



**Matematyczne echa
lessowych wąwozów**



Zobacz matematykę

Warsztaty z modelowania

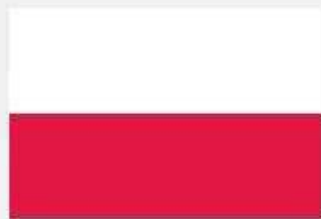
Modelowanie zjawisk przy pomocy rachunku prawdopodobieństwa i statystyki

dr Piotr Kowalski (UMCS), dr Tomasz Walczyński (UMCS)



**Ministerstwo Nauki
i Szkolnictwa Wyższego**

Projekt dofinansowany ze środków budżetu państwa, przyznanych przez Ministra Nauki
w ramach programu Społeczna odpowiedzialność nauki II



**DOFINANSOWANO
ZE ŚRODKÓW
BUDŻETU PAŃSTWA**

W RAMACH PROGRAMU
Społeczna odpowiedzialność nauki II
projekt: Matematyczne echa lessowych
wąwozów

DOFINANSOWANIE
[125 730 zł]

CAŁKOWITA WARTOŚĆ
[139 700 zł]

DATA PODPISANIA UMOWY
[08.2023]

Opis warsztatów

- Celem warsztatów jest zapoznanie uczestników z podstawowymi technikami modelowania zjawisk losowych, przy wykorzystaniu narzędzi rachunku prawdopodobieństwa i statystyki.
- Uczestnicy będą mieli okazję nauczyć się, jak formułować modele matematyczne, które pozwalają na opisanie i analizowanie wyników eksperymentów losowych, a także jak oceniać jakość tych modeli.
- Poszczególne części warsztatów będą obejmowały:
 - 1. Modelowanie rozkładu wyników rzutu monetą i kostką do gry
 - 2. Modelowanie zależności dwóch cech – regresja
 - 3. Wizualizacja wyników

1. Modelowanie rozkładu wyników rzutu monetą i kostką do gry

- Warsztaty rozpoczniemy od omówienia podstawowego zagadnienia z rachunku prawdopodobieństwa – modelowania wyników rzutu monetą oraz kostką do gry.
- Uczestnicy przeprowadzą eksperymenty losowe i będą analizować rozkłady prawdopodobieństwa dla różnych wyników tych eksperymentów.
- Zostanie wprowadzony tzw. schemat Bernoulliego, który jest podstawą dla wielu bardziej złożonych modeli statystycznych.
- Podczas tej części warsztatów szczególną uwagę poświęcimy następującym zagadnieniom:
 - **Zmienne losowe** – wyjaśnienie pojęcia zmiennej losowej i jej roli w modelowaniu procesów losowych.
 - **Charakterystyki zmiennych losowych** – obliczanie podstawowych parametrów zmiennych losowych.
- Formalne zapisanie modeli umożliwi przeprowadzenie symulacji komputerowych, w których uczestnicy będą mogli zmienić parametry i zaobserwować, jak wpływają one na wyniki.
- Do realizacji tych symulacji wystarczą podstawowe arkusze kalkulacyjne, co pozwala na szybkie i wygodne eksperymentowanie bez potrzeby zaawansowanego oprogramowania.

2. Modelowanie zależności dwóch cech – regresja

- W drugiej części warsztatów przejdziemy do bardziej zaawansowanego zagadnienia, jakim jest modelowanie zależności między dwiema zmiennymi za pomocą metod regresji.
- Użyjemy do tego rzeczywistych danych na temat zależności między inteligencją a rozmiarem mózgu u ludzi.
- W ramach tej części warsztatów uczestnicy poznają:
 - Jak za pomocą metody najmniejszych kwadratów wyznaczyć liniowy model zależności między dwiema zmiennymi.
 - Jak ocenić dopasowanie modelu, wykorzystując współczynniki dopasowania oraz testy statystyczne.
 - Jakie czynniki mogą wpływać na jakość uzyskanego modelu i jak uniknąć typowych błędów w analizie danych.

3. Wizualizacja wyników

- Ważnym elementem modelowania statystycznego jest odpowiednia wizualizacja wyników.
- Uczestnicy warsztatów poznają jak wizualizować dane oraz wyniki analiz statystycznych przy pomocy wykresów i diagramów.
- Poprawnie przygotowane wykresy pomagają nie tylko w lepszym zrozumieniu danych, ale także w prawidłowej interpretacji wyników, co może zapobiec wyciąganiu błędnych wniosków na podstawie surowych danych liczbowych.

