Practical 4
Inferential Statistic 3

Introduction

In this practical session, we will be doing exercises on chi-square test and non-parametric analysis.

Chi-Square Test (χ^2)

This is the most frequent statistical analysis that is tested for during examination. So make sure that you really understand it. This analysis is done to test for association between two qualitative variables.

Observed data would be sorted accordingly in a contingency table. Then the expected value table is calculated, using the rows and column totals, as illustrated below;

Observation Table

	+	-	
+	a	b	g
-	c	d	h
	e	f	n

Expected Value Table

	+	-	
+	eg/n	fg/n	g
-	eh/n	fh/n	h
	e	f	n

Chi-square is calculated by summing up (observed – expected)^2/expected for each cell.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

$$df = (r - 1)(c - 1)$$

Do the following exercise. Then compare with the answer for Q.4c from Practical 2.

1. The rate of SGA for mothers exposed to cigarette smoke ("passive smoker") is 89/156. The rate of SGA for mothers not exposed to cigarette smoke is 20/61.

Observation table

	SGA	Normal	
Passive Smoker	89	67	156
Non-Smoker	20	41	61
	109	108	217

a. Complete the table of expected values below;

	SGA	Normal	
Passive Smoker			156
Non-Smoker			61
	109	108	217

b. What is the null hypothesis?

c. What is the rate of SGA for passive smokers and non-smokers? Is there any difference?

d. Conduct the appropriate statistical test to prove your hypothesis. Discuss your findings.

Chi-Square Test with Yates's Continuity Correction

Using the chi-squared distribution to interpret Pearson's chi-squared statistic requires one to assume that the discrete probability of observed binomial frequencies in the table can be approximated by the continuous chi-squared distribution. This assumption is not quite correct, and introduces some error.

To reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2 × 2 contingency table.

The effect of Yates's correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5 and sample size larger than 40.

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Please do the following exercise.

2. Based on the earlier findings, a cohort study was conducted to identify whether underweight mothers can lead to small for gestational age (SGA). It is believed that underweight mothers were of higher risk of getting SGA babies. A total of 25 underweight mothers and 25 normal weight mothers from first trimester were followed up until delivery.

	SGA Babies	**Normal Babies**	**Total**
Underweight	8	17	25
Normal	1	24	25
Total	9	41	50

a. Complete the table of expected values below;

	SGA Babies	**Normal Babies**	**Total**
Underweight			25
Normal			25
Total	9	41	50

b. What is the null hypothesis?

c. What is the rate of SGA for underweight mothers and normal weight mothers? Is there any difference?

d. Conduct the appropriate statistical test to prove your hypothesis. Discuss your findings.

Fisher's Exact Test

Fisher's Exact Test is conducted to test the association between 2 dichotomous qualitative variables and has a small sample size; less than 20 or less than 40 and at least one of the expected values is less than 5. The formula is as follows;

	+	-	
+	a	b	g
-	c	d	h
	e	f	n

$$\text{probability } p = \frac{e!f!g!h!}{n!a!b!c!d!}$$

3. From the earlier SGA case-control study, 23 of the respondents had miscarriages in the past. By analysing this group of patients with poor obstetric history, is there an association between exposure to cigarette smoke and SGA? Based on the following contingency table, conduct the appropriate statistical test.

	SGA	Normal	
Passive Smoker	10	7	17
Non-Smoker	0	6	6
	10	13	23

a. What is the p value?

b. What conclusion can you make from the above results?

Wilcoxon Rank-Sum Test

This test is the non-parametric equivalent of the Student's t test, but conducted on not normally distributed data. It is used to test for the association between a qualitative dichotomous variable with a quantitative variable. The method is simple, just sort the data in an ascending order, rank them, sum up the ranks according to groups and compare the value with the table of critical values for Wilcoxon Rank-Sum Test.

4. For practise, do the following exercise. The data is a subset of the earlier study. We are trying to see whether there is any association between exposure to cigarette smoke and the weight of the baby. Since the sample size is quite small, the appropriate test is a non-parametric analysis.

    a. What is the null hypothesis?

    b. Conduct the appropriate statistical test to prove your hypothesis. Discuss your findings.

Non-Smoker (n=15)		**Passive Smoker (n=15)**	
Birth Weight	**Rank**	**Birth Weight**	**Rank**
4.20		3.76	
3.96		3.60	
3.70		3.55	
3.61		3.48	
3.26		3.25	
3.15		3.06	
3.12		3.05	
3.00		2.55	
2.98		2.47	
2.84		2.46	
2.81		2.45	
2.57		2.45	
2.44		2.43	
2.43		2.30	
2.10		2.09	

Wilcoxon Signed-Rank Test

This test is the non-parametric equivalent of the paired t test, but conducted on data not normally distributed. It is conducted to test whether there is any association between 2 quantitative variables which are repeated measures on the same individual, of the same thing, at different times. As indicated by the name, the calculation depends on the sign and relative magnitude of the data, not of the real value of the data.

5. For practise, do the following exercise. The data is a subset of the earlier study (Hb<10 instead of Hb<11). We are trying to see whether the intervention of haematinics can increase the level of haemoglobin of the anaemic mothers. Since the sample is quite small, the appropriate test is a non-parametric analysis.

    a. What is the null hypothesis?

    b. Conduct the appropriate statistical test to prove your hypothesis. Discuss your findings.

	Hb1	Hb2	Hb Diff	Absolute Diff	Rank
1	9.8	9.0			
2	9.8	9.6			
3	9.3	9.5			
4	9.5	10.0			
5	9.5	10.0			
6	9.6	11.0			
7	9.7	12.0			

Research Project 4

Each lecturer will demonstrate how to analyse the data using the computer and advice the students on how to interpret the results. For homework, the students will complete the analysis and prepare a PowerPoint presentation for the final practical session.