Politecnico di Milano
STATISTICS (079086), 2009/2010, Prof. A.Barchielli
Problem set n. 10

11th June 2010

Exercise 1 The following table summarizes the distribution of the number of defects found on a random sample of 30 steel plates having equal dimension.

Number of defects	Observed absolute frequencies
0	1
1	4
2	6
3	7
4	5
5	3
6	2
7	1
8	1

(a) Verify, with a test of significance level 5%, the hypothesis that the sample is extracted from a population having Poisson distribution with parameter 2.

(b) Verify, with a test of significance level 5%, the hypothesis that the sample is extracted from a population having Poisson distribution.

Solution

(a) By computing the theoretical absolute frequencies $[30 \cdot P(X = k) = 30 \cdot e^{-\lambda} \frac{\lambda^k}{k!}$] we obtain:

classes	abs obs freq.	abs theoret freq.
0	1	4.06
1	4	8.12
2	6	8.12
3	7	5.41
4	5	2.70
5	3	1.08
6	2	0.36
7	1	0.10
8	1	0.03

We unify the classes having theoretical absolute frequency smaller then 5 and we obtain:

classes	abs obs freq.	abs theoret freq.
$0, 1$	5	12.18
2	6	8.12
$3, 4, 5, 6, 7, 8$	19	9.69

The test statistic is $Q = \sum_k \frac{(\text{abs obs freq. - abs theoret freq.})^2}{\text{abs theoret freq.}}$, and its value is:
$q = \frac{(7.18)^2}{12.18} + \frac{(2.12)^2}{8.12} + \frac{(9.31)^2}{9.69} = 13.73$.
Since, under the null hypothesis, $Q \sim \chi^2(k-1) = \chi^2(2)$ e $q > \chi^2_{0.05;2} = 5.99$, we reject the null hypothesis.

(b) First we estimate the parameter λ:

$$\hat{\lambda} = \frac{0 \cdot 1 + 1 \cdot 4 + 2 \cdot 6 + 3 \cdot 7 + 4 \cdot 5 + 5 \cdot 3 + 6 \cdot 2 + 7 \cdot 1 + 8 \cdot 1}{30} = 3.3$$

As in point (a) we have:

classes	abs obs freq.	abs theoret freq.
0	1	1.11
1	4	3.65
2	6	6.02
3	7	6.63
4	5	5.47
5	3	3.61
6	2	1.98
7	1	0.94
8	1	0.39

and, by unifying suitably the classes,

classes	abs obs freq.	abs theoret freq.
$0, 1, 2$	11	10.78
3	7	6.63
4	5	5.47
$5, 6, 7, 8$	7	6.91

We have $q = 0.066$. Since $Q \sim \chi^2(k-1-1) = \chi^2(2)$ and $q < \chi^2_{0.05;2} = 5.99$, we accept that our data follows a Poisson distribution with parameter (3.3).

Exercise 2 The cylindrical mechanical pieces produced by a production plant must satisfy the following requirements. The length L (in cm) of the piece must satisfy: $19.70 \leq L \leq 20.30$; the diameter D (in cm) of the piece must satisfy: $1.98 \leq D \leq 2.26$. We have checked 2000 pieces, obtaining the results reported in the following table:

	$D < 1.98$	$1.98 \leq D \leq 2.26$	$D > 2.26$	
$L < 19.70$	26	150	8	
$19.70 \leq L \leq 20.30$	124	1320	160	
$L > 20.30$	10	186	16	

1. Verify with a suitable test wether the length and the diameter of the mechanical pieces are independent.

2. Verify with a suitable test wether the length of the mechanical is normally distributed with mean 20 cm and variance $0.018\,\text{cm}^2$.

3. Construct a bilateral confidence interval, of approximate level 95%, for the proportion of pieces satisfying the requirement on length and diameter.

Solution.
We compute the marginal counts and use the following notation:

	$D < 1.98$	$1.98 \leq D \leq 2.26$	$D > 2.26$	
$L < 19.70$	26 (N_{11})	150 (N_{12})	8 (N_{13})	184 (N_{1L})
$19.70 \leq L \leq 20.30$	124 (N_{21})	1320 (N_{22})	160(N_{23})	1604 (N_{2L})
$L > 20.30$	10 (N_{31})	186 (N_{32})	16 (N_{33})	212 (N_{3L})
	160 (N_{1D})	1656 (N_{2D})	184 (N_{3D})	2000 (n)

1. We want to verify the null hypothesis H_0: "L, D are independent" versus the alternative H_1: "L, D are not independent". Since we have a large sample of grouped data, we can use an asymptotic chi-squared independence test. We thus compute the statistics

$$Q_1 = \sum_{i,j=1}^{3} \frac{\left(N_{ij} - \frac{N_{iL}N_{jD}}{n}\right)^2}{\frac{N_{iL}N_{jD}}{n}} = 2000\left(\frac{26^2}{160 \times 184} + \frac{150^2}{1656 \times 184} + \cdots + \frac{16^2}{184 \times 212}\right) - 2000 \simeq 18.74.$$

Notice that all counts are bigger than 5, so that the distribution of Q_1 can be approximated by a chi-squared distribution with $(3-1)(3-1) = 4$ degrees of freedom. Since $18.74 > \chi^2_{0.999,4}$, there is a strong empirical evidence against the null hypothesis.

