UNIVERSITY OF LONDON

PHM102

MSc/POSTGRADUATE DIPLOMA/POSTGRADUATE CERTIFICATE EXAMINATION

PUBLIC HEALTH

Basic Statistics for Public Health and Policy

June 2020 examination

Your answers must be YOUR OWN WORK without any help from others.

Read the “Instructions for completing LSHTM DL examinations (June 2020)” document **before** you start answering these questions.

You must answer **ALL** three questions on this paper.

Type all your answers to each question in the correct answer document provided.

You must upload all your answers to the EMS **within 48 hours** of downloading this exam paper.

Recommended time limit for completion of this paper: **2 hours and 15 minutes**.

You are advised to spend the first 15 minutes of this exam reading the question paper and planning your answers.

A pre-programmable calculator may be used to answer the questions in this paper.

When calculations are needed to answer a question, please show your working in full. ALL calculations MUST be typed into the answer documents. Do NOT upload any extra documents.

Statistical tables and a formulae sheet are provided for your use at the end of this paper. An accessible Word version of the formulae sheet is provided in the download zip file.

QUESTIONS START ON THE NEXT PAGE

QUESTION 1

Word limit: up to 1000 words per whole question (equivalent to approximately 2 pages of typed text with size 11 font)

A study in Finland (Räikkönen *et al*. 2017) examined maternal liquorice consumption during pregnancy and its effects on pubertal outcomes in children.

Liquorice consumption is thought to be higher in Finland than in other European countries due to the popularity of salmiakki, a salty liquorice snack. Liquorice contains a chemical called glycyrrhizin which has been shown in animal models to alter the timing of puberty.

Participants were recruited from an urban community-based cohort of women and their healthy, singleton infants born in 1998 in Helsinki, Finland. The average age of the children at the time of the study was 12.5 years. Low maternal consumption of glycyrrhizin during pregnancy was defined as ≤249 mg/week. High maternal consumption of glycyrrhizin during pregnancy was defined as ≥500 mg/week.

One of the measures taken to estimate age of puberty was a child's height. Analyses of pubertal development were conducted separately in girls and boys because adolescent girls are, on average, ahead of boys in pubertal development. 451 of the participants in the study were girls. Their height measurements are shown in Table 1.1 below.

Table 1.1: Girls' mean height by level of maternal consumption of glycyrrhizin in liquorice during pregnancy, Helsinki, Finland, 2009–2011

	Maternal glycyrrhizin consumption during pregnancy	
	Low (0-249mg/week)	High (≥500mg/week)
Mean height cm (SD)	155.4 (7.58)	159.1 (9.59)
n	327	51

(a) Do girls born to mothers with low consumption of glycyrrhizin during pregnancy or girls born to mothers with high consumption of glycyrrhizin during pregnancy have the larger variation in height? Explain your answer. **(10%)**

(Question 1 continues on the next page)

Question 1 continued

(b)

i. What is the difference in mean height between the low consumption and high consumption groups? **(5%)**

ii. Calculate the standard error of this difference. **(15%)**

iii. Use these answers to calculate the 95% confidence interval for this difference. **(10%)**

iv. What evidence does this confidence interval provide, that in the population from which the sample was taken, girls whose mothers consumed high amounts of glycyrrhizin in pregnancy have on average a greater height at aged 12.5 years than girls whose mothers consumed low amounts of glycyrrhizin in pregnancy? **(10%)**

(c) Carry out a statistical test for the difference in height between girls whose mothers consumed low amounts of glycyrrhizin in pregnancy and girls whose mothers consumed high amounts of glycyrrhizin in pregnancy. Remember to state your null hypothesis, show your working and interpret your results. **(30%)**

When considering their results, the researchers also looked at data from another cohort of women and similarly aged children in Finland. This cohort was recruited from a rural area. In the rural cohort the mean difference in heights between girls whose mothers consumed zero-low amounts of glycyrrhizin in pregnancy and girls whose mothers consumed high amounts of glycyrrhizin in pregnancy was 4.9cm 95% CI (1.92, 11.64cm).

