Aplikace měr variability pro nominální proměnné ve vícerozměrné statistické analýze

Hana Řezanková
VŠE Praha, ÚI AV ČR

Obsah

- Míry variability pro nominální proměnné
- Využití pro hodnocení intenzity závislosti
- Využití k měření podobnosti objektů charakterizovaných nominálními proměnnými
- Využití v metodách shlukové analýzy a hodnocení shlukování kategoriálních dat
- Využití v rozhodovacích stromech

Míry variability

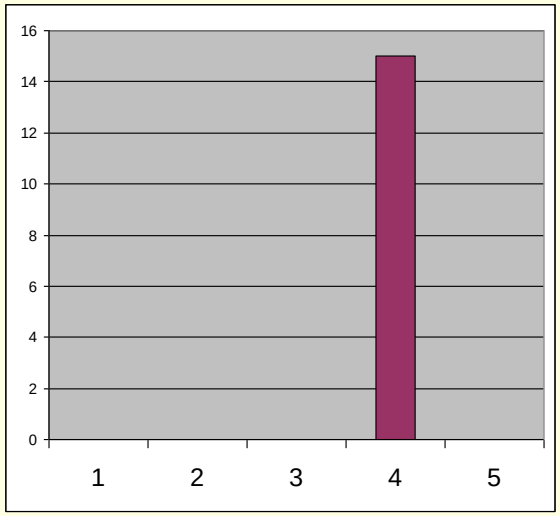
a) variační poměr $v = 1 - p_{\mathrm{Mo}} = 1 - n_{\mathrm{Mo}} / n$

b) nominální rozptyl (*Giniho míra, mutabilita, Gibbs: M1*)