2. We want to verify H_0 : "$L \sim \mathcal{N}(20, 0.018)$" versus the alternative H_1 : "$L \not\sim \mathcal{N}(20, 0.018)$". We can use an asymptotic goodness-of-fit test. We compute the following quantities:

	$L < 19.70$	$19.70 \leq L \leq 20.30$	$L > 20.30$
N_{iL}	184	1604	212
p_{0i}	0.0125	0.975	0.0125
np_{0i}	25	1950	25

where $p_{01} = P_{H_0}(L < 19.70) = \Phi\left(\frac{19.70-20}{\sqrt{0.018}}\right) = \Phi(-\sqrt(5)) \simeq 1 - \Phi(2.24) \simeq 0.0125$, $p_{03} = P_{H_0}(L > 20.30) = 1 - \Phi(\sqrt(5)) = p_{10}$ and $p_{02} = 1 - p_{01} - p_{02} = 0.975$. The statistics is given by

$$Q_2 = \sum_{i=1}^{3} \frac{(N_{iL} - np_{0i})^2}{np_{0i}} = \frac{184^2}{25} + \frac{1604^2}{1950} + \frac{212^2}{25} - 2000 \simeq 2471.39.$$

Notice that $np_{01} > 5$ for all $i = 1, 2, 3$, so that the distribution of Q_2 can be approximated by a chi-squared distribution with 2 degrees of freedom, i.e. an exponential distribution with mean 2. The p-value is thus given by $\mathrm{e}^{-1946.059/2} \simeq 0$: there is a very strong empirical evidence against H_0.

3. Let θ be the proportion of mechanical pieces satisfying the requirement of length and diameter. We have $\widehat{\theta} = 1320/2000 = 0.66$, and an asymptotic confidence interval of level 95% is given by

$$\widehat{\theta} \mp z_{0.975}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{2000}} \quad : \quad (0.64, 0.68).$$

Exercise 3** In a small town in the United States in the month of June last year there has been 35 black-out. The lengths (in minutes) of the black-out are registered in the following table:

Length of black-out	Number of black-out
less than 1 min	12
from 1 to 2 mins	13
from 2 to 3 mins	5
more than 3 mins	5

a) Verify, at significance level 1%, the null hypothesis that the length of the black-out is exponentially distributed (with mean 2).

b) Compute the p-value of the test.

The following table instead gives the daily number of black-out (during the 30 days of the month of June):

Number of black-out per day	Number of days
0	8
1	12
2	8
3	1
4 or more	1

c) Verify, at significance level 5%, the null hypothesis that the daily number of black-out has Poisson distribution.

d) Compute the p-value of the test.

Solution

a) The theoretical frequencies under H_0 are

Length of black-out	Theoretical frequencies n_i
$(0,1]$	$35 \times (1 - e^{-0.5}) = 35 \times 0.393 = 13.8$
$(1,2]$	$35 \times (e^{-0.5} - e^{-2\times 0.5}) = 35 \times 0.239 = 8.3$
$(2,3]$	$35 \times (e^{-2\times 0.5} - e^{-3\times 0.5}) = 35 \times 0.145 = 5.1$
$(3,\infty)$	$35 \times (1 - 0.393 - 0.239 - 0.135) = 35 \times 0.223 = 7.8$

This gives

$$Q = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{n_i} = 3.82$$

This value must be compared with $\chi^2_{0.01,3} = 11.34$: the null hypothesis is not rejected at level 1%.

b) The p-value falls in the interval (10%,50%). The data are thus compatible with the null hypothesis.

c) First, the parameter λ of the Poisson distribution must be estimated by the sample mean of the data, obtaining $\hat{\lambda} = 1.167$. The theoretical frequencies under H_0 are thus

Number of black-out per day	Theoretical frequencies n_i
0	$30 \times e^{-1.167} = 30 \times 0.311 = 9.34$
1	$30 \times e^{-1.167} 1.167 = 30 \times 0.363 = 10.9$
2	$30 \times e^{-1.167} \frac{1.167^2}{2} = 30 \times 0.212 = 6.36$
3	$30 \times e^{-1.167} \frac{1.167^3}{6} = 30 \times 0.082 = 2.47$
4 or more	$30 \times (1 - 0.311 - 0.363 - 0.212 - 0.082) = 30 \times 0.031 = 0.93$

