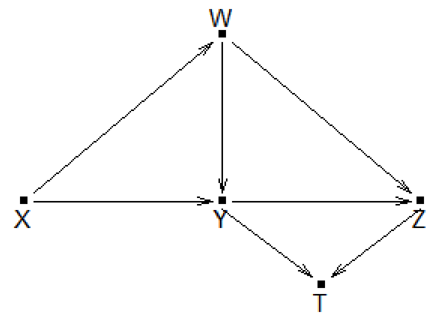
STATISTICS 265 – Spring 2018 – Homework 3

Handed out: Tuesday May 15, 2018
Due: Thursday May 31, 2018

1. **Graphs.** Use the graph below to answer the following questions.



   (a) Name all of the parents of Z.
   (b) Name all of the ancestors of Z.
   (c) Name all of the children of W.
   (d) Name all of the descendents of W.
   (e) Identify all paths (not necessarily directed) between X and Z.
   (f) Identify all directed paths between X and T.
   (g) Use the product decomposition of probability distributions to identify the joint probability distribution of the five variables represented as nodes (T, W, X, Y, Z).

2. **Structural Causal Models.** Suppose we specify the following structural model with $U = (U_x, U_y, U_z)$ as exogenous variables, $V = (X, Y, Z)$ as endogenous variables, and functions (f_x, f_y, f_z). Assume that all of the exogenous variables are independent of each other with expected value 0.

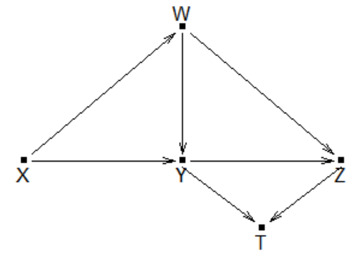
$$f_x : X = U_x$$

$$f_y : Y = (X/3) + U_y$$

$$f_z : Z = (Y/16) + U_z$$

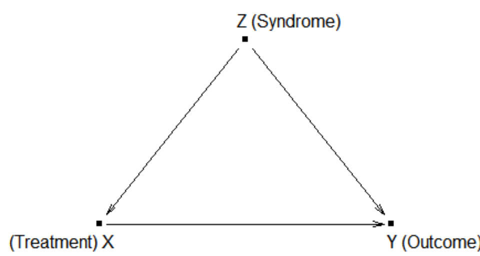
   (a) Draw the graph that corresponds to this model.
   (b) Find the expected value of X, Y and Z.
   (c) Find the expected value of Z given that $Y = 3$.
   (d) Find the expected value of Z given that $X = 3$.
   (e) The answer to (c) does not change if you are also told that $X = 1$. First, verify this by finding the expected value of Z given that $X = 1$ and $Y = 3$. Second, explain why this is true.

3. **d-Separation.** This question explores the topic of d-separation using the graph from problem 1 which is reproduced below. For purposes of this question you can assume that the error terms $(U_x, U_w, U_y, U_z, U_t)$ which are not shown are mutually independent (and can thus be ignored).



   (a) Are W and T (marginally) independent? Explain.
   (b) Are W and T conditionally independent given Z? Explain. If so, confirm this result using the joint probability distribution from 1(g) (i.e., show that $f(w, t|z) = f(w|z)f(t|z)$).
   (c) Are W and T conditionally independent given Y, Z? Explain. If so, confirm this result using the joint probability distribution from 1(g) (i.e., show that $f(w, t|y, z) = f(w|y, z)f(t|y, z)$).
   (d) Find a set of variables that d-separates X and T. Explain your reasoning. What does this tell us about the conditional independence of X and T in the data (i.e., when are they conditionally independent).

4. **Interventions.** Assume that we are studying a population of patients with a particular disease for which a treatment X may be helpful. It turns out that some of these patients are impacted by a syndrome Z that impacts survival Y but also impacts the patients ability to tolerate the treatment X.



The graph above describes the relationship between the syndrome Z, the treatment/drug X and the outcome Y (death or survival) for the population of patients. Suppose that a fraction of the population r suffers from the syndrome (i.e., have $Z = 1$) with the remaining proportion $1 - r$ not having the syndrome (i.e., $Z = 0$). Let $X = 1$ represent a patient taking the drug and $X = 0$ represent a patient not taking the drug. And let $Y = 1$ indicate that a patient dies and $Y = 0$ indicate that a patient survives. Assume that patients without the syndrome die with probability p_1 if they don't take the drug and die with probability p_2 if they do take the drug. Patients with the syndrome die with probability p_3 if they don't take the drug and die with probability p_4 if they do take the drug. The complication is that having the syndrome makes it uncomfortable to to take the potentially life-saving drug. Assume that patients with the syndrome take the drug with probability q_2 and patients without the syndrome take the drug with probability q_1.

   (a) Find the joint distribution $P(x, y, z)$ for all x, y, z (8 values) in terms of the parameters $r, p_1, p_2, p_3, p_4, q_1, q_2$.
   (b) Calculate the difference in death probabilities between takers and nontakers of the drug, $P(y = 1|x = 1) - P(y = 1|x = 0)$.
   (c) Calculate the difference in death probabilities between takers and nontakers of the drug for those with $z = 1$ (having the syndrome), $P(y = 1|x = 1, z = 1) - P(y = 1, |x = 0, z = 1)$.
   (d) Calculate the difference in death probabilities between takers and nontakers of the drug for those with $z = 0$ (not having the syndrome).
   (e) Find a combination of parameter values that exhibit Simpson's paradox (i.e., where (c) and (d) show negative effect (lower death rate) but (b) doesn't.
   (f) Compute $P(y \mid do(x))$ for all values of x and y. (Recall this requires one to average over the parent(s) of X.)
   (g) Compute the average treatment effect $P(y = 1 \mid do(x = 1)) - P(y = 1 \mid do(x = 0))$. How does this quantity differ from the quantity computed in (b) above? Which is more relevant in assessing the effectiveness of the treatment? Explain.