STAT 337 – Introduction to Biostatistics

Spring 2021 – Final Exam

Instructor: Dina Dawoud

Time: Available from 7am EST on August 10th and must be submitted to crowdmark by 10pm EST on August 10th. Once started must be completed and submitted within 3.5 hours.

Lecture Section: 001

Exam type: Restricted Open Book

Number of Exam Pages: 23

INSTRUCTIONS:

- There are **7 questions**. Marks are indicated for each part. The exam is out of **73 marks**.
- You may print the exam and type your solutions in the allotted space, use a PDF editor to add your solutions, or complete your solutions entirely separate from the exam PDF. Explanations and solutions requiring sentences MUST be TYPED out. Formulas can be typed (preferred) or handwritten (where a screen shot of your handwritten formulas can be pasted into your solution document). Submissions that are not typed will receive a 5% penalty.
- Submit your solutions to each question separately using the Crowdmark link you received via email. Be sure to give yourself plenty of time to submit before the deadline.
- ***NEW*** Submit your solutions as ONE pdf document to the learn dropbox as well before the deadline.
- Complete and submit the academic integrity statement to Crowdmark as Question 0.

PERMISSIBLE MATERIALS LIST:

- All items posted to the Spring 2021 STAT 337 LEARN page (including course notes, videos, assignments, and assignment solution keys)
- The course textbook
- Previous posts made to the Spring 2021 STAT 337 Piazza discussion forum (no new posts will be permitted during the exam)
- Your own written notes and assignments assembled during the term.
- A calculator or mathematical software (Excel, R, SAS, etc) used solely as a calculator.

Question 0

Academic Integrity Statement

Please complete the following academic integrity statement. You must submit this statement along with your examination or your examination will not be graded. If you do not have a printer or can't fill out the statement electronically you can copy and paste the text or write it out by hand.

I (Print full name) _____ , pledge that the work submitted herein is entirely my own. I have only accessed materials from the permissible materials list (on the exam coversheet) and I received no additional assistance on this final exam.

I understand that violating the exam regulations is an offence under Policy 71.

In addition, I have not and will not discuss the contents of this examination with my fellow classmates until after 10:00pm on Tuesday August 10th, 2021. I will not share the exam paper or the exam questions at any point in the future.

I act in a responsible manner because I have academic integrity and out of respect to myself (full name)_____, my fellow students in STAT 337, my professor, Dina Dawoud, and the institution of the University of Waterloo.

Signed,

_____ _____

Signature Date

_____

UW ID Number

Question 1 (10 marks):

a) Consider the following mortality data for two study populations (A and B), stratified according to twenty four-year age categories from 0 to 75 years and above.

	Population A		**Population B**		**Reference Population**	
Age group (years)	**Mid-year population**	**Deaths**	**Mid-year population**	**Deaths**	**Mid-year population**	**Death rate per 1000**
0 - 24	18,000	35	13,000	30	11,000	1.94
25 – 49	11,000	60	7,000	50	17,000	5.45
50 – 74	9,000	370	11,000	400	20,000	41.11
75 and above	3,000	250	4,000	380	3,000	83.33
Total	41,000	715	35,000	860	51,000	

i. (1 mark) What is the annual overall crude death/ mortality rate per 1000 for each population (separately)?

ii. (3 marks) Use indirect standardisation to compute the overall SMR for each population (separately) making use of the reference population. For one selected SMR value provide a precise written interpretation of your estimate.

iii. (1 mark) Compute the indirect adjusted rate (IAR) for each region.

iv. (1 mark) Does population A or population B experience a more favourable mortality experience? Cite the rates on which you base your decision.

b) Consider a prospective cohort study looking at the development of acute injuries among young elite athletes. The study includes six participants, and the goal of the study is to follow each participant for 1 year. Each participant is free of acute injuries at the start of the study. Over the course of the follow-up, some participants develop acute injuries (a participant can be classified multiple times), some drop out of the study, and some are followed till the end with no injury. Here are the details of each participant:

- Participant 1 develops an acute injury after 7 months in the study and remains injured till end of follow-up.
- Participant 2 develops an acute injury after 2 months in the study, recovers after a month and develops another acute injury after 6 months in the study. The participant recovers from the second acute injury after 3 months and no other acute injury is recorded till end of follow-up.
- Participant 3 completes the follow-up period with no acute injury.
- Participant 4 withdraws from the study after 4 injury free months.
- Participant 5 develops an acute injury after 1 month in the study, recovers after 2 months and remains injury free till the end of follow-up.
- Participant 6 develops an acute injury after 7 months in the study and then drops out after 9 months in the study (we have no record of recovery).

i. (1 mark) What is the point prevalence of acute injury at 2 months and at 8 months?

ii. (1 mark) What is the cumulative incidence of acute injury at 6 months?

iii. (2 marks) What is the incidence density of acute injury over the duration of the study? Use a person-time calculation in the denominator.

