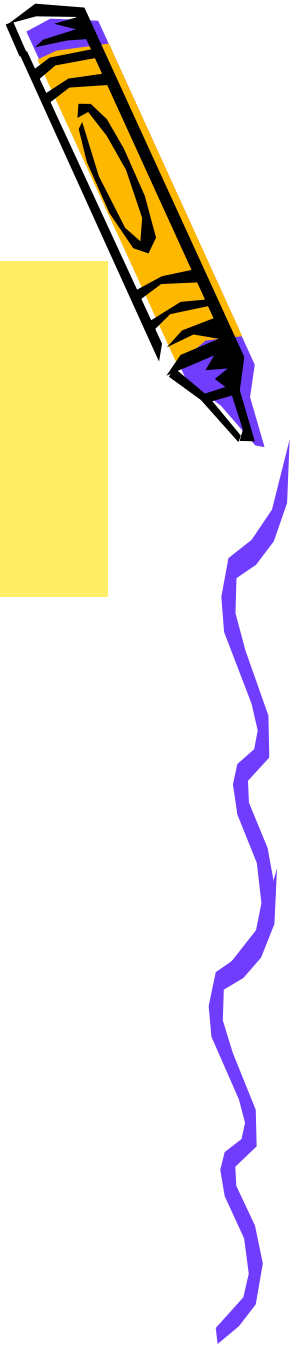


Ekonometria

Wykład 3

dr hab. Małgorzata Radziukiewicz, prof. PSW Biała Podlaska

Określenie badanego zjawiska

- I etap budowy modelu to sprecyzowanie zakresu badania i w związku z tym podjęcie decyzji, która zmienna bedzie traktowana jako zmienna objaśniana, a jakie zmienne odgrywać będą w modelu rolę zmiennych obajśniających.

Określenie badanego zjawiska

- W I etapie badania możliwe są dwa przypadki:
  1. Teoria ekonomii dostarcza wiadomości na temat badanego zjawiska (procesu) i czynników kształtujacych wielkość zmiennej objaśnianej Y;
  2. Teoria ekonomii nie stanowi o badanym zjawisku (procesie).

Określenie badanego zjawiska

- W praypadku 1:

„zapotrzebowanie", czyli potrzeba wyjaśnienia mechanizmu kształtowania się pewnego zjawiska bądź procesu ekonomicznego definiuje zmienną objaśnianą Y i listę głównych czynników (zmiennych obajśniających X_i);

W rezultacie ustala się zbiór potencjalnych zmiennych objaśniających $V = \{ X_1, X_2, ...X_k \}$;

W zbiorze V wyróżnia się:

- zmienne mierzalne,
- zmienne niemierzalne (np. dobrobyt, jakość wyrobu, kwalifikacje, płeć),
- zmienne zero-jedynkowe.

Określenie badanego zjawiska

- W praypadku 2, kiedy teoria ekonomii nie stanowi o badanym zjawisku, wybieramy jako zmienne obajśniające te zmienne:

- które **są silnie skorelowane ze zmienną obajśnianą i słabo skorelowane między sobą**;
- **mają** również **interpretowalny związek ze zmienną obajśnianą Y.**

Dane statystyczne

II-gi etap budowy modelu to zebranie danych statystycznych, na podstawie których będzie można oszacować model (oszacować parametry strukturalne i parametry stochastycznej struktury modelu).

Należy pamiętać o tym, iż równania zbudowane w etapie I-ym są o tyle abstrakcją, że ich parametry są nieznane i że poszczególne równania postulują jedynie pewne relacje między zmiennymi, nie informując jednak w sposób konkretny o liczbowych proporcjach w jakich zmiana wartości jednej ze zmiennych objaśniających wpływa na zmianę wielkości zmiennej objaśnianej.

Dążeniem ekonometryka jest zgromadzenie możliwie bogatego oraz dokładnego materiału statystycznego, gdyż im większa jest liczba obserwacji statystycznych, tym większa jest dokładność, z jaką można oszacować parametry modelu.

Z drugiej strony trzeba pamiętać o tym, że zbytnie powiększanie liczby obserwacji wiąże się z rozszerzeniem odcinka czasu, z którego pochodzą dane, a to z kolei stwarza niebezpieczeństwo utraty porównywalności danych, jeżeli w rozpatrywanych zjawiskach zachodzą w czasie jakieś istotne zmiany strukturalne.

Rodzaje i źródła danych

- **Źródła danych**:
- opisy zasad funkcjonowania obiektu:
  - przepisy;
  - regulaminy wewnętrzne;
  - dane dotyczące procesów technologicznych.
- bieżąca rejestracja zdarzeń (np. rejestr kosztów w przedsiębiorstwie);
- sprawozdania (np. ze sprzedaży, zatrudnienia, wydatkowania dochodów);
- spisy (np. maszyn, zapasów);
- zapisy wyników badań specjalnych (np. jakość wyrobów produkowanych przez firmę).

Dane statystyczne w badaniach

Do najważniejszych źródeł danych należy przede wszystkim:

- ewidencja gospodarcza
- sprawozdawczość
- badania ankietowe

Dane statystyczne w badaniach

- **Ewidencja gospodarcza** jest jest podstawowym źródłem informacji ekonomicznej w jednostkach gospodarczych.
- Rodzaje ewidencji gospodarczej:
- **ewidencja operatywna** – system bieżących zapisów służących do obserwacji, pomiaru, rejestracji i grupowania poszczególnych zjawisk związanych z działalnością jednostki gospodarującej;
- **ewidencja księgowa** – księgowość prowadzona w sposób systematyczny na podstawie danych ewidencji operatywnej, rejestruje w odpowiednich przekrojach (metodą bilansową) wyrażone wartościowo dane liczbowe dotyczące występujących w jednostce gospodarującej: stanu środków, źródeł ich pochodzenia, ruchu środków oraz wyników działalności;
- **ewidencja statystyczna** – na podstawie ewidencji operatywnej i księgowej dostarcza informacji w postaci rozmaitych wskaźników ekonomicznych charakteryzujących działalność gospodarczą.

Podstawowe klasyfikacje sprawozdawczości:

- **częstotliwość sprawozdań:**
  - sprawozdawczość **operatywna** (sprawozdania zestawiane z dużą częstotliwością np. dzienne, tygodniowe, dekadowe),
  - sprawozdawczość **okresowa** (sprawozdania sporządzane na odpowiednie okresy np. miesiące, kwartały, lata),
  - sprawozdawczość **sporadyczna** (sprawozdania zestawiane doraźnie dla celów odbiorcy).

