Zadania ze statystyki cz.2 I rok socjologii grup dyspozycyjnych

Zadanie 1.

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

|  |  |
| --- | --- |
|  |  |
| **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) oraz ustalić wielkość błędu standardowego (reszt) tego modelu. 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.

Zadanie 4

Poniższa tabela prezentuje rozkład wydatków na kulturę i wieku głowy gospodarstwa domowego. Należy wyrangować te cechy i następnie obliczyć współczynnik korelacji rang Spearmana.

|  |  |
| --- | --- |
| **y** | **x** |
| 300 | 32 |
| 250 | 28 |
| 320 | 29 |
| 200 | 44 |
| 280 | 52 |
| 140 | 60 |
| 230 | 40 |

Zadanie 5

Tabela poniżej prezentuje szereg korelacyjny dwóch cech: x – liczba godzin przeznaczonych na naukę; y – ocena ze sprawdzianu. Proszę zbadać siłę i kierunek związku miedzy tymi zmiennymi, wykorzystując współczynnik korelacji rang Spearmana.

|  |  |
| --- | --- |
| x | y |
| 3 | 3 |
| 4 | 3,5 |
| 5 | 4 |
| 6 | 4 |
| 6 | 4,5 |
| 7 | 3,5 |
| 7 | 4 |
| 7 | 4,5 |
| 9 | 5 |

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.



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

 

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?





**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

 

gdzie:

  ; 

* 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?





  



Inne sposoby ustalania wartości współczynnika korelacji r Pearsona





**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.

 

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.

 

* 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.

 

Najlepsze estymatory współczynników a i b obliczamy





Regresja liniowa – błąd standardowy pomiaru

* Po wyprowadzeniu funkcji liniowej, każdemu X możemy przypisać dwie wartości Y - wartość empiryczną (zbadaną) *yi* oraz teoretyczną *y’*(wynikającą z danej funkcji y=ax+b).
* Błąd standardowy modelu regresji „e”, to różnice między wartościami empirycznymi a teoretycznymi.



Współczynnik korelacji rang Spearmana

* Jest szczególnym przypadkiem współczynnika korelacji liniowej Pearsona. Różnica między tymi współczynnikami jest tym większa, im większa jest krzywoliniowość badanego związku.
* Dedykowany jest związkom krzywoliniowym.
* Opiera się na rangach wartości zmiennej niezależnej i zależnej.
* Gdy każda ze zmiennych posiada wartości niepowtarzalne (nie ma dwóch obserwacji o takich samych wartościach dwóch wśród zmiennych), a tym samym rangi w obu zmiennych są niepowtarzalne, korelację między zmiennymi obliczymy”:

Nadawanie rang – rangowanie cech

* Rangowanie to inaczej uporządkowywanie cech ze względu na wartości cechy w taki sposób, że każdej obserwacji nadaje się numer porządkowy, wskazujący na miejsce w strukturze wyrażające natężenie danej cechy.



Rangowanie zmiennej – rangi wiązane

* Jeżeli w szeregu występują jednakowe pomiary (obserwacje, które miały identyczne wartości danej zmiennej) przypisane różnym obserwacjom, to każdej z takich obserwacji musimy przypisać jednakową rangę (numer porządkowy).
* Sytuację, w której co najmniej dwie różne obserwacje otrzymały taką samą rangę nazywamy wiązaniem, taki rozkład szeregiem powiązanym rangami.
* W przypadku takich rozkładów musimy posługiwać się współczynnikami dla zmiennych powiązanych rangami.



* Sytuację gdy obserwacje posiadają powtarzające się wartości zmiennej X lub Y, a tym samym powtarzają się rangi tych zmiennych, nazywamy pomiarami **powiązanymi rangami**, lub **rangami wiązanymi**.
* W takiej sytuacji musimy uwzględnić ilość obserwacji, które są powiązane rangami.









Przykład

Zmienna X prezentuje dochód na osobę w rodzinie, zmienna Y przeciętne wydatki na używki. Czy wysokość dochodu jest związana z kwotą przeznaczaną na używki?

 



Ponieważ pomiary są niepowtarzalne, to również rangi są niepowiązane, a zatem stosujemy:







Przykład 2

Zmienna X prezentuje dochód na osobę w rodzinie, zmienna Y przeciętne wydatki na używki. Czy wysokość dochodu jest związana z kwotą przeznaczaną na używki?

Tym razem występują powtarzalne pomiary (np. dwa razy kwota 1200 i dwa razy kwota 1400 oraz wydatki 300 zł ). Występują więc rangi wiązane.



Występują rangi wiązane, a zatem rhos obliczamy:



Uwzględniając obserwacje powiązane rangami w każdej zmiennej (X i Y)















A zatem:



