University of Edinburgh
School of Mathematics

Applied Statistics

Problem sheet 1 **Semester 2, 2020–2021**

Topics: χ^2 goodness-of-fit test, Independence and Homogeneity tests

Assessed question (duplicated on Gradescope):

1. The table below comes from one of the first studies of the link between lung cancer and smoking. In 20 hospitals in London, patients admitted with lung cancer in the preceding year were queried about their smoking behavior. For each of the 709 patients admitted, researchers studied the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. The 709 cases in the first column of table are those having lung cancer and the 709 controls in the second column are those not having it. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

	Lung cancer case	Control case
Smoker	688	650
Nonsmoker	21	59
Total	709	709

   (a) What is the name of the test to indicate whether there is an association between smoking and lung cancer? State the test hypotheses. [3 Marks]

   (b) Find the expected cell counts under H_0. [2 Marks]

   (c) For your hypotheses, state the associated test statistic and its corresponding approximate distribution, assuming that H_0 is true. Is the null hypothesis rejected? [2 Marks]

   (d) Apply the test using `R` and report the p-value. [2 Marks]

   (e) Only based on this table; can you claim that smoking causes lung cancer? [1 Mark]

Additional questions (not part of the assessment):

2. Show that Pearson's χ^2 test statistic, $X^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}$, can be written as,

$$X^2 = \sum_{k=1}^{K} \frac{O_k^2}{E_k} - n.$$

3. (From *simpleR - Using R for Introductory Statistics*, J. Verzani, 2002.) The letter distribution of the 5 most popular letters in the English language is known to be approximately as below (of course, the true distribution is for all 26 letters. This is simplified down to look just at these 5 letters),

letter	E	T	N	R	O
frequency	29	21	17	17	16

That is when either E,T,N,R,O appear, on average 29 times out of 100 it is an E and not the other 4. This information is useful in cryptography to break some basic secret codes. Suppose a text is analysed and the number of E,T,N,R and O's are counted. The following data is observed:

letter	E	T	N	R	O
observed freq	100	110	80	55	14

We want to test to see if the letter proportions for this text are following the first table, i.e., P(E)=0.29, P(T)=0.21, P(N)=0.17, P(R)=0.17, P(O)=0.16, or are different.

   (a) What is the name of the test that should be conducted here? State the test hypotheses.

   (b) Find the expected cell counts if H_0 is correct.

   (c) For your hypotheses, state the associated test statistic and its corresponding approximate distribution, assuming that H_0 is true. Is the null hypothesis rejected?

   (d) Apply the test using `R` and report the p-value.

4. The below table shows the number of individuals who survived/died when the Titanic sank in 1912 based on their status, corresponding to crew member, passenger booked in first, second and third-class staterooms.

	Crew	First	Second	Third	Total
Alive	212	202	118	178	710
Dead	673	123	167	528	1491
Total	885	325	285	706	2201

   (a) For a given individual chosen at random on the Titanic, calculate the probability that they survived.

(b) Calculate the probability that an individual chosen at random on the Titanic was a passenger.

(c) We would like to test whether survival is related to status on the ship. State the associated hypotheses.

(d) Assuming that the null hypothesis is true, calculate the expected values of the contingency table.

(e) The analyst conducts the following test in R:

```
> titanic <- matrix(c(212,673,202,123,118,167,178,528),ncol=4)
> chisq.test(titanic)
Pearson's Chi-squared test
data: titanic
X-squared = 187.79, df = 3, p-value < 2.2e-16
```

State, and justify, any conclusions that can be drawn.

5. The following data on the severity of a car crash is tabulated for the cases where the passenger had a seat belt; or did not:

	No injury	minimal	minor	major
Seatbelt	12813	647	359	42
No seatbelt	65963	4000	2642	303

We want to determine whether the two categorical variables are independent.

(a) What is the name of the test and state the test hypotheses.

(b) Apply the test using R. Report the value of the test statistic, its approximate distribution, and the p-value.

(c) Is the null hypothesis rejected?

(d) Using R find the expected cell counts if H_0 is correct.