Podstawowe klasyfikacje sprawozdawczości:

- **odbiorcy sprawozdań:**
  - **sprawozdawczość wewnętrzna** – odbiorcami są zarządzający jednostką
    - na potrzeby zarządzania tym obiektem;
    - na potrzeby prognozowania.
  - **sprawozdawczość zewnętrzna** – dane, których zakres nie zależy do obiektu a dotyczą
    - otoczenia bliższego (np. dostawcy, klienci, konkurenci, pośrednicy, oferty sprzedaży);
    - otoczenia dalszego (instytucje krajowe i międzynarodowe typu administracyjnego i gospodarczego np. sejm – ustawy, uchwały, banki – stopy %, giełdy – kursy akcji, wprowadzanie do obrotu nowych spółek, GATT i EWG – umowy celne).

Podstawowe klasyfikacje sprawozdawczości: (c.d.)

- **obowiązek sporządzania** sprawozdań:
  - sprawozdawczość **obligatoryjna** – sporządzana na mocy odpowiednich przepisów prawa;
  - sprawozdawczość **fakultatywna** – na wewnętrzne potrzeby jednostki gospodarczej.

Podstawowe klasyfikacje sprawozdawczości: (c.d.)

- **przedmiot sprawozdawczości**:
  - sprawozdawczość **rzeczowa** – obejmuje dane liczbowe dotyczące rzeczowych mierników działalności jednostki wyrażone w jednostkach naturalnych;
  - sprawozdawczość **finansowa** – dane liczbowe dotyczące wartościowych mierników działalności gospodarczej (dane z ewidencji księgowej).

Klasyfikacja danych wykorzystywanych w badaniach (m.in. ekonometrycznych)

- - dane **dynamiczne**;

szeregi czasowe – dane statystyczne dotyczą wielu okresów czasu

- - dane **przekrojowe**;

Dane o charakterze statycznym, bowiem ilustrują wyniki badania pewnej zbiorowości w jednym momencie lub okresie czasu (np. BGD)

- - dane **dynamiczno-przekrojowe**.

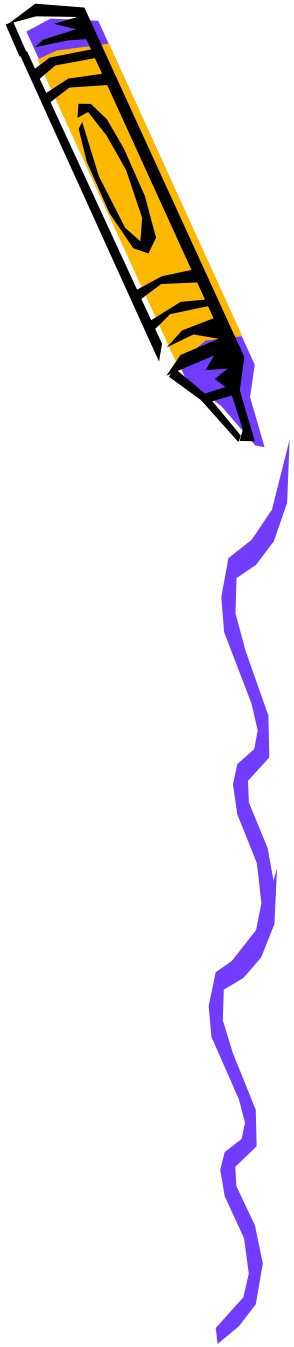
- BDD są powtarzane co roku i w rezultacie są to dane przekrojowe w kolejnych latach

Klasyfikacja danych c.d.

- Z punktu widzenia skali jednostek, do których się odnoszą:

- dane **mikroekonomiczne** (przedmiotem zainteresowania są pewne prawidłowości ilościowe zachodzące na szczeblu najmniejszych podmiotów występujących w gospodarce narodowej np. przedsiębiorstw, gospodarstw domowych, konsumentów;
- -dane **makroekonomiczne** ilustrujące zjawiska w skali gałęzi całej gospodarki narodowej, w skali regionu (województwa, powiatu, gminy) czy w skali całego kraju).

Dane statystyczne c.d.



Przykłady danych statystycznych

- Jednowymiarowy szereg czasowy.
- Wielowymiarowy szereg czasowy.
- Jednowymiarowy szereg przekrojowy.
- Wielowymiarowy szereg przekrojowy.
- Szereg przekrojowo-czasowy.

Przykłady

Tablica 1. Średni miesięczny dochód na 1 osobę w gospodarstwach domowych ogółem

Wyszczególnienie		1994	1995	1996	1997	1998	1999	2000
dochód na osobę w złotych	ceny bieżące	258,14	337,80	443,90	533,74	590,57	641,07	697,55
	ceny stałe 1994 r.	258,14	260,40	293,20	307,09	304,47	307,73	303,30
	ceny stałe 2000 r.	593,69	612,69	674,32	706,26	700,24	707,74	697,55

Przykłady

Tablica 2. Wybrane dane dotyczące województwa dolnośląskiego (1995 - 2001)

Dolnośląskie	bezrobotni			pracujący w tysiącach			
Lata	ogółem	w wieku <=25	ludność	ogółem	w rolnictwie	w przemyśle i budownictwie	w usługach
1995	189	47	2988278	1039	124	396	520
1996	157	38	2986884	1085	135	389	561
1997	176	32	2985381	1117	130	404	582
1998	161	36	2982128	1177	112	426	639
1999	193	50	2977611	1120	108	375	638
2000	284	66	2972667	972	98	321	554
2001	290	68	2101654	914	99	288	527