(d)

i. Is this difference in mean height significantly different (at $p<0.05$) from the difference in mean height in the urban cohort? **(10%)**

ii. Was the sample size of the rural cohort larger or smaller than that in the urban cohort? You can assume that the standard deviations were similar. **(10%)**

Note: it is not necessary to carry out more calculations for question (d), but you should give reasons for your answers.

(Question 2 starts on the next page)

QUESTION 2

Word limit: up to 1000 words per whole question (equivalent to approximately 2 pages of typed text with size 11 font)

Researchers in Bangladesh were studying the effect of changing lifestyles on non-communicable diseases, and they sought to identify whether eating a healthy diet could help men to maintain a healthy weight. Using data from a nationally representative random sample of men aged 35-70 years in Bangladesh to test this hypothesis, the researchers examined the relationship between body mass index (BMI) and alternate healthy eating index score (AHEI).

BMI, measured in kg/m^2, uses one's height and weight to determine if their weight is healthy. For South Asians, a BMI between 18.0 and 22.9 kg/m^2 is considered a healthy weight. A BMI below 18.0 kg/m^2 is considered underweight. A BMI of 23.0 kg/m^2 or higher is considered as overweight or obese.

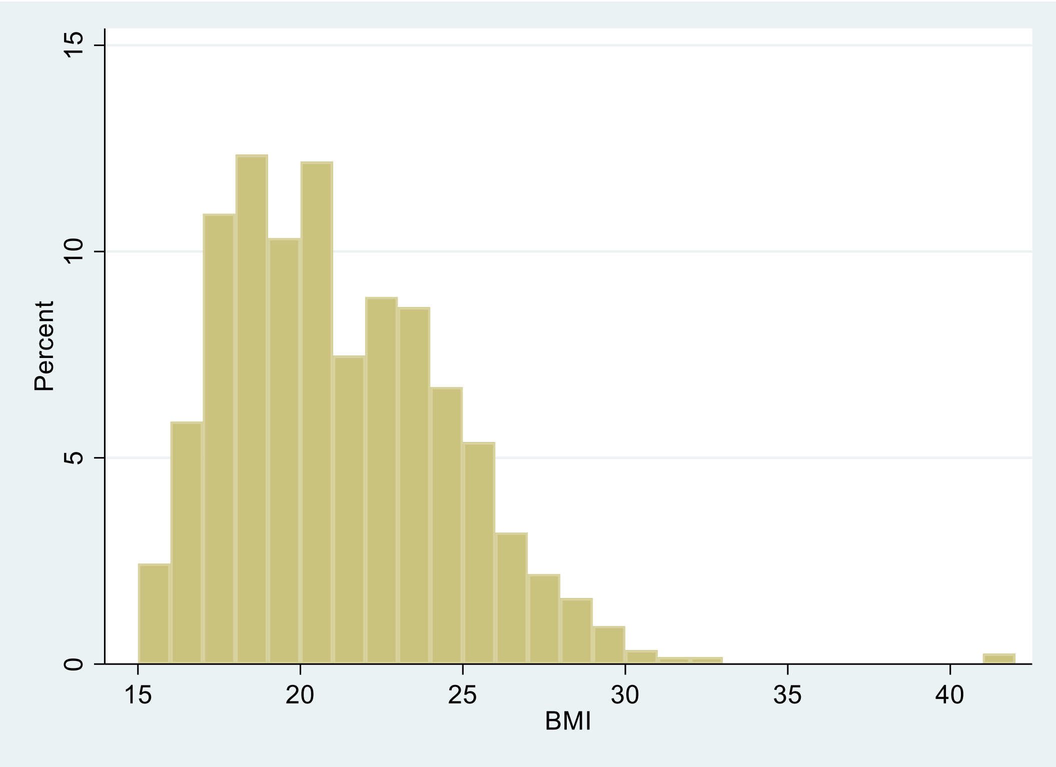
AHEI is a measure of dietary quality based on evidence-based recommendations. It ranges from a score of 0 points (lowest quality diet) to 110 points (highest quality diet).

