state-based models are shown in italics. The reference is the top. |A|=|B|=|Z|=2

#	Model	T	%I	df	L^2	p
1	ABZ	- - -	100%	7	--	1.000
2	*AB:Z:a0BZ*	*0.0002*	*100%*	*6*	*0.3*	*0.603*
3	*AB:Z:a0b1Z*	*0.0696*	*61%*	*5*	*120.3*	*0.000*
4	*AB:Z:a0b0Z*	*0.0876*	*51%*	*5*	*151.4*	*0.000*
5	AB:AZ:BZ	0.1478	17%	6	255.5	0.000
6	AB:BZ	0.1482	17%	5	256.2	0.000
7	*AB:Z:a1b1Z*	*0.1610*	*10%*	*5*	*278.4*	*0.000*
8	*AB:Z:a1b0Z*	*0.1720*	*3%*	*5*	*297.4*	*0.000*
9	AB:AZ	0.1777	0%	5	307.2	0.000
10	AB:Z	0.1780	0%	4	307.6	0.000

(a1) The model # for the best variable based model is ______ (2 pts).

(a2) The model # for the best state based model is ______ (2 pts)

(b) Assume that the AB projection of the ABZ data is nearly what I would expect if A and B were independent, and so I want a (non-standard-RA) model where the hypothesis of independence for A and B is included in the model. The model will have a relation written as $AB_{A:B}$ and I want the model $AB_{A:B}$:AZ:BZ, which means that q(AZ) = p(AZ) and q(BZ) = p(BZ), but q(AB) = p(A)*p(B), *instead of* q(AB) = p(AB).

df($AB_{A:B}$:AZ:BZ) = ______ (2 pts)

6. [10 points]
(a) (4 pts) The set-theoretic mapping below maps values of A, B, C onto values of Z.

A	B	C	Z
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	0

(a1) The ABCZ mapping is (circle one):

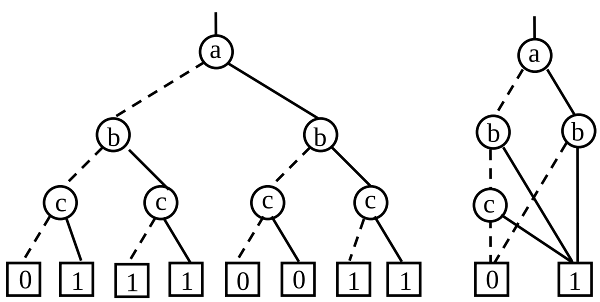
i. decomposable (without constraint loss)
ii. not decomposable (without constraint loss)

(a2) If you chose *decomposable,* the simplest structure equivalent to ABCZ is ___________________________. Its *predicting* components are (circle one):

i. all deterministic ii. all stochastic iii. some deterministic, some not

If you chose *not decomposable,* the # of extra tuples of the 1st decomposition is ____

(b) *Introduction to binary decision diagrams*: The set-theoretic relation R = {001, 010, 011, 110, 111} can be represented by the tree on the *left side* of the figure below, where lines coming out of and down from a variable indicate its possible values, where a dashed line means 0 and a solid line means 1. In the square boxes at the bottom of the tree, 1 indicates that the tuple is in the relation; 0 indicates it is not. So reading down the tree on its left side, the three dashed lines culminating in a boxed 0 mean that 000 (i.e., a=0, b=0, c=0) is not in the relation. But 001 is. Etc. By information-preserving operations one can transform the tree on the left to the graph on the right. Reading the graph downwards in the same way, we see that 000 is not in the relation, but 001 is; also 01* is in the relation, as is 11*, but 10* is not. The right hand graph – called a binary decision diagram – specifies a 'compressed' relation, R_{BDD} = {001, 01*, 11*}.



(b1) Suppose I don't know R, but I am given R_{BDD}. To get the set of abc tuples implied by R_{BDD}, I replace 01* by {010, 011} and 11* by {110, 111}. These replacements expand R_{BDD}, and together with 001, make R_{BDD} = R. The *theoretical justification* for making these replacements is (circle one; 2 pts)

i. model is loopless ii. model has loops
iii. law of subscripting iv. law of distribution
v. minimum entropy vi. maximum entropy

(b2) The compressed relation R_{BDD} is a set-theoretic analog of information-theoretic (circle one; 2 pts)

i. k-systems analysis	ii. Bayesian networks
iii. latent variable modeling	iv. state-based modeling

(c) I have two separate frequency distributions AB and BC which I *think* are samples of the same population, but I am not sure. I am interested in the relationship between A and C, so I want to merge these distributions by composition into $ABC_{AB:BC}$ and then take the AC projection of the result of this composition. What do I need to do first to establish whether such a merging is possible, i.e., statistically legitimate? (2 pts)

Discrete Multivariate Modeling (SySc 551/651) Professor M. Zwick
FINAL EXAM Spring 2021 (6/7/2021)

60 points; 2 hours. $\Gamma(a, b,...) = – a \log a – b \log b$ -...