Przykłady

Tablica 3. Wybrane dane dotyczące województwa dolnośląskiego

dolnośląskie	przeciętne			
dane kwartalne	zatrudnienie w sektorze przedsiębiorstw w tysiącach	miesięczne wynagrodzenie brutto w sekt. przedsiębiorstw w złotych	realne miesięczne wynagrodzenie brutto w sekt. przedsiębiorstw w złotych	zatrudnienie w przemyśle w tysiącach
I kw 1999	400684	1628,36	1628,36	236069
II kw.	399960	1756,60	1756,60	232261
III kw.	397724	1757,38	1757,38	230136
IV kw.	393748	1909,74	1909,74	221754
I kw. 2000	370318	1820,05	1809,11	215484
II kw.	371430	1978,23	1965,64	214584
III kw.	373292	1959,38	1950,69	208632
IV kw.	371424	2131,17	2129,36	210484
I kw. 2001	358038	2008,66	1997,26	206364
II kw.	356176	2050,01	2036,40	204086
III kw.	354455	2185,33	2175,02	200098
IV kw.	352591	2299,38	2297,58	198340
I kw.2002	339607	2077,24	2065,16	189978
II kw.	337801	2108,76	2095,45	187042
[illegible]	337165	2217,44	2207,65	187667
[illegible]	336815	2373,15	2371,10	186957

Gromadzenie danych

Wiadomości te są upowszechniane poprzez:

- mass media;
- literaturę specjalistyczną – dzienniki i czasopisma poświęcone problematyce gospodarczej („Gazeta Bankowa", Życie Gospodarcze", „Business", „Puls Business'u", „Polska XXI") oraz pisma koncentrujące się na gospodarce („Przegląd Statystyczny", „Economic Forecasts" i inne).
- agendy rządowe:

  GUS https://www.stat.gov.pl;

  Ministerstwo Finansów https://www.mf.gov.pl;

  Ministerstwo Rodziny, Pracy i Polityki Społecznej https://www.mpips.gov.pl

  Ministerstwo Inwestycji i Rozwoju https://www.miir.gov.pl

  Narodowy Bank Polski https://wwww.nbp.pl

- organizacje przedsiębiorców Business Center Club, Lewiatan
- instytuty naukowe – IPiSS, PIE, szkoły wyższe

Wybór zmiennych objaśniających

- W specyfikacji zmiennych chodzi o wybór właściwych zmiennych objaśniających tj. takich, których łączny wpływ na kształtowanie się zmiennej objaśnianej jest na tyle znaczny, że umożliwia praktyczne zastosowanie modelu do analizy danego zjawiska oraz przewidywania kierunków jego rozwoju

Dobór zmiennych obajśniających

może być dokonany na podstawie:

1. informacji a priori, a więc np. na podstawie teorii ekonomii (**metoda delficka**);

2. przy zastosowaniu jednej z procedur wybierających optymalny zbiór zmiennych z ustalonej listy potencjalnych zmiennych:

- badanie pojemności nośników informacji **metodą Hellwiga**,
- **metodą grafu**.

Dobór zmiennych obajśniających

- Warunkiem wstępnym do tego, by dana zmienna X_i mogła być uznana za zmienną objaśniającą w modelu, jest jej wystarczające zróżnicowanie;
- Zmienną obajśniajacą nie może być zmienna, której poszczsególne obserwacje nie różnią się między sobą (są stałe lub *quasi*-stałe);
- Do mierzenia zróżnicowania wykorzystuje się klasyczny współczynnik zmienności:

$$V(x) = \frac{s(x)}{\bar{x}} \cdot 100\%$$

gdzie: s(x) odchylenie standardowe zmiennej X_i $\bar{X}$ – średnia arytmetyczna

Zwykle obiera się krytyczną wartość współczynnika zmienności V^* (np. $V^* = 0{,}1$). Zmienne spełniające nierówność $V_i < V^*$ uznaje się za mało zróżnicowane

Podstawą wyboru zmiennych objasniających do modelu ekonometrycznego jest

analiza korelacji

- Poszczególne jednostki populacji mogą być badane:
  - ze względu na **jedną cechę**;
  - jednocześnie ze względu na **dwie** lub **więcej cech**.

Przykład 1.

Gospodarstwa domowe mogą być badane nie tylko ze względu na wysokość miesięcznych dochodów, lecz również ze względu na liczbę osób w gospodarstwie, wiek głowy gospodarstwa, wysokość miesięcznych wydatków, liczbę osób pracujących, czy stosowany lek (wielkość dawki) ma wpływ na stan zdrowia itp.

- Inaczej mówiąc możemy badać populację ze względu na **m** cech. Wektor cech zapisujemy:

$$\mathbf{x} = [x_1, x_2, \ldots, x_m]$$

Przykład 2.

- Studenci statystyki PSW w Białej Podlaskiej byli badani ze względu na wagę (x). Teraz mogą być badani nie tylko ze względu na wagę (x_1), lecz również według wzrostu (x_2), wieku (x_3), płci (x_4), charakteru studiów (dzienne, zaoczne) (x_5) itp.

- Poszczególne cechy mogą być:
  - od siebie odizolowane;
  - wzajemnie ze sobą powiązane.

- Dział statystyki zajmujący się badaniem związków między kilkoma cechami (zmiennymi) nosi nazwę **teorii współzależności.**

- **Wykrycie zależności między cechami nie jest łatwe, nawet jeśli ich występowanie wydaje się oczywiste.**

  Przykład 3.

  - chociaż dany lek jest bardzo dobry, to jednak nie dla każdej osoby będzie skuteczny;
  - chociaż dane gospodarstwo ma wysoki dochód, to nie koniecznie musi dużo wydawać na dobra luksusowe, itp..

- **Występowanie zależności można wykryć tylko przez obserwację większej liczby przypadków**.

  Przykład 4.

  - chorzy, którzy zażywają skuteczny lek są częściej wyleczeni, niż ci, którzy go nie przyjmują;
  - gospodarstwa z wysokimi dochodami wydają przeciętnie więcej na dobra luksusowe niż ubogie gospodarstwa;
  - określona liczba studentów poświęca tę samą ilość czasu na przygotowanie się do egzaminu, ale uzyskane wyniki są różne;
  - działki zasilamy tą samą dawka nawozu, ale w efekcie możemy mieć różne plony itp..

