Modeling categorical relationships

NHH
TECH3

Categorical relationships

- Qualitative measurements
- Counts/frequencies
- Observed vs expected counts
- Univariate or multivariate
- Example: Website traffic across weekdays

Weekday	Observed visits
Monday	120
Tuesday	100
Wednesday	110
Thursday	130
Friday	140
Saturday	90
Sunday	80
Total	770

NHH
TECH3

Pearson's chi-squared test

$$\chi^2 = \sum_i \frac{(observed_i - expected_i)^2}{expected_i}$$

- H_0: All weekdays have equal website traffic
- H_A: All weekdays do not have equal website traffic
- Under H_0:

$$expected_i = \frac{Total\ traffic}{Number\ of\ days} = \frac{770}{7} = 110$$

NHH TECH3

Pearson's chi-squared test

Weekday	Observed visits	Expected visits	Squared deviations	χ^2
Monday	120	110	100	0.91
Tuesday	100	110	100	0.91
Wednesday	110	110	0	0
Thursday	130	110	400	3.64
Friday	140	110	900	8.18
Saturday	90	110	400	3.64
Sunday	80	110	900	8.18
Total	770	770	2800	25.5

NHH
TECH3

Pearson's chi-squared test

$$\chi^2 = \sum_i \frac{(observed_i - expected_i)^2}{expected_i} = 25.5$$

$$\chi^2 \sim \chi^2_k \qquad (k = 7 - 1)$$

P-value:

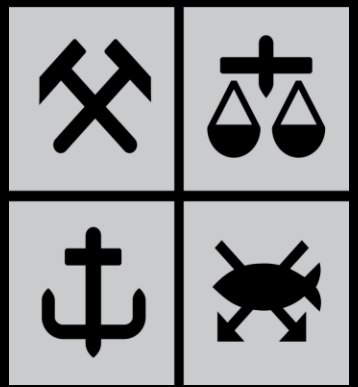
$$P(\chi^2 > 25.5 | H_0) = 0.00028$$

```
from scipy import stats
print(1-stats.chi2.cdf(x=25.5, df=7-1))
0.0002758
```

NHH
TECH3

NHH

TECH3



Sondre Hølleland
Geir Drage Berentsen