PSK - zajęcia zaplanowane na 14 & 21.05

Analiza statystyczna wyników symulacji.

Nsbp. Badania symulacyjne przeprowadza się najczęściej po to, aby ustalić **wartość** θ pewnej **wielkości** związanej z określonym modelem stochastycznym, wtedy podstawowe do wykonania tej symulacji:

Niech wynikiem takiej symulacji jest zmienna losowa X, której wartość oczekiwana $EX = \theta$.

Takie przebiegi symulacji powtarza się w sposób niezależny aż do pewnej ilości, K- przebiegów.

Mamy do dyspozycji niezależne zmienne losowe $X_1, X_2, \ldots, X_K$ o tym samym rozkładzie co X.

Wtedy (na mocy mocnego prawa wielkich liczb Bernoulliego) średnia

$$\overline{X(\omega)} = \frac{1}{K}\left(X_1(\omega) + \ldots + X_K(\omega)\right) \approx \theta$$

jest estymatorem tej wartości θ.

Pojawia się problem: kiedy zatrzymać proces symulacji, czyli jak ustalić K, aby pomyłka popełniona

będzie adekwatne do badanej sytuacji metodę symulacyjną.
Idąc dalej, czy dla danej liczy $\alpha \in (0,1)$
da się znaleźć tenże <u>przedział losowy</u> ufności
przez dwie statystyki

$$Z_1 = f_1(X_1, \dots, X_k) < Z_2 = f_2(X_1, \dots, X_k),$$

(*) tj. $P(\text{Kwadr: } Z_1(\omega) < \theta < Z_2(\omega)) \geq 1 - \alpha$.

<u>Średnia z próby jako estymator</u>

Niech X oznacza zadany rozkład. Dla $n \geq 2$,
niech $X_1, X_2, \dots, X_n$ - zmienne losowe, które są

(i) niezależne

(ii) <u>każda</u> ma rozkład <u>jak</u> X

Jeśli $EX = \theta$, $\text{Var}(X) = \sigma^2$, to

$E(X_j) = \theta$, $\text{Var}(X_j) = \sigma^2$, j

a wektor $(X_1, X_2, \dots, X_n)$ nazwiemy PRÓBĄ.

Przez ŚREDNIĄ Z PRÓBY rozumiemy

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

Pokazuje jak wykorzystać $\bar{X}$ do oszacowania (nieznanej) wartości θ.

Zauważmy, że

$$E\bar{X} = E\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n} EX_j = \frac{n\theta}{n} = \theta$$

Obliczmy teraz średnie kwadratowe odchylenie pomiędzy wartością $\bar{X}$ a wartością θ, czyli

$$E\left[(\bar{X}-\theta)^2\right] = \mathrm{Var}(\bar{X}), \text{ bo } E\bar{X} = \theta$$

Ale z niezależności $X_1, \ldots, X_n$:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) =$$

$$= \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{Var}(X_j) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n},$$

co oznacza, że jeśli n jest dostatecznie duże, to $E\left[(\bar{X}-\theta)^2\right]$ jest tak małe jak chcemy, czyli $\bar{X}$ w tym sensie dobrze estymuje θ.

Co więcej, korzystając z nierówności Czebyszewa możemy oszacować miarę rozproszenia wartości $\bar{X}$ względem θ. Dokładniej, dla $c>0$

$$P\left(\omega\in\Omega: |\bar{X}(\omega)-\theta| \geq c\sqrt{\operatorname{var}\bar{X}}\,\eta\right) \leq \frac{1}{c^2}$$

$$\| \quad$$

$$P\left(\omega\in\Omega: |\bar{X}(\omega)-\theta| \geq \frac{c\sigma}{\sqrt{m}}\,\eta\right) \leq \frac{1}{c^2},$$

co równoważnie można zapisać

$$P\left(\omega\in\Omega: |\bar{X}(\omega)-\theta| < \frac{c\sigma}{\sqrt{m}}\,\eta\right) > 1-\frac{1}{c^2}$$

$$\| \quad$$

$$(**)\quad P\left(\omega\in\Omega: \theta\in\left(\bar{X}(\omega)-\frac{c\sigma}{\sqrt{m}},\ \bar{X}(\omega)+\frac{c\sigma}{\sqrt{m}}\right)\right) > 1-\frac{1}{c^2}$$

co pozwala określić nam przedział ufności, o którym wspominaliśmy w (*) (sk. 2).

