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. $\Gamma(a, b,...) = -a \log a - b \log b - ...$

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 "$\Gamma(a, d) + \Gamma(b, c) - \Gamma(a, b, c)$". (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 $\Delta df = df(XY) - df(X_0Y_0)$? (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)