- **Zaprezentowane w przykładzie 3 związki cech (zmiennych) są stochastyczne.**

Współzależność zjawisk

- **współzależność funkcyjna** – zmiana wartości jednej zmiennej (X) powoduje ściśle określoną zmianę drugiej zmiennej (Y). Oznacza to, że zmiennej X odpowiada tylko jedna wartość zmiennej Y np. pole kwadratu jest funkcją jego boku, czyli $P = a^2$ (wszystkie kwadraty o boku a maja takie samo pole);
- **współzależność stochastyczna** – wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem jest **zależność korelacyjna.**

- Stochastyczny związek cech można prezentować tabelarycznie.
- Tablicę ujmującą ten związek nazywa się **tablicą korelacyjną** (*łac. corelatio*: współzależność, wzajemny stosunek).
- przyjmujemy zasadę: Y – cecha zależna; X – cecha niezależna (lub odwrotnie), a więc mówiąc o związku cech, rozumiemy **związek 2-óch cech**.
- W tablicy korelacyjnej mamy s + r szeregów rozdzielczych warunkowych oraz 2 szeregi rozdzielcze główne (brzegowe).
- Wszystkie rozkłady są jednowymiarowe (zastosowanie mają uprzednio poznane statystyczne miary opisu dotyczące jednej cechy)

x_i \ y_i	y_1	y_2	…	y_s	$\sum_j$
x_1	n_{11}	n_{12}	…	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2s}	$n_{2\bullet}$
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rs}	$n_{r\bullet}$
$\sum_i$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet s}$	n

x przyjmuje r wariantów - i = 1,2,3,4…r (odmiany cechy niezależnej)
y przyjmuje s wariantów - j = 1,2,3,4,…s (odmiany cechy zależnej)

Przykład 5.
Wydajność pracy Y (w tys. sztuk wyrobów na osobę) oraz staż pracy X (w latach) pracowników w zakładzie A przedstawia tablica 1.
Tablica 1.

- n_{ij} – liczba jednostek, które posiadają jednocześnie wariant x_i cechy X oraz wariant y_j cechy Y

x_i \ y_i	1 - 3	3 - 5	5 - 7	7 - 9	Razem
0 - 2	6	4	-	-	10
2 - 4	2	10	-	-	12
4 - 6	-	8	16	12	36
6 - 8	-	4	18	20	42
Razem	8	26	34	32	100

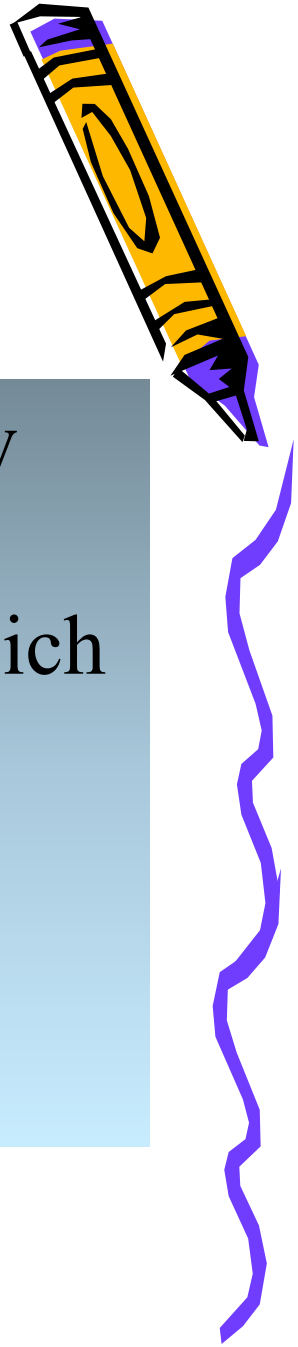
- I tak np. liczbę 20 (znajdująca się w dolnym prawym rogu) można interpretować jako liczbę osób o wydajności w granicach 7 – 9 tys. sztuk wyrobów i o stażu pracy od 6 do 8 lat.

- Tablica korelacyjna, którą budujemy zazwyczaj według uporządkowania cechy niezależnej (X), może być także czytana „odwrotnie”, jeśli zamiana cech ma sens z merytorycznego punktu widzenia.

  Przykład 6.

  Interesuje nas związek między liczbą osób w gospodarstwie domowym a spożyciem mleka.

  W tym przypadku liczba osób wpływa na spożycie mleka, ale nie na odwrót. Zatem spożycie mleka będzie zmienną zależną (Y) a liczba osób w gospodarstwie zmienną niezależną (X).



- Poza tabelaryczną prezentacją związków stochastycznych (w postaci tablicy korelacyjnej) istnieją graficzne sposoby ich obrazowania.