---

1. [10 points]
(a) For *any* ABC data (choose one; 2 pts):

i. $T(A:B) \leq T_C(A:B)$ ii. $T(A:B) \geq T_C(A:B)$ iii. Either i or ii could be true

(b) The statement, "$H(Z) – H(Z|AB) = T(B:Z) + T_B(A:Z)$" is (choose one; 2 pts):

i. Always true ii. Always false iii. True or false, depending on the data

(c) Consider data ABZ where $T(A:Z) > 0$ and $T_B(A:Z) = 0$. Ignore the possible relevance of variables other than A,B, and Z. The second of these two equations can be interpreted as saying (choose all interpretations of this equation that *could* be true; 2 pts)

i. A directly predicts B which directly predicts Z, so A predicts Z indirectly
ii. B directly predicts A which directly predicts Z, so B predicts Z indirectly
iii. A directly and separately predicts both B and Z
iv. B directly and separately predicts both A and Z.
v. A and B both directly predict Z via a 3-way interaction effect
vi. A does not predict Z at all, either directly or indirectly
vii. B does not predict Z at all, either directly or indirectly

(d) Data ABZ is fit with the AZ:BZ model. True or False? (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 this projected AB has $T(A:B) > 0$.

(e) Suppose one has an AB frequency table, where $|A|=|B|=2$, and in this 2x2 table there is one systematic zero, i.e., an (A_i, B_j) state whose frequency *must* be zero (like a frequency of pregnant males). True or false? (2 pts): The independence model with this *additional* systematic zero constraint necessarily has T=0.

2. [16 points]

(a) Consider the following *observed* contingency data table (the p's) for a *directed* system, with sample size N and a + b +...+ h=1. None of parameters (a...h) is 0.

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

Let the calculated contingency table (the q's) for model AB:BZ be, with i + j +…+ r = 1.

	Z:	0		1	
	B:	0	1	0	1
A:	0	i	j	k	l
	1	m	n	o	r

(a1) Fitting the AB:BZ model involves maximizing an expression, subject to a set of constraints. Which expression is maximized? (choose one; 2 pts)

i. Γ(a, b, …h) ii. Γ(i, j, …r)
iii. Γ(a/i, b/j, …h/r) iv. Γ(i/a, j/b,…r/h) v. none of the above

(a2) State the constraint equations in two parts, as follows. (Be sure the equations are mutually independent, i.e., not redundant.) First, state the maximum number of constraints that would be associated with AB if the model consisted only of this one relation. Second, state the additional constraints added by BZ. The general constraint that p's or q's sum to 1 should be taken for granted and not here specified.

(a2.1) In terms of a…r, state the AB constraints. (1pt)

(a2.2) In terms of a…r, state the *additional* (linearly independent) BZ constraints. (2 pts)

(b1) Write an expression in terms of a...h for $q_{AB:BZ}$(A=1,B=0,Z=1). (2 pts)

(b2) In terms of a...r and using the p log [p/q] expression for T, write an expression for the first two terms in T(AB:BZ). (1 pt)

(c) Given a directed system with variables A, B, C, and Z, the information distance I(ABC:ABZ $\rightarrow$ ABC:BZ) = (choose one, 2 pts)

i. H(Z|A) – H(Z|AB) v. H(Z|A) + H(Z|AB) ix. none of the above
ii. H(Z|B) – H(Z|AB) vi. H(Z|B) + H(Z|AB)
iii. H(Z|AB) – H(Z|A) vii. H(Z|AB) + H(Z|A)
iv. H(Z|AB) – H(Z|B) viii. H(Z|AB) + H(Z|B)

(d) Let I(m) = the *normalized* information for model m, which is 1 for data, m_0, and 0 for independence model, m_{ind}. Write an expression for I(m) in terms of some or all of the quantities $H(m_0)$, H(m), $H(m_{ind})$ and any necessary numerical constants. (2 pts)

(e) Suppose I have a state-based model $A:B:A_1B_2$ for an AB neutral system, where A and B have cardinality 3. Let q_1 be the calculated distribution for $A:B:A_1B_2$. I want to pick a second state to add to this model; then the model will be $A:B:A_1B_2:A_iB_j$ for some i and j. I have eight possible choices for this second state. Call the distribution I get after choosing the second state q_2. I should pick that state among the three possible choices that maximizes (choose one; 2 pts)

i. Σ p log $[q_1/q_2]$	ii. Σ p log $[q_2/q_1]$
iii. Σ q_1 log $[q_1/q_2]$	iv. Σ q_1 log $[q_2/q_1]$
v. Σ q_2 log $[q_1/q_2]$	vi. Σ q_2 log $[q_2/q_1]$

(f) True or false (2 pts): A state that is chosen for a state-based model must have a probability (in the data or a projection) *greater* than the probability it would have if all states (in the data or projection) were equally probable.