The last three classes are joint together in order to have theoretical frequencies all greater than 5:

Number of black-out per day	Theoretical frequencies n_i	Observed frequencies N_i
0	$30 \times e^{-1.167} = 30 \times 0.311 = 9.34$	8
1	$30 \times e^{-1.167} 1.167 = 30 \times 0.363 = 10.9$	12
2 or more	$30 \times (1 - 0.311 - 0.363) = 30 \times 0.326 = 9.76$	11

This gives

$$Q = \sum_{i=1}^{k} \frac{(n_i - N_i)^2}{n_i} = 0.46$$

This value must be compared with $\chi^2_{0.05,1} = 3.84$: the null hypothesis is not rejected at level 5%.

d) The p-value of the test is about 50%. The data are thus compatible with the null hypothesis.

Exercise 4** A clinical study has been done on 200 patients suffering from diabetic retinopathy. An eye, randomly chosen between the right and left one, is treated, while the other one is observed without treatment. T_1 represents the time from an initial time 0 up to the blindness of the treated eye and T_2 up to the blindness of the non-treated eye. The times T_1, T_2 are both expressed in years and the collected data have been grouped and are reported in the following table:

$T_1 \setminus T_2$	$(0,6]$	$(6,7]$	$(7,\infty)$	
$(0,6]$	20	20	40	
$(6,7]$	20	20	10	
$(7,10]$	15	10	15	
$(10,\infty)$	5	10	15	

1. Verify with a suitable hypothesis test whether the exponential density of mean 5.8 fits the data of the time T_1.
2. Verify with a suitable hypothesis test of significance level $\alpha = 5\%$ whether the times T_1, T_2 are independent.

Solution.
We complete the table of the data with the marginals:

$T_1 \setminus T_2$	$(0,6]$	$(6,7]$	$(7,\infty)$	
$(0,6]$	20	20	40	$\boxed{80}$
$(6,7]$	20	20	10	$\boxed{50}$
$(7,10]$	15	10	15	$\boxed{40}$
$(10,\infty)$	5	10	15	$\boxed{30}$
	$\boxed{60}$	$\boxed{60}$	$\boxed{80}$	$\boxed{200}$

1. We use a goodness-of-fit χ^2 test for H_0: "T_1 has an exponential density with mean 5.7" versus the alternative hypothesis H_1 : "T_1 has not an exponential density with mean 5.7". Under H_0 the probabilities that T_1 belongs to each of the 4 classes are

$$\begin{aligned}
p_{01} &= P_{H_0}(T_1 \le 6) = 1 - \mathrm{e}^{-6/5.8} \simeq 0.6446,\\
p_{02} &= P_{H_0}(6 < T_1 \le 7) = \mathrm{e}^{-6/5.8} - \mathrm{e}^{-7/5.8} = 1 - p_{01} - p_{03} - p_{04} \simeq 0.0563,\\
p_{03} &= P_{H_0}(7 < T_1 \le 10) = \mathrm{e}^{-7/5.8} - \mathrm{e}^{-10/5.8} \simeq 0.1208,\\
p_{04} &= \mathrm{e}^{-10/5.8} \simeq 0.1783,
\end{aligned}$$

and the Pearson statistics takes the value

$$\begin{aligned}
Q_1 = \sum_{i=1}^{4} \frac{(N_{i\cdot} - np_{0i})^2}{np_{0i}} &= \sum_{i=1}^{4} \frac{N_{i\cdot}^2}{200p_{0i}} - 200\\
&= \frac{1}{200}\left(\frac{80^2}{0.6446} + \frac{50^2}{0.0563} + \frac{40^2}{0.1208} + \frac{30^2}{0.1783}\right) - 200 \simeq 163.13
\end{aligned}$$

(there is no need of of a further grouping of the data because $0.0563 \times 200 = 11.26 > 5$). The p-value of the test is $1 - F_{\chi^2_{4-1}}(163.13) \simeq 0$ (consider that $F_{\chi^2_3}(16.266) \simeq 0.999$). Therefore, there is a very strong empirical evidence against the null hypothesis of exponentially distributed data with the given mean.

2. We perform a χ^2 test of independence between T_1 and T_2. The test statistics is

$$Q_2 = 200 \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{N_{ij}^2}{N_{i\cdot} N_{\cdot j}} - 200 = 200 \left(\frac{20^2}{80 \times 60} + \frac{20^2}{80 \times 60} + \cdots + \frac{15^2}{30 \times 80} \right) - 200 \simeq 15.451$$

and, at level 5%, we reject the hypothesis of independence if $Q_2 > \chi^2_{(4-1)(3-1)}(95\%)$. But we have $\chi^2_6(95\%) \simeq 12.592$ and $15.451 > 12.592$, and we reject the hypothesis of independence between T_1 and T_2.