

Badanie zależności zjawisk losowych na opisanych rozkładem normalnym

1. Badanie wzajemnej korelacji dwu zmiennych losowych – współczynnik korelacji liniowej Pearsona.

1.1 Zależność stochastyczna dwu zmiennych losowych

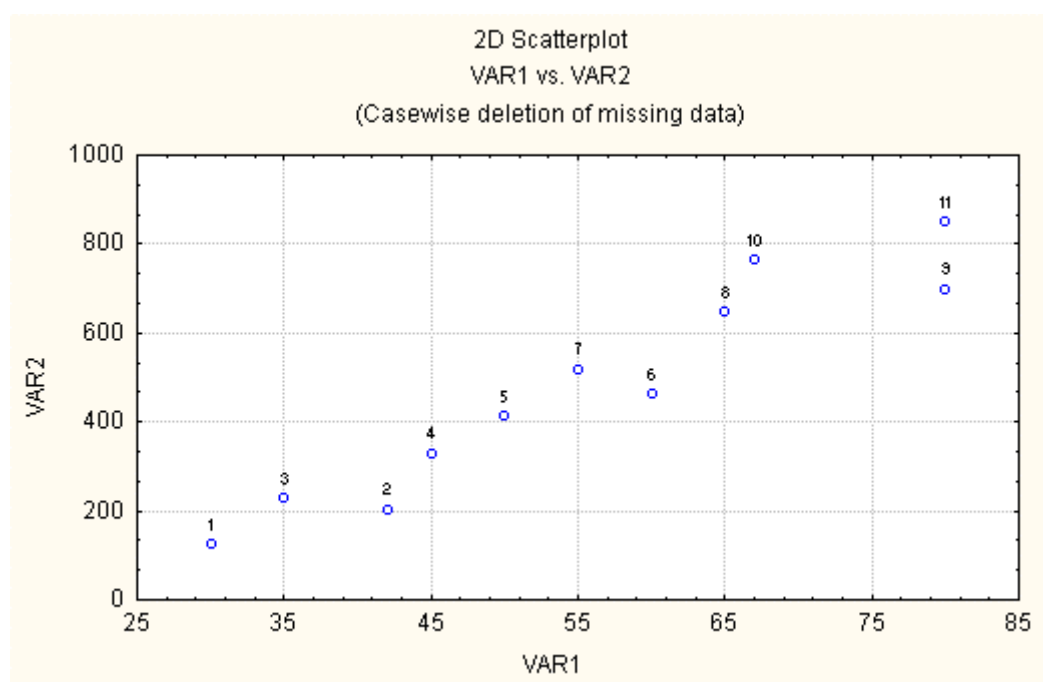
Symbolem **Y** oznaczamy zmienną **zależną** (objaśnianą), zaś symbolem **X** zmienną **niezależną** (objaśniającą). Zależność **stochastyczna** występuje wtedy gdy wraz ze zmianą wartości jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej.

Szczególnym przypadkiem zależności stochastycznej jest zależność **korelacyjna (statystyczna)**. Polega ona na tym, że określonym wartościom jednej zmiennej odpowiadają ściśle określone wartości oczekiwane (średnie) drugiej zmiennej.

Badanie zależności korelacyjnej ma sens jedynie wtedy, gdy między zmiennymi istnieje więź przyczynowo-skutkowa, dająca się logicznie wytłumaczyć.

Wykres korelacyjny

Jeżeli chcemy zbadać zależność stochastyczną cech statystycznych X oraz Y opisanych liczbowo, to należy zaobserwować próbkę losową składającą się z n **par** obserwacji (X_i, Y_i) , $i=1, \dots, n$. Wyniki obserwacji należy nanieść na wykres korelacyjny (pole rozrzutu, wykres XY).



W przypadku braku zależności (korelacji) pomiędzy analizowanymi zjawiskami losowymi punkty na wykresie korelacyjnym powinny ułożyć się w beładną (przypadkową) chmurę.

Jeżeli punkty na wykresie korelacyjnym układają się wg jakiegoś wzoru (np. wzdłuż linii prostej), to należy przypuszczać, że analizowane zjawiska są wzajemnie zależne (skorelowane).

1.2 Współczynnik korelacji liniowej Pearsona

W celu zmierzenia siły korelacji liniowej dwu cech statystycznych wykorzystujemy **współczynnik korelacji liniowej Pearsona**.

Obserwujemy n par liczb (x_i, y_i) , $i=1, 2, \dots, n$ stanowiące realizacje par *zmiennych losowych* (X, Y) opisujących analizowane cechy statystyczne.

Wyznaczamy oszacowanie **kowariancji** zmiennych losowych X i Y :

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

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

Wyznaczamy empiryczne **odchylenia standardowe** $s(x)$ oraz $s(y)$

Współczynnik korelacji liniowej Pearsona wyznaczamy z zależności:

$$r = r_{xy} = r_{yx} = \frac{Cov(x, y)}{s(x)s(y)}$$

Interpretacja: $r=0$ - brak zależności **liniowej**; $r = 1$ - **dodatnia** zależność liniowa; $r = -1$ - **ujemna** zależność liniowa.