Question 2 (17 marks):

Parts a) to f) Refer to the following (lighted edited) abstract [Maurice Zeegers et al, *American Journal of Epidemiology*, Vol 153, No. 1, pp 38-41.]

"Alcohol Consumption and Bladder Cancer Risk: Results from the Netherlands Cohort Study"

Although several epidemiologic studies have been conducted on alcohol consumption and bladder cancer risk, the risk according to quantity and type of alcohol consumed is not clear. The authors investigated these associations in a large [redacted] on diet and cancer among 120,852 subjects in the Netherlands aged 55–69 years at baseline (1986). Subjects completed a questionnaire on risk factors for cancer, including alcohol consumption. Follow-up for incident cancer was established by record linkage to cancer registries. The case-[redacted] analysis was restricted to a follow-up period of 6.3 years and was based on 594 cases with bladder cancer and 3,170 subcohort members. The authors corrected for age and smoking in multivariable analyses. The incidence rate ratios for men who consumed <5, 5-<15, 15-<30, and ≥30 grams of alcohol per day were 1.49, 1.52, 1.16, and 1.63 compared with non-drinkers, respectively (p for trend =0.13). Alcohol consumption from beer, wine, and liquor was associated with moderately elevated risks, although most were not statistically significant. The incidence rate ratios for women varied around unity. The results of this study do not suggest any important association between alcohol consumption and bladder cancer risk.

a) (1 mark) Identify the study design. Explain why you selected this design.

b) (2 marks) Consider the following data reported in the paper.

Exposure	**Cases**	**Total Person Years**
Male (E+)	517	9555
Female (E-)	87	9748

Estimate the relative risk of bladder cancer for males compared with females. Provide a precise written interpretation of your estimate.

c) (2 marks) The authors state that “sex modified the association between alcohol consumption and bladder cancer incidence.” In your own words describe what it means for the variate “sex” to be an effect modifier in this study.

d) (2 marks) The authors state that “Additional correction for other potential confounders did not change the risk estimates substantially.” One confounder investigated was smoking status (amount and duration). In terms of the model findings describe, in your own words, what it means that smoking status was not found to be a confounder.

e) (2 marks) In general, if X1 is a confounder for the association between X2 and D, does that necessarily mean that X2 is a confounder for the association between X1 and D? Explain your reasoning.

f) (1 mark) The authors state “In our study, we found that the risk of bladder cancer increased slightly according to the quantity of alcohol consumed, irrespective of the type of alcoholic beverage, …”. Referring to Hill’s criteria for causation which criteria is being described here?

Part g) and h) refers to the study by "Moseid et al. The prevalence and severity of health problems in youth elite sports: A 6-month prospective cohort study of 320 athletes. *Scand J Med Sci Sports*, 2018; 28:1412-1423."

In the paper the authors look to investigate associations between type of sports played (endurance, technical and team sports) and health problems reported.

g) (3 marks) Consider the following data reported in the paper.

	All health problems		
	Illness (D-)	**Injury (D+)**	**Total**
Males (E+)	31	74	105
Females (E-)	18	32	50
Total	49	106	155

Estimate the relative risk of injury for males compared to females, together with a 95% confidence interval. Use your confidence interval to test the null hypothesis of no association. Provide a precise written interpretation of your estimate and confidence interval.

h) The researchers categorised the sports in to three categorises: Endurance sports, Technical sports and Team sports. Below you will find the illness status of males vs. females under each sports category (numbers not as quoted in paper).

