**Research Methods MEDP5321 Assignment 2**

**Assignment questions**

|  |  |  |
| --- | --- | --- |
| 1. | Explain the importance of bias in the context of observational studies and study design. Give two examples of the types of bias which may affect observational studies. | (35) |
| 2. | Explain why randomised clinical trials (RCTs) are less likely to be prone to bias compared to observational studies. Describe the key aspects of RCTs which minimise bias. | (25) |
| 3. | A research team collected data on 25 individuals who had experienced severe knee pain over the last 12 months. Their age distribution was assessed and summarised below.    (a) Describe the type of distribution seen above and explain what type of summary statistics should be used to describe these data correctly. | (20) |
|  | (b) The research team wished to assess the strength of the correlation between age and knee pain, with pain levels measured on a scale from 0-99 and able to take any value. Explain what analytical approach would be most appropriate and why. | (25) |
|  | (c) A replication study was repeated 6 months later on a different sample. A scatter plot was produced, correlation analysis conducted and results summarised below:      Explain how you would interpret these findings. | (30) |
|  | d) What additional information would be helpful to know in order to draw appropriate conclusions? | (20) |
|  | e) Suppose you had identified three other confounders and collected data on these variables as part of an extended analysis. Assume that the main outcome variable was pain level and age was the main exposure variable. What type of analysis would be appropriate for this example? Explain how you would carry out the analysis. | (30) |
| 4. | The table below summarises the number of subjects recruited into a case control study examining whether residing in an area with ‘high’ or ‘normal’ levels of air pollution affected the risk of being hospitalised for respiratory illness.   |  |  |  |  | | --- | --- | --- | --- | | **Air pollution** | **Whether hospitalisation was required** | | | | Yes (case) | No (control) | Total | | High | 44 | 106 | 150 | | Normal | 25 | 120 | 145 |   a) Calculate an appropriate estimate in order to compare the probability of being hospitalised in those from a highly polluted area compared to an area with normal levels of pollution. Interpret this estimate. | (25) |
|  |  |  |
|  | b) What additional analysis might be useful to conduct in order to confirm whether any association between air pollution and respiratory disease was a real causal relationship? | (20) |
|  |  |  |
|  | | |
| 5. | Individuals aged over 50 years were randomly selected to take part in a study examining blood pressure and the association with heavy smoking. The research question of interest was whether blood pressure was notably lower in those who had been heavy smokers in the past. Diastolic blood pressure (‘dbp’) measured in mmHg was recorded for each person. A binary variable indicated whether subjects classified themselves as a heavy smoker, defined as those who smoked at least 20 cigarettes per day for at least 5 years.  Output from Stata is provided below;    a) Explain what sort of test has been carried out and whether this was a valid statistical approach.  b) Interpret the results from the output above.  c) Identify two other limitations associated with this type of statistical analysis  d) Suppose an alternative exposure variable for smoking was created which classified individuals according to ‘low’, ‘moderate’ and ‘high’ levels of smoking. How might differences in blood pressure be assessed across these three groups? | (20)  (40)  (20)  (20) |
| 6. | Information was collected on systolic blood pressure (SBP) (mmHg) which represented the primary outcome variable for a study investigating the association with weight (kg), current age (years) and sex (males coded as 0, females coded as 1).    Stata output from a multiple linear regression model is summarised below. Assume a significance level of 1% was chosen.    a) Explain using the output above how each variable affects blood pressure.  b) Explain what the ‘\_cons’ term refers to in the output above.  c) Explain what the “R-squared” value represents and what would happen to it if additional explanatory covariates were added into the regression model above.  d) What error has been made in relation to the output for the confidence intervals? | (40)  (10)  (30)  (20) |

|  |  |  |
| --- | --- | --- |
| 7. | Discuss the dangers of limiting presentation of statistical analyses to P-values. Within your answer, suggest alternative ways of presenting results instead of using P-values and consider the following aspects:   * effect estimation versus hypothesis testing, * precision of estimates, * statistical significance versus clinical significance, * critical values (level of significance), * sample size, * power. | (30)  (20)  (20)  (10)  (10)  (10) |