UWAGA

Zerowa wartość współczynnika korelacji liniowej Pearsona nie oznacza braku zależności pomiędzy zmiennymi losowymi. Mówimy w takim przypadku o braku korelacji pomiędzy badanymi zmiennymi losowymi.

1.3 Wnioskowanie statystyczne o współczynniku korelacji liniowej Pearsona

Dokładny rozkład prawdopodobieństwa współczynnika korelacji liniowej Pearsona jest znany w przypadku gdy zmienne losowe X oraz Y mają rozkłady (rozkłady brzegowe) normalne.

W takim przypadku wykorzystujemy statystykę Z , której zaobserwowaną wartość wyznaczamy ze wzoru

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

Przedział ufności dla tej statystyki na poziomie ufności β możemy wyznaczyć już dla niewielkich licznosci próbki i wynosi on

$$\left\{ z - y_{(1+\beta)/2} \frac{1}{\sqrt{n-3}}, z + y_{(1+\beta)/2} \frac{1}{\sqrt{n-3}} \right\}$$

gdzie $y_{(1+\beta)/2}$ jest kwantylem rzędu $(1+\beta)/2$ w rozkładzie standaryzowanym normalnym $N(0,1)$ (Tablice, funkcja MS Excel).

Korzystając z powyższego przedziału ufności możemy weryfikować hipotezy statystyczne o współczynniku korelacji liniowej Pearsona.

Jeżeli przedział ufności dla statystyki Z obejmuje zero, to nie ma podstaw do odrzucenia hipotezy (na poziomie istotności $\alpha=1-\beta$) o braku korelacji liniowej zmiennych losowych X oraz Y .

W przypadku gdy zmienne losowe X oraz Y nie mają rozkładu normalnego, wnioskowanie statystyczne o współczynniku korelacji liniowej Pearsona możliwe jest tylko w przypadku asymptotycznym

Zaobserwowana w próbie wartość r jest realizacją zmiennej losowej (statystyki) ρ , która dla dużych próbek ($n > 120$) ma w przybliżeniu rozkład normalny. W takim przypadku dwustronny przedział ufności na poziomie ufności β dla statystyki ρ dany jest zależnością:

$$\left\{ r - y_{(1+\beta)/2} \frac{1 - r^2}{\sqrt{n}}, r + y_{(1+\beta)/2} \frac{1 - r^2}{\sqrt{n}} \right\}$$

gdzie $y_{(1+\beta)/2}$ jest kwantylem rzędu $(1+\beta)/2$ w standaryzowanym rozkładzie normalnym.

Korzystając z powyższego asymptotycznego przedziału ufności możemy weryfikować hipotezy statystyczne o współczynniku korelacji liniowej Pearsona. Jeżeli przedział ufności dla statystyki ρ obejmuje zero, to nie ma podstaw do odrzucenia hipotezy (na poziomie istotności $\alpha = 1 - \beta$) o braku korelacji liniowej zmiennych losowych X oraz Y .

Przykład: Obliczyć współczynniki korelacji dla danych giełdowych

WIG20	FORTE	BRE
1448.6	12.6	89
1451.8	13.2	91.5
1449.6	13	90
1451.8	13.4	91
1449.6	14.1	92
1489.3	14.2	92
1489.3	14.2	93
1537.9	14	95
1551.9	13.7	97
1554.4	13.3	97

$$X - \text{WIG20} \quad \bar{X} = 1487.42 \quad s(X) = 42.54945$$

$$Y - \text{FORTE} \quad \bar{Y} = 13.57 \quad s(Y) = 0.52735$$

$$Z - \text{BRE} \quad \bar{Z} = 92.75 \quad s(Z) = 2.61964$$

$$\frac{1}{10} \sum_{i=1}^{10} x_i y_i = \frac{1}{10} (1448.6 \cdot 12.6 + \dots + 1554.4 \cdot 13.3) = 20191.567$$

$$\frac{1}{10} \sum_{i=1}^{10} x_i z_i = \frac{1}{10} (1448.6 \cdot 89 + \dots + 1554.4 \cdot 97) = 138063.82$$

$$\text{Cov}(XY) = 7.2776 \quad \rho(XY) = 0.324$$

$$\text{Cov}(XZ) = 105.615 \quad \rho(XZ) = 0.948$$

Wartości akcji spółki BRE są **silnie liniowo skorelowane** z wartościami indeksu WIG20. Wartości akcji spółki FORTE są **słabo liniowo skorelowane** z wartościami indeksu WIG20.

2. Analiza regresji

2.1 Badanie zależności dla przypadku gdy wartości zmiennej losowej zależą od wartości innej zmiennej (zmiennych). Wyznaczanie liniowej funkcji regresji

W wielu przypadkach spotykanych w praktyce interesuje nas zależność **obserwowanej zmiennej losowej (zmiennej zależnej) Y** od wartości jakie przyjmuje inna zmienna (*nie koniecznie losowa*), zwana **zmienną niezależną X** . Zmienną zależną Y nazywamy czasami **zmienną objaśnianą**, a zmienną niezależną X nazywamy wówczas **zmienną objaśniającą**. Interesują nas zazwyczaj przypadki gdy zależność ta **ma postać liniową**

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

gdzie ε jest zmienną losową (zakłóceniem) o zerowej wartości oczekiwanej i stałej wariancji.