Endurance sports

	All health problems		
	Illness (D-)	**Injury (D+)**	**Total**
Males	8	20	28
Females	4	9	13
Total	12	29	41

$\widehat{RR} = 1.03$

Technical sports

	All health problems		
	Illness	**Injury**	**Total**
Males	4	18	22
Females	2	9	11
Total	6	27	33

$\widehat{RR} = 1$

Team sports

	All health problems		
	Illness	**Injury**	**Total**
Males	30	25	55
Females	3	23	26
Total	33	48	81

$\widehat{RR} = 0.32$

i. (2 marks) This term, we used the Mantel-Haenzsel method to calculate a single pooled odds ratio for the association between the exposure and the disease. The Mantel-Haenszel method can also be used to calculate a single pooled estimate of the relative risk. For *k* stratum the formula is:

$$RR_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i(c_i + d_i)}{n_i}}{\sum_{i=1}^{k} \frac{c_i(a_i + b_i)}{n_i}}$$

Calculate the Mantel-Haenszel relative risk of injury in males vs. females while controlling for type of sports.

ii. (2 marks) Show that the formula for the Mantel-Haenszel relative risk can be written as weighted average of the stratum-specific relative risk with $w_i = c_i(a_i + b_i)/n_i$.

$$RR_{MH} = \frac{\sum_{i=1}^{k} RR_i w_i}{\sum_{i=1}^{k} w_i}$$

Question 3 (13 marks)

a) (5 marks) The table below provides some hypothetical results from four unrelated case-control studies. For each study a crude odds ratio (OR_c) and two stratum-specific odds ratios (OR_1 and OR_2) were calculated. The stratums were defined by the two levels of a binary third variable.

For each study, fill in the missing cells of the table indicating whether the results suggest that confounding is *present or absent* and whether the effect modification is *present or absent*.

For one study of your choosing explain your reasoning behind your decisions.

Study	**Crude OR** OR_C	**Stratum-Specific ORs** OR_1	OR_2	**Confounding**	**Effect Modification**
1	0.80	2.10	6.10		
2	2.75	1.20	6.35		
3	2.95	2.95	2.99		
4	3.50	1.05	1.01		

Parts b) to d) refer to the following:

Between 1969 and 1971, the Collaborative Group for the study of Stroke in Young Women conducted a case-control study in 12 university hospitals of cerebrovascular disease and oral contraceptive (OC) use in non-pregnant women ages 14- to 44-years. Data were *matched* according to the age, sex, and race of subjects. Data for hemorrhagic stroke with data for *neighborhood* controls are:

		Control		**Total**
		OC^+	**OC^-**	
Case	**OC^+**	5	30	35
	OC^-	13	107	120
	Total	18	137	155

b) (2 marks) Estimate the Odds Ratio for the association between exposure and outcome. Provide a precise written interpretation of your estimate.

c) (4 marks) Use McNemar's Test to test the significance of the association between exposure and outcome. Be sure to clearly state the null and alternative hypotheses, give the formula for the test statistic, calculate its value, and find the p-value. What is the conclusion of the test?
Recall $P\left[\chi^2_{(1)} > 2.91\right] = 0.10;\ P\left[\chi^2_{(1)} > 3.84\right] = 0.05;\ P\left[\chi^2_{(1)} > 6.63\right] = 0.01$

d) (2 marks) Name the bias that most closely matches the situations described below: Options to select from: Information Bias, Interviewer bias, lead-time bias, length-time bias, misclassification bias, recall bias, record bias, selection bias, survival bias.

i) A case is more motivated to participate than a control, and thus more likely to report past exposures accurately.

ii) A case-control study being performed obtains detailed information on cases from a medical database, where as information on controls is collected through a self-reported survey.

iii) An interviewer learns to distinguish between cases and controls and subsequently differs slightly between them in how they ask the questions.

iv) A history of binge drinking is defined as having had five or more drinks in one day for males or four or more drinks in one day for females at least once during the past year. The analyst incorrectly uses the five drink threshold for both males and females in the study. Since there are more male cases than females this distorts the association between a history of binge drinking and the risk of AMI.

Question 4 (6 marks)

The following question is based on the study from "Jamil et al. Pilates improves pain, function and quality of life in patients with chronic low back pain: a randomised controlled trial. *Clinical Rehabilitation*. 2015. Vol. 29(I) 59-68."

An exert from the paper's abstract is given below:

Objective: To assess the effectiveness of Pilates method on patients with chronic non-specific low back pain (LBP).

Method: A randomized controlled trial was carried out in sixty patients with a diagnosis of chronic non-specific LBP. Patients were randomly assigned to one of two groups: Experimental Group (EG) that maintained medication treatment with use of NSAID and underwent treatment with the Pilates method and Control Group (CG) that continue medication treatment with use of NSAID and did not undergo any other intervention. A blinded assessor performed all evaluations at baseline (T0), after 45, 90, and 180 days (T45, T90 and T180) for: pain (VAS), function (Roland Morris questionnaire), quality of life (SF-36), satisfaction with treatment (Likert scale), flexibility (sit and reach test) and NSAID intake.

