HMP 5114: Biostatistics

Attempt all questions in section A and any 3 from section B

SECTION A: Attempt all questions in this section (10 marks each)

1. Attempt question (a) to (h) in this section

   a. Serum cholesterol was measured in a sample of 24 subjects before and after a six week diet. Data: mean difference 9.3 with standard deviation 16.8. Is the mean cholesterol lower than before? (1 mark)

   b. An epidemiologist wants to know if the prevalence of a certain disease has changed. From historical data it is known that the prevalence in the past was 0.20, based on a population wide study. A random sample is drawn from the population with size n=100 and sample prevalence p is determined. Question: Based on the data, what is the evidence of a change if 25 cases were observed? (1 mark)
      i. How large should the study have been to have a high power, let's say 90%, against an increase of 5%? (1 mark)

   c. In a family of 4 children, what is the probability of selecting 3 boys? (1 mark)
      i. What is the expected value of the number of boys in this family? (1 mark)

   d. Suppose that from experiment or literature it is known that $\pi_0 = 0.80$. Suppose further that an increase in cure probability of $\pi_1 - \pi_0 = 0.1$ is considered both clinically relevant and realistic. How many patients are needed for power $1 - \beta = 0.85$. (1 mark)

   e. Suppose that age in a population being considered has a normal distribution with a mean of 50 years and a standard deviation of 10. What proportion of the sample has age greater than 85 years? (1 mark)

f. A sample of 20 people were drawn from a population of persons with Angina pectoris with a mean total cholesterol of 5.81mmol/l and standard deviation of 1.2, Calculate the standard error of the mean and approximate 95% confidence interval of the mean. (1 marks)

g. Suppose X=24hr total energy expenditure (MJ/day) compared between lean and obese women

group	*mean*	*sd*	*n*
lean	8.066	1.238	13
obese	10.298	1.389	9

h. Are the two means different? Use exact test and give 95%CI of the difference in means. (2 mark)

SECTION B: Attempt any 3 questions (20 marks each)

2. In **comparing means of two populations**, the data in the table below are form Charles Darwin's study of cross- and self-fertilization (Darwin C. (1876): The effect of cross-fertilization in the Vegetable Kingdom). Pairs of seedlings of the same age, one produced by cross-fertilization and the other by self-fertilization, were grown together so that the members of each pair were reared under nearly identical conditions. The data are the final height of each plant after a fixed period of time. The question is whether cross-fertilized reach higher final height than self-fertilized plants.

Descriptive Statistics

	N	Mean	Std. Deviation
Cross-fertilized	15	20.1933	3.61613
Self-fertilized	15	17.5867	2.03816
difference	15	2.6067	4.71282
Valid N (listwise)	15		

i. Which 2 non-parametric tests are appropriate for analysis of these data? (2 marks)
   a. Perform one of these tests. (3 marks)
   b. Formulate the null hypothesis (1 marks)
   c. Give the value of the test statistic and bounds for the p-value (3 marks)

ii. Test the null hypothesis that cross-fertilized and self-fertilized plans reach on average the same final height. (5 marks)
   a. Give also a 90% confidence interval for the mean difference in final height. (4 marks)
   b. What is the underlying assumption for this test and confidence interval? (2 marks)

3. Dichotomous outcome: Sixty-five pregnant women at high risk of pregnancy-induced hypertension participated in a randomized controlled clinical trial comparing 100mg of aspirin daily and a matching placebo during the 3rd trimester of pregnancy. The observed numbers with hypertension are shown in the following table.

	Hypertension		
	yes	No	Total
Aspirin	5	29	34
placebo	10	21	31
Group Total	15	50	65

i. Is the risk of hypertension in aspirin treated women significantly lower than in placebo treated women? (1 marks)

a. Which test do you use? (1 marks)
b. What is the P-value? (3 marks)

ii. Give the **estimate** and **approximate 95% confidence interval** for the following of hypertension between aspirin and placebo treated women
   a. Difference in risk (3 marks)
   b. Risk ratio (3 marks)
   c. Odds ratio (3 marks)

iii. Suppose a new study is planned. What sample size is approximately needed in order to have a power of 80% if the risk of hypertension is 0.1 lower in aspirin treated women ($\alpha = 0.5$)? (6 marks)

4. A two-period two-treatment cross-over trial was carried out in 67 patients with cerebrovascular deficiency. Each patient was treated with either active treatment (A) in the first and with placebo (B) in the second period, or with placebo in the first and active treatment in the second period. The order in which the treatment were applied was randomized. The outcome variable was the assessment of ECG, scored as normal or abnormal. The table below gives the data. The variables are defined as: PATNO=patient number; ORDER= order in which the treatment were applied (AB or BA); ECG_A=assessment of ECG under treatment A (0=abnormal, 1=normal); ECG_B= assessment of ECG under treatment B (0=abnormal, 1=normal); DIF=ECG_A-ECG_B.