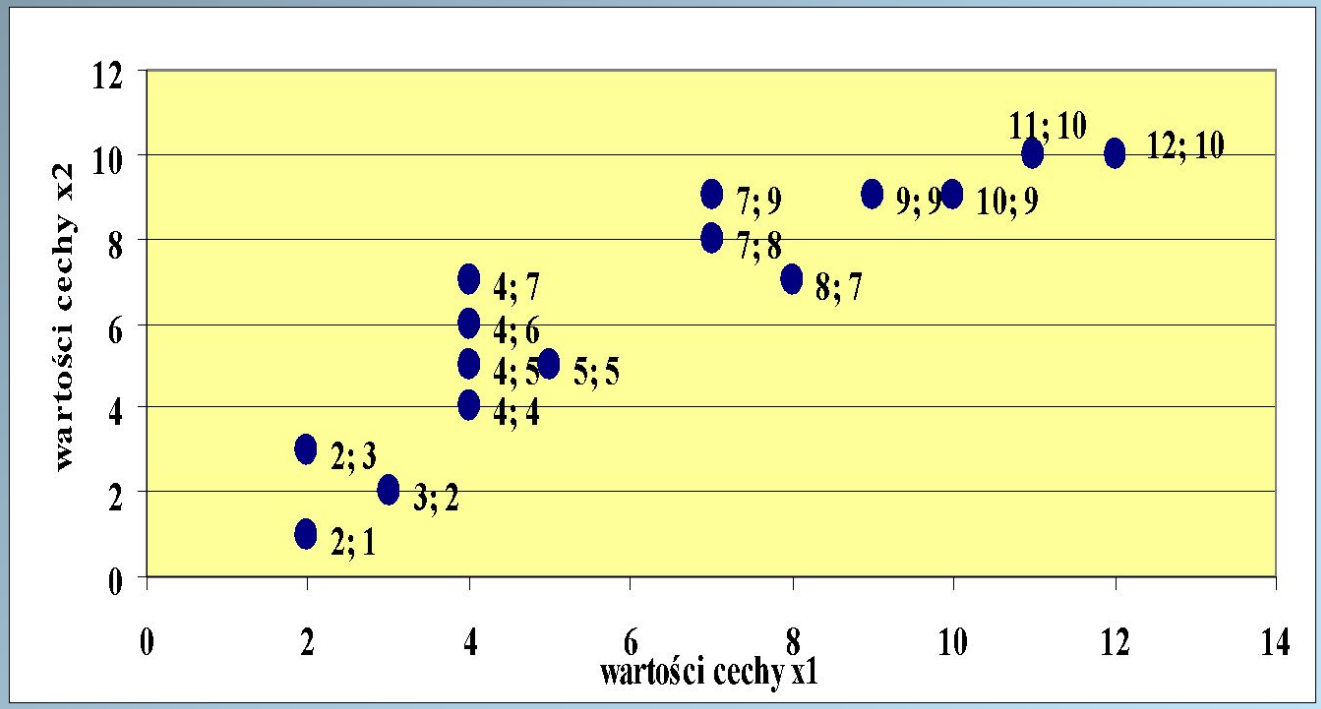
Badanie populacji na 2 cechy

- *Przykład 7.*
- Załóżmy, że populacja studentów (n = 15) jest opisywana za pomocą dwóch cech **(x_1)** i **(x_2)**, tzn. **m = 2, n = 36.**
- Wtedy macierz obserwacji ma wymiary **n** x **m** (36 x 2), a **i - ta** obserwacja opisywana jest parą liczb $\mathbf{x_{i1}}$ oraz $\mathbf{x_{i2}}$.
- W układzie współrzędnych odpowiada to punktowi $\mathbf{p_i} = [\mathbf{x_{i1}}, \mathbf{x_{i2}}]$. Mamy więc 15 punktów.

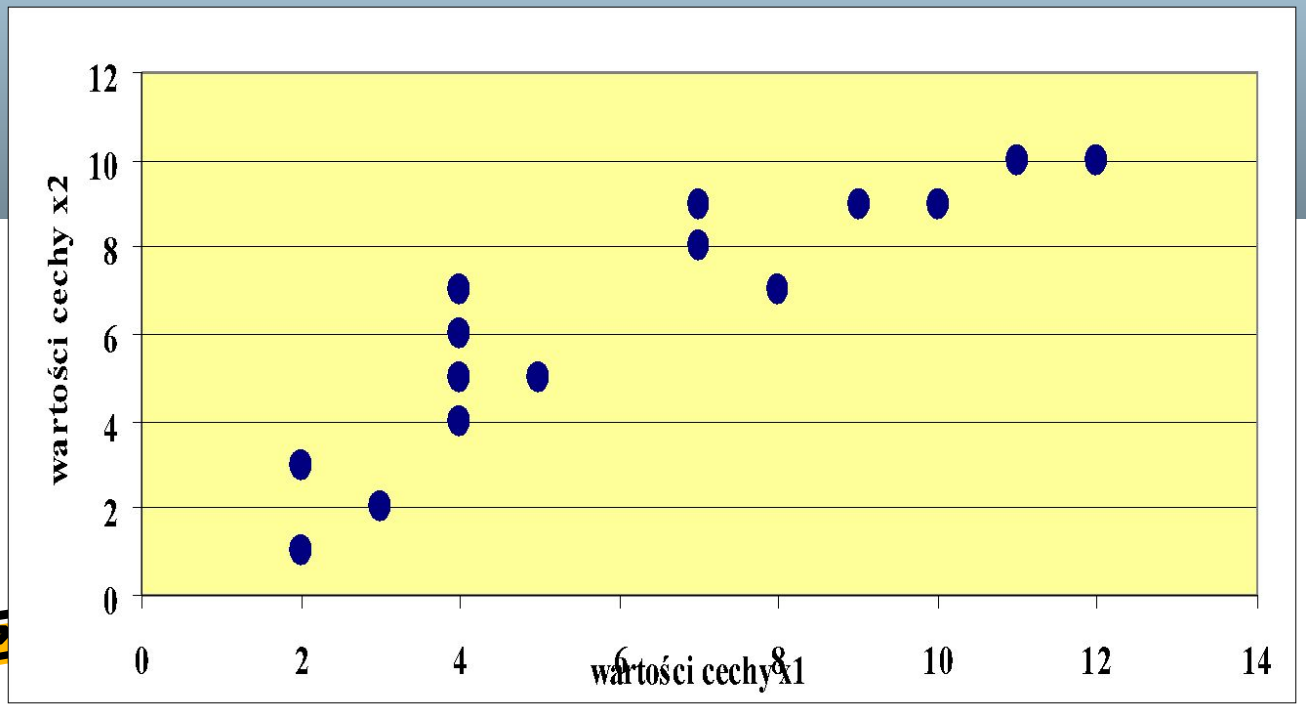
Tablica 2. Wartości cech odpowiadające poszczególnym obserwacjom (i)

Numer obserwacji i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Wartość cechy x_1	2	2	3	4	4	4	4	5	7	7	8	9	10	11	12
Wartość cechy x_2	1	3	2	4	5	6	7	5	8	9	7	9	9	10	10

