BN001, BN009, BN012, BN117, BN121, BN903

Year 2

Hypothesis Testing – Worksheet 4

Question 1

In a study on the effects of smoking, a group of students counted the number of incidents of respiratory illness during each of the four seasons, among non-smokers (NS), light smokers (LS) and heavy smokers (HS). Determine whether there is a link between smoking and respiratory health across the four seasons. The results are shown in the following table.

	Winter	Spring	Summer	Autumn
NS	9	8	4	6
LS	28	22	6	7
HS	43	20	8	23

Solution

The test proceeds as follows:

Framing the Hypotheses: Strictly speaking, we are testing the independence of the level of smoking of the subjects and the seasons, as measured by the level of respiratory illnesses.

- The null hypothesis is that there is no link between the level of smoking of the subjects and the seasons, in other words they are independent. The alternative hypothesis is that there is a link, that they are not independent.
- We will take the level of significance as 0.05 (or 0.01).

- The number of degrees of freedom is (4 – 1)(3 – 1) = 6.
- Using the tables, the critical value for these parameters is 12.59 for 0.05 (or 16.81 for 0.01); this means

  $$P[\chi^2 > 12.59] = 0.05 \text{ (or } P[\chi^2 > 16.81] = 0.01).$$

- The values in the table above are the ***observed*** values. The ***expected*** values for a given cell are the two corresponding subtotals multiplied, and then divided by the overall total. So for example, HS, winter, the calculation is

  $$80 \text{ x } 94 / 184 = 40.870.$$

- The sample value is 7.830, not higher than the critical value for 0.05. The null hypothesis is therefore not rejected at 0.05 level of significance and we conclude there is no connection between the level of smoking of the subjects and the seasons, as measured by the number of illnesses. They are independent according to this data.

Question 2

A Study was carried out into what sports were played by residents of three environments; Dublin city, the Commuter Belt towns and Rural areas. A random sample of subjects were classified by where they live, and asked what sport they play on a regular basis. Determine whether there is a link between type of sport played and the urban/rural environment. The results are shown in the following table.

	GAA	Soccer	Rugby	Other
Dublin	39	31	23	12
Commuter	48	45	26	34
Rural	18	22	18	4

Solution

The test proceeds as follows:

Framing the Hypotheses: Strictly speaking, we are testing the independence of the type of environment the subjects live in and the sport they play. In this test, there is no parameter or number we have a value for, so our null and alternative hypothesis are just statements.

- The null hypothesis is that there is no link between the type of environment the subjects live in and the sport they play, in other words they are independent. The alternative hypothesis is that there is a link, that they are not independent.
- We will take the level of significance as 0.05 (or 0.01).
- The number of degrees of freedom is (4 – 1)(3 – 1) = 6.
- Using the tables, the critical value for these parameters is 12.59 for 0.05 (or 16.81 for 0.01); this means

  $$P[\chi^2 > 12.59] = 0.05 \text{ (or } P[\chi^2 > 16.81] = 0.01).$$

- The values in the table above are the ***observed*** values. The ***expected*** values for a given cell are the two corresponding subtotals multiplied, and then divided by the overall total. So for example, Commuter belt, soccer, the calculation is

  $$153 \text{ x } 98 / 320 = 46.856.$$

- The sample value is 13.468, higher than the critical value for 0.05, but lower than the critical value for 0.01 (16.81). The null hypothesis is therefore not rejected at 0.05 level of significance and we conclude there is a connection between the type of environment the subjects live in and the sport they play. They are not independent according to this data. Note that it is not rejected at 0.01.

Question 3

A trial was conducted on the effects of cigarette smoking and pollution levels on the occurrence of respiratory infections. Subjects were selected at random from 5 cities, listed in order of the recorded levels of pollution, 1 to 5, and classified by smoking as in the previous question. The table below records the number of incidents of respiratory infection in the subjects. Determine whether the data suggests the smokers are suffering more illnesses because of pollution.

	City 1	*City 2*	*City 3*	*City 4*	*City 5*
HS	131	241	252	222	303
OS	52	54	68	63	85
NS	23	18	25	29	34

Question 4

A company are trying to establish if there is a link between the type of fault reported in production and the production method. The number of faulty components produced is recorded during one week, counted by production method and type of fault, shown in the table shown. Determine whether or not the 4 faults arise roughly equally from the 3 production methods.

	Fault A	Fault B	Fault C	Fault D
Method 1	126	154	169	178
Method 2	152	151	157	153
Method 3	175	142	145	125

Solution

The test proceeds as follows:

Framing the Hypotheses: Strictly speaking, we are testing the independence of the type of fault and the production method. In this test, there is no parameter or number we have a value for, so our null and alternative hypothesis are just statements.

- The null hypothesis is that there is no link between the type of fault and the production method, in other words they are independent. The alternative hypothesis is that there is a link, that they are not independent.
- We will take the level of significance as 0.01 (or 0.05).
- The number of degrees of freedom is (4 – 1)(3 – 1) = 6.
- Using the tables, the critical value for these parameters is 16.81 (or 12.59 for 0.05); this means

$$P[\chi^2 > 16.81] = 0.01 \text{ (or } P[\chi^2 > 12.59] = 0.05).$$

- The values in the table above are the ***observed*** values. The ***expected*** values for a given cell are the two corresponding subtotals multiplied, and then divided by the overall total. So for example, fault C, method 3, the calculation is

471 x 587/1827 = 151.328.

- The value found is 18.277, higher than the critical value of 16.81. The null hypothesis is therefore rejected, and we conclude there is a connection between the type of fault and the method of production. They are not independent according to this data.