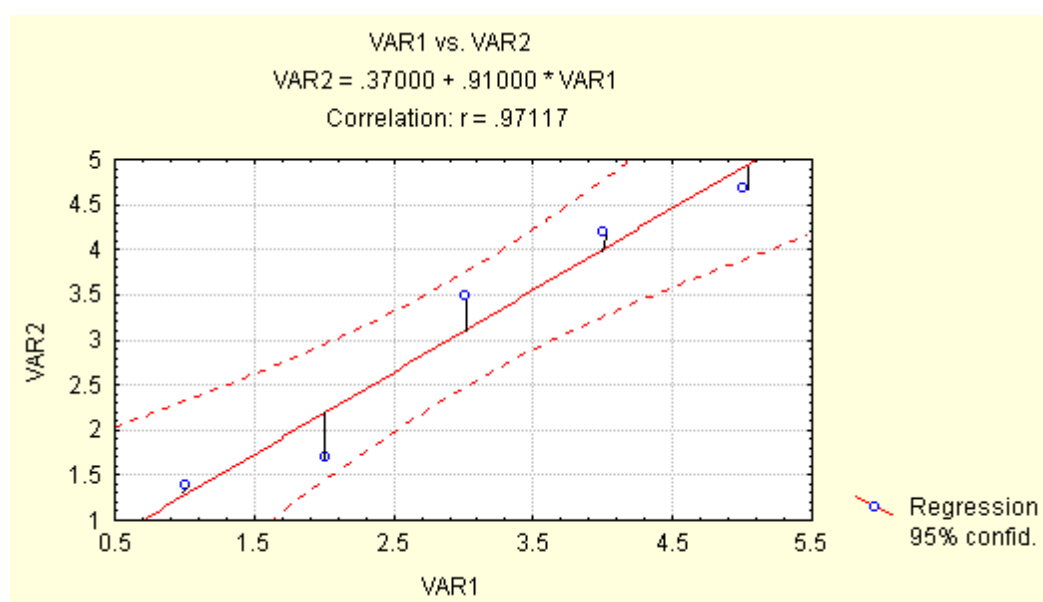
W przedstawionych powyżej przypadkach **regresji liniowej** konieczna jest **estymacja nieznanych parametrów modelu β_0 oraz β_1** na podstawie **obserwacji** par (X, Y) .

Wykorzystujemy do tego celu tzw. metodę **najmniejszej sumy kwadratów błędów** (nazywaną często potocznie *metodą najmniejszych kwadratów*).

Na podstawie obserwacji n par (X_i, Y_i) , $i=1, \dots, n$ poszukujemy takie **estymatory** b_0 , b_1 nieznanych parametrów modelu β_0 oraz β_1 , by zminimalizować wartość sumy:

$$S = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

Uzyskujemy w ten sposób taką prostą $Y=b_1X+b_0$, że zostanie **zminimalizowana** suma **kwadratów odległości** pomiędzy zaobserwowanymi punktami (X_i, Y_i) , a wyznaczoną prostą.



Minimalizacja S ze względu na b_0 oraz b_1 przebiega następująco:

- 1) Wyznaczamy pochodne funkcji S ze względu na b_0 oraz b_1 i przyrównujemy je do zera uzyskując tzw. *układ równań normalnych*.
- 1) Rozwiązujemy *układ równań normalnych* ze względu na b_0 oraz b_1 uzyskując następujące rozwiązanie:

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \left[\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right) \right] / n}{\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 / n} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

oraz

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Oszacowane równanie regresji zmiennej Y względem zmiennej X przyjmuje teraz postać

$$\hat{Y} = b_1 \cdot X + b$$

Można też zauważyć, że równanie regresji można również zapisać w następujący sposób:

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

W przypadku gdy interesuje nas zależność *odwrotna*, tzn. gdy funkcja regresji jest postaci

$$X = a_1 Y + a_0$$

estymatory a_0 oraz a_1 mają postać:

$$a_1 = \frac{\sum_{i=1}^n X_i Y_i - \left[\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right) \right] / n}{\sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i \right)^2 / n} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

oraz

$$a_0 = \bar{X} - a_1 \bar{Y}$$

Jak widać, funkcja regresji zmiennej X względem zmiennej Y *nie jest odwrotnością* funkcji regresji zmiennej Y względem zmiennej X .

Przykład

W firmie handlowej analizowano wydajność $n=20$ handlowców. Celem badania było ustalenie zależności pomiędzy wysokością kwoty zawartych przez danego handlowca w ciągu ostatniego roku transakcji a jego stażem pracy. Wyniki badania przedstawiają się następująco:

Lp.	Staż(X)	Obrót(Y)	Lp.	Staż(X)	Obrót(Y)
1	1.250	172.000	11	3.000	215.000
2	1.000	158.000	12	3.500	222.000
3	1.000	184.000	13	4.000	219.000
4	2.000	175.000	14	4.750	225.000
5	2.500	185.000	15	4.000	228.000
6	2.000	201.000	16	4.500	240.000
7	2.000	197.000	17	4.000	210.000
8	2.750	209.000	18	5.000	226.000
9	3.000	200.000	19	5.500	238.000
10	3.250	189.000	20	5.000	243.000