1.1. Modelowanie rzutu monetą

- Otwieramy arkusz **Rzut monetą**
- W grupach dwuosobowych wykonujemy 2 doświadczenia po 5 rzutów monetą.
- Wyniki rzutów zapisujemy do arkusza i wyliczamy numery serii.
- Wyliczamy liczbę wystąpień (oraz procent wystąpień) orłów i reszek w doświadczeniu 1, w doświadczeniu 2 oraz w obu doświadczeniach razem.
- Wizualizujemy otrzymane wyniki na odpowiednich wykresach (wykres kołowy i wykres kolumnowy).
- Wyznaczamy długości dla poszczególnych serii oraz liczbę serii z daną długością.

1.2. Zmienna losowa dyskretna

- Prawie wszystkie wielkości, z którymi mamy do czynienia na co dzień, mają (mniej lub bardziej) charakter losowy, np. wzrost lub waga pierwszej osoby spotkanej po wyjściu z domu, ocena otrzymana na najbliższym egzaminie, cena kostki masła w najbliższym sklepie.
- Zdarzeniom losowym często przyporządkujemy pewne wielkości liczbowe, które charakteryzują interesujące nas cechy tych zdarzeń.
- Jeśli rzucamy monetą tak długo, aż pojawi się orzeł, to liczba wykonanych rzutów jest wielkością liczbową przyporządkowaną zdarzeniu losowemu.
- **Zmienna losowa** jest odpowiednikiem cechy statystycznej.
- Cecha statystyczna przyjmuje w badanej zbiorowości wartości $x_1, x_2, \dots, x_n$ z określonymi częstościami empirycznymi (są to rozkłady empiryczne).
- Zmienna losowa przyjmuje wartości (np. ze zbioru skończonego) z określonymi prawdopodobieństwami (są to rozkłady teoretyczne).

1.2. Zmienna losowa dyskretna

- **Zmienną losową** nazywamy dowolną funkcję (mierzalną) $X: \Omega \rightarrow \mathcal{R}$ o wartościach rzeczywistych, określoną na zbiorze zdarzeń elementarnych Ω .
- Rozważmy prosty eksperyment losowy polegający na jednokrotnym rzucie symetryczną monetą.

- Przestrzeń zdarzeń elementarnych Ω (przestrzeń prób) składa się z następujących 2 wyników:

$$\omega_1 = (O), \omega_2 = (R), \Omega = \{\omega_1, \omega_2\}$$

- W tym przypadku na przestrzeni zdarzeń elementarnych Ω możemy zdefiniować np. zmienną losową X określającą liczbę orłów w jednym rzucie:

$$X(\omega_1) = 1, X(\omega_2) = 0.$$

- Każdej zmiennej losowej odpowiada funkcja rozkładu, określająca prawdopodobieństwa występowania poszczególnych wartości zmiennej.

1.2. Zmienna losowa dyskretna

- Zmienne losowe dzielimy na: **dyskretne** (skokowe) i ciągłe (absolutnie ciągłe).
- **Zmienna losowa dyskretna** X (o skończonym zbiorze wartości) ma zbiór wartości $x_1, x_2, \dots, x_n$, przy czym prawdopodobieństwa $p_1, p_2, \dots, p_n$ spełniają warunki:
 - $0 < p_i \leq 1, i = 1, 2, \dots, n$
 - $p_1 + p_2 + \dots + p_n = 1$
- **Rozkład zmiennej losowej dyskretnej** zapisujemy w postaci
 - $P(X = x_i) = p_i, i = 1, 2, \dots, n$lub w postaci tabelki:

x_i	x_1	x_2	$\dots$	x_n
p_i	p_1	p_2	$\dots$	p_n

1.2. Zmienna losowa dyskretna

- **Zmienna losowa X o rozkładzie zero-jedynkowym**, z parametrem $p \in (0,1)$, ma następujący rozkład:
 - $P(X = 1) = p$
 - $P(X = 0) = 1 - p = q$ ($p + q = 1, p > 0, q > 0$)
- **Zmienna losowa X o rozkładzie (dyskretnym) jednostajnym**, z parametrem $n \in \mathbb{N}_+$, dla n różnych wartości $x_1, x_2, \dots, x_n$, ma następujący rozkład:
 - $P(X = x_1) = P(X = x_2) = \dots = P(X = x_n) = \frac{1}{n}$

1.2. Zmienna losowa dyskretna

- Weźmy pod uwagę prosty eksperyment losowy polegający na trzykrotnym rzucie symetryczną monetą.
- Przestrzeń zdarzeń elementarnych Ω (przestrzeń prób) składa się z następujących 8 wyników:
 $\omega_1 = (O, O, O)$, $\omega_2 = (O, O, R)$, $\omega_3 = (O, R, O)$, $\omega_4 = (O, R, R)$, $\omega_5 = (R, O, O)$, $\omega_6 = (R, O, R)$,
 $\omega_7 = (R, R, O)$, $\omega_8 = (R, R, R)$
- W tym przypadku, zdarzeniami losowymi są wszystkie podzbiory zbioru $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}$ (wszystkich podzbiorów jest $2^8 = 256$).
- Np. podzbiór odpowiadający zdarzeniu otrzymania dokładnie dwóch orłów jest równy $\{\omega_2, \omega_3, \omega_5\}$.

