Measuring relationship of variables

Relation of qualitative variables

- **Test statistic:** $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{\left(n_{ij}-n_{ij}^*\right)^2}{n_{ij}^*}\right]$$
- **Expected frequency under the assumption of independence:** n_{ij}^*
- **Degrees of freedom (in case of estimation):**

$$\nu = k - b - 1 = rc - (r-1) - (c-1) - 1 = (r-1)(c-1)$$

- **Condition: sufficiently large sample:** $n_{ij}^* \geq 10(5)$
- **Right-tailed critical region:** $c_f = \chi^2_{1-\alpha}$
- **A significant relationship is not necessarily a strong relationship (Cramér's V)**

Relation of qualitative variables

$$n_{ij}^{*} = \frac{n_{i.} \cdot n_{.j}}{n} = n \cdot \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} = n \cdot p_{i.} \cdot p_{.j}$$

$$40{,}83 = \frac{175 \cdot 56}{240} = 240 \cdot \frac{175}{240} \cdot \frac{56}{240} = 240 \cdot 0{,}7292 \cdot 0{,}2333$$

Employment Type	Credit rating		Total
	Good	Bad	
Non-earner	40,83	15,17	56,00
Pensioner	29,17	10,83	40,00
Employed	51,04	18,96	70,00
Entrepreneur	53,96	20,04	74,00
Total	175,00	65,00	240,00

Relation of qualitative variables

$\alpha \leftarrow \chi^2$

Type	Credit rating	n_{ij}	n_{ij}^*	$\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*}$
Non-earner	Good	35	40,83	0,833
Pensioner		35	29,17	1,167
Employed		55	51,04	0,307
Entrepreneur		50	53,96	0,290
Non-earner	Bad	21	15,17	2,244
Pensioner		5	10,83	3,141
Employed		15	18,96	0,826
Entrepreneur		24	20,04	0,782
Total		240,00	240,00	9,590

$$\boxed{\frac{(35-40{,}83)^2}{40{,}83}}$$

$$\boxed{\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-n_{ij}^*\right)^2}{n_{ij}^*}}$$

$$\chi^2 = \frac{(35-40{,}83)^2}{40{,}83} + \frac{(35-29{,}17)^2}{29{,}17} + \ldots + \frac{(24-20{,}04)^2}{20{,}04} = 9{,}59$$

Relation of qualitative variables

$$\boxed{\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-n_{ij}^*\right)^2}{n_{ij}^*}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2 - 2n_{ij}\cdot n_{ij}^* + \left(n_{ij}^*\right)^2}{n_{ij}^*}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{n_{ij}^2}{n_{ij}^*} - \frac{2n_{ij}\cdot \cancel{n_{ij}^*}}{\cancel{n_{ij}^*}} + \frac{\left(n_{ij}^*\right)^{\cancel{2}}}{\cancel{n_{ij}^*}}\right) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{n_{ij}^2}{n_{ij}^*} - 2n_{ij} + n_{ij}^*\right)$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{ij}^*} - 2\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij} + \sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}^* = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{ij}^*} - 2n + n$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{\frac{n_{i.}\cdot n_{.j}}{n}} - n = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2\cdot n}{n_{i.}\cdot n_{.j}} - n = \boxed{n\cdot\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i.}\cdot n_{.j}} - 1\right)}$$

Relation of qualitative variables

Type	Credit rating		Total
	Good	Bad	
Non-earner	35	21	56
Pensioner	35	5	40
Employed	55	15	70
Entrepreneur	50	24	74
Total	175	65	240

$$\chi^2 = n \cdot \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.} \cdot n_{.j}} - 1 \right)$$

$$\chi^2 = 240 \cdot \left(\frac{35^2}{175 \cdot 56} + \frac{35^2}{175 \cdot 40} + \frac{55^2}{175 \cdot 70} + \frac{50^2}{175 \cdot 74} + \frac{21^2}{65 \cdot 56} + \frac{5^2}{65 \cdot 40} + \frac{15^2}{65 \cdot 70} + \frac{24^2}{65 \cdot 74} - 1 \right) =$$

$$\chi^2 = 240 \cdot (1{,}039959 - 1) = 240 \cdot 0{,}039959 = 9{,}59$$

Relation of qualitative variables

Hypotheses:

$H_0 : P_{ij} = P_{i.} \cdot P_{.j}$ **(independence)**

$H_1 : P_{ij}$ is not equal to $P_{i.} \cdot P_{.j}$ in every case

Test statistic:

$$\chi^2 = 9{,}59$$

Critical value:

$$c_f = \chi^2_{0,95}\left[(r-1)(c-1) = 3 \cdot 1 = 3\right] = 7{,}81$$

Decision:

$$c_f = 7{,}81 < 9{,}59$$

We reject H_0; the relationship between client type and credit rating is significant at the 5% significance level.