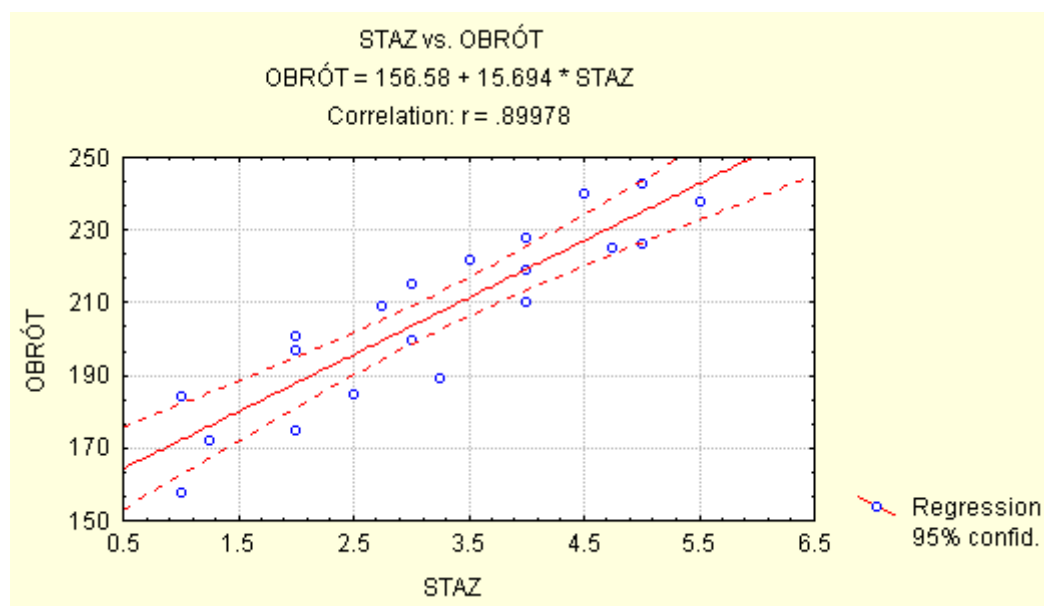
Po podstawieniu do wzorów na b_0 oraz b_1 uzyskujemy:

$$b_1=15.6941 \quad \text{oraz} \quad b_0=156.5789$$

Tak więc oszacowanie równania liniowej funkcji regresji Y względem X ma postać:

$$\hat{Y} = 15.6941 \cdot X + 156.5789$$

Równanie to możemy wykorzystać do **predykcji** (przewidywania) nieznanej wartości obrotu Y dla znanej wartości stażu pracy X .



Własności statystyczne oszacowanej funkcji regresji liniowej

Statystyczne własności liniowej funkcji regresji określa się zazwyczaj przyjmując że zmienne losowe ϵ_i są wzajemnie niezależne i mają rozkład normalny o zerowej wartości oczekiwanej i jednakowej (nieznanej) wariancji. Przedziały ufności na poziomie ufności $1-\alpha$ mają wówczas postać

$$\left(b_1 \pm \frac{t_{n-2, 1-\alpha/2} \cdot s}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right)$$

$$\left(b_0 \pm \frac{t_{n-2, 1-\alpha/2} \cdot s \cdot \sqrt{\sum_{i=1}^n X_i^2}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right)$$

gdzie

$$s = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}$$

Przykład (kontynuacja)

Dla danych z przykładu mamy: $s=10.83$. Dla $\alpha=0.10$ mamy $t_{18,0.95}=1,734$.

Wobec tego przedział ufności dla parametru β_1 wynosi:

$$15.5941 \pm 1.734 \cdot 1.794 \approx (12.48, 18.7)$$

zaś dla parametru β_0 wynosi:

$$156.5789 \pm 1.734 \cdot 6.23 \approx (145.78, 167.38)$$

2.2 Weryfikacja hipotezy o poprawności przyjętego modelu liniowej regresji

Przyjmijmy, że chcemy zweryfikować, czy rzeczywiście występuje liniowa zależność regresyjna pomiędzy badanymi wielkościami. Weryfikujemy hipotezę o braku zależności liniowej, a więc

H: $\beta_1=0$ przy alternatywie **K:** $\beta_1 \neq 0$.

Wykorzystuje się do tego celu statystykę:

$$F = \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{s^2} = t^2$$

Przy słuszności hipotezy zerowej statystyka ta ma rozkład **F-Snedecora** o parze stopni swobody (1,n-2). Hipotezę **H** *odrzucaamy* (tzn. przyjmujemy, że zależność rzeczywiście występuje) gdy wartość statystyki **F** jest większa od kwantyla rzędu $1-\alpha$ w rozkładzie F-Snedecora o parze stopni swobody (1,n-2). Możemy tu również wykorzystać to, że kwantyl ten jest równy kwadratowi kwantyla rzędu $1-\alpha$ w rozkładzie t-Studenta o **n-2** stopniach swobody.

Przykład (kontynuacja)

Dla danych z przykładu $F=76,538$. Dla $\alpha=0,05$ z tablicy kwantyli rozkładu t-Studenta odczytujemy $t_{18,0.95}=1,737$. Po podniesieniu tej wartości do kwadratu otrzymujemy wartość krytyczną statystyki F , która wynosi 3.017. Zaobserwowana wartość F jest więc znacznie większa od wartości krytycznej, a więc hipotezę $\beta_1=0$ odrzucamy. Tak więc istnieje zależność liniowa pomiędzy badanymi cechami.