Problem z wyliczaniem miary (**) polega na tym, że na ogół wariancja σ^2 nie jest znana.

Trzeba ją zatem oszacować (estymować).

Bezpośrednim odpowiednikiem $\sigma^2 = E(X-\theta)^2$ jest estymator oznaczany jako

$$S^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2.$$

Ma on jednak pewne wady, mianowicie

$$ES^2 \neq \sigma^2.$$

Jeśli dokonamy jego korekty i weźmiemy

$$\frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2,$$

to otrzymamy estymator oznaczany $\hat{S}^2$, taki że

$$E\hat{S}^2 = \sigma^2 \longrightarrow \text{WARIANCJA Z PRÓBY}$$

Wtedy $\hat{S} = \sqrt{\hat{S}^2} \longrightarrow$ DYSPERSJA Z PRÓBY

Mamy dwie wykluczające (**) . Pomyślmy, w której (**):

(i) chcemy, aby

$$1 - \frac{1}{c^2} \geq 0{,}95 \equiv \quad (c > 0)$$

$$\frac{5}{100} \geq \frac{1}{c^2} \equiv c^2 \geq 20 \equiv$$

czyli $c \geq \sqrt{20}$

(ii) długość przedziału

$\left(\bar{X} - \frac{c\sigma}{\sqrt{n}}, \bar{X} + \frac{c\sigma}{\sqrt{n}}\right)$, czyli

$\frac{2c\sigma}{\sqrt{n}}$ ma być mniejsza od zadanej liczby

$2d$ $(d > 0)$, czyli

$$\frac{2c\sigma}{\sqrt{n}} \leq 2d \equiv \frac{c\sigma}{\sqrt{n}} \leq d$$

$$\equiv \sqrt{n} \geq \frac{c\sigma}{d} \equiv$$

$$n \geq \frac{c^2\sigma^2}{d^2} \geq 20\frac{\sigma^2}{d^2} \approx 20\frac{\hat{S}^2}{d^2}$$

Zatem dla $n \geq 20 \frac{\hat{s}^2}{d^2}$, warunek (**) jest spełniony, przy wymaganiach zadanych w (i), (ii).

Uwaga. Prawdopodobieństwo wystąpienia w (**),

$$P(\omega \in \Omega : \theta \in (\bar{X}(\omega) - \frac{c\sigma}{\sqrt{n}}, \bar{X}(\omega) + \frac{c\sigma}{\sqrt{n}}))$$

można oszacować lepiej, aniżeli wynika to z nierówności Czebyszewa.

W tym celu zauważmy, że

$$\frac{\bar{X} - E\bar{X}}{\sqrt{\mathrm{var}(\bar{X})}} = \frac{\bar{X} - \theta}{\frac{\sigma}{\sqrt{n}}}$$ ma rozkład dla

odpowiednio dużych n ($n \geq 200$) z C.T.G.

$$\frac{\bar{X} - \theta}{\frac{\sigma}{\sqrt{n}}} \in N(0,1).$$

Ponieważ

$$\theta \in (\bar{X}(\omega) - \frac{c\sigma}{\sqrt{n}}, \bar{X}(\omega) + \frac{c\sigma}{\sqrt{n}}) \equiv$$

$$|\bar{X}(\omega) - \theta| < \frac{c\sigma}{\sqrt{n}} \equiv$$

$$\left|\frac{\bar{X}(\omega) - \theta}{\frac{\sigma}{\sqrt{n}}}\right| < c \text{, m.j.c dla takich } n,$$

$$P(\{\omega\in\Omega: \theta \in (\bar{X}(\omega) - \tfrac{c\sigma}{\sqrt{n}}, \bar{X}(\omega) + \tfrac{c\sigma}{\sqrt{n}})\}) =$$

$$= P(\{\omega\in\Omega: \left|\frac{\bar{X}(\omega) - \theta}{\frac{\sigma}{\sqrt{n}}}\right| < c\}) \simeq$$

$$\Phi(c) - \Phi(-c) = 2\Phi(c) - 1,$$

gdy Φ oznacza dystrybuantę $\mathcal{N}(0,1)$

Jeśli teraz ma być spełnione założenie (1)

$$2\Phi(c) - 1 = 0{,}95 \iff$$

$$\Phi(c) = 0{,}975 \equiv \underline{\underline{c = 1{,}96}}$$

(tablice rozkładu $\mathcal{N}(0,1)$!)