Quick summary of the data

	Order AB			Order BA			
	0	1		0		1	
	0	0	1	0	1	0	1
Total	**6**	**6**	**22**	**9**	**2**	**4**	**18**

Some descriptive statistics for the whole group are given in the table below.

	N	Std. Deviation		
	Statistic	Statistic	Std. Error	Statistic
ecg_A	67	.75	.054	.438
ecg_B	67	.63	.060	.487
Diff	67	.12	.050	.409
Valid N (listwise)	67			

The data are analyzed under the assumption that the treatment effect does not depend on the order in which the placebo and active treatment were given (i.e. no "carry-over" effect). Furthermore it is assumed that other is no difference between the two periods with respect to the chance of response (no "period" effect).

i. Give the cross table that adequately describe the data. (3 marks)
   a. Give the percentage response under the placebo and active treatment. (2 marks)

ii. Give an estimate of treatment effect (i. e. percentage response under the active minus placebo treatment). (2 marks)
   a. What is the corresponding standard error? (1 marks)
   b. Compute also an approximate 95% confidence interval (2 marks)

iii. Test with an approximate test the hypothesis that there is no treatment effect (provide the value of the test statistic and the corresponding p-value) (3 marks)
   a. Use also an exact test for this hypothesis (provide the bounds for the p-value) (3 marks)

Some statistics per treatment order group are given below.

AB

	N	Mean		
	Statistic	Statistic	Std. Error	Std
ecg_A	34	.82	.066	.387
ecg_B	34	.65	.083	.485
Diff	34	.18	.066	.387
Valid N (listwise)	34			

BA

	N	Mean		
	Statistic	Statistic	Std. Error	Std
ecg_A	33	.67	.083	.479
ecg_B	33	.61	.086	.496
Diff	33	.06	.075	.429
Valid N (listwise)	33			

One assumption underlying the above analysis was that the treatment effect did not depend on the order in which placebo and active treatment were given.

i. Describe a simple approximate way to test this hypothesis. (1 marks)
ii. Carry out this test, give the p-value and state your conclusion. (3 marks)

5. To demonstrate t-test, the Dutch PCB/Dioxin study was set up to investigate adverse health effects of perinatal exposure to PBC's and dioxins. One of the topics that was studied was the influence of breast feeding on the development of the child. In this problem we look at the difference between breast-fed and formula-fed infants for two characteristics: age of the mother and LPCB, using a 2-sample t-test. The variable LCPB stands for the exposure to PCB's, qualified as the sum of four PCB congeners, log-transformed.
   The SPSS output is given below.

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
Age	breast-fed	88	28.99	4.314	.460
	formula-fed	95	28.88	3.179	.326

Lpcb	breast-fed	88	.3486	.16258	.01733
	formula-fed	95	.3996	.18772	.01926

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age	Equal variances assumed	4.216	.041	.187	181	.852	.104	.557	-.995	1.204
	Equal variances not assumed	XXXX	XXXX	.185	159.237	.853	.104	.564	-1.009	1.218
Lpcb	Equal variances assumed	1.010	.316	????	????	????	????	.02605	????	????
	Equal variances not assumed	XXXX	XXXX	-1.970	180.203	.050	-.05104	.02591	-.10217	.00008

i. Which are the assumptions underlying a 2 sample t-test? (2 marks)

ii. Comment on the validity of these assumptions for this application (2 marks)

iii. Is the difference in mean age statistically significantly different between breast-feeding and formula-feeding mothers? (5 marks)

   a. Give the p-value and formulate your conclusion (3 marks)

iv. Fill in the cells with a question mark in the result of the t-test for LPCB. (6 marks)

   a. What is your conclusion about the difference in LPCB between the groups? (2 marks)