2.3 Określanie procentu zmienności wyjaśnianego przez równanie regresji

Procent zmienności, który można wyjaśnić przez równanie regresji wyznaczamy obliczając wskaźnik:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Jeżeli zmienność jest *całkowicie* wyjaśniana przez równanie regresji (wszystkie punkty pomiarowe leżą na prostej regresji) powyższy wskaźnik przyjmuje wartość 1. Jest to sytuacja idealna. Im mniejsza jest wartość wskaźnika R^2 , tym gorzej dana liniowa funkcja regresji opisuje zależność badanych wielkości.

Uwaga: W przypadku powtarzających się wartości zmiennej niezależnej (X) wskaźnik R^2 będzie zawsze przyjmował wartości mniejsze od 1.

Przykład (kontynuacja)

Dla danych z przykładu mamy $R^2=0.8998$. Oznacza to, że wyznaczona przez nas funkcja regresji wyjaśnia 89.98% zmienności danych.

3. Analiza zależności w przypadku liczby zmiennych większej od dwu

3.1 Współczynniki korelacji cząstkowej i wielokrotnej (wielorakiej)

Przyjmijmy, że analizie poddane zostaje m zmiennych X_1, X_2, \dots, X_m opisujących dany obiekt. W szczególnym przypadku możemy wśród nich wyróżnić jedną zmienną zależną (objaśnianą) $Y=X_1$ i $m-1$ zmiennych niezależnych (objaśniających) X_2, X_3, \dots, X_m .

Założmy, że wzajemne zależności pomiędzy obserwowanymi zmiennymi opisane są **macierzą** P , której elementami są współczynniki korelacji pomiędzy poszczególnymi zmiennymi

$$P = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1m} \\ r_{21} & 1 & r_{23} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & 1 \end{bmatrix}$$

W pewnych przypadkach może nas interesować związek pomiędzy dwiema zmiennymi (np. zmienną X_i oraz zmienną X_j) z **wyłączeniem wpływu pozostałych zmiennych**. Do opisu zależności tego typu wykorzystujemy **współczynnik korelacji cząstkowej**

$$r_{ij.k..z} = - \frac{P_{ij}}{\sqrt{P_{ii}P_{jj}}}$$

gdzie P_{ij} jest dopełnieniem algebraicznym macierzy P .

Dopełnienie algebraiczne P_{ij} wyznacza się wykreślając w macierzy P i -ty wiersz oraz j -tą kolumnę. Następnie oblicza się **wyznacznik** tak uzyskanej macierzy i mnoży się go przez współczynnik $(-1)^{i+j}$.

Na przykład, w przypadku trzech zmiennych X_1, X_2, X_3 , gdy interesuje nas związek pomiędzy zmiennymi X_1 oraz X_2 przy wyłączeniu wpływu zmiennej X_3 uzyskujemy

$$r_{12.3} = -\frac{P_{12}}{\sqrt{P_{11}P_{22}}} = -\frac{\begin{vmatrix} - & r_{21} & r_{23} \\ & r_{31} & 1 \end{vmatrix}}{\sqrt{\begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} \cdot \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix}}} = \frac{r_{12} - r_{23}r_{13}}{\sqrt{(1-r_{23}^2)(1-r_{13}^2)}}$$

W przypadku większej liczby zmiennych do obliczeń wykorzystuje się metody komputerowe.

Gdy interesuje nas związek pomiędzy *jedną* zmienną *objaśnianą* (np. X_1) a *pozostałymi* zmiennymi *objaśniającymi* X_2, X_3, \dots, X_m wykorzystujemy **współczynnik korelacji wielokrotnej (wielorakiej)** wyznaczany ze wzoru

$$R_{1.23\dots m} = \sqrt{1 - \frac{\det D}{\det R}}$$

gdzie symbol **det** oznacza wyznacznik macierzy, macierz D jest macierzą współczynników korelacji zmiennej objaśnianej i zmiennych objaśniających, zaś R jest macierzą współczynników korelacji *pomiędzy zmiennymi objaśniającymi*.

Również i w tym przypadku dla większej liczby zmiennych do obliczeń wykorzystuje się metody komputerowe.

3.2 Regresja wielokrotna (wieloraka)

Uogólnieniem regresji liniowej dla dwu zmiennych jest **regresja wielokrotna (wieloraka)** stosowana w przypadku, gdy poszukujemy liniowego związku pomiędzy zmienną objaśnianą Y a m zmiennymi objaśniającymi w postaci

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m + \varepsilon$$

Estymatory parametrów powyższego równania regresji uzyskujemy w analogiczny sposób jak w przypadku dwu zmiennych. Wymaga to rozwiązania układu m równań liniowych. Algorytm poszukiwania estymatorów równania regresji wielokrotnej wygodnie jest zapisać korzystając z notacji macierzowej. Przyjmijmy, że dysponujemy n obserwacjami Y_i , $i=1, \dots, n$ zmiennej objaśnianej Y wraz z odpowiadającymi im wektorami $(X_0=1, X_1, X_2, \dots, X_m)$ zmiennych objaśniających. Wprowadźmy następujące oznaczenia **macierzy**

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X' = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{11} & X_{21} & \dots & X_{m1} \\ \vdots & \vdots & \dots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{mn} \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Wektor b estymatorów współczynników równania regresji wielokrotnej $(\beta_0, \beta_1, \dots, \beta_m)$ wyznacza się ze wzoru:

$$b = (X'X)^{-1} X'Y$$

Podane poprzednio rozwiązanie dla przypadku jednej zmiennej objaśniającej jest szczególnym przypadkiem powyższego wzoru. W praktyce do wyznaczenia wartości składowych wektora b wykorzystuje się komputerowe pakiety statystyczne.

