Statystyka Inżynierska

dr hab. inż. Jacek Tarasiuk

AGH , WFiIS 2013

Wykład 3

Dyskretne i ciągłe rozkłady jednowymiarowe, Podstawy estymacji

Dwuwymiarowa, dyskretna funkcja rozkładu prawdopodobieństwa, Rozkłady brzegowe i warunkowe, Kowariancja i współczynnik korelacji, Korelacja i niezależność, Prosta regresji, Regresja ortogonalna, Populacja, Próba losowa, Estymator, Estymacja

Zmienna losowa dwuwymiarowa

Przez zmienną losową dwuwymiarową będziemy rozumieli taką parę funkcji $X(e)$ i $Y(e)$ określonych na przestrzeni zdarzeń elementarnych, że dla każdej pary liczb rzeczywistych x , y można określić prawdopodobieństwo zaistnienia zdarzenia A takiego że:

$$P(A) = P(X < x, Y < y)$$

Dwuwymiarowa zmienna losowa jest **zmienną dyskretną** jeśli składowe X i Y mają skończony lub przeliczalny zbiór wartości.

Rozkład i dystrybuanta

Rozkładem prawdopodobieństwa dwuwymiarowej dyskretnej zmiennej losowej nazywamy zbiór prawdopodobieństwa wszystkich zdarzeń ($X=x_i$, $Y=y_j$):

$$P\left(X = x_i , Y = y_j\right) = p_{ij}$$

Gdzie $i=1,2$ ... oraz $j=1,2$...

Z warunków normalizacji wynika, że: $\sum_i \sum_j p_{ij} = 1$

Dystrybuantą takiej zmiennej losowej będzie funkcja:

$$F\left(x, y\right) = P(X < x, Y < y)$$

Rozkłady brzegowe

Rozkłady brzegowe w przypadku dwuwymiarowej dyskretnej zmiennej losowej definiujemy jako:

$$p_{i\bullet} = \sum_k p_{ik} \qquad p_{\bullet j} = \sum_k p_{kj}$$

Rozkłady brzegowe określają prawdopodobieństwo, że jedna ze zmiennych losowych przyjmie jakąś wartość, niezależnie od tego jaką wartość przyjmuje druga zmienna losowa.

$$P(X = x_i) = p_{i\bullet} \qquad P(Y = y_j) = p_{\bullet j}$$

Rozkłady warunkowe

Rozkłady warunkowe w przypadku dwuwymiarowej dyskretnej zmiennej losowej definiujemy jako:

$$P\left(X = x_i \mid Y = y_k\right) = \frac{p_{ik}}{p_{\bullet k}}$$

$$P\left(Y = y_j \mid X = x_k\right) = \frac{p_{kj}}{p_{k\bullet}}$$

Rozkłady brzegowe określają prawdopodobieństwo, że jedna ze zmiennych losowych przyjmie jakąś wartość, w sytuacji gdy druga przyjmuje konkretną wartość (podaną w warunku).

Charakterystyki

Momentem zwykłym rzędu r+s dwuwymiarowej zmiennej losowej (X,Y) nazywamy wartość oczekiwaną iloczynu zmiennych losowych X^rY^s:

$$\alpha_{rs} = E\left(X^r Y^s\right) = \sum_i \sum_j x_i^r y_j^s p_{ij}$$

Gdy s=0 i r=1 to α_{10}=EX , czyli jest to wartość oczekiwana zmiennej X.

Gdy s=1 i r=0 to α_{01}=EY , czyli jest to wartość oczekiwana zmiennej Y.

Punkt (α_{10}, α_{01}) nazywamy **środkiem masy rozkładu prawdopodobieństwa** dwuwymiarowej zmiennej losowej (X,Y).

Gdy E(XY)=EX·EY **to zmienne X i Y są niezależne**.

Charakterystyki

Momentem centralnym rzędu r+s dwuwymiarowej zmiennej losowej (X,Y) nazywamy wartość oczekiwaną iloczynu zmiennych losowych $(X-EX)^r(Y-EY)^s$:

$$\mu_{rs} = E\left(\left(X - EX\right)^r \left(Y - EY\right)^s\right)$$

$$\mu_{rs} = \sum_i \sum_j \left(x - EX\right)^r \left(y - EY\right)^s p_{ij}$$

Gdy s=0 i r=2 to $\mu_{20}=D^2X$, czyli jest to wariancja zmiennej X.