1.2. Zmienna losowa dyskretna

- W tym przypadku na przestrzeni zdarzeń elementarnych Ω możemy zdefiniować np. zmienną losową X określającą całkowitą liczbę orłów w trzech rzutach:
 - $\omega_1 = (O, O, O), \omega_2 = (O, O, R), \omega_3 = (O, R, O), \omega_4 = (O, R, R), \omega_5 = (R, O, O), \omega_6 = (R, O, R), \omega_7 = (R, R, O), \omega_8 = (R, R, R)$
 - $X(\omega_1) = 3, \quad X(\omega_2) = 2, \quad X(\omega_3) = 2, \quad X(\omega_4) = 1, \quad X(\omega_5) = 2, \quad X(\omega_6) = 1, \quad X(\omega_7) = 1, \quad X(\omega_8) = 0.$
- Zatem zmienna losowa X przyjmuje wartości: 0, 1, 2, 3.
- Ponadto zakładając, że wszystkie zdarzenia elementarne są jednakowo prawdopodobne (moneta jest symetryczna), otrzymujemy następujący rozkład prawdopodobieństwa zmiennej losowej X :
- $P(X = 0) = P(\omega \in \Omega: X(\omega) = 0) = P(\{\omega_8\}) = \frac{1}{8}, P(X = 1) = P(\omega \in \Omega: X(\omega) = 1) = P(\{\omega_4, \omega_6, \omega_7\}) = \frac{3}{8}$
- $P(X = 2) = P(\omega \in \Omega: X(\omega) = 2) = P(\{\omega_2, \omega_3, \omega_5\}) = \frac{3}{8}, P(X = 3) = P(\omega \in \Omega: X(\omega) = 3) = P(\{\omega_1\}) = \frac{1}{8}$

1.2. Zmienna losowa dyskretna

- Podstawowe charakterystyki (rozkładów) zmiennych losowych dyskretnych:
 - **wartość oczekiwana** (wartość przeciętna, wartość średnia) μ (m , EX) wyraża przeciętną wartość zmiennej losowej:

$$\mu = EX = x_1p_1 + x_2p_2 + \dots + x_np_n$$

- **wariancja** (dyspersja) σ^2 (σ^2X , D^2X , $Var X$) wyraża przeciętny kwadrat odchyłeń od wartości oczekiwanej:

$$\sigma^2 = \sigma^2X = E(X - \mu)^2 = (x_1 - \mu)^2p_1 + (x_2 - \mu)^2p_2 + \dots + (x_n - \mu)^2p_n$$

- **odchylenie standardowe** σ (σX , DX) wyraża przeciętne odchylenie od wartości oczekiwanej:

$$\sigma = \sigma X = \sqrt{\sigma^2}$$

1.2. Zmienna losowa dyskretna

- **Dystrybuenta** (skumulowana funkcja rozkładu prawdopodobieństwa) $F(x)$, zmiennej losowej X , jest niemalejącą funkcją, wyznaczoną jednoznacznie przez rozkład zmiennej losowej X , określoną wzorem $F(x) = P(X \leq x), x \in \mathcal{R}$.
- Dla zmiennych losowych dyskretnych:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p_i$$

1.3. Symulacja rzutu monetą

- Otwieramy arkusz **Symulacja rzutu monetą**
- W grupach dwuosobowych wykonujemy symulację 10, 100 i 1000 rzutów monetą.
- Wyliczamy liczbę wystąpień (oraz procent wystąpień) orłów i reszek w pierwszych 10, 100 i 1000 rzutów.
- Wyznaczamy długości dla poszczególnych serii oraz liczbę serii z daną długością.
- Zmieniamy wartość parametru $p \in (0,1)$ i analizujemy jego wpływ na otrzymane wyniki.