As a first step, the researchers produced the histogram on the next page (Figure 2.1).

(Question 2 continues on the next page)

Question 2 continued

Figure 2.1: BMI in a nationally representative sample of men aged 35-70 years in Bangladesh



(a) What type of variable is BMI? **(5%)**

(b) Based on the histogram, what is the approximate range of BMI values observed in this national sample of men aged 35-70 years in Bangladesh? **(10%)**

(c) Based on the histogram, is the percentage of men who are considered underweight greater or less than the percentage of men who are considered overweight or obese? Please justify your answer. **(15%)**

(Question 2 continues on the next page)

Question 2 continued

The researchers then performed a linear regression analysis of body mass index (bmi) and alternate healthy eating index score (aheiscore). The Stata output from this analysis follows:

```
                                          Number of obs   =1,159
                                          F (1, 1157)      =6.10
                                          Prob > F        =0.0126
                                          R-squared       =0.0578
                                          Adj R-squared   =0.0469
                                          Root MSE        =9.4194

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-------------+----------------------------------------------------------------
   aheiscore |   .0666926   .0221073     3.02   0.003     .0233177    .1100676
       _cons |   18.75784   .8201063    22.87   0.000     17.14877     20.3669
------------------------------------------------------------------------------
```

(d) Write out the following sentence, choosing between the alternatives in each of the square brackets to provide the correct interpretation of the analysis. **(15%)**

> "The regression analysis shows that [*BMI* **or** *AHEI score*] [*increased* **or** *decreased*] by [*0.07* **or** *18.76* **or** *0.02* **or** *3.02* **or** *22.87*] [*kg/m^2* **or** *points*] for each [*kg/m^2* **or** *point*] increase in [*BMI* **or** *AHEI score*]."

(e) Is it likely that the association between the two variables has arisen by chance? Please justify your answer. **(15%)**

(f) Calculate the BMI predicted from this regression analysis for a Bangladeshi male with an AHEI score of 40 points. Show your working. **(15%)**

(g) As shown in linear regression output, the R-squared is equal to 0.0578. Interpret this result. **(10%)**

(h) Measures of physical activity, calorie intake and age were entered into a multiple linear regression analysis, giving a coefficient for AHEI score of 0.035 with a 95% confidence interval of [-0.009 to 0.079]. Interpret this result. **(15%)**

(Question 3 starts on the next page)

QUESTION 3

Word limit: up to 1000 words per whole question (equivalent to approximately 2 pages of typed text with size 11 font)

The Rankin score is a measure of disability after stroke. It ranges from 0 (no symptoms) to 6 (dead). A hospital has recorded the Rankin score on admission of all patients admitted alive to hospital with a stroke over a four-year period. Shown below is a tabulation of Rankin score on admission versus death in hospital (Table 3.1). Note all alive patients have Rankin scores 0-5.

Table 3.1: Survival and disability (by Rankin score) at admission in stroke patients

SURVIVAL IN HOSPITAL	RANKIN SCORE					TOTAL
	≤1 no symptoms/ no significant disability	2 slight disability	3 moderate disability	4 moderately severe disability	5 severe disability	
Discharged alive	1	10	76	88	37	212
Died in hospital	1	3	1	13	21	39
TOTAL	2	13	77	101	58	251

(a) Calculate the percentage of patients who were discharged alive. Give your answer to one decimal place. **(2%)**

(b) If survival in hospital were independent of Rankin score on admission, what would be (to 2 decimal places) the expected number of patients who:

i. Had Rankin score <=1 on admission and were discharged alive? **(2.5%)**

ii. Had Rankin score <=1 on admission and died in hospital? **(2.5%)**

(c) Your task is to assess evidence against the null hypothesis that survival in hospital is independent of Rankin score on admission. You have decided to use a chi-squared test of independence. Is this test valid using the table above? Explain your answer in one sentence. **(10%)**

