Tema de casa nr 5 mspi

Dimitriu Gabriel ISC

15.06.2000

1 Prezentarea succinta a metodele statistice utilizate

1.1 Inferenta asupra mediei

In acest proiect se foloseste numai inferentele de medie pentru $sigma^2$ necunoscut deoarece nu cunoastem dispersia, ea urmind sa fie evaluata la un moment ulterior, dar folosim si estimarea si testul de medie.

1.1.1 Estimare

Acesta inferenta vrea sa determine media dintr-un esantion de date i.i.d. de volum N date $[x_1, x_2, \ldots, x_N]$ rezultind $\overline{x}$ estimatia care este de distributie gaussiana $\overline{X} = N(\mu, \sigma^2/N)$.

Se construieste

$$T = \sqrt{N}\frac{\overline{X} - \mu}{S}$$

unde S este un estimator al dispersiei:

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

Sau pentru estimatie avem:

$$t = \sqrt{N}\frac{\overline{x} - \mu}{s}$$

cu

$$s = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Se demonstreaza ca T are legea de repartitie Student cu N-1 grade de libertate daca variabla aleatoare X masurata este Gaussiana. De observat ca pentru $N > 30$ legea Student se comporta ca o lege Gauss.

Se spune ca $P(|T| < t_{\alpha/2}) = 1 - \alpha \Rightarrow |t| < t_{\alpha/2}$ in $(1 - \alpha) * 100\%$ *din cazuri*.

Deci

$$\|\sqrt{N}\frac{\overline{x}-\mu}{s}\| < t_{alpha/2}$$

Iar intervalul de incredere este

$$[\overline{x}-\varepsilon, \overline{x}+\varepsilon]$$

cu $\varepsilon = t_{\alpha/2}\frac{s}{\sqrt{N}}$

Daca nu se specifica N trebuie sa se faca iteratii pentru aflarea lui.

1.1.2 Testare

Se presupune ca datele de intrare provin din populatii gaussiene deci $[x_1, x_2, ..., x_N] \in N(\mu, \sigma^2)$ si sunt i.i.d.

Se pun ipotezele $H_0 : \mu = \mu_0$ si $H_1 : \mu \neq \mu_0$.

Se alege α care este pragul de semnificatie, calculez $t = \sqrt{N}\frac{\overline{x}-\mu_0}{s}$ care are adistrbutia Student.

Accept testul daca$t \in (-t_{\alpha/2}, t_{\alpha/2})$, adica ipoteza H_0 si daca $t \notin (-t_{\alpha/2}, t_{\alpha/2})$ accept ipoteza H_1.

Eroare de ordin I este α iar eroarea de ordin II este β.

1.2 Inferente asupra dispersiei

In acest caz se va folosi inferetele asupra dispersiei dar cu media necunoscute. Caz in care se restring ipotezele la populatii gaussiene.

1.2.1 Estimare

Avem datele $[x_1, x_2, ..., x_N)$ de distributie $X \sim N(\mu, \sigma^2)$

Trebuie cunoscuta legea de distribuie a variabilelor de intrare deoarece dispersia dispersiei este dependenta de tipul populatiei astfel ca estimatiile sunt:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Iar pentru populatii gaussiene avem:

$$(N-1)s^2 = \sum_{i=1}^{N}(x_i - \overline{x})^2$$

Atentie acesta suma de patrate de distributii gauss nu sunt independente intre ele, deoarece au acceasi medie $\overline{x}$. Deci nu se poate supe direct ca este legea χ^2 de N grade de libertate. Se poate demonstra ca acesta este o lege de tip χ^2 de N-1 grade de libertate si de parametru σ^2.

Avem:

$$(N-1)\frac{S^2}{\sigma^2} = Y = \chi^2_{N-1;1} \in (y_{1-\alpha/2}, y_{\alpha/2})$$

valabila cu probabilitatea $1-\alpha$.

Aleg$1-\alpha$ N este dat si voi determina $y_{\alpha/2}\ si\ y_{1-\alpha/2}$.