Results: The groups were homogeneous at baseline. Statistical differences favoring the EG were found with regard to pain ($P < 0.001$), function ($P < 0.001$) and the quality-of-life domains of functional capacity ($P < 0.046$), pain ($P < 0.010$) and vitality ($P < 0.029$). Statistical differences were also found between groups regarding the use of pain medication at T45, T90 and T180 ($P < 0.010$), with the EG taking fewer NSAIDs than the CG.

Conclusions: The Pilates method can be used by patients with LBP to improve pain, function and aspects related to quality of life (functional capacity, pain and vitality). Moreover, this method has no harmful effects on such patients.

In addition, according to the paper "Ninety-seven patients with non-specific low back pain were screened and 60 patients who fulfilled the eligibility criteria were selected … Three drop-outs occurred throughout the study (one in the control group and two in the experimental group) … Regarding the compliance ninety-six percent of patients completed all Pilates sections."

a) (1 mark) Based on the information above, how many individuals would be included in each group for an intention-to-treat analysis?

b) (1 mark) now suppose you wanted to conduct a per-protocol analysis. What would be the sample size for each group? (if needed you can round numbers up to the nearest whole).

c) (2 marks) Suppose that some of the participants did not comply with their assigned treatment and/or switched to the other treatment. How would this affect the results of an intention-to-treat analysis? Explain your reasoning.

d) (2 marks) The authors indicate that the person performing the evaluations at time points T0 etc was blinded to which treatment the subject was receiving. Would it be possible for this randomized trial to be triple-blinded? If so, explain why in 1-2 sentences. If not, explain which group(s) could not be blinded and why.

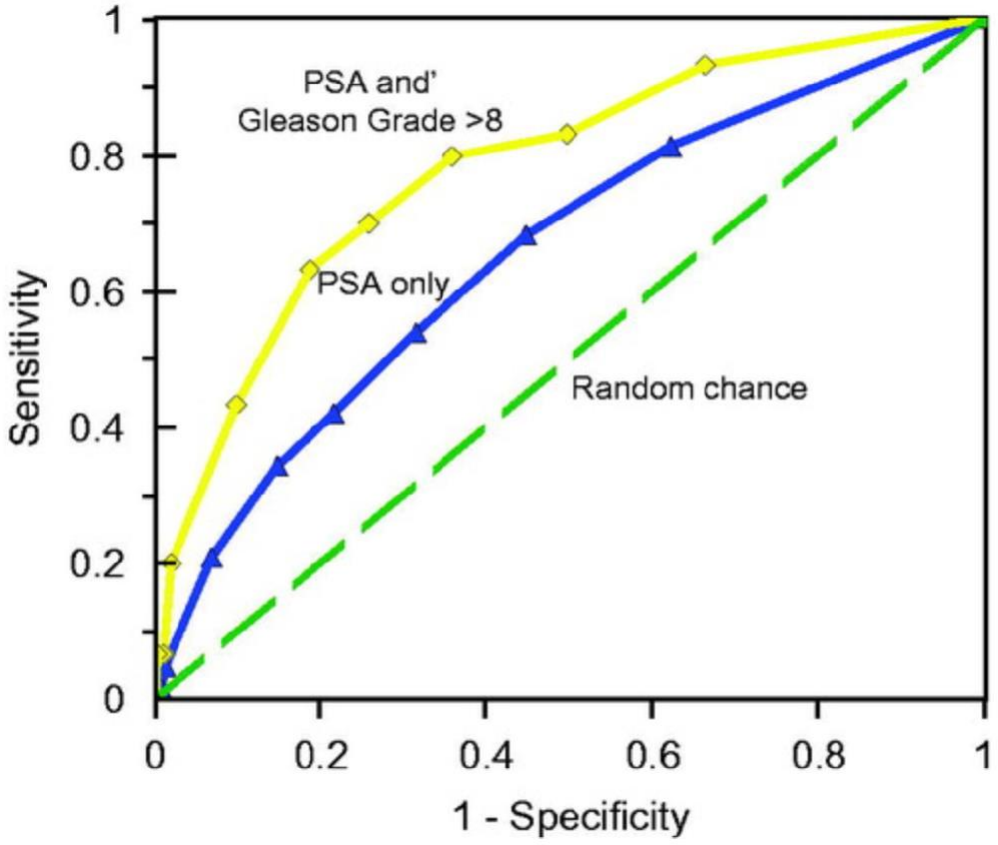
Question 5 (7 marks)