(Question 3 continues on the next page)

Question 3 continued

(d) For the modified table, the observed values (O) and the expected values (E) under the null hypothesis for each cell are shown in Table 3.2 below. Some values are missing and are denoted by A, B, C and D. Calculate:

i. A **(5%)**

ii. B **(5%)**

iii. C **(5%)**

iv. D **(5%)**

v. The chi-squared test statistic, correct to 2 decimal places. **(10%)**

Table 3.2: Expected values under null hypothesis of independence (lowest categories combined)

SURVIVAL IN HOSPITAL	RANKIN SCORE		
	≤3 no symptoms/ no significant disability / slight disability moderate disability	4 moderately severe disability	5 severe disability
Discharged alive	O=**A** E=77.7052 $((O-E)^2/E)$ =**B**	O=88 E=85.3068 $((O-E)^2/E)$ = 0.0850	O=37 E=48.9881 $((O-E)^2/E)$ =2.9337
Died in hospital	O=**C** E=14.2948 $((O-E)^2/E)$ =**D**	O=13 E=15.6932 $((O-E)^2/E)$ = 0.4622	O=21 E=9.0120 $((O-E)^2/E)$ = 15.9468

(e)

i. Determine the degrees of freedom for the chi-squared test. **(5%)**

ii. Determine the p-value for this test. **(5%)**

iii. State your conclusion from this test. **(5%)**

(Question 3 continues on the next page)

Question 3 continued

(f) A chi-squared test for trend is conducted to obtain evidence about a linear trend in the percentages or proportions of patients in each Rankin score group who died in hospital. The chi squared test for trend statistic is calculated to be 23.74.

i. State the null hypothesis you are testing here. **(5%)**

ii. Determine the degrees of freedom for this test. **(3%)**

iii. Determine the p-value for this test. **(10%)**

iv. State your conclusion. **(5%)**

(g) In a second hospital where similar data were collected, the Rankin scores and survival of n=121 patients were recorded. As in the first hospital in part (f) (above) a chi-squared test for trend was carried out, with a resulting test statistic of 12.516. However, the second hospital only has access to the Standard Normal table, and not to the Chi-squared distribution table.

i. Explain how to find the p-value using the Standard Normal Table. **(10%)**

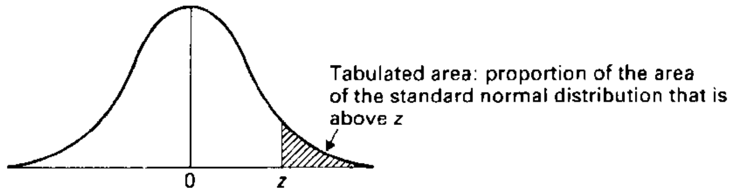
ii. Calculate this p-value and state your conclusion for this test. **(5%)**

END OF PAPER

Statistical tables and a formulae sheet are provided on the next pages.

Table A1 Areas in tail of the standard normal distribution.

Adapted from Table 3 of White *et al.* (1979) with permission of the authors and publishers.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.000.42	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003

Table A3 Percentage points of the t distribution.
Adapted from Table 7 of White *et al*. (1979) with permission of authors and publishers.

