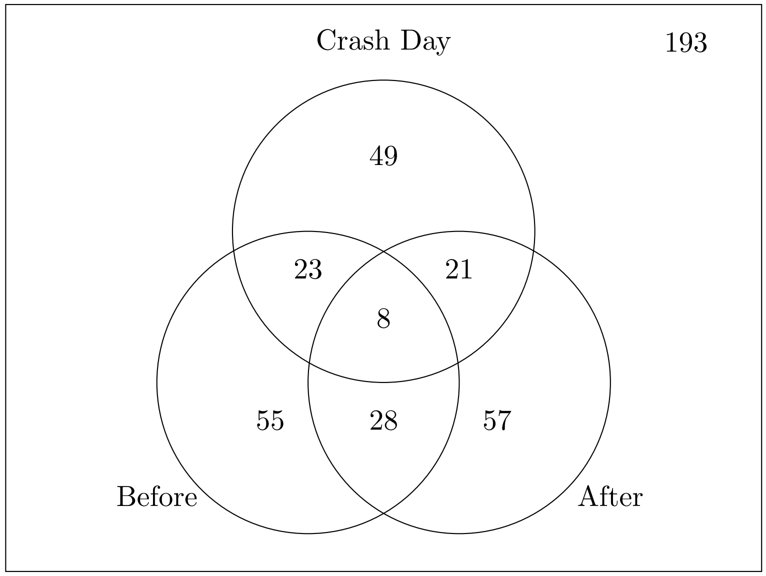
STAC51 (Winter 2020): Final Exam
April 21st 4pm - April 22nd, 2020 10pm
All relevant work must be shown for full marks

Note: In any question, if you are using `R`, all `R` codes and `R` outputs must be included in your answers. You should assume that the reader is not familiar with `R` and so explain all your findings, quoting necessary values from your outputs. **Please note that academic integrity is fundamental to learning and scholarship. You cannot discuss any answers with anybody else.** Answers can be handwritten but the `R` codes and outputs should be printed. **You will provide only one PDF file for your final submission. Multiple file submissions will not be accepted.** There are 3 questions with further subsections and you have to answer all of them to receive full marks. **The question paper has 4 pages**. Make sure you check all of them.

Late Submissions: Late submissions will not be marked. You will get a '0' in Final. This will be strictly followed. You will not be able to submit the final even 1 minute after the deadline. No extensions. If you miss the deadline for any reason you need to defer the exam.

Total Marks = 90

Best of luck!



1. The numbers in the Figure above indicate the weather (overcast or not) of 434 location-matched triplets of days, one day on which a traffic accident took place, and two control days without an accident (the day before the accident and the day after the accident). This dataset could be analyzed as a 1:2-matched case-control. The Venn diagram presentation of the data is rather unconventional. In matched case-control studies the data could alternatively be presented in 2×2-tables

	exposed	unexposed
case	a_i	b_i
control	c_i	d_i

for each matched set $i = 1, \ldots, 434$. We denote $n_i = a_i + b_i + c_i + d_i$.

   (a) **[10 Marks]** We note that there are six types of location-specific 2×2-tables with the same exposure-case configuration. List these tables (i.e. different combinations of the numbers a_i, b_i, c_i and d_i) and their counts.

   (b) **[6 Marks]** The null hypothesis assumes that there is no relationship between being a case and being exposed. Under the null hypothesis the distribution of the cell count a_i conditional on the row and column marginals is hypergeometric. Find $E(a_i \mid a_i + c_i)$ and $\mathrm{Var}(a_i \mid a_i + c_i)$ under the null.

   (c) **[6 Marks]** Test the null hypothesis of no association between weather and accidents using the Cochrane-Mantel-Haenszel (CMH) test statistic, given by

$$\frac{\left(\sum_{i=1}^{434} a_i - \sum_{i=1}^{434} E(a_i|a_i + c_i)\right)^2}{\sum_{i=1}^{434} \mathrm{Var}(a_i \mid a_i + c_i)},$$

   which is asymptotically distributed as χ^2 with one degree of freedom.
   Note: $\chi^2_{0.95}(1) = 3.84$.

(d) **[4 Marks]** Recall that for 1:1 matching there exist 4 unique types of CMH tables. For 1:2 matching there exist 6 unique types of tables. If we have a 1:k matched case control study how many, unique types of tables exist? Here $k < \infty$.

(e) **[10 Marks]** Let's assume we have the following table is a triplet specific contingency table from a 1:2 matched case control study.

	exposed	unexposed	Total
case	a	b	1
control	c	d	2
Total	$a+c$	$b+d$	3

The odds of being exposed in the case groups is θ times of the odds of being exposed in the control group. Also, let's assume that $P(a=1) = \dfrac{\theta\Omega}{1+\theta\Omega}$ and $P(c=1) = \dfrac{\Omega}{1+\Omega}$

Show that, $P(a = 1 \mid a + c = 1) = \dfrac{\theta}{2+\theta}$

(Hint: The 2 in the denominator comes from 2 controls).

2. For this question you need to use the `warpbreaks` dataset from the `datasets` package. That is you need to run the following code,

```
## Run this code to get the veteran dataset ##
library(datasets)
data(warpbreaks)
```

You can find the details about the dataset by using '`?warpbreaks`' code. We are interested in the count of warp breaks per loom (i.e., variable = '`breaks`') by wool and tension level.

(a) **[8 Marks]** Execute a Poisson regression to estimate the mean number of breaks by wool type and tension level.

(b) **[8 Marks]** Execute a negative binomial regression to estimate the mean number of breaks by wool type and tension level.

(c) **[6 Marks]** Compare the models using the AIC values. Interpret the dispersion parameter of the negative binomial regression. Which model performed better?

3. For this question you have to simulate a dataset.

(a) **[5 Marks]** Perform the following simulations.

- Generate 500 random values from $X_1 \sim \text{Uniform}[0,1]$, $X_2 \sim \text{Uniform}[0,1]$, $X_3 \sim \text{Uniform}[0,1]$, $X_4 \sim \text{Uniform}[0,1]$, $X_5 \sim \text{Uniform}[0,1]$
- Generate, $f(\boldsymbol{X}) = 4[\sin(\pi x_1 x_2) + 8(x_3 - 0.5)^3 + 1.5x_4 - x_5 - 0.77]$. Here, $\pi = 3.14.....$
- Generate $Y \sim \text{Bernoulli}\left(p(\boldsymbol{X}) = \dfrac{\exp(f(\boldsymbol{X}))}{1+\exp(f(\boldsymbol{X}))}\right)$

(b) **[10 Marks]** Fit a logistic regression where Y is the outcome and $X_1, X_2, ..., X_5$ are the predictors. Show the coefficients table. Produce the ROC curve. State the AUC value and interpret.

(c) **[10 Marks]** Now instead of using the original $X_1, X_2, ..., X_5$ as predictors, transform the variables in such a way that they resembels the individual terms in $f(\boldsymbol{X})$. That, is create new variables from $X_1, X_2, ..., X_5$ in such a way that $f(\boldsymbol{X})$ is transformed to a linear

predictor. Now run a logistic regression using the new variables. Show the coefficients table. Produce the ROC curve. State the AUC value.
(**Hint:** You have to create 4 new variables from $X_1, X_2, ..., X_5$)

(d) **[7 Marks]** Compare your results in (b) and (c): how did your coefficients and AUC change from (b) to (c)? Explain why you think this happened.