- Źródło: dane fikcyjne

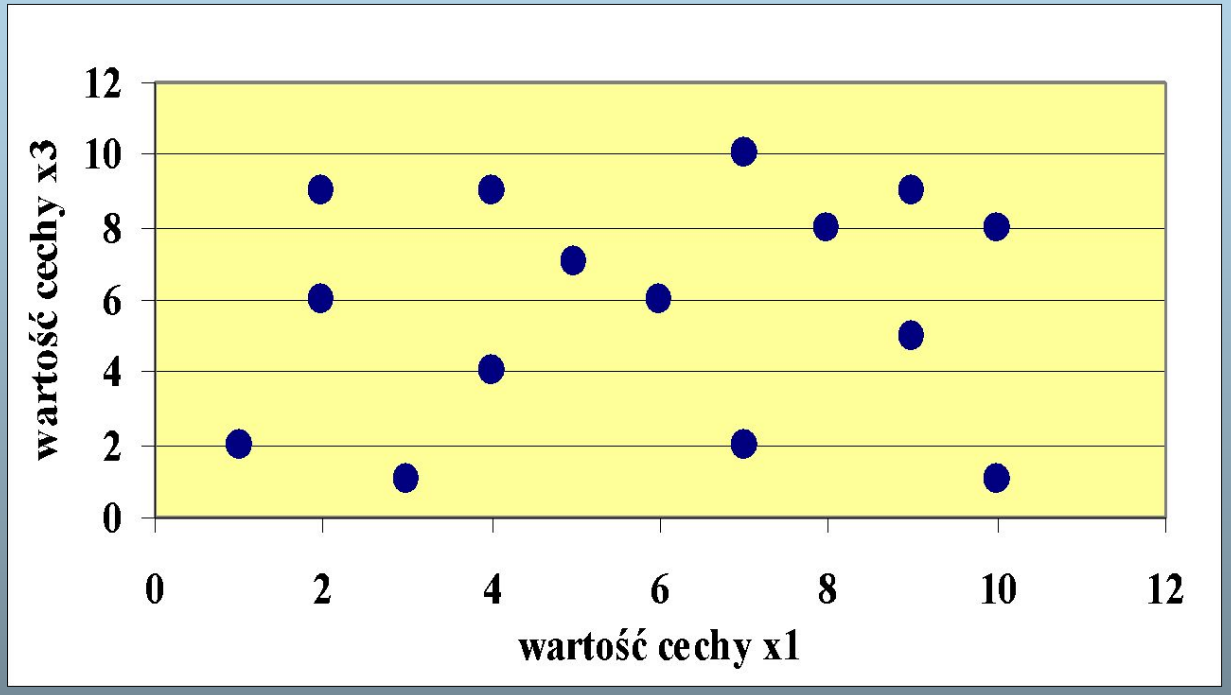

wartości cechy x2
wartości cechy x1
0
2
4
6
8
10
12
14
11; 10
12; 10
7; 9
9; 9
10; 9
7; 8
8; 7
4; 7
4; 6
4; 5
5; 5
4; 4
2; 3
3; 2
2; 1

- Z rys.1 widać wyraźnie, iż „na ogół” im większa wartość cechy (x_1), tym większą wartość przyjmuje cecha (x_2) i odwrotnie.

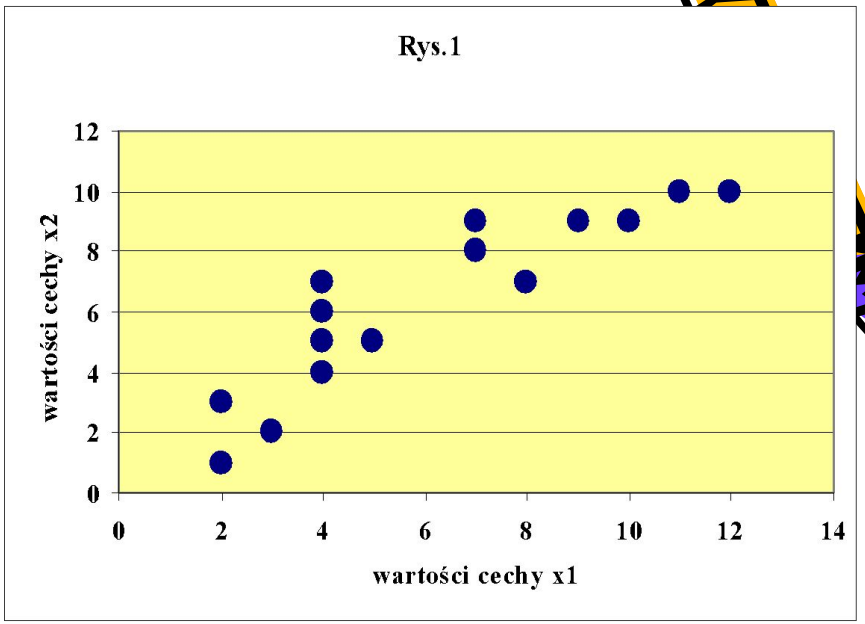


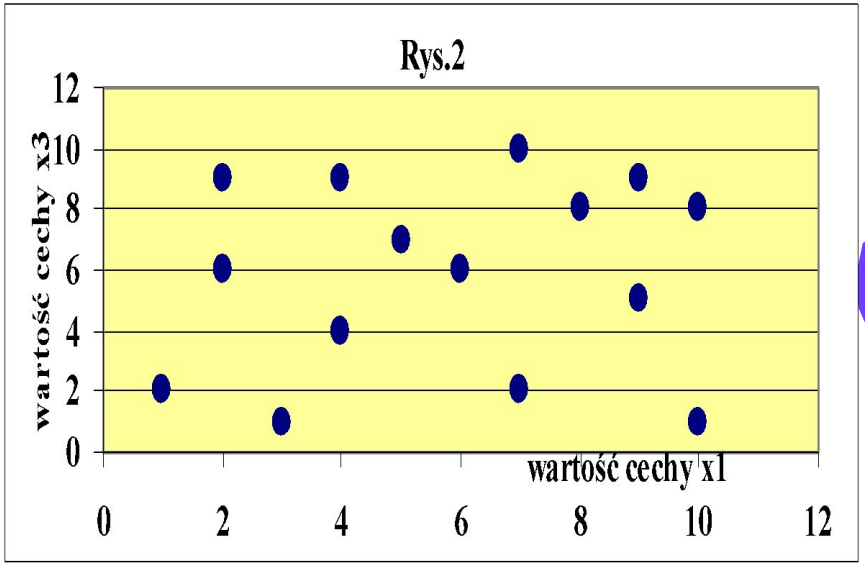
Przykład 4.

- Załóżmy, że obecnie populacja studentów (n = 15) jest opisywana za pomocą dwóch innych cech **(x_1) i (x_3).** Wyniki próby 15-elementowej badane ze względu na te cechy prezentują się na poniższym rysunku 2:
- Rys.2.

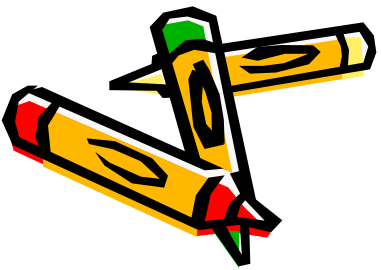


- Z rys.2 , w odróżnieniu od rys.1, nie widać wyraźnie, aby wartości cechy x_1 i x_3 były w jakiś sposób ze sobą powiązane.
- „Na oko” można tylko stwierdzić, iż cechy x_1 i x_2 (rys.1)są zapewne ze sobą ściślej powiązane niż cechy x_1 i x_3 (rys.2).
- **Pytanie 1?** – Jak ocenić i zmierzyć siłę związku dwóch cech?