	One-sided P value								
	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	Two-sided P value								
d.f.	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62
2	0.82	1.89	2.92	4.30	6.96	9.92	14.09	22.33	31.60
3	0.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92
4	0.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.36	4.03	4.77	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96
7	0.71	1.42	1.90	2.36	3.00	3.50	4.03	4.78	5.41
8	0.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04
9	0.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59
11	0.70	1.36	1.80	2.20	2.72	3.11	3.50	4.02	4.44
12	0.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32
13	0.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22
14	0.69	1.34	1.76	2.14	2.62	2.98	3.33	3.79	4.14
15	0.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07
16	0.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02
17	0.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.96
18	0.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92
19	0.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88
20	0.69	1.32	1.72	2.09	2.53	2.84	3.15	3.55	3.85
21	0.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82
22	0.69	1.32	1.72	2.07	2.51	2.82	3.12	3.50	3.79
23	0.68	1.32	1.71	2.07	2.50	2.81	3.10	3.48	3.77
24	0.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.74
25	0.68	1.32	1.71	2.06	2.48	2.79	3.08	3.45	3.72
26	0.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71
27	0.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69
28	0.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67
29	0.68	1.31	1.70	2.04	2.46	2.76	3.04	3.40	3.66
30	0.68	1.31	1.70	2.04	2.46	2.75	3.03	3.38	3.65
40	0.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55
60	0.68	1.30	1.67	2.00	2.39	2.66	2.92	3.23	3.46
120	0.68	1.29	1.66	1.98	2.36	2.62	2.86	3.16	3.37
∞	0.67	1.28	1.65	1.96	2.33	2.58	2.81	3.09	3.29

Table A5 Percentage points of the X^2 distribution.
Adapted from Table 8 of White *et al*. (1979) with permission of the authors and publishers.

d.f. - 1. In the comparison of two proportions (2 x 2 X^2 or Mantel-Haenszel X^2 test) or in the assessment of a trend, the percentage points give a two-sided test. A one-sided test may be obtained by halving the P values. (Concepts of one- and two-sidedness do not apply to larger degrees of freedom, as these relate to tests of multiple comparisons.)

	p-value							
d.f	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	17.12	21.06	2368	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	26.04	30.81	33.92	36.78	40.29	42-80	48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	69.33	77.58	85.53	90.53	95.02	100.43	104.22	112.32
80	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
100	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

PUBLIC HEALTH:

PHM102 Basic Statistics for Public Health and Policy

The formulae in the table below have been produced using MS Equation Editor. An accessible Word version is included within the downloaded zip file from which you can copy and paste any formulae into your answer document, editing them as needed in order for you to show your working.

Context	Standard Error	95% CI	z, t, or chi-squared
Single mean (large sample)	$SE(\bar{x}) = \frac{s}{\sqrt{n}}$	$\bar{x} \pm 1.96 SE(\bar{x})$	$z = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$
Single mean (small sample)	As above	$\bar{x} \pm t_{0.05} SE(\bar{x})$ (1)	$t = \frac{\bar{x} - \mu_0}{SE(\bar{x})}$ (1)
Single proportion (Different standard error formulae for test and confidence interval)	$SE(p) = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$		$z = \frac{p - \pi_0}{SE(p)}$
	$SE(p) = \sqrt{\frac{p(1-p)}{n}}$	$p \pm 1.96 SE(p)$	
Comparing two means (large sample)	$SE(diff) = \sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$	$(\bar{x}_1 - \bar{x}_2) \pm 1.96 SE(diff)$	$z = \frac{\bar{x}_1 - \bar{x}_2}{SE(diff)}$
Comparing two means (small sample)	$SE(diff) = s_p \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$ (2)	$(\bar{x}_1 - \bar{x}_2) \pm t_{0.05} SE(diff)$ (3)	$t = \frac{\bar{x}_1 - \bar{x}_2}{SE(diff)}$ (3)
Comparing two proportions (Different standard error formulae for test and confidence interval)	$SE(diff) = \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}$ Where $\bar{p}$ = overall proportion		$z = \frac{p_1 - p_2}{SE(diff)}$
	$SE(diff) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$(p_1 - p_2) \pm 1.96 SE(diff)$	
Testing association in an r X c table	-	-	$X^2 = \sum \frac{(O-E)^2}{E}$ (4)

Notation

$\bar{x}$ is a sample mean; p is a sample proportion; n is the sample size

$\bar{x}_i$, p_i, and n_i are sample means, proportions, and sizes in group i.

μ_0 = A known reference mean; π_0 = A known reference proportion;

Notes

(1) Student t-distribution with degrees of freedom = n − 1.

