**Note**: If r = c = 2 we have a  $2 \times 2$  table, and another **exact** test can be used which does not rely on the large sample approximation

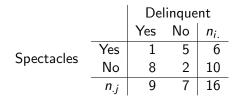
#### Fisher's Exact Test

- another test for independence of assignment of the row and column factor levels
- test statistic and null distribution are complicated (based on the hypergeometric distribution)
- SPSS computes test statistic and *p*-value.

**Example:** Juvenile Delinquency and Spectacle Wearing.

Is there any association between the two factors ?

- A : Spectacle Wearing (Yes/No)
- *B* : Juvenile Delinquent (Yes/No)



# **Example:** Juvenile Delinquency and Spectacle Wearing.

**Chi-squared Test:** 

$$X^2 = 6.112$$

Compare with Chi-squared((r-1)(c-1)) = Chi-squared(1); we have

 $Chi-squared_{0.05}(1) = 3.841$ 

and a *p*-value of 0.013. Therefore we reject  $H_0$ .

**Fisher's Exact Test:** *p*-value is 0.035 (1-sided) or 0.024 (2-sided).

Thus we reject  $H_0$  and we have evidence of association between the factors.

## **Case-Control Studies**

Nonparametric Statistics Categorical Data Single Population Tests

A **case-control** study is an observational study where participants are selected for the study with regard to their **disease status**.

- a sample of cases (disease sufferers)
- a sample of controls (healthy patients)

We investigate the possible association between disease status and a factor that takes two levels. A  $2 \times 2$  table of counts is formed for all combinations of disease status/factor level.

### Example: BCG Vaccination and Leprosy.

Disease Status : Leprosy Sufferer (Yes/No) Factor : Vaccination Scar (Yes/No)

		Disease Status					
		Case	Control				
		Yes	No	n <sub>i.</sub>			
Scar	Yes	101	554	655			
	No	159	446	605			
	n <sub>.j</sub>	260	1000	1260			

Is there an association ? Does vaccination induce leprosy ?

The Chi-squared test is potentially not valid here because of the design. An alternative test statistic is based on the **odds ratio** 

$$\mathsf{O.R.} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \widehat{\psi}$$

say. The test statistic is

$$Z = \frac{\log \widehat{\psi}}{\text{s.e.}(\log \widehat{\psi})}$$

where

s.e.
$$(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Т

$$Z = \frac{\log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

#### Under

 $\ensuremath{\textit{H}_{0}}\xspace$  : No association between factor and disease status

it follows that

$$Z \sim N(0,1)$$

Here log means In or natural log.

#### Example: BCG Vaccination and Leprosy.

Nonparametric Statistics Categorical Data Single Population Tests

$$n_{11} = 101, n_{12} = 554, n_{21} = 159, n_{22} = 446$$

Therefore

$$\widehat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 0.511 \qquad \log \widehat{\psi} = -0.671$$

and

s.e.
$$(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = 0.142$$

so

$$Z = \frac{-0.671}{0.142} = -4.717$$

For a text at  $\alpha =$  0.05, the two-sided critical value is  $\pm 1.96,$  so we

**Reject** 
$$H_0$$
.

#### Example: Smoking and Lung Cancer.

Nonparametric Statistics Categorical Data Single Population Tests

$$n_{11} = 647, n_{12} = 622, n_{21} = 2, n_{22} = 27$$

Therefore

$$\log \widehat{\psi} = \log \frac{647 \times 27}{2 \times 622} = 2.642$$

and

s.e.
$$(\log \widehat{\psi}) = \sqrt{\frac{1}{647} + \frac{1}{2} + \frac{1}{622} + \frac{1}{27}} = 0.735$$

so

$$Z = \frac{2.642}{0.735} = 3.590$$

For a text at  $\alpha =$  0.05, the two-sided critical value is  $\pm 1.96,$  so we

#### **Reject** $H_0$

and report evidence for association.

# Single Population Tests

Nonparametric Statistics Categorical Data Single Population Tests

We seek non-parametric or distribution-free tests for hypotheses relating to single samples, the equivalents of one-sample Z- or T-tests, which rely on the **normality** of the samples.

Normally these tests are formulated in terms of  $\ensuremath{\textit{ranks}}$  of the data to give

#### **Rank Tests**

For example, if the data are

 $0.24 \quad 3.16 \quad 1.97 \quad 2.10 \quad 0.92$ 

we sort them into ascending order, and assign ranks in order

	0.24	0.92	1.97	2.10	3.16
Rank	1	2	3	4	5

The tests depend on the behaviour of statistics computed in terms of the ranks, and rely on a **large sample** justification.