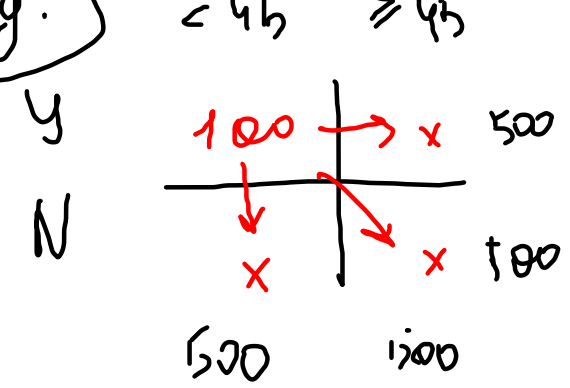
Consider the following table:

| Desire to travel abroad | Age: Under **45** | Age: **45** or older |
|---|---|---|
| Yes | 50% | 50% |
| No | 50% | 50% |
| Column totals | 100% | 100% |
| Number of participants | 500 | 500 |

(a) Is there significant association between variables?
(b) Measure the strength of association.

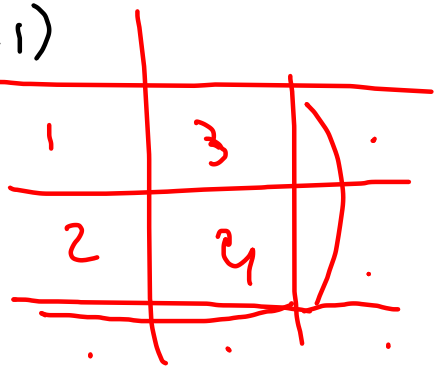
E.g.




$H_0$: no association

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(2-1)\times(2-1)}$$

$$(n_r - 1) \times (n_c - 1)$$




$O_i$

| | < 45 | ≥ 45 | |
|---|---|---|---|
| Y | 250 | 250 | = 500 |
| N | 250 | 250 | |

$E_i$

| | < 45 | ≥ 45 |
|---|---|---|
| Y | 250 | 250 |
| N | 250 | 250 |

$$E_{11} = \frac{500 \cdot 500}{1000} = \frac{n_r n_c}{N}$$

$\alpha = 0{,}05$

$\chi^2_{obs} = 0 < \chi^2_{crit,\,1,\,0{,}95}$

$\approx 8$

$\Rightarrow$ not rejected $H_0$

$$\chi^2_n = \sum_{i=1}^{n} \xi^2 \qquad \xi \sim N(0,1)$$

max $\chi^2$ :

500 0

0 500

$$\chi^2 = \frac{2{,}0^2}{250} \cdot 4 = 1000$$

# Effect size

Cont: $\hat{\rho}_{pearson}$, $\hat{\rho}_{spearman}$

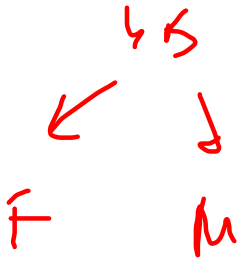
Cat: $\hat{\phi} = \sqrt{\frac{\chi^2}{N}} \in [0, 1]$