(2) $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$ (pooled estimate of common standard deviation)

(3) Degrees of freedom = $n_1 + n_2 - 2$.

(4) O= observed cell count; E=expected count if null hypothesis were true; degrees of freedom = (r−1) X(c−1), where r=n. of rows, c=n. of columns.

(5) If proportions are expressed as percentages, then replace '1' with 100' in above formulae.

Ref BA ss:2005.06.25, revised 2011-09-22

PHM102 Exam report for students 2019-20 (June 2020 exam)

General comments:

Due to Covid-19 students did not attend exam centres to sit their exam. Students downloaded the exam paper at home between 1st June and 22nd June and uploaded their answers within 48 hours. The format of the exam differed from previous years, with students answering all three questions, rather answering three questions from a choice of four.

Each of the three questions were answered to a very high standard. Calculations were generally well done. Some question parts were particularly challenging, and attempted responses varied in quality.

The following sub-questions scored lowest. These sub-questions tended to require interpretation rather than calculation and were expected to be challenging. This has also been observed in previous years:

1b) iv) – a good answer here would include a comment on the relationship between the confidence interval and the strength of the statistical evidence. Students tended to either interpret the confidence interval in its own right (which is not what was required here), and/or incorrectly assess the strength of the statistical evidence.

1di) – many students correctly noted an overlap in confidence intervals and answered the question in these terms, but a better way to answer this question was to note that the mean in one group lay within the confidence interval of the other group.

2a) – some students incorrectly identified BMI as a categorical variable because certain measurements of BMI are classified as underweight/normal/obese – but the measurement itself is numerical/quantitative and continuous.

2c) – a good answer here required demonstration of calculation of the area under the curve. An observation of the shape of the distribution was fine, but more detail was needed for higher marks.

2h) - most students were able to correctly interpret the confidence interval for a regression coefficient but, as in 1biv), many did not correctly comment (or comment at all) on the relationship between the confidence interval and the strength of statistical evidence, which is what was required for higher marks.

3c) - this was a conceptual question on the validity of the chi-squared test (see CAL session BS09 p35). To get good marks here students needed to have noticed that the data did not meet two of the validity criteria – many students only identified one criterion, or incorrectly concluded that the chi-squared test was valid based on sample size.

3gi) To get full marks here an answer needed to state both that the standard normal distribution squared is chi-squared AND with 1df. Students commonly omitted the 1df.

Percentage/0-5 grade conversion -0-24 = 0, 25-49 = 1, 50-59 = 2, 60-74 = 3, 75-89 = 4, 90+ = 5

Question 1

(a) **(10%)**

The high exposure group has the greater variation **(5%)** – because of higher standard deviation of 9.59cm versus 7.58cm **(5%).**

(b)

i. **(5%)**

159.1 – 155.4 = 3.7cm (**5%**. Award full marks for 155.4 – 159.1 = -3.7cm. Deduct 2% if units are omitted).

ii. **(15%)**

$$SE(diff) = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$SE(diff) = \sqrt{\left(\frac{7.58^2}{327} + \frac{9.59^2}{51}\right)}$$

$$SE(diff) = \sqrt{1.979004}$$

$SE(diff) = 1.406771 = 1.41cm$ **(15%)**

(Award 7% if final answer incorrect but numerical substitutions correctly made in formula and squares and square root correctly placed. Deduct 2% for omitting units).

iii. **(10%)**

95% CI = $(\bar{x}_1 - \bar{x}_2) \pm 1.96SE(diff)$

= 3.7 $\pm$ $1.96(1.406771)$

= 3.7 $\pm$ 2.757271

= 0.94, 6.46cm **(10%)**

Question 1 continued

Award full marks if student has calculated 1 b) i) as -3.7cm and has calculated a 95% CI of -3.7$\pm$ 2.757271= -6.46, -0.94cm. Deduct 2% if units are omitted. Allow for reasonable rounding in calculations.

iv. **(10%)**

The difference between mean heights of girls in the zero-low and high exposure groups is statistically significant at the 0.05 level **(2.5%)**. There is at least quite strong evidence against the null hypothesis that the heights of the girls in both maternal consumption groups are equal **(2.5%** - other similar wording referring to strength of evidence allowed**)** because the confidence interval does not include the null value, zero **(5%)**.

