NON-PARAMETRIC STATISTICAL TESTS
TESTS OF INDEPENDENCE
using the Pearson's test

The problem:

We consider a 2-D RV (X, Y) of the discrete type (or categorical type) and we want to test the hypothesis:
are the two variables independent of each other?
Suppose: X has been divided (classified) into r intervals (classes) and Y has been divided (classified) into c intervals (classes)
We form a sample consisting on n $X - Y$ pairs; n_{ik} — the number (frequency) of sample elements with X belonging to the i-th class and Y belonging to the k-th class. Let's denote the marginal frequencies:

$$n_{i\cdot} = \sum_{k=1}^{c} n_{ik} \quad n_{\cdot k} = \sum_{i=1}^{r} n_{ik} \quad n = \sum_{k=1}^{c}\sum_{i=1}^{r} n_{ik}$$

In a similar way we may introduce ,,straight" and ,,marginal" probabilities:

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

$$p_{i\cdot} = \mathcal{P}(X \in < class >_i; \text{any } Y) \quad p_{\cdot k} = \mathcal{P}(\text{any } X; Y \in < class >_k)$$

$$\sum_{i}^{r} p_{i\cdot} = \sum_{k}^{c} p_{\cdot k} = \sum_{i,k} p_{ik} = 1$$

The problem, cntd.

We may visualise the situation with the aid of the following table (*contingency table, two-way table*):

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

X↓	Y	c classes →			
r classes	1	2	$\ldots$	c	
1	n_{11}	n_{12}	$\ldots$	n_{1c}	$\sum = n_{1.}$
2	n_{21}	n_{22}	$\ldots$	n_{2c}	$\sum = n_{2.}$
$\vdots$	$\ldots$	$\ldots$	n_{ik}	$\ldots$	$\ldots$
r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	$\sum = n_{r.}$
	$\sum = n_{.1}$	$\sum = n_{.2}$	$\ldots$	$\sum = n_{.c}$	$= \boldsymbol{n}$

(Summing the cell frequencies across the rows gives the marginal row frequencies $n_{i.}$, and summing the cell frequencies down the columns gives the marginal column frequencies $n_{.k}$.)

The problem, cntd.

The $X - Y$ independence hypothesis is consistent with the statement: $\boxed{p_{ik} = p_{i\cdot} \times p_{\cdot k}}$. On the other hand, we have (it's not hard to show):

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

Consequently, the χ^2 statistic is:

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

What about the number of DoF? From the data we have to estimate $r - 1 + c - 1 = r + c - 2$ parameters (r $p_{i\cdot}$ and c $p_{\cdot k}$ – but they are linked by two normalisation identities: $\sum p = 1$). Thus the number of DoF is: the number of inedpendent data – the number of estimated parameters. We have:

$$\text{No of DoF } = rc - 1 - (r + c - 2) = (r - 1)(c - 1).$$

Note: the number of *independent* data is $n - 1$ as n probabilities (class frequencies) p_{ik} are again normalised: $\sum_{i,k} p_{ik} = 1$.