Formular de Analiza a Datelor si ACP (cu Explicatii)

Extras din documente

16 noiembrie 2025

1 Document 1: Analiza Preliminara a Datelor

Varianta (caz discret): Media ponderata a patratelor abaterilor de la medie.

$$\sigma^2 = \sum_{x \in R} (x - \mu)^2 \cdot f(x) \tag{1}$$

Varianta (pentru esantion / repartitie uniforma): Media patratelor abaterilor de la medie.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{2}$$

Abaterea standard: Radacina patrata a variantei; masoara împrastierea în unitatile variabilei.

$$\sigma = \sqrt{\sigma^2} \tag{3}$$

Coeficientul de variatie: Raportul dintre abaterea standard si medie; o masura relativa a împrastierii.

$$C_v = \frac{\sigma}{\mu} \tag{4}$$

1.1 Indicatori de Forma

Asimetria (Skewness): Masoara gradul de simetrie a distributiei (bazata pe momentul 3).

$$S = \frac{MC_3}{\sigma^3} \tag{5}$$

Aplatizarea (Kurtosis): Masoara 'boltirea' distributiei (bazata pe momentul 4).

$$K = \frac{MC_4}{\sigma^4} \quad \text{sau} \quad K = \frac{MC_4}{\sigma^4} - 3 \tag{6}$$

1.2 Teste de Concordanta si Independenta

Statistica testului χ^2 de concordanta: Masoara diferenta dintre frecventele observate (fa) si cele asteptate (fe).

$$\chi^2_{\text{Calculat}} = \sum_{i=1}^{m} \frac{(fa_i - fe_i)^2}{fe_i} \tag{7}$$

Statistica testului Smirnov-Kolmogorov: Diferenta maxima absoluta dintre functia de repartitie empirica (Fe) si cea teoretica (F).

$$D = \max_j |Fe(x_{(j)}) - F(x_{(j)})| \tag{8}$$

Statistica testului χ^2 de independenta (frecvente absolute): Testeaza daca exista o asociere între doua variabile calitative.

$$\chi^2_{\text{Calculat}} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - ne_{ij})^2}{ne_{ij}} \tag{9}$$

Statistica χ^2 de independenta (frecvente relative): Varianta de calcul bazata pe proportii (frecvente *f*).

$$\chi^2_{\text{Calculat}} = T \cdot \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} \tag{10}$$

1.3 Relatia dintre Variabile Cantitative

Modelul de regresie liniara simpla: Relatia dintre *y* si *x*, incluzând un termen de eroare *e*.

$$y_i = ax_i + b + e_i \tag{11}$$

Covarianta: Masoara tendinta a doua variabile de a se modifica împreuna.

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{12}$$

Coeficientii regresiei liniare: *a* este panta (calculata cu covarianta) si *b* este interceptul.

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{si} \quad b = \overline{y} - a\overline{x} \tag{13}$$

Descompunerea variantei: Varianta totala = Varianta explicata de model + Varianta reziduala.

$$\text{Var}(y) = \text{Var}(ax + b) + \text{Var}(e) \tag{14}$$

Coeficientul de determinare (R^2): Proportia din varianta lui *y* care este explicata de *x*.

$$R^2(x, y) = \frac{\text{Var}(ax + b)}{\text{Var}(y)} = \frac{\text{Cov}(x, y)^2}{\text{Var}(x)\text{Var}(y)} \tag{15}$$

Coeficientul de corelatie liniara (R): Masoara puterea si directia legaturii liniare (de la -1 la 1).

$$R = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \tag{16}$$

2 Document 2: Analiza în Componente Principale (Spatiul Instantelor)

Definitia componentei principale (combinatie liniara): Componenta *k* este o suma ponderata (combinatie liniara) a variabilelor originale *X*.

$$C_k = a_{1k}X_1 + a_{2k}X_2 + \cdots + a_{mk}X_m \tag{17}$$

Solutia (Problema valorilor si vectorilor proprii): Solutia a_1 este vectorul propriu al matricei de covarianta $\frac{1}{n}X^tX$.

$$\frac{1}{n}X^tXa_1 = \lambda a_1 \tag{18}$$

Varianta explicata este egala cu valoarea proprie: Varianta maxima explicata de componenta este λ, valoarea proprie corespunzatoare vectorului a_1.

$$\frac{1}{n}a_1^tX^tXa_1 = \lambda \tag{19}$$

3 Document 3: Analiza în Componente Principale (Evaluare)

3.1 Deducere (Spatiul Variabilelor)

Varianta explicata de axa k: Varianta (inertia) componentei *k* este egala cu valoarea proprie α_k.

$$\text{Varianta}(C_k) = \alpha_k \tag{20}$$

Suma R^2 este egala cu valoarea proprie: Varianta explicata (α_k) este, de asemenea, suma corelatiilor la patrat dintre componenta *k* si toate variabilele originale.

$$\sum_{j=1}^{m} R^2(C_k, X_j) = \alpha_k \tag{21}$$

Contributia instantei *i* la varianta axei *j*: Arata ce procent din varianta totala a axei *j* este 'generat' de punctul *i*.

$$\beta_{ij} = \frac{1}{n} \cdot \frac{c_{ij}^2}{\alpha_j} \tag{22}$$

Comunalitatea variabilei *Xj* (pentru primele *s* componente): Cât la suta din varianta variabilei originale *Xj* este 'capturata' de primele *s* componente.

$$\text{Comunalitate}(X_j) = \sum_{k=1}^{s} R(X_j, C_k)^2 \tag{23}$$

Corelatii factoriale (Factor Loadings) - vectorul de corelatii: Calculeaza corelatia dintre componenta *k* si *toate* variabilele originale. Aceasta este formula folosita în problema din imagine.

$$R_k = a_k\sqrt{\alpha_k} \tag{24}$$