(c) (**30%)**

Null hypothesis: in the population from which this sample is drawn, there is no difference in the mean heights of girls whose mothers consumed zero-low amounts of glycyrrhizin during pregnancy and girls whose mothers consumed high amounts of glycyrrhizin in pregnancy (**10% -** allow any similar wording. Deduct 5% if the student does not refer to a 'true' or 'population' difference).

$$z = \frac{\bar{x}_1 - \bar{x}_2}{SE(diff)}$$

$$z = \frac{3.7}{1.406771}$$

$z \ = 2.63$ **(5%)**

$p = 0.00427\ (one\ sided)$

$p = 0.00427 \times 2 = 0.00854 = 0.008\ (two\ sided)$ (**5%**, Deduct 3% if one sided value only is presented or if student states p<0.05).

There is strong evidence against the null hypothesis of no true difference in mean height in the population from which these samples are drawn (**10%**. Allow any similar wording. Deduct 5% if assessment of the strength of evidence is missing or incorrect).

Question 1 continued

(d)

i. **(10%)**

The mean from the rural study is not significantly different (at $p<0.05$) from the mean from the urban study **(5%)** because the 95% CI for the rural study includes the mean difference in heights found in the urban study (3.7cm) **(5%)**.

ii. **(10%)**

The second study sample must have been smaller **(5%)** because the CI for the difference in mean heights is wider in the rural sample than it is in the urban sample **(5%)**.

Note: students are told in the question that they can assume standard deviations are similar, so a discussion of the effect of SD on width of CI is not needed here for full marks.

Question 2

(a) **(5%)**

Quantitative or numerical **(2.5%)** continuous **(2.5%)** variable.

(b) **(10%)**

The observed BMI values range approximately from 15.0 kg/m^2 **(5%)** to 42.0 kg/m^2 **(5%)**. Deduct 2.5% for incorrect answers within 1kg/m^2 of these values. Note that some students will miss the histogram bar corresponding to the outlier observations greater than 40 kg/m^2.

(c) **(15%)**

The percentage of men who are considered underweight is less than the proportion of men who are considered overweight or obese **(5%)**.

The cumulative percentage of men with BMI values below 18.0 kg/m^2 is approximately 19%, obtained by summing the heights of the first three bars of the histogram (2 + 6 + 11) **(5%)**; while the cumulative percentage of men with BMI values of 23.0 kg/m^2 or above is 30% (to the nearest whole number), obtained by summing the heights of the histogram bars for 23.0 kg/m^2 and higher **(5%)**. Accept any other valid justifications and calculations within 1% of the stated values for full marks. Award half of available marks if calculations are >1% of the stated values. Note that the precise cumulative percentage of men with BMI values below 18.0 kg/m^2 is 19.0%, and 29.9% for the cumulative percentage of men with BMI values of 23.0 kg/m^2 or above.

(d) **(15%)**

"The regression analysis shows that ***BMI*** (2.5%) ***increased*** (2.5%) by ***0.07*** (2.5%) ***kg/m²*** (2.5%) for each ***point*** (2.5%) increase in ***AHEI score*** (2.5%)."

(e) **(15%)**

It is unlikely that the association between the two variables has arisen by chance **(5%)**, because the p-value for the regression coefficient [p=0.003] provides strong evidence against the null hypotheses of no linear association OR because the confidence interval for the regression coefficient [0.02-0.11] does not include the null value of 0 (**10%** for either justification).

