CHI-SQUARE TEST – SOLUTIONS

1. You believe that people who die from overdoses of narcotics die rather young. To test this theory you have obtained the following distribution of # of deaths from overdoses:

Age interval	15–19	20–24	25–29	30–34	35–39	40-44	45–49
Number of deaths	40	35	32	10	13	13	4

Total: 147.

An appropriate H0 hypothesis would be that equal numbers die in all seven age groups (i.e. $147/7 = 21$). Perform the Pearson's test to check whether H0 cannot be rejected.

We calculate the usual chi-square statistic. $\chi^2 = \sum_{i=1}^{3} \frac{(O_i - T_i)^2}{T_i}$, where O_i and T_i are observed and theoretical numbers, respectively. This value is 57.90, much greater than $\chi^2_{0.95;6} \approx 12.6$. The H0 hypothesis should be rejected.

2. A sample, of the size equal to 200, has been taken from a population whose property follows an unknown distribution. The 200 results (frequencies) have been grouped into 10 classes of equal width (0.5) — they are given in the two first columns of the table below. We form a conjecture: the distribution is uniform over the interval [45,50]. Verify this hypothesis (with $\alpha = 0.05$).

class midpoint	n_i exp	$n \cdot \pi_i$ theory	$(n_i - n \cdot \pi_i)^2$	$(n_i - n\pi_i)^2/n\pi$
45.25	23	20		
45.75	19	20		
46.25	25	20		
46.75	18	20		
47.25	17	20		
47.75	24	20		
48.25	16	20		
48.75	22	20		
49.25	20	20		
49.75	16	20		
$\sum n_i = 200$; $\sum n\pi_i = 200$				

Hint. The Pearson test. Uniform distribution means that all the theoretical frequencies must be equal each to other (the 200 values are evenly distributed over 10 intervals). Fill the remaining columns of the table, calculate the χ^2 value. The hypothesis cannot be rejected.

3. Let the result of a random experiment be classified by two attributes – eye color and hair color. One of the attributes, eye color, can be divided into mutually exclusive and exhaustive (filling the whole event space) events:

$X1$ – blue eyes; $X2$ – brown eyes; $X3$ – grey eyes; $X4$ – black eyes; $X5$ – green eyes

The other attribute can be also divided into four mutually exclusive and exhaustive events:

$Y1$ – black hair; $Y2$ – brown hair; $Y3$ – black eyes; $Y4$ – red hair.

The experiment is performed by observing $n = 500$ people and each of them are categorized according to eye color and hair color. Let $Xi \bigcap Yj$ be the event that a person with eye color Xi; $i = 1, 2, 3, 4, 5$ and hair color Yj; $j = 1, 2, 3, 4$. Let n_{ij} be the observed frequency of event $Xi \bigcap Yj$ and $p_{ij} = n_{ij}/N$ – its probability, where N is the total number of events.

The situation (outcome of the experiment) looks like this

EYES↓ 5 classes	HAIR 1	4 classes → 2	3	4	
1	50	87	5	8	$\sum = n_{1\cdot} = 150$
2	40	69	60	11	$\sum = n_{2\cdot} = 180$
3	15	13	42	5	$\sum = n_{3\cdot} = 75$
4	5	27	17	1	$\sum = n_{4\cdot} = 50$
5	15	4	1	25	$\sum = n_{5\cdot} = 45$
	$\sum = n_{\cdot 1}$ $= 125$	$\sum = n_{\cdot 2}$ $= 200$	$\sum = n_{\cdot 3}$ $= 125$	$\sum = n_{\cdot 4}$ $= 50$	$= \boldsymbol{N}$ $= 500$