4. Nieparametryczne testy statystyczne

4.1 Badanie losowości danych statystycznych

W wielu zastosowaniach statystyki zakłada się, że wyniki pomiarów są wynikami czysto przypadkowymi. Na przykład, w analizie regresji zakłada się, że reszty w modelu regresji są wzajemnie niezależnymi wielkościami losowymi o jednakowym rozkładzie.

Problem: jak zbadać losowość ciągu obserwacji?

Problem zajętości krzeseł w barze (Feller).

Pytanie: która z zaistniałych sytuacji wygląda na przypadkową (losową)?

**ZZZZZZZZZZWWWWWWWWWW
ZWZWZWZWZWZWZWZWZW
ZZWZWWWZZZWZZWZWZWWW**

Do analizy losowości wykorzystujemy pojęcie **serii**.

Seria jest sekwencją takich samych elementów, przed i po których występują inne elementy lub nie ma żadnego.

ZZ W Z WWW ZZZ W ZZ W Z W Z WWW

W powyższym przypadku mamy **12** serii.

Hipotezę o losowości obserwacji odrzucamy gdy liczba serii R jest zbyt mała ($R \leq C_1$) lub zbyt duża ($R \geq C_2$). Wartości krytyczne C_1 oraz C_2 znajduje się analizując rozkład prawdopodobieństwa liczby serii (podany np. w tabelicy 8, dodatku C, książki A.Aczela „Statystyka w zarządzaniu”) dla pary parametrów (n_1, n_2) , gdzie n_1 jest liczbą elementów jednego rodzaju, a n_2 jest liczbą elementów drugiego rodzaju.

W rozpatrywanym przykładzie, o braku losowości świadczyła by liczba serii mniejsza od 7 lub większa od 14 (prawdopodobieństwo każdego z tych zdarzeń wynosi 0,019). Tak więc nie ma podstaw by twierdzić, że zaobserwowany wynik nie jest wynikiem losowym.

Dla dużej liczby losowych obserwacji (zarówno n_1 jak i n_2) rozkład liczby serii jest asymptotycznie rozkładem normalnym o wartości oczekiwanej

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

oraz odchyleniu standardowym

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

W rezultacie zmienna losowa

$$Z_R = \frac{R - E(R)}{\sigma_R}$$

ma standaryzowany rozkład normalny $N(0,1)$. Hipotezę o losowości odrzucamy gdy zaobserwowana wartość statystyki Z_R jest albo większa od kwantyla $y_{1-\alpha/2}$ albo mniejsza od kwantyla $y_{\alpha/2} = -y_{1-\alpha/2}$.

4.2 Testy rangowe

Istnieje wiele testów statystycznych służących do weryfikacji hipotez w przypadku gdy dane statystyczne opisane są rozkładem normalnym. W przypadku gdy założenie to nie jest spełnione, tzn. gdy dane opisane są innymi rozkładami prawdopodobieństwa, bardzo często nie dysponujemy odpowiednimi testami statystycznymi.

Potrzebne więc są testy statystyczne, które **nie zależą od rozkładu prawdopodobieństwa** analizowanych danych. Do takich procedur, zwanych **testami nieparametrycznymi**, należą **testy rangowe**.

Założmy, że obserwujemy n wartości próbki losowej (X_1, X_2, \dots, X_n) . Wartości te porządkujemy w następujący **ciąg niemalejący** (tzw. szereg rozdzielczy):

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)}$$

Numer kolejnej wartości elementu próby w ciągu uporządkowanym (powyżej: liczba w nawiasie) **nazywamy rangą tego elementu**. Testy rangowe wykorzystują wyłącznie informacje o rangach poszczególnych obserwacji, a nie o ich konkretnych wartościach.

Przykład: Obserwujemy próbkę losową o wartościach (17,12,19,5,7). Po uporządkowaniu dane te tworzą szereg rozdzielczy (5,7,12,17,19). Tak więc zaobserwowanej próbce przypisujemy wektor rang (4,3,5,1,2).

4.2.1 Porównywanie rozkładów prawdopodobieństwa w dwu populacjach (próbki niezależne): test Wilcoxon-Manna-Whitneya

Jeżeli nie są spełnione warunki testu t -Studenta, to do porównywania rozkładu danych statystycznych w dwu populacjach możemy skorzystać z rangowego testu Wilcoxon-Manna-Whitneya.

Uwaga: Istnieją różne *równoważne* wersje tego testu zwane testem Wilcoxon oraz testem Manna-Whitneya.