$$\mathrm{nomvar} = 1 - \sum_{u=1}^{K} p_u^2 = \sum_{i=1}^{K} (p_u (1 - p_u))$$

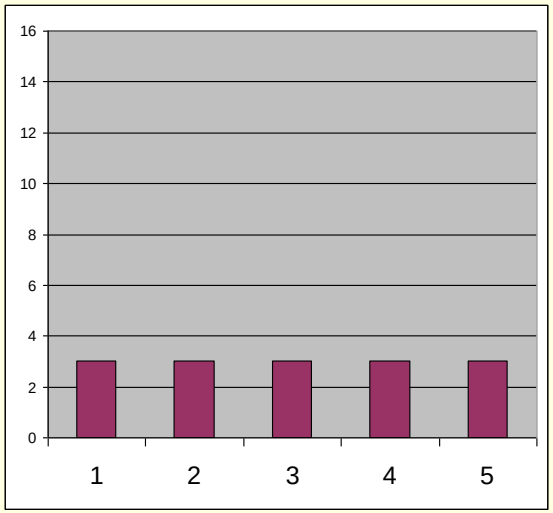
Corrado Gini (1884–1965), 1914

2 extrémní případy



$$1 - p_{\mathrm{Mo}} = 0$$

$$1 - \sum_{u=1}^{K} p_u^2 = 0$$



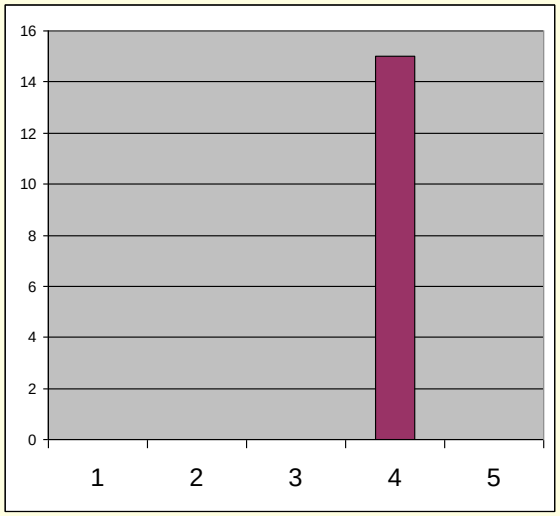
$$v = (K - 1)/K$$

$$\mathrm{nomvar} = (K - 1)/K$$

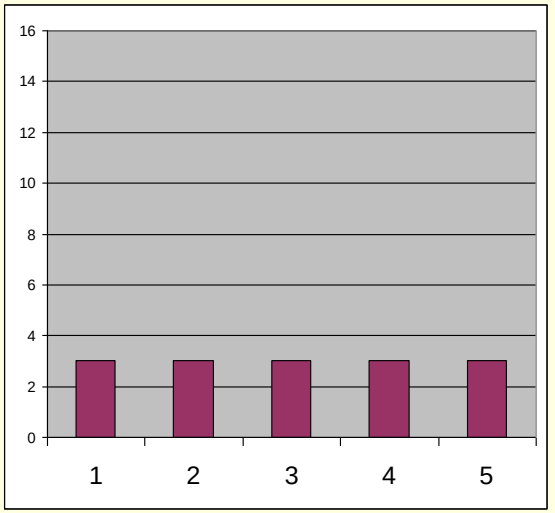
Míry variability

c) entropie $H = -\sum_{u=1}^{K} p_u \ln p_u$

2 extrémní případy



$H = 0$



$H = \ln k$

Závislost nominálních proměnných

Goodmanovo-Kruskalovo λ

- *vychází z principu analýzy rozptylu a poměru determinace*

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}$$

$$S_{Y|X} = \frac{var(Y) - \sum_{r=1}^{R} p_{r+} var(Y|x_r)}{var(Y)}$$

variační poměr

$$v = 1 - p_{\mathrm{Mo}}$$

$$var(Y) = var(Y,X) + var(Y|X)$$

Závislost nominálních proměnných

Goodmanovo-Kruskalovo τ

- *vychází z principu analýzy rozptylu a poměru determinace*

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}$$

$$S_{Y|X} = \frac{var(Y) - \sum_{r=1}^{R} p_{r+} var(Y|x_r)}{var(Y)}$$

nominální rozptyl

$$\text{nomvar} = 1 - \sum_{s=1}^{S} p_s^2$$

$$var(Y) = var(Y,X) + var(Y|X)$$

Závislost nominálních proměnných

Informační koeficient (koeficient neurčitosti, nejistoty)

- *vychází z principu analýzy rozptylu a poměru determinace*

$$S_{Y|X} = \frac{var(Y,X)}{var(Y)} = \frac{var(Y) - var(Y|X)}{var(Y)}$$

$$S_{Y|X} = \frac{var(Y) - \sum_{r=1}^{R} p_{r+} var(Y|x_r)}{var(Y)}$$

entropie

$$H = -\sum_{s=1}^{S} p_s \ln p_s$$

$$var(Y) = var(Y,X) + var(Y|X)$$

Měření podobnosti objektů

Míra VM (variable mutability)

$$S_{\mathrm{VM}}(x_{ic}, x_{jc}) = \begin{cases} \dfrac{K_c}{K_c - 1}\left(1 - \sum_{u=1}^{K_c} p_u^2\right) & \text{jestliže } x_{ic} = x_{jc} \\ 0 & \text{jinak} \end{cases}$$

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S(x_{ic}, x_{jc})}{m}$$

Míra VE (variable entropy)

$$S_{\mathrm{VE}}(x_{ic}, x_{jc}) = \begin{cases} -\dfrac{1}{\ln K_c} \sum_{u=1}^{K_c} p_u \ln p_u & \text{jestliže } x_{ic} = x_{jc} \\ 0 & \text{jinak} \end{cases}$$

balíček *nomclust* pro R software (Z. Šulc, H. Řezanková)
Comprehensive R Archive Network (CRAN)
http://CRAN.R-project.org/

Měření odlišnosti shluků

Míra D_{GK} (založená na Giniho koeficientu)

$$D_{\mathrm{GK}}(C_h, C_{h'}) = G_{hh'} - (G_h + G_{h'})$$

$$G_h = \sum_{c=1}^{m} \left(1 - \sum_{u=1}^{K_c} p_{hcu}^2 \right)$$

Míra D_{EN} (založená na entropii)

$$D_{\mathrm{EN}}(C_h, C_{h'}) = H_{hh'} - (H_h + H_{h'})$$

$$H_h = \sum_{c=1}^{m} \left(- \sum_{u=1}^{K_c} \left(p_{hcu} \ln p_{hcu} \right) \right)$$

T. Löster: *Hodnocení výsledků metod shlukové analýzy*. Disertační práce. VŠE, Praha 2011.

Další metody shlukové analýzy

Algoritmus COOLCAT (2002)

$$\overline{H}(\breve{C}) = \sum_{h=1}^{k}\left(\frac{n_h}{n}\sum_{c=1}^{m}\left(-\sum_{u=1}^{K_c} p_{hcu}\ln p_{hcu}\right)\right) \qquad \rightarrow \min$$

Barbará, D., Li, Y., Couto, J.: COOLCAT: An entropy-based algorithm for categorical clustering. *Proceedings of the 11th International Conference on Information and Knowledge Management*, ACM Press, McLean, VA 2002, 582–589.

Hodnocení shlukování

$$WCM(k) = \sum_{h=1}^{k} \frac{n_h}{n \cdot m} \sum_{c=1}^{m} \frac{K_c}{K_c - 1} \left(1 - \sum_{u=1}^{K_c} p_{hcu}^2 \right)$$

$$WCE(k) = \sum_{h=1}^{k} \frac{n_h}{n \cdot m} \sum_{c=1}^{m} \frac{1}{\ln K_c} \left(- \sum_{u=1}^{K_c} \left(p_{hcu} \ln p_{hcu} \right) \right)$$

$$PSFM(k) = \frac{(n-k)(WCM(1) - WCM(k))}{(k-1)WCM(k)}$$

$$PSFE(k) = \frac{(n-k)(WCE(1) - WCE(k))}{(k-1)WCE(k)}$$

Řezanková, H., **Löster, T.**, Húsek, D. (2011). Evaluation of categorical data clustering. In Mugellini, E, Szczepaniak, P. S., Pettenati, M. C. et al. (Eds.), In *Advances in Intelligent Web Mastering 3*. Berlin: Springer Verlag, 173-182.

Další využití

Využití v rozhodovacích stromech CART

pro výběr vysvětlující proměnné, pomocí jejíchž kategorií lze rozdělit hodnoty vysvětlované proměnné do skupin s nejmenší entropií, resp. nejnižší hodnotou Giniho indexu

Děkuji Vám za pozornost