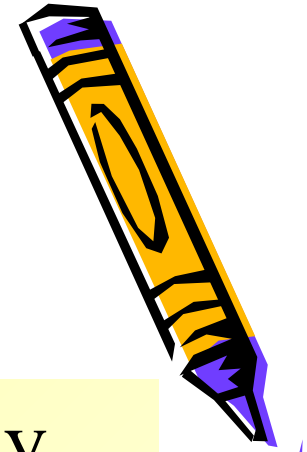
- Metoda pozwalająca na ocenę i mierzenie siły związku cech stanowi przedmiot **analizy korelacji**.



Uwaga!

- Badanie związków korelacyjnych ma sens jedynie tylko wtedy, gdy między zmiennymi istnieje więź przyczynowo-skutkowa, dająca się logicznie wytłumaczyć.
- Analiza związków między zjawiskami powinna być dwukierunkowa: jakościowa i ilościowa.
- Zawsze na podstawie analizy merytorycznej należy uzasadnić logiczne występowanie związku a dopiero potem można przystąpić do określania kierunku i siły zależności.

- Badanie korelacji między zmiennymi (szeregami)
- Zestawienie kilku szeregów=szukanie wzajemnych związków i porównanie wartości liczbowych cech w tych szeregach= wykrycie określonych prawidłowości
- Zmienna=szereg liczbowy=wartości liczbowe cech w szeregu



- Parametrem wykorzystywanym do oceny siły i kierunku zależności pomiędzy zmiennymi jest współczynnik korelacji, zwany również **współczynnikiem korelacji Persona**.

Współczynnik korelacji Pearsona

- $\mathbf{r_{xy}}$ jest miernikiem związku liniowego między dwiema cechami (zmiennymi) mierzalnymi
- jest wyznaczany poprzez standaryzację kowariancji
- **kowariancja** (wariancja wspólna cech x i y) jest średnią arytmetyczną iloczynu odchyleń wartości liczbowych tych cech (zmiennych) x i y od ich średnich arytmetycznych

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n \cdot S(x) \cdot S(y)}$$

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{S(x) \cdot S(y)}$$

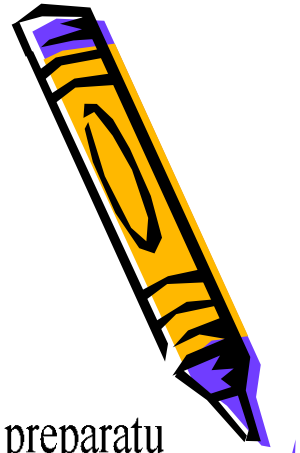
$$\mathrm{cov}(x,y) = \mathrm{cov}(y,x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$$

- Współczynnik korelacji jest symetryczny, tzn. $\mathbf{r_{xy} = r_{yx}}$ i przyjmuje wartości z przedziału <-1,1>.
- Równy jest zeru, gdy między cechami nie zachodzi liniowa zależność.
- Moduł (wartość bezwzględna) współczynnika korelacji równy jest jedności, gdy pomiędzy cechami zachodzi związek funkcyjny.
- Im wartość modułu współczynnika korelacji jest bardziej zbliżona do jedności, tym zależność między badanymi cechami jest silniejsza.
- Znak współczynnika charakteryzuje kierunek zależności.
- Jeżeli współczynnik korelacji jest dodatni, wówczas wzrost wartości jednej cechy powoduje wzrost wartości drugiej cechy (ewentualnie spadek wartości jednej cechy powoduje spadek wartości drugiej cechy).
- W przypadku ujemnej wartości współczynnika korelacji możemy stwierdzić, iż wzrost wartości jednej cechy powoduje spadek wartości drugiej cechy.

Inna postać współczynnika korelacji Pearsona

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- W analizach statystycznych przyjmuje się, że jeżeli współczynnik korelacji wynosi:
  - mniej niż 0,2 - brak związku liniowego między badanymi cechami;
  - 0,2 – 0,4 → zależność liniowa wyraźna, lecz niska;
  - 0,4 – 0,7 → zależność umiarkowana;
  - 0,7 – 0,9 → zależność znacząca;
  - powyżej 0,9 → zależność bardzo silna.
- Kwadrat współczynnika korelacji nazywamy **współczynnikiem determinacji R^2 .**



Przykład 6.

Na 10 doświadczalnych krzewach porzeczki sprawdzono wpływ pewnego preparatu ochronnego, podawanego w różnych dawkach koncentracji X (w %), na zdrowotność owoców Y (w kg zdrowych zebranych owoców). Uzyskano dane (zawarte w tablicy 2):

Tablica 2.

Dawka preparatu (x_i)	0,5	1,0	1,5	2,0	2,5	3,0	4,0	5,0	6,0	8,0
Zbiory owoców (y_i)	0,8	1,0	1,0	1,3	1,6	1,5	2,0	3,0	4,7	7,0

a) za pomocą współczynnika korelacji liniowej wyznaczyć kierunek i siłę związku.

Tablica 3

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
0,5	0,8	-2,85	-1,59	4,53	8,12	2,53
1,0	1,0	-2,35	-1,39	3,27	5,52	1,93
1,5	1,0	-1,85	-1,39	2,57	3,42	1,93
2,0	1,3	-1,35	-1,09	1,47	1,82	1,19
2,5	1,6	-0,85	-0,79	0,67	0,72	0,62
3,0	1,5	-0,35	-0,89	0,31	0,12	0,79
4,0	2,0	0,65	-0,39	0,25	0,42	0,15
5,0	3,	1,65	0,61	1,01	2,72	0,37
6,0	4,7	2,65	2,31	6,12	7,02	5,34
8,0	7,0	4,65	4,61	21,44	21,62	21,25
33,5	23,9	x	x	41,14	51,50	36,10

$$\bar{x} = \frac{\sum x}{n} = \frac{33,5}{10} = 3,35 \qquad [\%]$$

$$\bar{y} = \frac{\sum y}{n} = \frac{23,9}{10} = 2,39 \qquad [\text{kg}]$$

$$S(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{51,5}{10}} = \sqrt{5,51} = 2,27 \qquad [\%]$$

$$S(y) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{36,1}{10}} = \sqrt{3,61} = 1,9 \qquad [\text{kg}]$$

$$r_{xy} = r_{yx} = \frac{\operatorname{cov}(x,y)}{S(x)S(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S(x)S(y)} = \frac{41,14}{2,27 \cdot 1,9 \cdot 10} \approx 0,95$$

Między zbiorem zdrowych owoców a dawka koncentratu preparatu ochronnego występuje silna zależność korelacyjna.