Dupa prelucrari avem:

$$\sigma^2 \in (\frac{(N-1)*s^2}{y_{\alpha/2}}, \frac{(N-1)*s^2}{y_{1-\alpha/2}})$$

In care avem erorile:

$$\varepsilon_- = \frac{(N-1)*s^2}{y_{\alpha/2}}$$

$$\varepsilon_+ = \frac{(N-1)*s^2}{y_{1-\alpha/2}}$$

Daca $N > 800$ functia χ^2 se simetrizeaza si $\varepsilon_- = \varepsilon_+$.

1.2.2 Testare

Se foloses aceleasi ipoteze statistice ca si pentru estimare.

Ipoteze de test: $H_0 : \sigma^2 = \sigma_0^2$ si $H_1 : \sigma^2 \neq \sigma_0^2$. Se alege α si se calculeaza

$$y_calc = \frac{(N-1)*s^2}{\sigma_0^2}$$

Daca $y_calc \in$ zona de acceptare se trece testul daca nu se resping datele.

1.3 Testul de concordanta Kolmogorov Smirnov

Este un test de concordanta mai puteric decit testul χ^2 si se foloseste numai pentru legi de repartitie continue. Este un test neparametric (liber de repartitie).

Ca variabile de intrare avem N observatii i.i.d. $[x_1, x_2, ..., x_N]$ ordonate crescator: $x_1 \leq x_2 \leq ... \leq x_N$.

Concordanta se face cu o lege continua specificata F(x) de parametri cunoscuti sau necunoscuti.

Formulam ipotezele statistice:

H_0: legea de repartitie este F(x)

H_1: legea de repartitie nu este F(x)

Se alge pragul α de semnificatie statistica.

Se calculeaza functia de repartitie empirica.

$$F_N(x) = P\{X \leq x\} = \frac{m_x}{N}$$

$F_N(x)$ este o curba in trepte care se suprapune peste curba teoretica si se calculeaza:

$$\Delta = \max_{-\infty < x < \infty} |F(x) - F_N(x)|$$

$$P\{\Delta < \Delta_\alpha\} = P\{\sqrt{N}\Delta < \sqrt{N}\Delta_\alpha\} = 1 - \alpha$$

Pentru $N > 80$ avem

$$\Delta_\alpha \simeq \sqrt{\frac{1}{2N}\ln\frac{2}{\alpha}}$$

Din datele experimentale evaluam

$$\Delta = \sup_{-\infty < x < \infty} |F(x) - F_N(x)|$$

Daca $\Delta < \Delta_\alpha$ acceptam H_0.
Daca $\Delta > \Delta_\alpha$ respingem datele ca fiind semnificative.
Zona de acceptare test este $(0, \Delta_\alpha)$.
Zona de respingere test este $(\Delta_\alpha, 0)$.

2 Prezentarea metodei propuse

Intii vom face o estimare de medie asupra generatorului de numere aleatoare cu insumarea a N secvente uniforme. Cu valoarea estimata fom face un test de medie asupra generetorului preimplementat astfel incit se va observa daca mediile difera. Se mai putea face si cu o diferenta de medie pe baza a doua multimi de date experimentale.

Apoi vom face o estimare de dispersie asupra generatorului de numere aleatoare cu insumarea a N secvente uniforme. Cu valoarea estimata vom face un test de dispersie asupra generatorului preimplementat astfel incit se va observa daca dispersiile acestora difera. Se mai putea face si cu o diferenta de dispersie pe baza a doua multimi de date experimetale.

In final se va face un test Kolmogorov-Smirnov pentru a vedea daca generatorul de numere aleatoare cu insumarea a N secvente uniforme este un generetor aleator gaussian ca forma.

3 Concluzii

Generatorul de date aleatoare obtinut prin insumare este identic cu increderea 95.0% daca se ia N=5-6. De exemplu cu N=3 nu este trecut testul de dispersie.