Gdy s=2 i r=0 to $\mu_{02}=D^2Y$, czyli jest to wariancja zmiennej Y.

Gdy s=1 i r=1 to $\mu_{11}=\mathrm{cov}(X,Y)$ co nazywamy kowariancją zmiennych losowych X,Y.

Pomiędzy powyższymi wielkościami istnieją bardzo użyteczne związki:

$D^2X=E(X^2)-(EX)^2$ $D^2Y=E(Y^2)-(EY)^2$

$\mathrm{cov}(X,Y)=E(XY)-EX\cdot EY$

WYKŁAD 3

Charakterystyki

Współczynnikiem korelacji ρ dwuwymiarowej zmiennej losowej (*X*,*Y*) nazywamy:

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{D^2 X \cdot D^2 Y}}$$

Oczywiście D^2X ani D^2Y nie mogą być równe zeru.

Jeżeli ρ=1 to znaczy, że zmienne Y i X związane są ścisłą relacją Y=aX+b, przy czym a jest dodatnie. Podobnie, gdy ρ=-1, tylko a jest wówczas ujemne.

Im wartość ρ jest bliższa 1 lub -1 tym „częściej” zmienne losowe X i Y spełniają relację liniowej zależności.

Mówimy wówczas, że zmienne są silnie lub słabo **skorelowane**.

Wartość współczynnika **ρ=0** oznacza **brak korelacji**.

Korelacja i niezależność

Jeżeli E(XY)=EX·EY to zmienne X i Y są niezależne.

Ponieważ kowariancja: cov(X,Y)=E(X,Y)-EX·EY to jeżeli zmienne są niezależne, to ich kowariancja jest równa 0, w konsekwencji współczynnik korelacji ρ=0, czyli zmienne są nieskorelowane.

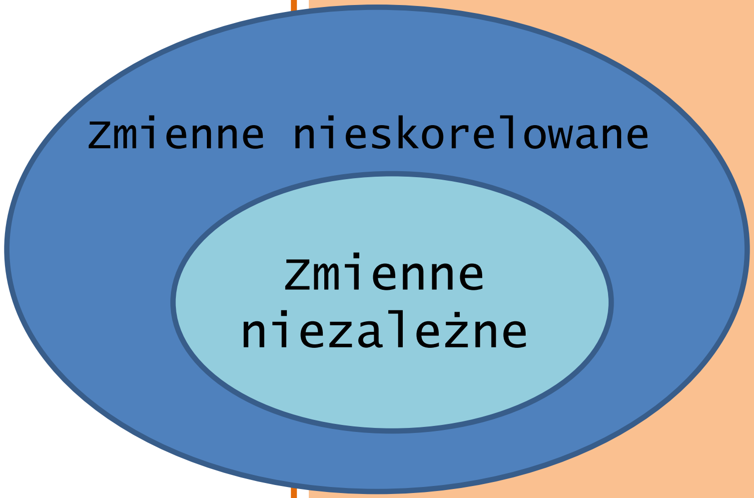
Jeżeli zmienne losowe są niezależne to są nieskorelowane, ale jeżeli są nieskorelowane to nie muszą być niezależne!

Jeżeli zmienne losowe są skorelowane, to są również zależne.

Jeżeli zmienne losowe X i Y są niezależne, to zachodzą również związki:

$$p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$$

$$F(x,y) = F_X(x) \cdot F_Y(y)$$



Prosta regresji

Prostą regresji drugiego rodzaju zmiennej losowej Y względem zmiennej losowej X nazywamy prostą o równaniu y=ax+b, której współczynniki a i b są tak dobrane, aby średnie odchylenie kwadratowe zmiennej losowej Y od zmiennej losowej aX+b było najmniejsze:

$$E\left[Y-\left(aX+b\right)\right]^2=\min$$

Można wykazać, że współczynniki takiej prostej dadzą się wyliczyć jako:

$$a=\rho\sqrt{\frac{D^2Y}{D^2X}}=\rho\frac{\sigma_Y}{\sigma_X}$$

$$b=EY-a\cdot EX$$

Regresja ortogonalna

Dla dowolnych zmiennych X i Y istnieje zawsze przekształcenie liniowe, sprowadzające te zmienne do postaci, w której współczynnik korelacji pomiędzy zmiennymi równa się zeru.