3. [16 points]

(a) True or false? (2 pts) A model with %Information captured of 95% or higher will always be statistically significant with a p-value of 0.05 or lower.

(b) Rejecting a null hypothesis when the p-value ≤ 0.05 (or less than some smaller cutoff) is reasonable for (2 pts)

i.. the reference model being the top but not the bottom
ii. the reference model being the bottom but not the top
iii. *both* reference being the top *and* reference being the bottom
iv. *neither* reference being the top *nor* reference being the bottom

(c) Suppose the reference is *independence*. For some m_j, one calculates $\Delta df(m_j \rightarrow m_{ind})$ and $\Delta L^2(m_j \rightarrow m_{ind})$. One also decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some L^2_c and that the model $\Delta L^2 > L^2_c$. One would then say that (choose; 2 pts)

i. $m_j = m_0$, so one must use the bottom, m_{ind}, as one's model
ii. $m_j \neq m_{ind}$, so one is happy with m_j (*or* one *tries* to go higher in the lattice)
iii. $m_j \neq m_{ind}$, so is *un*happy with m_j and *must* go lower in the lattice
iv. $m_j \neq m_{ind}$, so is *un*happy with m_j and *must* go higher in the lattice

(d) Suppose the reference is the *data*. For some m_j, one calculates $\Delta df(m_0 \rightarrow m_j)$ and $L^2(m_j)$. One also decides on some α_c. Say that going into the Chi-square table with Δdf and α_c one gets some $L^2{}_c$ and suppose that the model $L^2(m_j) > L^2{}_c$. One would then usually (choose; 2 pts)

i. not reject the null, and consider m_j as possibly a good model
ii. not reject the null, and choose the data, m_0, as one's model
iii. reject the null, and consider models higher in the lattice than m_j
iv. reject the null and consider models lower in the lattice than m_j

(e) A test is given to detect a disease. A 'negative' test result means that this condition is not detected, i.e., the patient is judged to be free of the disease; a 'positive' result means that the condition is detected, i.e., the patient is judged to have the disease. (So a 'negative' result is good and a 'positive' result is bad.) Assume that the null hypothesis for the test result is 'negative.' The actual condition of the patient and the test conclusions are summarized in these frequencies: TN = true negatives, FP = false positives, FN = false negatives, TP = true positives. The frequency marginals of the actual distribution are N = TN + FP, P = FN + TP.

	Test	(-)	(+)	
Actual	(-)	TN	FP	N
	(+)	FN	TP	P

Which of these is true (choose one; 2 pts)?

i. FP and FN are both Type I errors
ii. FP and FN are both Type II errors
iii. FP are Type I errors; FN are Type II errors
iv. FP are Type II errors; FN are Type I errors
v. none of the above

(f) and (g) have no connection to above question (e).

(f) <u>Reference = top</u>.
(f1) A Type I error will probably result in a model that is (choose one; 2 pts)

i. more complex than necessary   ii. too simple to fit the data

(f2) A Type II error will result in a model that is (choose one; 1 pt)

i. more complex than necessary   ii. too simple to fit the data

(g) <u>Reference = bottom</u>.
(g1) A Type I error will result in a model that is (choose one; 2 pts)

i. less complex than statistically justified   ii. too complex to be statistically justified

(g2) A Type II error will probably result in a model that is (choose one; 1 pt)

i. less complex than is statistically justified   ii. too complex to be statistically justified

4. [12 points]

(a) The table below shows variable-based (italics) and state-based (not italics) analyses. # is the model index, and $\#_{incr}$ is the reference for p_{incr}, which is an *incremental* p-value. p_{cum} is always relative to the bottom reference. I is information captured in the model, normalized so that it's 1 for the data and 0 for the independence model.