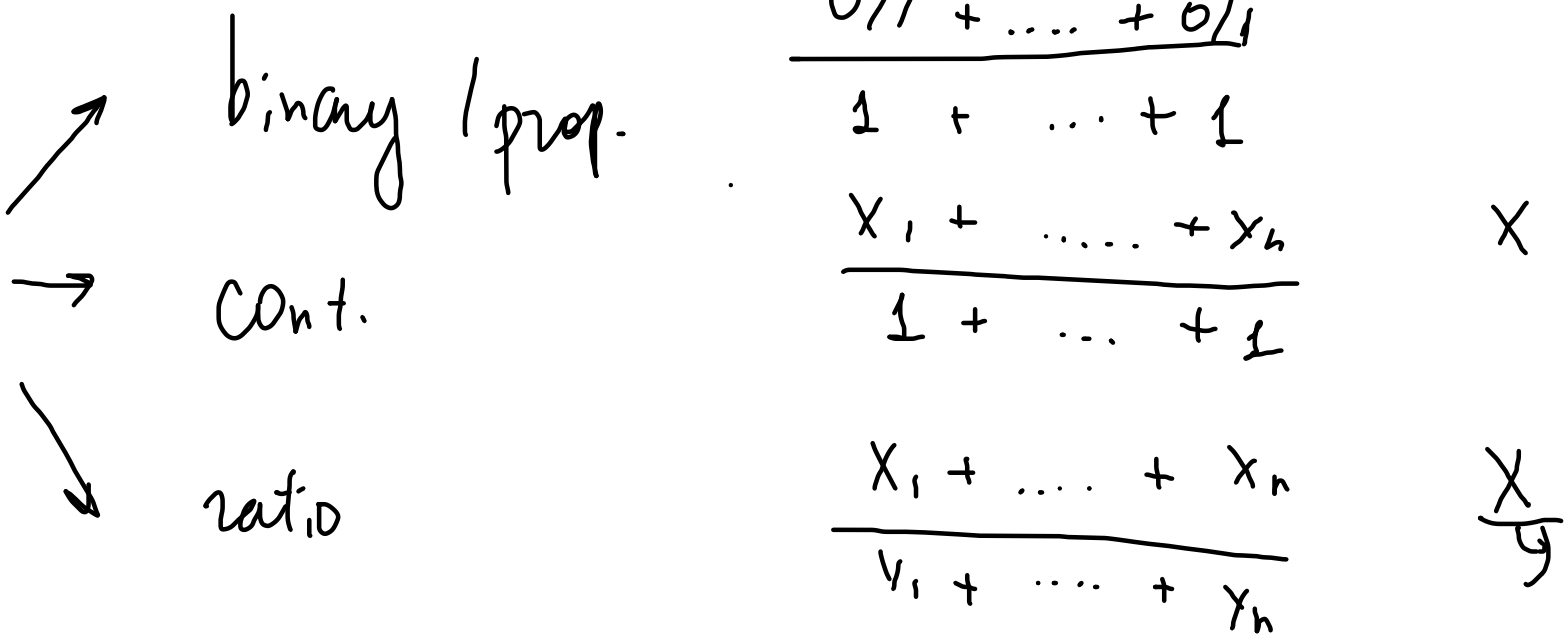
$\phi = 0,1$ ; $\phi = 0,9$

Let's introduced gender as the third variable.

| | Gender | | | |
|---|---|---|---|---|
| | Male age | | Female age | |
| Desire to travel abroad | Under **45** | **45** or older | Under **45** | **45** or older |
| Yes | 60% | 40% | 35% | 65% |
| No | 40% | 60% | 65% | 35% |
| Column totals | 100% | 100% | 100% | 100% |
| Number of participants | 300 | 300 | 200 | 200 |

(c) Is there significant association between variables within the subgroups?
(d) Measure the strength of association.
(e) Calculate Pearson residuals for Female under 45.



$2 \times 2 \times 2 = 8$

Let's introduced gender as the third variable.

| | Gender | | | |
|---|---|---|---|---|
| | Male age | | Female age | |
| Desire to travel abroad | Under 45 | 45 or older | Under 45 | 45 or older |
| Yes | 60% | 40% | 35% | 65% |
| No | 40% | 60% | 65% | 35% |
| Column totals | 100% | 100% | 100% | 100% |
| Number of participants | 300 | 300 | 200 | 200 |

(c) Is there significant association between variables within the subgroups?
(d) Measure the strength of association.
(e) Calculate Pearson residuals for Female under 45.

| $O_{ij}$ | $<45$ | $\geq 45$ |
|---|---|---|
| Y | 70 | 130 |
| N | 130 | 70 |

| $E_{ij}$ | $<45$ | $\geq 45$ |
|---|---|---|
| Y | 100 | 100 |
| N | 100 | 100 |