Przekształcenie to ma postać:

$$X^* = (X - EX)\cos\varphi + (Y - EY)\sin\varphi$$

$$Y^* = -(X - EX)\sin\varphi + (Y - EY)\cos\varphi$$

przy czym:

$$tg(2\varphi) = \frac{\operatorname{cov}(X, Y)}{D^2X - D^2Y}$$

$$y - EY = tg\,\varphi(x - EX)$$

Próba losowa i populacja

- Badania wyczerpujące (całkowite) i niewyczerpujące (częściowe).
- Populacja generalna i populacja próbna.
- Próba reprezentatywna:
  - losowy wybór z populacji generalnej
  - dostateczna liczność próby
- Oceny parametru populacji generalnej dokonujemy na podstawie pomiarów w populacji próbnej.

Prawdziwość oceny parametru populacji generalnej na podstawie próby jest zdarzeniem losowym.

Wartość tego prawdopodobieństwa zależy od wielkości próby

oraz

od dokładności oceny.

Parametr rozkładu i jego estymator

Parametr	Estymator

Wartość oczekiwana

$$EX = \sum_i x_i \cdot p_i \qquad \overline{X} = \frac{1}{n}\sum_i X_i$$

Wariancja

$$D^2 X = \sum_i \left(x_i - EX\right)^2 \cdot p_i \qquad S^{*2} = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$$

Prawdopodobieństwo sukcesu (wskaźnik struktury)

$$p = \frac{k}{n} \qquad \hat{p} = \frac{k}{n}$$

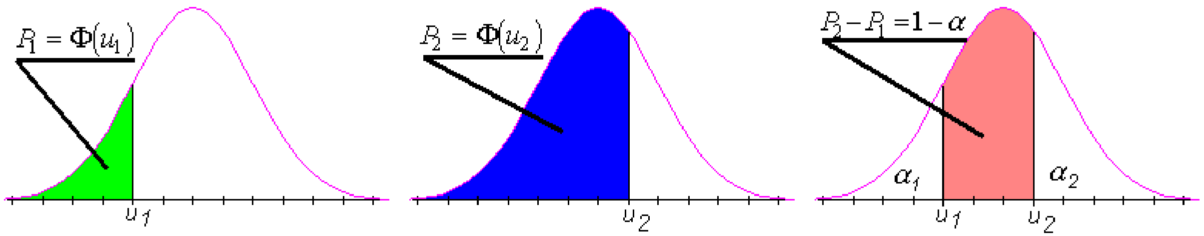
Estymacja przedziałowa

Jeśli X ma rozkład $N(\mu,\sigma)$, to $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$ ma rozkład $N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$

Po znormalizowaniu, zmienna $U=\frac{\overline{X}-\mu}{\sigma}\cdot\sqrt{n}$ będzie podlegać rozkładowi N(0,1).

Możemy znaleźć nieskończenie wiele przedziałów, w których zmienna ta znajdzie się z prawdopodobieństwem 1-α :

$$P(u_1<U<u_2)=1-\alpha=\Phi(u_2)-\Phi(u_1)$$



$$\alpha_1+\alpha_2=\alpha \qquad u_1=u(\alpha_1) \qquad u_2=u(1-\alpha_2)$$

$$P\left(u(\alpha_1)<\frac{\overline{X}-\mu}{\sigma}\cdot\sqrt{n}<u(1-\alpha_2)\right)=1-\alpha$$

$$P\left(\overline{X}-u(1-\alpha_2)\cdot\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}-u(\alpha_1)\cdot\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

Estymacja przedziałowa

Tylko jeden z tych przedziałów będzie symetryczny. Taki dla którego: $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, a wówczas możemy powiedzieć, że przedział:

$$\left(\overline{X} - u\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + u\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}} \right)$$

z prawdopodobieństwem 1-α pokrywa prawdziwą wartość oczekiwaną μ dla badanego rozkładu.

Jeśli nie znamy wariancji σ^2 to musimy ją estymować:

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

i wówczas przedział wygląda następująco:

$$\left(\overline{X} - u\left(1 - \frac{\alpha}{2}\right) \cdot \frac{S^*}{\sqrt{n}}, \overline{X} + u\left(1 - \frac{\alpha}{2}\right) \cdot \frac{S^*}{\sqrt{n}} \right)$$

Oba powyższe przedziały to tzw. przedziały ufności dla wartości oczekiwanej rozkładu na poziomie ufności 1- α.