Wraz ze wzrostem dawki preparatu rośnie zbiór zdrowych owoców.

Metoda Z.Hellwiga

Badanie pojemności nośników informacji metodą Z.Hellwiga

Wprowadzamy podstawowe pojęcia:

- **nośnikiem informacji nazywamy zmienną objaśniającą** (m zmiennych objaśniających czyli m nośników informacji);

Spośród rozpatrywanych wstępnie nośników informacji można utworzyć **k= 2^m - 1** ich **kombinacji**, czyli zestawów zmiennych objaśniających.

- **pojemnością indywidualną nośnika informacji X_j** wchodzącego w skład k-tej kombinacji nazywamy wyrażenie:

$$h_{kj} = \frac{r_j^2}{\sum_{i \in I_k} |r_{ij}|} \quad (i,j = 1,2,\ldots m) \qquad (1.1)$$

gdzie:

h_{kj} - pojemność indywidualna j-tej zmiennej objaśniającej w k-tej kombinacji,

r_j - współczynnik korelacji j-tej zmiennej objaśniającej X_j ze zmienną objaśnianąY

r_{ij} - współczynnik korelacji i-tej i j-tej zmiennej objaśniającej,

$R_o = (r_j)$ - macierz współczynników korelacji między Y a każdą z X_j;

$R = (r_{ij})$ - macierz współczynników korelacji między X_i i X_j;

m - liczba zmiennych objaśniających w modelu,

K_k - k-ta kombinacja gdzie subskrypt k = 1,2,...,2^m - 1;

I_k - zbiór numerów zmiennych tworzących k-tą kombinację;

r_{yx} - współczynnik korelacji liniowej Pearsona

$$r_{yx} = \frac{\sum xy - n\bar{x}\,\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

- **pojemnością integralną nośników informacji** H_k nazywamy wyrażenie będące **sumą pojemności indywidualnych** w ramach każdej z kombinacji:

$$H_k = \sum_{j \in I_k} h_{kj} \qquad (1.2)$$

Istota metody - wybór takiej kombinacji zmiennych objaśniających, której pojemność integralna jest największa.

I_{opt} : H_{opt} = max { H_k : k=1,2,...,2^m - 1 }

Metoda grafu

Metoda grafu

W metodzie tej przyjmuje się następujące założenia:

a). zmienne objaśniające winny być nie skorelowane między sobą, tzn. $r_{ij} \approx 0$ dla wszystkich kombinacji wskaźników i,j gdzie i≠j (i,j=2,3,...,m);

b). Zmienne objaśniające winny pozostawać w związku korelacyjnym ze zmienną objaśnianą, tzn. $r_{ij} \neq 0$ (j=2,3,...,m), przy czym m oznacza liczbę wszystkich zmiennych, w tym numerem 1 oznaczona jest zmienna objaśniana.

1. Efektem gromadzenia informacji statystycznych jest macierz **M** wyników obserwacji (**n obserwacji**) zmiennej objaśnianej i zmiennych objaśniających:

$$M = [\,Y; X\,] = \begin{bmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1m} \\ y_2 & x_{21} & x_{22} & \dots & x_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ y_n & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \qquad (1)$$

2. Z powyższych danych budujemy symetryczną macierz **W** współczynników korelacji r_{ij} o wym. mxm.

3. Z macierzy **W** (wyłączając 1-ą kolumnę i 1-y wiersz) tworzymy macierz współczynników korelacji zmiennych objaśniających łączonych parami ($\mathbf{R}=\{r_{ij}\}_{(m-1)x(m-1)}$).

4. W odniesieniu do współczynników korelacji r_{ij} weryfikujemy **hipotezę** $\mathbf{H_0: (r_{ij}=0,\ i \neq j)}$.

Sprawdzianem tej hipotezy jest statystyka:

$$t_e = \frac{r_{ij}}{\sqrt{1 - r_{ij}^2}} \sqrt{n-2} \qquad (2)$$

która.

porównana z $\mathbf{t_t}$ odczytaną z tablic rozkładu t-Studenta przy zadanym poziomie istotności α i danej wielkości próbki (n) pozwala na orzeczenie o słuszności hipotezy.

Jeśli $t_e > t_t$ hipotezę H_0 odrzucamy,

jeśli $t_e \leq t_t$ hipotezę utrzymujemy, co ozn. $r_{ij} \approx 0$.

W praktyce korzystając ze wzoru:

$$r^* = \sqrt{\frac{\frac{t_t^2}{n-2}}{1+\frac{t_t^2}{n-2}}} \qquad (3)$$

ustalamy wartość krytyczną $\mathbf{r^*}$ powodującą odrzucenie hipotezy H_0.

Wszystkie współczynniki korelacji, dla których zachodzi relacja: $|\mathbf{r_{ij}}| \leq \mathbf{r^*}$ **(i≠j)** zastępujemy w macierzy **R** zerami. Otrzymaną w ten sposób macierz oznaczymy jako **G**.

5. Wykorzystując macierz **G** budujemy **graf, w których wierzchołkami są zmienne, a łukami współczynniki korelacji $r_{ij} \neq 0$**. W rezultacie mogą powstać grafy spójne i wierzchołki izolowane, co oznacza, że obok grup zmiennych skorelowanych występują zmienne nie skorelowane z żadną z pozostałych.

6. Określamy **stopień każdego węzła grafu k,** tzn. liczbę łuków, którymi jest on związany z innymi wierzchołkami. Dla wierzchołków izolowanych k=0.

7. W każdym grafie spójnym wyróżniamy wierzchołek o maksymalnym k. Jeżeli w danym grafie jest kilka takich wierzchołków, to wybieramy wierzchołek tej zmiennej, dla której $|\mathbf{r_{1j}}| = \mathbf{max}$.

Ostatecznie jako zmienne objaśniające pozostawimy zmienne reprezentujące wierzchołki izolowane oraz wyróżnione z grafów spójnych zgodnie z p. 7.