Dlatego wtedy

$$|\bar{X}(\omega) - \theta| < \frac{\sigma}{\sqrt{n}} = 1{,}96\frac{\sigma}{\sqrt{n}},$$

gdy σ jest estymowane statystyką $\hat{S} = \sqrt{S^2}$

Zatem warunek na miarę rozproszenia ma postać:

$$1{,}96\frac{\hat{S}}{\sqrt{n}} \leq d, \text{ gdy } d -$$

akceptowalne rozproszenie.

Jeśli n - liczba powtórzeń symulacji spełnia teraz powyższy warunek, to dla tego n kończymy proces symulacji i mamy $\bar{X} \approx \theta$ z prawdop. 0,95.

Przykład 1

Załóżmy, że przedsiębiorstwo analizy pewnej firmy ubezpieczeń F. O F wiadomo, że po godzinie 17.00 nie powinny pojawiać się nowi klienci.

Chcemy oszacować oczekiwany czas, w którym klient F opuści ją do godziny 17.00.

Wymagamy, aby poziom tego oszacowania był na poziomie 95% oraz że szacunkowa wartość średnia różniła się od rzeczywistej mniej niż o 15 sekund.

Aby to zrobić, musimy generować (w wyniku symulacji) zdarzenia odnoszące się do omawianej sytuacji. Ze względu na wymagania C.T.G. liczba tych k powinna być ≥ 100 oraz powinna spełniać nierówność

$$1{,}96 \frac{\hat{S}}{\sqrt{k}} \leq 15$$, gdzie $\hat{S}$ mierzona

jest w sekundach. Wtedy $\theta \approx \overline{X}$.

Podsumujmy:

Z teoretycznego punktu widzenia potrafimy obliczyć liczbę cykli symulacji n, aby zgodnie z oczekiwaniami estymatorki wartości średniej θ zmiennej losowej X byłaby wynikiem tej symulacji.

Z praktycznego punktu widzenia (chodzi o przetwarzanie na maszynie) byłoby lepiej, dojść do tej wartości n dochodząc metodą rekurencyjną, a nie tak jak polecenie powyższa metoda – poprzez generowanie za każdym razem tylu powtórzeń, aż warunek $1{,}96\frac{\hat{S}}{\sqrt{k}} < d$ będzie spełniony.

Kolejne powtórzenia:
- ... NIE
- ... NIE
- } NIE
- ... TAK

Pokażemy jak to można zrobić!!

Niech:

(i) X - wynik symulacji

$EX = \theta$, $\mathrm{var}(X) = \sigma^2$

(ii) $X_1, X_2, \ldots, X_j, \ldots$

niezależny ciąg zm. losowych o rozkł. X oraz niezależnych.

Dla każdego $j \geq 2$, weźmy

$$\bar{X}_j = \frac{1}{j} \sum_{i=1}^{j} X_i \quad \left(= \frac{X_1 + X_2 + \ldots + X_j}{j} \right)$$

oraz $\hat{S}_j^2 = \frac{1}{j-1} \sum_{i=1}^{j} (X_i - \bar{X}_j)^2$,

Zauważmy, że:

Z definicji $(j=1)$: $\hat{S}_1^2 = 0$, oraz

$$\bar{X}_{j+1} = \frac{1}{j+1}\sum_{i=1}^{j+1} X_i = \frac{1}{j+1}\left(\sum_{i=1}^{j} X_i + X_{j+1}\right)$$

$$= \frac{1}{j+1}\sum_{i=1}^{j} X_i + \frac{X_{j+1}}{j+1} =$$

$$= \left(\frac{j}{j+1}\right)\frac{1}{j}\sum_{i=1}^{j} X_i + \frac{X_{j+1}}{j+1} =$$

$$= \left(1 - \frac{1}{j+1}\right)\bar{X}_j + \frac{X_{j+1}}{j+1}, \quad \text{czyli}$$

$$\boxed{\bar{X}_{j+1} = \bar{X}_j + \frac{X_{j+1} - \bar{X}_j}{j+1}} \quad (\#)$$