Test the hypothesis that Xi and Yj are independent events. We visualise the situation with the aid of the following table(cf. lecture)

$$\boxed{p_{ik} = \mathcal{P}(X \in < class >_i; Y \in < class >_k)}$$

X↓ EYES	Y 1	HAIR → 2	3	4	
1	n_{11}	n_{12}	n_{13}	n_{14}	$\sum = n_{1\cdot}$
2	n_{21}	n_{22}	n_{23}	n_{24}	$\sum = n_{2\cdot}$
$\vdots$	$\dots$	$\dots$	n_{ik}	$\dots$	$\dots$
5	n_{51}	n_{52}	n_{53}	n_{54}	$\sum = n_{5\cdot}$
	$\sum = n_{\cdot 1}$	$\sum = n_{\cdot 2}$	$\dots$	$\sum = n_{\cdot c}$	$= \boldsymbol{n}$

(Summing the cell frequencies across the rows gives the marginal row frequencies $n_{i\cdot}$, and summing the cell frequencies down the columns gives the marginal column frequencies $n_{\cdot k}$.) The $X - Y$ independence hypothesis is consistent with the statement: $\boxed{p_{ik} = p_{i\cdot} \times p_{\cdot k} \quad \text{or} \quad n_{ik} = n_i.n_{\cdot k}/n}$. On the other hand, we have :

$$p_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad p_{\cdot k} = \frac{n_{\cdot k}}{n}$$

Consequently, the χ^2 statistic is:

$$\chi^2 = n \sum_{i=1}^{5} \sum_{k=4}^{c} \frac{(n_{ik} - n_i.n_{\cdot k}/n)^2}{n_i.n_{\cdot k}}. \tag{1}$$

From the date in the (first) table we have:
$p_{1\cdot} = \frac{150}{500} = 0.3$ $\quad p_{2\cdot} = \frac{180}{500} = 0.36$ $\quad p_{3\cdot} = \frac{75}{500} = 0.15$ $\quad p_{4\cdot} = \frac{50}{500} = 0.1$ $\quad p_{5\cdot} = \frac{45}{500} = 0.09$
$p_{\cdot 1} = \frac{125}{500} = 0.25$ $\quad p_{\cdot 2} = \frac{200}{500} = 0.4$ $\quad p_{\cdot 3} = \frac{125}{500} = 0.25$ $\quad p_{\cdot 4} = \frac{50}{500} = 0.1$
we calculate the $n_{i\cdot}$, $n_{\cdot k}$, substitute into (1) – the value of chi-square is approximately 220. It's much higher then the limiting value $\chi^2_{0.95;12} = 21.26$. We must reject the $H0$ hypothesis about independence of X and Y. By the way – the # of degrees of freedom is $(5-1) \times (4-1) = 12$ (cf. lecture).

4. Often frequency data are tabulated according to two criteria, with a view toward testing whether the criteria are associated. Consider the following analysis of the 157 machine breakdowns during a given period.

	MACHINE A	B	C	D	Total per shift
Shift 1	10	6	13	13	41
Shift 2	10	12	19	21	62
Shift 3	13	10	13	18	54
Total per machine	33	28	44	52	157

We are interested in whether the same percentage of breakdown occurs on each machine during each shift or whether there is some difference due perhaps to untrained operators and/or other factors peculiar to a given shift.

Solution: the above formula gives the *observed* numbers of breakdowns – O_{ij}; where $x = 1, 2, 3$ and $y = 1, 2, 3\,4$.

For independent shifts/machines we shoul have $\boxed{p_{ik} = p_{i\cdot} \times p_{\cdot k} \quad \text{or} \quad T_{ik} = n_{i\cdot}n_{\cdot k}/n}$, where T_{ik} are theoretical numbers of cases in each of the table cells:

	A	B	C	D
Shift 1	8.62	7.3	11.5	13.57
Shift 2	13.03	11.06	17.38	20.54
Shift 3	11.35	9.63	15.13	17.88

We calculate the usual chi-square statistic.
$\chi^2 = \sum_{i=1}^{3} \sum_{k=4}^{4} \frac{(O_{ik} - T_{ik})^2}{T_{ik}}$.
It's value is 2.17 The # of degrees of freedom is $(4-1) \times (3-1) = 6$ (cf. lecture).$\chi^2_{0.95;6} = 12.6$. We must NOT reject the $H0$ hypothesis about independence of machine and shift in determining incidence of breakdown.