Reference: Moore, David S., and George P. McCabe (1989). Introduction to the Practice of Statistics. Original source: Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970-1972, Her Majesty's Stationery Office, London, 1978.

Description: Data summarizes a study of men in 25 occupational groups in England. Two indices are presented for each occupational group. The smoking index is the ratio of the average number of cigarettes smoked per day by men in the particular occupational group to the average number of cigarettes smoked per day by all men. The mortality index is the ratio of the rate of deaths from lung cancer among men in the particular occupational group to the rate of deaths from lung cancer among all men.

Number of cases: 25

Variable Names:
Smoking: Smoking index (100 = average)
Mortality: Lung cancer mortality index (100 = average)

SYSTAT file name: smoking_mortality.syz

Operations we will do:
1) standardize to z-scores (from the data menu: standardize)
2) scatterplot the z-scores; smooth (smoother) and influence plot (options) (from Graph/scatterplot)
3) calculate the Pearson correlation, gamma, Spearman's rank-order correlation, and Guttman's measure of monotonicity (from Analyze/Correlations/Simple)

5,387 school children from Caithness, Scotland (data from R. A. Fisher in the 1930s).

	Hair Color					
	Fair	Red	Medium	Dark	Black	Totals
Eye Color						
Light	688	116	584	188	4	1580
Blue	326	38	241	110	3	718
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Totals	1455	286	2137	1391	118	5387

In predicting column category from knowing the row category, errors in prediction are as follows:

Row:
Light: (1580 - 688)/1580 = 892/1580 = .56;
Blue: (718 - 326)/718 = 392/718 = .55;
Medium: (1774 - 909)/1774 = 865/1774 = .49;
Dark: (1315 - 681)/1315 = 634/1315 = .48 .

Overall Error of Prediction:
(.56)(1580/5387) + (.55)(718/5387) + (.49)(1774/5387) + (.48)(1315/5387) =
(892 + 392 + 865 + 634)/5387 = 2783/5387 = .5166 = $P_{error|row}$

Error in prediction without being told the row category (and predicting "medium" hair color based on the largest column frequency):
P_{error} = (5387 - 2137)/5387 = 3250/5387 = .6033

$\lambda_{hair|eye}$ = (.6033 - .5166)/.6033 = .144

So, the proportional reduction in error in predicting hair color from eye color is 14.4% (the absolute reduction in error is 8.7%.

Rows (eye color) is considered an independent variable used to predict the dependent variable of columns (hair color).

We note the the usual Pearson Chi-square statistic has a value of 1,240.0 on 12 degrees of freedom. This indicates a "significant" dependence but because the statistic increases directly in proportion to the sample size (which is very large here), the statistic tells us nothing about the strength of association.

Operations we do on the file eye_hair_fisher.syz

Use the command line:

```
USE G:\eye_hair_fisher
XTAB
FREQUENCY COUNT
PLENGTH LONG
TABULATE EYE$*HAIR$
```

(You will get everything you might want all at once from the two-way table) –