Podobnie

$$\boxed{\hat{S}^2_{j+1} = \left(1 - \frac{1}{j}\right)\hat{S}^2_j + (j+1)\left(\bar{X}_{j+1} - \bar{X}_j\right)^2} \quad (\#\#)$$

Zależności $(\#)$ i $(\#\#)$ dają nam te rekurencje.

Przykład 2

Przypuśćmy, że $X_1 = 5$, $X_2 = 15$, $X_3 = 9$.

Wtedy $(\#)$ i $(\#\#)$ mają postać:

$\bar{X}_1 = 5$

$\bar{X}_2 = \bar{X}_1 + \frac{X_2 - \bar{X}_1}{2} = 5 + \frac{14-5}{2} = 5 + \frac{9}{2} = \frac{19}{2}$

$\hat{S}_2^2 = (1-\frac{1}{2})\hat{S}_1^2 + (1+1)(\bar{X}_2 - \bar{X}_1)^2$

$= 2(\frac{19}{2} - 5)^2 = \frac{81}{2}$

itd.

Sytuacja przedstawiona wyżej wygląda niemal prościej, kiedy wartości danych są binarne, czyli

X ma rozkład: $\begin{array}{c|c|c} & 0 & 1 \\ \hline & 1-p & p \end{array}$, wtedy

$EX = p$, $\operatorname{Var} X = p(1-p)$,

gdzie p jest przedmiotem estymacji ($\theta = p$)

Zatem, chcąc potrafimy wygenerować ciąg niezależnych zm. losowych X_i o rozkładzie każda $\begin{array}{c|c|c} & 0 & 1 \\ \hline & 1-p & p \end{array}$

Jeśli mamy $X_1, X_2, \ldots, X_n$ o ustalonej j.w., to

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ jest estymatorem $p = EX$,

w takim razie dla $\sigma^2 = Var(X) = p(1-p)$,

$\bar{X}_n(1-\bar{X}_n)$ jest estymatorem σ^2.

Stąd, w takim (dyskretnym) przypadku procedura jest następująca:

1). Ustalamy akceptowalny margines błędu d

2) Generujemy co najmniej 100 wartości wyników symulacji (CCTG)

3) Kontynuujemy aż do chwili k, kiedy

$$\frac{\bar{X}_k(1-\bar{X}_k)}{k} \leq d^2$$

Przykład 3.

Przypuśćmy, że obserwujemy proces obsługi klienta firmy F jak w przykładzie 1.

Interesuje nas prawdop. że o 17.30 była obsługiwana klient.

Aby to uzyskać mierzymy w kolejnych dniach obserwacji sytuację firmy. Robimy to za pomocą symulacji.

Niech $X_i := \begin{cases} 1, & i\text{-tego dnia o 17.30 } \exists \text{ klient} \\ 0, & \text{nie ma żadnego} \end{cases}$

Dla $i \geq 100$ bierzemy takie $k = i, 21$

$$\frac{p_k(1-p_k)}{k} \leq d, \text{ gdzie:}$$

d - akceptowalna wielkość rozproszenia

$p_k = \overline{X_k}$ - frekwencja dych dni, że klient był w F o 17.30.

ZADANIE 1

Zaprogramować rekurencyjnie $(\#)$ i $(\#\#)$.

ZADANIE 2

Aby estymować wielkość θ, wygenerujemy 20 wielkości zmiennych losowych o średniej θ:

102, 112, 131, 107, 114, 95, 133, 145, 139, 117
93, 119, 124, 122, 136, 141, 119, 122, 151, 143.

Sprawdzić czy ta ilość jest dostateczna, jeśli z pewnością 99% oszacujemy wartość estymowaną różni się od rzeczywistej o nie więcej niż 0,5.