Measuring relation

Cramér’s V association coefficient

$$0 \le C = \sqrt{\frac{\chi^2}{\underbrace{N \cdot \min\{(r-1),(c-1)\}}_{\text{max value of } \chi^2}}} \le 1$$

ANOVA

Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_M = \mu$$

There is no relationship between variable Y and the characteristic that distinguishes the populations.

$$H_1 : \exists\, j, \quad \mu_j \neq \mu$$

There is a stochastic relationship between the two variables.

ANOVA

Test statistic:

$$F = \frac{SSB/(M-1)}{SSW/(n-M)} = \frac{s_b^2}{s_w^2}$$

Mean between sum of squares:

$$S_b^2 = \frac{SSB}{M-1}$$

Mean within sum of squares :

$$S_w^2 = \frac{SSW}{n-M}$$

Distribution of the test statistic: $F(\nu_1 = M-1, \ \nu_2 = n-M)$

Acceptance and critical range :

H^2 interpretation

- What proportion of the variance in the Y characteristic is explained by the grouping characteristic (interpretable in percentage, a distribution ratio-type measure).
- To what extent does the grouping characteristic reduce the uncertainty in the inference about the membership of the Y characteristic (PRE)?

H

$$H = \sqrt{H^2}$$

H **cannot be expressed as a percentage, but is solely used to assess the strength of the relationship.**

Limits

Limits: $0 \leq H^2, H \leq 1$

$$\boxed{H^2 = \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_W^2}{\sigma_W^2 + \sigma_B^2}} \qquad \sigma^2 \neq 0$$

$$H^2 = H = 0, \text{ if } \quad \forall \bar{Y}_j = \bar{Y} \rightarrow \sigma_B = 0 \rightarrow \sigma_W = \sigma$$

$$H^2 = H = 1, \text{ if } \quad \forall Y_{ij} = \bar{Y}_j \rightarrow \sigma_W = 0 \rightarrow \sigma_B = \sigma$$

The Strength of the Linear Relationship - Covariance (C)

If X and Y are independent of each other. : $$\frac{\sum X_i Y_i}{N} = \overline{X}\,\overline{Y}$$

The measure of covariance (the joint variability of X and Y) :

$$C(X,Y) = \frac{\sum X_i Y_i}{N} - \overline{X}\,\overline{Y} = \frac{\sum d_{X_i} d_{Y_i}}{N}$$

where: $$d_{X_i} = X_i - \overline{X} \qquad d_{Y_i} = Y_i - \overline{Y}$$

$C(X,Y)$

$$\boxed{\sum_{i=1}^{N} d_{X_i} d_{Y_i}} = \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{N} (X_i Y_i - X_i \cdot \overline{Y} - Y_i \cdot \overline{X} + \overline{X} \cdot \overline{Y}) =$$

$$= \sum_{i=1}^{N} X_i Y_i - \overline{Y} \sum_{i=1}^{N} X_i - \overline{X} \sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} \overline{X}\overline{Y}$$

$$= \sum_{i=1}^{N} X_i Y_i - \overline{Y} \cdot N \cdot \overline{X} - \cancel{\overline{X} \cdot N \cdot \overline{Y}} + \cancel{N \cdot \overline{X} \cdot \overline{Y}} = \boxed{\sum_{i=1}^{N} X_i Y_i - N \cdot \overline{Y} \cdot \overline{X}}$$

$$C(X,Y) = \frac{\sum X_i Y_i}{N} - \overline{X}\overline{Y} = \frac{\sum X_i Y_i - N \cdot \overline{X}\overline{Y}}{N} = \frac{\sum d_{X_i} d_{Y_i}}{N}$$

STRENGTH OF THE LINEAR RELATIONSHIP – COVARIANCE (C)

- $X = Y$ in special case $\quad C(Y,Y) = \frac{\sum d_{Y_i} d_{Y_i}}{N} = \frac{\sum d_{Y_i}^2}{N} = \sigma_Y^2$
- The $C(X,Y)$ alone indicates the existence and direction of the relationship between X and Y
- Its sign is determined by the sign of the $\sum d_{X_i} d_{Y_i}$ product sum
- $C(X,Y) = 0 \rightarrow$ no linear correlation
- $C(X,Y) > 0 \rightarrow$ positive correlation
- $C(X,Y) < 0 \rightarrow$ negative correlation

$$0 \leq \left| C(X,Y) \right| \leq \sigma_X \sigma_Y$$

LINEAR CORRELATION COEFFICIENT

$$r(X,Y)=\frac{C(X,Y)}{\sigma_X\sigma_Y} \qquad C(X,Y)=\frac{\sum d_{X_i}d_{Y_i}}{N}$$

$$r(X,Y)=\frac{\dfrac{\sum d_{X_i}d_{Y_i}}{N}}{\sqrt{\dfrac{\sum d_{X_i}^2}{N}}\sqrt{\dfrac{\sum d_{Y_i}^2}{N}}}=\frac{\sum d_{X_i}d_{Y_i}}{\sqrt{\sum d_{X_i}^2\sum d_{Y_i}^2}}$$

$$-1\le r\le 1$$