Parts a) to c) refer to the following:

Consider two methods used for detecting prostrate cancer:

1. A prostate specific antigen (PSA) test only.
2. A PSA test and Gleason score with a cut off of 8 or more.

Receiver operating curves (ROC) for both methods are displayed below



Where, AUC of PSA only is 0.678 and AUC of PSA + Gleason Grade is 0.827

a) (1 mark) Consider the ROC curve for the PSA and Gleason Grade >8 test, which combination along the curve would most reliably indicate no prostrate cancer among those that do not have the disease? Explain why.

b) (2 marks) for each test identify (approximately) the point of maximal discrimination (based on the displayed cut off points). What is the (approximate) sensitivity and specificity of each test at this point?

c) (2 marks) Based on the AUC values which test is the better one and why? In your own words describe what the AUC value captures.

Part d) refers to the following:

Researchers develop a new test to screen for chronic kidney disease (CKD). They administer the test to 5500 apparently healthy people from six communities, A central laboratory conducts all of the testing. A total of 450 of the tested individuals have a positive test.

The researchers select s comparison group of 450 patients with established CKD from kidney clinics in the same communities. They compare mortality rates among people who have CKD detected by the new screening test to that of the patients who have CKD identified from the kidney clinics. The results are presented below.

	Number of people	**Median follow-up in years**	**Adjusted mortality rate per 100 person-years**
CKD detected by screening test	450	9.1	3.0
CKD from kidney disease clinics	450	5.0	5.5

The researchers conclude that the new screening test reduces death from CKD and should be used to screen asymptomatic people for this condition.

d) (2 marks) Which type of bias is most likely responsible for the observed difference in survival? (selection/referral bias, length-time bias and/or lead-time bias) Explain.

Question 6 (10 marks):

In pregnancy, women typically undergo screening to assess whether their fetus is likely to have Down Syndrome. The screening test evaluates levels of specific hormones in the blood. Screening test results are reported as positive or negative, indicating that a women is more or less likely to be carrying an affected fetus. Suppose that a population of N=4,810 pregnant women undergo the screening test and are scored as either positive or negative depending on the levels of hormones in the blood. In addition, suppose that each woman is followed to birth to determine whether the fetus was, in fact, affected with Down Syndrome. The results of the screening tests are summarized below.

Screening Test	**Down Syndrome**	**No Down Syndrome**	**Total**
Positive	8	350	358
Negative	2	4450	4452
Total	10	4800	4810

a) (7 marks) Based on the given data, estimate the following:

- Prevalence

- Sensitivity

- Specificity

- Positive Predictive Value

- Negative Predictive Value

- Likelihood Ratio for Positive Test

- Likelihood Ratio for Negative Test.

b) (3 marks) In the context of the study, provide a precise interpretation of the sensitivity, negative predictive value and likelihood ratio for positive test.

Question 7 (10 marks)

A (hypothetical) randomized controlled trial was conducted to evaluate the efficacy of a new drug for prevention of hypertension in patients with prehypertension (defined a systolic blood pressure between 120 mmHg and 139 mmHg or diastolic blood pressure between 80 mmHg and 89 mmHg). A total of 20 patients are randomized to receive the new drug or a currently available drug for treatment of high blood pressure. Participants were followed for up to 12 months, and the time to progression to hypertension is measured. The experiences of participants in each arm of the trial are given in the table below.

Months to Progression to Hypertension or Withdrawal (Free of Hypertension) by Treatment Group

Treatment Group (New Drug)	Hypertension	7	8	8	9		
	Withdrawal	8	8	9	11	12	12
Control Group (Currently Available Drug)	Hypertension	6	7	9	10	11	
	Withdrawal	8	9	11	11	12	

a) (6 marks) Estimate the survival functions (time to progression to hypertension) for each group using the Kaplan-Meier method. Your answer should include two tables. Show the precise calculations and formulas used for at least one of the rows of the Treatment Group table. Carry out your calculations to 4 decimal places.

b) (4 marks) Sketch (by hand) the Kaplan-Meier survival functions from part (a) for each group on the same plot. Give an estimate of the median time to progression to hypertension in each group.

LEFT BLANK FOR ROUGH WORK (WILL NOT BE MARKED)