1.4. Rozkład dwumianowy (rozkład Bernoulliego)

- Jeżeli wykonujemy n niezależnych doświadczeń, z których każde może zakończyć się sukcesem z prawdopodobieństwem p lub porażką z prawdopodobieństwem $q = 1 - p$, to liczba k sukcesów w ciągu n niezależnych doświadczeń (prób) jest zmienną losową o rozkładzie dwumianowym.
- Zmienna losowa X o rozkładzie dwumianowym przyjmuje wartości ze zbioru $\{0, 1, \dots, n\}$ z prawdopodobieństwami:

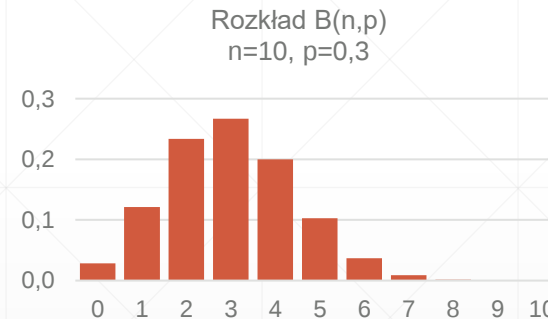
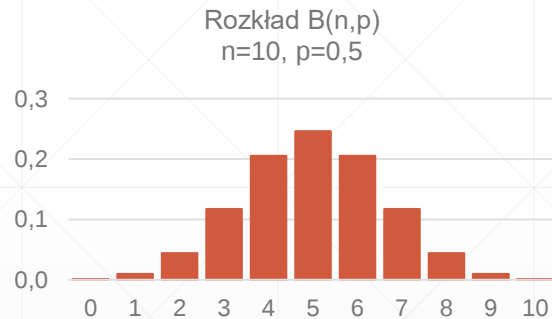
$$P(X = k) = \binom{n}{k} p^k q^{n-k}, k = 0, 1, \dots, n, q = 1 - p, 0 < p < 1,$$

$$\binom{n}{k} = \frac{n!}{k! (n - k)!}, n! = 1 \cdot 2 \cdot \dots \cdot n, 0! = 1$$

- Rozkład dwumianowy jest całkowicie określony przez liczbę niezależnych doświadczeń n oraz prawdopodobieństwo sukcesu p .

1.4. Rozkład dwumianowy (rozkład Bernoulliego)

- Symetria rozkładu zależy od wartości parametru p :
 - $p = 0,5$ — rozkład jest symetryczny; $p < 0,5$ — asymetria prawostronna; $p > 0,5$ — asymetria lewostronna.



- Wartość oczekiwana $\mu = EX = n \cdot p$
- Wariancja $\sigma^2 = n \cdot p \cdot q = n \cdot p \cdot (1 - p)$
- Odchylenie standardowe $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{n \cdot p \cdot (1 - p)}$

1.4. Rozkład dwumianowy (rozkład Bernoulliego)

- Jeżeli $X_i, i = 1, 2, \dots, n$ są niezależnymi zmiennymi losowymi o rozkładzie zerojedynekowym: $P(X_i = 1) = p, P(X_i = 0) = q = 1 - p$, to zmienna losowa $S_n = X_1 + X_2 + \dots + X_n$ ma rozkład dwumianowy z parametrami n i p .
- Rozkład dwumianowy znajduje szerokie zastosowanie w badaniach, takich jak: skuteczność leczenia (analiza liczby pacjentów, którzy zareagowali pozytywnie na dany lek w grupie pacjentów), wyniki testów psychologicznych (obliczanie prawdopodobieństwa, że osoba osiągnie określoną liczbę punktów w teście z odpowiedziami typu prawda/fałsz), badania rynkowe (szacowanie liczby klientów, którzy preferują określoną markę w badanej grupie osób).

1.4. Rozkład dwumianowy

- Otwieramy arkusz **Rozkład dwumianowy**
- Wpisujemy liczbę otrzymanych orłów w swoich dwóch doświadczeniach (próbkach).
- Wyliczamy liczbę wystąpień (oraz procent wystąpień) liczby orłów we wszystkich doświadczeniach i wizualizujemy je na odpowiednim wykresie słupkowym.
- Zapoznajemy się z symulacją rozkładu dwumianowego dla $n = 5$ i $p = 0,5$.
- Zmieniamy wartość parametru $p \in (0,1)$ i analizujemy jego wpływ na otrzymane wyniki.

1.5. Rzut kostką

- Otwieramy arkusz **Rzut kostką**
- Wykonujemy 12 rzutów sześcienną kostką do gry i zapisujemy wyniki.
- Wyliczamy liczbę wystąpień (oraz procent wystąpień) poszczególnych liczb oczek i wizualizujemy je na odpowiednim wykresie słupkowym.
- Wyznaczamy numery serii i obliczamy ich długości oraz zliczamy ile było serii o danej długości.

