Tutorial 10

Notation used in class	Notation used in the textbook
$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$ $i = 1, \ldots, r$	r_i $i = 1, \ldots, r$
$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$ $j = 1, \ldots, c$	c_j $j = 1, \ldots, c$

Question 1

Consider the problem of testing the *hypothesis of independence*:

$$H_0: \text{ There exist probabilities } p_{i\bullet} \text{ and } p_{\bullet j}, i = 1, \ldots, r, j = 1, \ldots, c, \text{such that}$$
$$p_{ij} = p_{i\bullet} p_{\bullet j} \quad \text{for all} \quad i = 1, \ldots, r, \quad j = 1, \ldots, c.$$

$$H_1: \text{ Not } H_0.$$

Here, p_{ij} represents the probability that an object or individual selected randomly from the population under study will belong to category i of argument 1 and category j of argument 2.

The data is represented by n_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, which counts the number of observations that fall in category i of argument 1 and category j of argument 2.

Note that the probabilities $p_{i\bullet}$ and $p_{\bullet j}$ must satisfy $\sum_{i=1}^{r} p_{i\bullet} = 1$ and $\sum_{j=1}^{c} p_{\bullet j} = 1$.

Consider estimating $p_{i\bullet}$ and $p_{\bullet j}$ under H_0. Show that the MLEs of $p_{i\bullet}$ and $p_{\bullet j}$ are given by:

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}, \quad i = 1, \ldots, r$$

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}, \quad j = 1, \ldots, c.$$

Question 2

(14.18) A study of the amount of violence viewed on television as it relates to the age of the viewer yielded the results shown in the accompanying table for 81 people. (Each person in the study was classified, according to the person's TV viewing habits, as a low-violence or high-violence viewer.) Do the data indicate that viewing of violence is not independent of age of viewer, at the 5% significance level?

	Age		
Viewing	16-34	35-54	55 and over
Low violence	8	12	21
High violence	18	15	7

Question 3

(14.26) A manufacturer of buttons wished to determine whether the fraction of defective buttons produced by three machines varied from machine to machine. Samples of 400 buttons were selected from each of the three machines, and the number of defectives were counted for each sample. The results are shown in the table below. Do these data present sufficient evidence to indicate that the fraction of defective buttons varied from machine to machine?

Machine number	Number of defectives
1	16
2	24
3	9

(a) R Test, using $\alpha = 0.05$, with a χ^2 test.

(b) R Test, using $\alpha = 0.05$, with a likelihood ratio test. (Refer to exercise 10.106, covered in Tutorial 5 Question 2.)

Question 4

(14.38) Counts on the number of items per cluster (or colony or group) must necessarily be greater than or equal to one. Thus, the Poisson distribution generally does not fit these kinds of counts. For modelling counts on phenomena such as number of bacteria per colony, number of people per household, and number of animals per litter, the *logarithmic series* distribution often proves useful. This discrete distribution has probability function given by

$$p(y \,|\, \theta) = -\frac{1}{\ln(1-\theta)} \cdot \frac{\theta^y}{y}, \quad y = 1, 2, 3, \ldots, \quad 0 < \theta < 1,$$

where θ is an unknown parameter.

(a) Show that the MLE $\hat{\theta}$ of θ satisfies the equation

$$\overline{Y} = \frac{\hat{\theta}}{-(1-\hat{\theta})\ln(1-\hat{\theta})}, \qquad \text{where } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

(b) R The data in the following table give frequencies of observation for counts on the number of bacteria per colony, for a certain type of soil bacteria.

Bacteria per colony	1	2	3	4	5	6	7+
Number of colonies observed	359	146	57	41	26	17	29

Test the hypothesis that these data fit a logarithmic series distribution. Use $\alpha = 0.05$. (Notice that the value $\overline{y}$ must be approximated because we do not have exact information on counts greater than six.)