University of Toronto
Dalla Lana School of Public Health
CHL 5210 – Categorical Data Analysis
Assignment #1 – due 23 October 2016

1. An investigator wishing to examine the adverse effects of a new anti-hypertension drug, collected data from three different hospitals. They reviewed charts for all in-patients who had an active prescription for either the new drug or the usual drug and recorded whether the patient complained of persistent dry cough. The data are summarized in the table below with the number of patients using each drug (n) and the number of patients reporting the adverse effect (y). Show that the investigator would arrive at a misleading conclusion if they were to base their analysis on aggregate counts, summed across the three hospitals. Support your argument with relevant statistics.

	New drug		Usual drug	
Hospital	**n**	**y**	**n**	**y**
Princeton-Plainsboro	625	166	48	12
County General	222	70	398	120
St. Eligius	150	63	549	220
aggregate	997	299	995	352

[4 points]

2. A study recruited 478,000 healthy adults. During the five-year follow-up period, the proportion diagnosed with colorectal cancer was 0.003649 among participants with daily intake of processed meat of 80 g or more and 0.002723 among participants with lower intake of processed meat. Evaluate and interpret the point estimates for the difference of proportions and the relative risk. Identify which of these two measures of association is more informative for this study and explain your assertion.

[3 points]

3. Exercise 2.16 in Agresti (2013).

[2 points]

4. The 2013-2014 Canadian Community Health Survey asked a sample of Toronto residents about their use of protective equipment while bicycling. Of the 447 people who responded that they had bicycled in the past 12 months, 175 reported always using a helmet. Calculate (show calculations) and interpret a 95% confidence interval, based on the score test, for the proportion of bicyclists wearing a helmet.

[3 points]

5. Using the Survey Documentation and Analysis (SDA) tool at the University of Toronto's Computing in the Humanities and Social Sciences (CHASS) online database collection, access the 2013-2014 dataset for the Canadian Community Health Survey (CCHS). Identify an example for each of the following: (a) binary variable, (b) nominal variable, (c) ordinal variable, and (d) a categorized continuous variable contained in the 2013-2014 CCHS dataset – provide both "Name" and "Label" for each variable. No points will be awarded for either the *bicycling in the past year* or *bicycle helmet use* variables.

[4 points]

6. A clinician-investigator wishes to run a very small pilot study (n=10) to describe whether their patients feel proud at least some of the time. It is known that the proportion is 0.90 in the general population. Realizing that they are proposing a very small sample, they have come to you for advice regarding the use of confidence intervals based on the Wald test, the "exact" binomial test, or a "plus-four" adjusted Wald test described in Agresti & Coull (1998). Using language that an intelligent non-statistician would understand, describe anticipated relative performance of these intervals for this specific proposal.

[5 points]

7. Exercise 3.11 in Agresti (2013).

[3 points]

8. Exercise 5.6 in Agresti (2013).

[3 points]

9. Exercise 5.34 in Agresti (2013).

[3 points]