Przyjmijmy, że mamy do dyspozycji rezultaty badania dwu niezależnych próbek losowych (X_1, X_2, \dots, X_n) oraz (Y_1, Y_2, \dots, Y_m) . Próbkę te nie muszą mieć jednakowej liczności. Na podstawie tych wyników tworzymy **łączny szereg rozdzielczy** dla połączonych obu próbek. Na podstawie tego szeregu wyznaczamy rangi obserwacji w próbce łącznej. W następnym kroku sumujemy wartości rang dla wszystkich elementów jednej z próbek. Suma ta będzie podstawą do budowy odpowiedniej statystyki testu rangowego.

Na przykład, w teście U Manna-Whitneya statystyka testowa ma postać

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

gdzie n_1 jest licznością pierwszej próbki, n_2 jest licznością drugiej próbki, a R_1 jest sumą rang elementów pierwszej próbki.

Dla małych liczności próbek ($n_1, n_2 < 10$) opracowano tablice rozkładu prawdopodobieństwa statystyki U (podane np. w tablicy 9, dodatku C, książki A. Aczela „Statystyka w

zarządzaniu”). Dla większych próbek statystyka U ma rozkład normalny o wartości oczekiwanej

$$E(U) = \frac{n_1 n_2}{2}$$

oraz odchyleniu standardowym

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

W rezultacie zmienna losowa

$$Z_U = \frac{R - E(U)}{\sigma_U}$$

ma standaryzowany rozkład normalny $N(0,1)$. Hipotezę o jednakowym rozkładzie porównywanych populacji odrzucamy gdy zaobserwowana wartość statystyki Z_U jest albo większa od kwantyla $y_{1-\alpha/2}$ albo mniejsza od kwantyla $y_{\alpha/2} = -y_{1-\alpha/2}$.

Uwaga 1. W niektórych podręcznikach (np. Koronackiego i Mielniczuka) opisywany jest test Wilcoxon, w którym statystyką testową jest suma rang w jednej z próbek.

Uwaga 2. W przypadku powtarzających się wartości obserwacji każdej z nich przypisujemy rangę będącą wartością średnią z rang odpowiadających tym wartościom. Na przykład, dla ciągu obserwacji (4,4,4,8) odpowiedni wektor rang wynosi (2,2,2,4).

4.2.2 Porównywanie rozkładów prawdopodobieństwa w wielu populacjach (próbki niezależne): test Kruskala-Wallisa

Nieparametrycznym odpowiednikiem analizy wariancji jest rangowy test Kruskala-Wallisa. Test ten służy do weryfikacji hipotezy o jednakowym rozkładzie prawdopodobieństwa cechy statystycznej w k populacjach, na podstawie analizy k niezależnych próbek losowych

$$(X_{i1}, X_{i2}, \dots, X_{i, n_i}) \quad i = 1, \dots, k.$$

Test Kruskala-Wallisa jest szczególnie czuły na różnicowanie *położenia* porównywanych rozkładów, czyli nadaje się do porównywania wartości średnich.

Na podstawie tych wyników tworzymy **łączny szereg rozdzielczy** dla połączonych k próbek. Na podstawie tego szeregu wyznaczamy rangi obserwacji w próbce łącznej. W następnym kroku sumujemy wartości rang dla wszystkich elementów każdej z badanych próbek. Sumy te, oznaczone przez R_1, R_2, \dots, R_k , będą podstawą do budowy następującej statystyki testu rangowego:

$$H = \frac{12}{n(n+1)} \left(\sum_{j=1}^k \frac{R_j^2}{n_j} \right) - 3(n+1)$$

gdzie n jest sumą licznosci wszystkich badanych próbek.

Dokładny rozkład prawdopodobieństwa statystyki H (w przypadku słuszności weryfikowanej hipotezy) jest podawany tylko dla bardzo małych licznosci próbek (<5). Jeżeli każda z badanych próbek liczy przynajmniej 5 elementów, to statystyka H ma w przybliżeniu rozkład chi-kwadrat o $k-1$ stopniach swobody. Hipotezę o równości wartości średnich w badanych populacjach odrzucamy na poziomie istotności α gdy zaobserwowana wartość statystyki H jest większa od kwantyla $\chi^2_{k-1;1-\alpha}$.

Uwaga. W przypadku powtarzających się wartości obserwacji każdej z nich przypisujemy rangę będącą wartością średnią z rang odpowiadającym tym wartościom.

4.2.3 Porównywanie rozkładów prawdopodobieństwa w dwu populacjach (próbki zależne): test rangowanych znaków Wilcoxona dla par obserwacji

W przypadku analizy próbek zależnych do weryfikacji hipotezy o zerowej wartości różnicy median (różnicy wartości średnich, w przypadku rozkładów symetrycznych) wykorzystujemy test rangowanych znaków Wilcoxona.

Zakładamy, że dysponujemy n parami obserwacji (x_i, y_i) , $i=1, \dots, n$, dla których wyznaczamy różnice $d_i = x_i - y_i$.

Z kolei, wyznaczamy wartości bezwzględne tych różnic $|d_i|$ i tworzymy z nich szereg rozdzielczy. Następnie sumujemy rangi różnic ujemnych (S^-) oraz rangi różnic dodatnich (S^+).