Goal: any relation

$H_0$

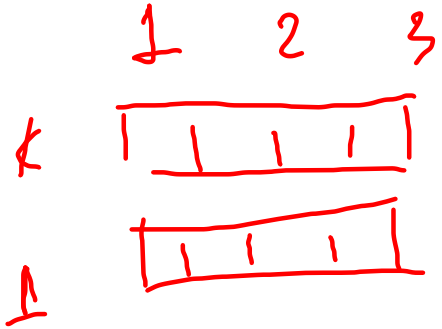
$$E_{ij} = n_j \cdot \frac{n_i}{N}$$

$$= n_j \cdot \frac{n_i}{N}$$

prop in column w.r.t. to rows

$$\chi^2_{obs.} = \frac{30^2}{100} \cdot 4 = 36 > \chi^2_{1,0,95}$$

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum r_{ij}^2$$




$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{30}{10} = 3,$$

$-3 < -1{,}96$

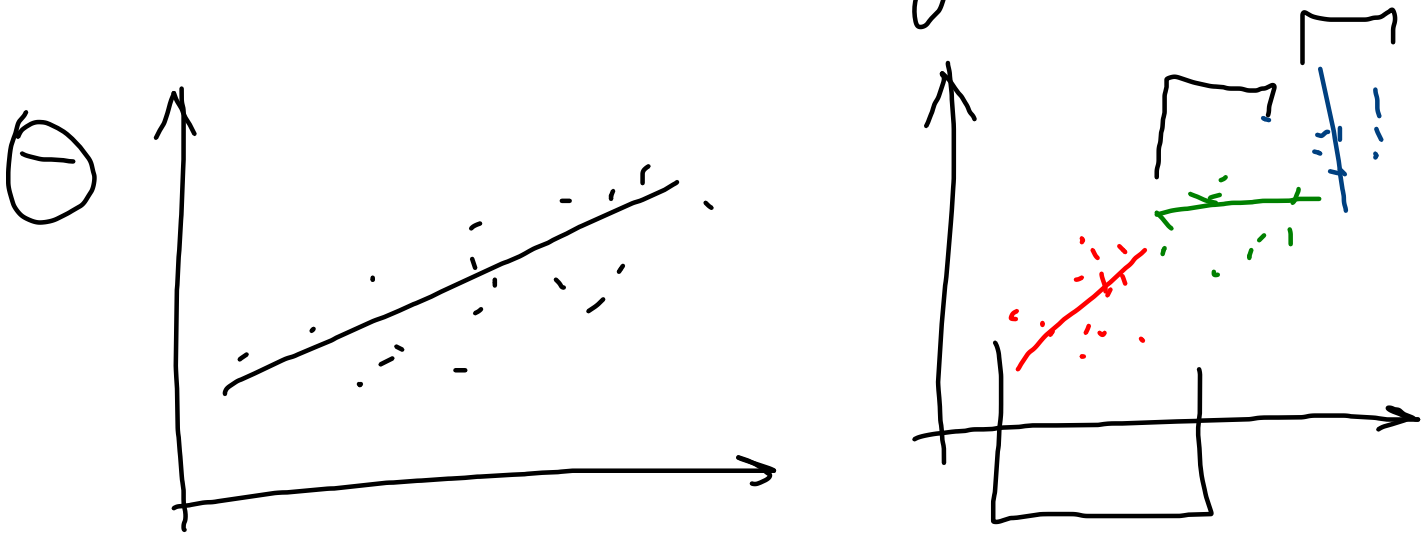
$r_{ij}$

| | |
|---|---|
| $-3$ | $3$ |
| $3$ | $-3$ |

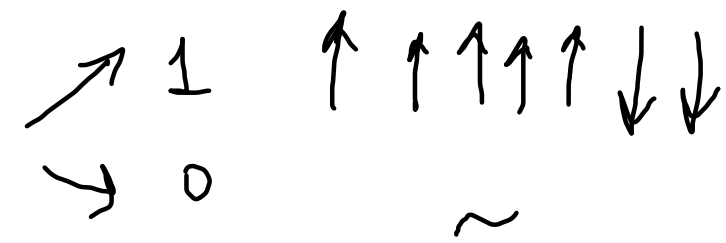
$i = 2$
$i = 1$

$r_{ij} \sim N(0,1)$

① Unobserved heterogeneity



② Seasonality



Holidays / One-time events

↳ no counterfactual

↳ matching

↳ $X_1, \ldots, X_k$

exogenous

$\perp$ with $D_t$

## ② Bucketization

$$\mathbb{E}(R_i) \approx \mathbb{E}(R_B)$$

$$R^1 = \frac{X_1^1 + \dots + X_n^1}{y_1^1 + \dots + y_n^1} \qquad \mathbb{E}(R_1)$$

$$R^B = \frac{X_1^1 + \dots + X_n^1 + X_1^2 + \dots + X_n^2 + \dots}{y_1^1 + \dots + y_n^1 + y_1^2 + \dots + y_L^2 + \dots}$$

$1, \dots, K \in [B]$ i.i.d.