Question 2 continued

(f) **(15%)**

bmi = 18.75784 + 0.0666926 x 40 (**10%** for correct equation)
= 21.425544
= 21.43 kg/m^2 (**5%**, deduct 2% for omitting units)

(g) **(10%)**

The R-squared value in the regression output indicates that 5.78% or a proportion of 0.0578 of the variation **(5%)** in BMI is explained by the linear regression on AHEI score **(5%) OR** 5.78% or a proportion of 0.0578 of variability in the data **(5%)** is explained by the linear association between BMI and AHEI score **(5%)**. Award full marks for any correct variation in wording. Deduct 2% if ‘linear’ not mentioned in answer.

(h) **(15%)**

After adjusting for physical activity, caloric intake and age, the magnitude/ slope of the association between BMI and AHEI score was reduced from 0.067 to 0.035 **(5%)** but still remained positive. After adjusting, there is only weak evidence against the null hypothesis of no association **(5%)**; p must be >0.05 because the 95% confidence interval for the adjusted association includes the null value of 0 **(5%)**.

Note that up to 5% may be awarded (if the maximum of 15% has not been achieved) for answers that discuss confounding, i.e. that physical activity, caloric intake and/or age appear to have confounded the crude or unadjusted association between BMI and AHEI score (3%), and a further 2% for any correct justification for this assessment of confounding.

Question 3

(a) (**2%)**

212/251 = 0.845 = 84.5% **(2%)**

(b)

$$E = \frac{row\ total\ \times\ column\ total}{overall\ total}$$

i. **(2.5%)**

$E = \frac{212 \times 2}{251}$ = 1.69 **(2.5%)**

ii. **(2.5%)**

$E = \frac{39 \times 2}{251}$ = 0.31 **(2.5%)**

(c) **(10%)**

No, this test is not valid **(2%)**. Because it requires most cells to have **expected** values > 5 and from b) we know at least 2/10, i.e. at least 20%, are less than 5 and at least one cell has expected value <1 **(8%)**.

(d)

i. **(5%)**

1 + 10 + 76 = 87 **(5%)**

ii. **(5%)**

$\frac{(87-77.7052)^2}{77.7052}$ = 1.1118 **(5%)**

Question 3 continued

iii. **(5%)**

1 + 3 + 1 = 5 **(5%)**

iv. **(5%)**

$\frac{(5-14.2948)^2}{14.2948}$ = 6.0437 **(5%)**

v. **(10%)**

$$X^2 = \sum \frac{(O-E)^2}{E}$$

= 1.1118 + 0.0850 + 2.9337 + 6.0437 + = 0.4622 + 15.9468

= 26.5832 = 26.58 **(10%)**

(e)

i. **(5%)**

Degrees of freedom = (r − 1) × (c − 1)

= (2 − 1) × (3 − 1) = 2 **(5%)**

ii. **(5%)**

$p<0.001$ from chi-squared table. **(5%)**

iii. **(5%)**

There is very strong evidence against the null hypothesis that in the population from which this sample of patients was drawn, Rankin score on admission and survival to discharge are independent. **(5%)**

Question 3 continued

(f)

i. **(5%)**

There is no true/population linear trend in death in hospital across the ordered Rankin score categories [deduct 2% if true/population or linear trend not stated] **(5%)**

ii. **(3%)**

Degrees of freedom = 1 (constant for a test for trend) **(3%)**

iii. **(10%)**

Using stated chi-squared value and 1df p<0.001. **(10%)**

iv. **(5%)**

There is very strong evidence against the null hypothesis of no true/population linear trend in death across Rankin score categories. Deduct 2% if 'true/population' or 'linear trend' is not stated. **(5%)**

(g)

i. **(10%)**

The standard normal distribution squared is chi-squared with 1df. Therefore, you would take the square root of the test statistics and compare with standard normal tables, with 2-sided alternative **(10%)**.

ii. **(5%)**

$\sqrt{12.516} = 3.54$, associated with a p-value of 0.0002 x 2 = 0.0004 **(2.5%)**

There is very strong evidence against the null hypothesis of no true linear trend in death across ordered Rankin score categories . **(2.5%)**

End of exam report