Statystyka testowa testu rangowanych znaków Wilcoxon ma postać

$$T = \min(S^-, S^+)$$

Dla małych licznosci próbek opracowano tablice wartości krytycznych statystyki T (podane np. w tablicy 10, dodatku C, książki A.Aczela „Statystyka w zarządzaniu”). W przypadku słuszności hipotezy o zerowej różnicy median porównywanych populacji (tzn. jednakowych medianach) i odpowiednio dużych próbek ($n > 25$) statystyka T ma w przybliżeniu rozkład normalny o wartości oczekiwanej

$$E(T) = \frac{n(n+1)}{4}$$

oraz odchyleniu standardowym

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

W rezultacie zmienna losowa

$$Z_T = \frac{R - E(T)}{\sigma_T}$$

ma standaryzowany rozkład normalny $N(0,1)$. Hipotezę o jednakowych medianach porównywanych populacji odrzucamy gdy zaobserwowana wartość statystyki Z_U jest albo większa od kwantyla $y_{1-\alpha/2}$ albo mniejsza od kwantyla $y_{\alpha/2} = -y_{1-\alpha/2}$.

4.2.3 Badanie zależności: test niezależności ρ –Spearmana

Analizujemy współzależność dwu zmiennych losowych X oraz Y o dowolnych rozkładach ciągłych. Próbkę losową składa się z n par obserwacji tych zmiennych losowych i ma postać: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Uporządkujemy wartości obu składowych (X oraz Y) uzyskując **ciągi rang**: (R_1, R_2, \dots, R_n) dla składowej X oraz (S_1, S_2, \dots, S_n) dla składowej Y .

Statystykę:

$$\rho = \frac{\sum_{i=1}^n \left(R_i - \frac{n+1}{2} \right) \left(S_i - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(R_i - \frac{n+1}{2} \right)^2 \sum_{i=1}^n \left(S_i - \frac{n+1}{2} \right)^2}}$$

nazywamy **współczynnikiem korelacji rang Spearmana**.

Inny zapis:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

gdzie

$$d_i = R_i - S_i, \quad i = 1, \dots, n$$

Gdy zgodność rang jest idealna, $\rho=1$ i świadczy to o *dodatniej korelacji* pomiędzy dwiema cechami. Jeśli uporządkowania obu cech

są dokładnie przeciwne mamy $\rho = -1$ i świadczy to o *ujemnej korelacji* pomiędzy dwiema cechami.

Uwaga: Badana korelacja *nie musi być liniowa*. Może to być dowolna zależność *monotoniczna* (rosnąca lub malejąca).

Gdy badane cechy są *niekorelowane* (również *niezależne*) rozkład statystyki ρ -Spearmana ma wartość oczekiwaną *zero* oraz odchylenie standardowe $\sigma_\rho = 1/\sqrt{n-1}$

Dla $8 \leq n < 30$ to rozkład statystyki

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

można aproksymować rozkładem **t-Studenta** o **n-2** stopniach swobody. Dla $n \geq 30$ rozkład statystyki ρ -Spearmana można aproksymować rozkładem normalnym $N(0, 1/\sqrt{n-1})$.

Powyższe własności możemy wykorzystać do weryfikacji hipotezy o *niezależności* badanych cech **X** oraz **Y**.

Nie ma podstaw do kwestionowania hipotezy o niezależności gdy (dla $8 \leq n < 30$)

$$t \in (-t_{n-2, 1-\alpha/2}, t_{n-2, 1-\alpha/2})$$

gdzie $t_{n-2, 1-\alpha/2}$ jest kwantylem rzędu $1-\alpha/2$ w **rozkładzie t-Studenta** o **n-2 stopniach swobody** (Tablice)

lub gdy (dla $n \geq 30$)

$$\rho \in \left(-\frac{y_{1-\alpha/2}}{\sqrt{n-1}}, \frac{y_{1-\alpha/2}}{\sqrt{n-1}} \right)$$

gdzie $y_{1-\alpha/2}$ jest kwantylem rzędu $1-\alpha/2$ w standaryzowanym rozkładzie normalnym. W przeciwnym przypadku hipotezę o niezależności **odrzucaamy**.

Przykład

Określić współczynnik korelacji rangowej ρ -Spearmana pomiędzy notowaniami spółki FORTE oraz indeksem giełdowym WIG20. Na poziomie istotności $\alpha=0.05$ sprawdzić hipotezę o wzajemnym nieskorelowaniu tych dwu wskaźników.

Dane:

WIG20	Ranga WIG	FORTE	Ranga FORTE
1448.60	1	12.60	1
1451.85	5	13.20	3
1449.60	2	13.00	2
1451.80	4	13.40	5
1449.65	3	14.10	8
1489.30	6	14.25	10
1489.35	7	14.20	9
1537.90	8	14.00	7
1551.90	9	13.70	6
1554.40	10	13.30	4

$$\rho = 1 - \frac{6 \cdot 96}{1000 - 10} = 1 - 0.582 = 0.418$$

$$t = \frac{0.418\sqrt{8}}{\sqrt{1 - 0.418^2}} = 1.30 < t_{8,0.975} = 2.06$$

Tak więc na poziomie istotności $\alpha=0.05$ nie ma podstaw (!) do kwestionowania hipotezy o wzajemnym nieskorelowaniu tych dwu wskaźników.