#	I	Δdf	p_{cum}	p_{incr}	$\#_{incr}$	Model
11	*1.000*	*8*	*0.0014*	*0.9684*	10	*A B Z*
10	1.000	7	0.0007	0.8736	9	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z:A_2B_2Z$
9	0.999	6	0.0003	0.5036	8	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z$
8	0.981	5	0.0002	0.3022	7	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z:B_3Z:A_3B_1Z$
7	0.939	4	0.0001	0.3695	5	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z:A_3B_3Z$
6	*0.502*	*4*	*0.0129*	*0.1098* *0.0148*	*3* *2*	*AB:AZ:BZ*
5	0.907	3	0.0000	0.0575	4	$AB:Z:A_2B_1Z:A_3Z:A_1B_2Z$
4	0.765	2	0.0001	0.0045	1	$AB:Z:A_2B_1Z:A_3Z$
3	*0.327*	*2*	*0.0161*	*0.0161*	*0*	*A B : A Z*
2	*0.169*	*2*	*0.1188*	*0.1188*	*0*	*A B : B Z*
1	0.445	1	0.0008	0.0008	0	$AB:Z:A_2B_1Z$
0	*0.000*	*0*	*1.00*	*1.0000*		*A B : Z*

(a1) Assume one wants $p_{cum} \le 0.05$ and $p_{incr} \le 0.05$ for every step on a path to the model,
(a1.1) The best variable-based model is #_____

(a1.2) The best state-based model is #____ (2 pts)

(a2) The table shows that state-based models are superior to variable-based models because (choose one; 2 pts)

i. for particular Δdf values, SB models have more information than VB models
ii. SB models sample the possible Δdf values more finely than VB models
iii. both i. and ii.
iv. neither i. and ii. (some *other* reason)

(b) True or false (2 pts): A model that captures some information in the data, and thus exhibits constraint that is absent in the independence model, will always have its %correct greater than the %correct of the independence model.

(c) Suppose that for model m = ABC:AZ:BZ, with binary Z, the *frequency* of (A_i, B_j) is too *small* for differences between $q(Z_0|A_iB_j)$ and $q(Z_1|A_iB_j)$ to be statistically significant. A reasonable strategy for this situation is to (choose one; 1 pt)

i. predict using the rule coming from the independence model, AB:Z
ii. choose a *child* of m with fewer IVs, e.g., ABC:AZ & predict using the rule for A_i
iii. choose a *parent* of m, e.g., ABC:AZ:BZ:CZ and predict using a rule from its q
iv. obtain the needed prediction rule from the data, ABZ

(d) Suppose an Occam fit output is this table, where Z_1 = healthy; Z_0 = not healthy. Prefer p-values at or less than 0.05.

A	B	C	N(AB)	$q(Z_0\|ABC)$	$q(Z_1\|ABC)$	rule	p-rule	p-margin
0	0	0	18	58.7	41.3	0	0.454	0.412
0	0	1	22	58.7	41.3	0	0.408	0.364
0	1	0	38	27.3	72.7	1	0.005	0.006
0	1	1	20	27.3	72.7	1	0.041	0.049
1	0	0	15	58.7	41.3	0	0.495	0.454
1	0	1	24	86.0	14.0	0	0	0
1	1	0	18	27.3	72.7	1	0.053	0.062
1	1	1	20	61.8	38.2	0	0.294	0.26
			175	49.1	50.9	1		

(d1) Which IV state of A predicts health as more likely than non-health? (1 pt)

(d2) Which IV state of B predicts health as more likely than non-health? (1 pt)

(d3) Which IV state of C predicts health as more likely than non-health? (1 pt)

(d4) Of the three individual variables, which one most predicts Z? (1 pt)

(d5) Which ABC state gives a prediction rule that you can be most confident in? (1 pt)

5. [6 points]

(a) True or false (1 pt): For the same variables, the Lattice of Structures for set-theoretic RA (SRA) is the same as the Lattice of Structures for information-theoretic RA (IRA).

(b) True or false (1 pt): Shannon entropy in IRA has exactly the same value as Hartley entropy in SRA if the IRA probability distribution is uniform.

(c) In set-theoretic RA, for XY:XZ:YZ, the calculated relation is (choose one; 2 pts)

i. $XY \cup XZ \cup YZ$

ii. $XY \cap XZ \cap YZ$

iii. $XY \otimes XZ \otimes YZ$

iv. $(XY \otimes Z) \cup (XZ \otimes Y) \cup (YZ \otimes X)$

v. $(XY \otimes Z) \cap (XZ \otimes Y) \cap (YZ \otimes X)$

vi. Not obtainable in closed form

(d) Consider the following AB, BC, and AC relations.

	B_0	B_1
A_0	0	.7
A_1	.3	0

	C_0	C_1
B_0	0	.3
B_1	.7	0

	C_0	C_1
A_0	.4	.3
A_1	.3	0

(d1) True or false: Every pair of these three relations is consistent in the common margins, i.e., AB and BC are consistent in B, AB and AC are consistent in A, and BC and AC are consistent in C. (1 pt) (d2) Treat AB, AC, BC set-theoretically, so that if a probability > 0,the tuple is present; if it is 0, the tuple is absent. For example, the AB tuples are {01, 10}. True or false: There exists an ABC set-theoretic relation that satisfies the AB, AC, and BC set-theoretic relations. (1 pt)