Zadania ze statystyki cz.3 I rok Praca socjalna

Zadanie 1.

Tabela poniżej prezentuje rozkład dwóch zmiennych: X - wielkości gospodarstw domowych oraz Y - miesięcznych wydatków na kulturę. Proszę ustalić, czy istnieje związek między wielkością gospodarstwa domowego a wydatkami na kulturę, jego kierunek oraz siłę. Proszę przedstawić relacje między wielkością gospodarstwa domowego a wydatkami na wykresie układu współrzędnych, gdzie na osi X zostanie zaprezentowany rozkład zmiennej: wielkość gospodarstwa domowego.

x	y
4	300
5	250
3	320
6	200
4	280
7	140
4	230

Zadanie 2.

W badaniach nad gospodarstwami domowymi dodatkowo zajęto się problemem związku wysokości wydatków na kulturę(zmienna Y) a wiekiem głowy gospodarstwa domowego (zmienna X) oraz wykształceniem, mierzonym liczba lat poświęconych na naukę (zmienna Z). Proszę ustalić, która zmienna, wiek głowy czy wykształcenie (X czy Z) jest silniej związane z wydatkami na kulturę. Proszę również ustalić charakter, kierunek oraz stopnień zdeterminowania związków (współczynnikiem determinacji) między zmiennymi X a Y oraz Z a Y.

y	x	z
300	35	21
250	37	12
320	28	17
200	41	11
280	55	15
140	63	8
230	42	13

Zadanie 3.
Proszę przeprowadzić analizę regresji (zbudować funkcję regresji liniowej) dla zmiennych z zadania 2 (mamy otrzymać dwie funkcje regresji dla pary zmiennych Y i X oraz Y i Z). Następnie proszę na podstawie funkcji regresji ustalić, jakie wydatki na kulturę będzie ponosić gospodarstwo z głowa rodziny w wieku 46 lat oraz jakie wydatki na kulturę będzie ponosić gospodarstwo z głową rodziny, która uczyła się przez 14 lat.

PODSUMOWANIE WYKŁADU

- Jeżeli relacja między dwoma zmiennymi ma charakter prostoliniowy, to siłę i kierunek tego związku możemy badać **kowariancją** oraz **współczynnikiem korelacji r Pearsona**.
- Kowariancja bada, czy dwie cechy wykazują się podobnym tempem zmiany wokół swoich średnich.

$$cov(XY) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- Kowariancja jest miarą statystyczną związku, która wskazuje czy cechy są ze sobą związane.
- Związek miedzy zmiennymi występuje, gdy kowariancja jest różna od zera

$$cov_{(x,y)} \neq 0$$

Przykład

- Badaczy interesował problem czy istnieje związek między wzrostem a wagą uczniów w pewnej szkole. W tym celu dokonano stosownych pomiarów. Poniższa tabela prezentuje pomiary pięciu uczniów. Czy występuje związek miedzy wzrostem a wagą?

wzrost (X)	waga (Y)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
120	30	-15	-4	60
135	32	0	-2	0
125	34	-10	0	0
150	36	15	2	30
145	38	10	4	40
135	**34**			130

$$cov(XY) = \frac{130}{5 - 1} = 32{,}5$$

Współczynnik korelacji liniowej r Pearsona

- Współczynnik korelacji r Pearsona standaryzuje kowariancje. Dzięki temu czyni ją niewrażliwą na jednostki pomiarowe cech.
- Współczynnik r Pearsona dzieli kowariancję przez iloczyn odchyleń standardowych obu zmiennych

$$r_{(x,y)} = \frac{cov(XY)}{s_x \times s_y}$$

gdzie:

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}} \quad ; \quad s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N - 1}}$$

- Wadą współczynnika r Pearsona jest ograniczenie jego zastosowanie do badania związków prostoliniowych.
- Siłę związku odczytujemy poprzez wartość współczynnika. Zakres w jakim może mieścić się wartość r Pearsona wynosi od -1 do +1.
- Wartość współczynnika r Pearsona równa 0 oznacza brak związku
- Kierunek związku określa znak przy współczynniku (dodatni lub ujemny).

Przykład
Jaka jest siła związku między wagą a wzrostem uczniów?

wzrost (X)	waga (Y)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
120	30	-15	-4	60
135	32	0	-2	0
125	34	-10	0	0
150	36	15	2	30
145	38	10	4	40
135	**34**			130

$$cov(XY) = \frac{130}{5-1} = 32{,}5$$

$$s_x = 12{,}75 \quad s_y = 3{,}16$$

$$r_{(x,y)} = \frac{32{,}5}{12{,}75 \times 3{,}16} = 0{,}8$$

Inne sposoby ustalania wartości współczynnika korelacji r Pearsona

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}}$$

$$r = \frac{\frac{1}{N}\sum(x_i y_i) - \sum \bar{x} \sum \bar{y}}{\sqrt{\left(\frac{1}{N}\sum(x_i - \bar{x})^2\right)\left(\frac{1}{N}\sum(y_i - \bar{y})^2\right)}}$$

Współczynnik determinacji związku

- Współczynnik determinacji związku jest miarą informującą o stopniu w jakim możemy wyjaśnić (przewidzieć) zmiany wartości zmiennej zależnej przez zmienną niezależną.
- Wyrażona procentem wskazuje na procent zmienności (wariancji) zmiennej zależnej wyjaśnionej zmiennością (wariancją) zmiennej niezależnej.

$$WD = r^2 \cdot 100\%$$

Koncepcja liniowości związku – regresja liniowa

- Jeżeli punkty na wykresie wyznaczające pozycje wszystkich badanych ułożyły się w taki sposób, że możliwe stało się poprowadzenie linii prostej pomiędzy lub przez te punkty, to badana zależność ma charakter liniowy i można ją opisać za pomocą funkcji liniowej w układzie współrzędnych.

$$y = ax + b$$

- Analiza regresji liniowej to procedura dopasowania linii prostej do danych, dzięki której każdej wartości X odpowiadałoby najlepsze dopasowanie wartości Y.
- Odszukanie „najlepiej pasującej prostej" odbywa się **metodą najmniejszych kwadratów**.
- Każdy pomiar naszej próby łączymy odcinkami prostopadłymi do osi x z hipotetyczną linią regresji . Mamy zatem dla każdej wartości cechy X dwie wartości zmiennej y: empiryczną y (pochodzącą z pomiarów) oraz y' -wartości zmiennej leżącej na prostej.
- Dążymy do tego aby różnice między y a y' były najmniejsze.

$$\sum (y - y')^2 = min$$

Najlepsze estymatory współczynników a i b obliczamy

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$