1.6. Symulacja rzutu kostką

- Otwieramy arkusz **Symulacja rzutu kostką**
- W grupach dwuosobowych wykonujemy symulację 10, 100 i 1000 rzutów symetryczną kostką sześcienną.
- Wyliczamy liczbę wystąpień (oraz procent wystąpień) poszczególnych liczb oczek w pierwszych 10, 100 i 1000 rzutów.
- Wyznaczamy długości dla poszczególnych serii oraz liczbę serii z daną długością.

2.1. Analiza regresji

- **Analiza regresji** jest kluczowym narzędziem statystycznym, które pozwala na badanie zależności między zmiennymi oraz prognozowanie wartości jednej zmiennej na podstawie wartości innych.
- Jest szeroko stosowana w różnych dziedzinach nauki, takich jak ekonomia, psychologia, biologia czy inżynieria, a jej podstawowym celem jest zrozumienie i modelowanie złożonych zjawisk.
- W ramach analizy regresji wyróżnia się dwa główne rodzaje: **regresję prostą** i **regresję wielokrotną (wieloraką)**.
- Regresja prosta dotyczy sytuacji, w której badamy zależność między dwiema zmiennymi — jedną niezależną i jedną zależną.
 - Pozwala to na ocenę, w jaki sposób zmiany w zmiennej niezależnej wpływają na zmienną zależną.
- Z kolei regresja wielokrotna rozszerza ten model, umożliwiając jednoczesne badanie wpływu wielu zmiennych niezależnych na jedną zmienną zależną, co prowadzi do bardziej złożonych i precyzyjnych analiz.
- Może nas interesować, czy między badanymi zmiennymi zachodzą jakieś zależności, jaka jest ich siła, jaki jest ich kształt i kierunek.

2.1. Analiza regresji

- **Współzależność** pomiędzy zmiennymi może być **funkcyjna** lub **stochastyczna (probabilistyczna)**.
- **Zależność funkcyjna** polega na tym, że zmiana wartości jednej zmiennej powoduje ściśle określoną (wg pewnego wzoru) zmianę drugiej zmiennej, czyli każdej wartości jednej zmiennej niezależnej (X) odpowiada tylko jedna, jednoznacznie określona wartość zmiennej zależnej (Y). Np. obwód kwadratu jest funkcją (liniową) jego boku $Ob = 4a$, a pole jest także funkcją (kwadratową) jego boku $P = a^2$.
- **Zależność stochastyczna** występuje wtedy, gdy wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem zależności stochastycznej jest zależność korelacyjna (statystyczna).
 - **Zależność korelacyjna (statystyczna)** polega na tym, że określonym wartościom jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennej. Zależność ta mówi nam jak zmieni się (średnio) wartość zmiennej zależnej (objaśnianej) Y w zależności od wartości zmiennej niezależnej (objaśniającej) X .

2.1. Analiza regresji

- Na podstawie analizy merytorycznej należy oczywiście najpierw **uzasadnić logiczne występowanie związku**, a dopiero potem przystąpić do określenia siły i kierunku zależności.
- Liczbowe stwierdzenie występowania współzależności nie zawsze oznacza występowanie związku przyczynowo-skutkowego między badanymi zmiennymi.
- Współwystępowanie dwóch zjawisk może również wynikać z bezpośredniego oddziaływania na nie jeszcze innego, trzeciego zjawiska.
- Prosty sposób wykrywania związku korelacyjnego między badanymi cechami jest **obserwacja szeregów statystycznych**, które zawierają informacje o tych cechach.
- Graficzny związek pomiędzy zmiennymi możemy zobaczyć na **wykresach rozrzutu**.
- **Analizę zależności powinno się rozpoczynać od sporządzenia właśnie takich wykresów.**
- Wzrokowa ocena umożliwia często określenie siły i rodzaju zależności.

2.1. Analiza regresji

- Załóżmy, że zbiorowość jest badana ze względu na dwie zmienne X i Y , a realizacje tych zmiennych w próbie n -elementowej zestawione są w postaci dwóch szeregów szczegółowych.
- Punkty, odpowiadające poszczególnym wartościom cech, tworzą **korelacyjny wykres rozrzutu** (w prostokątnym układzie współrzędnych: na osi odciętych zaznaczamy wartości pierwszej zmiennej, a na osi rzędnych wartości drugiej zmiennej).
- Bardzo rzadko zdarza się, aby zaznaczone punkty leżały dokładnie na linii prostej (wówczas jest pełna korelacja liniowa).
- Częściej spotykana konfiguracja składa się z wielu zaznaczonych punktów leżących mniej więcej wzdłuż konkretnej krzywej (najczęściej linii prostej).
- Gdy korelacja staje się coraz mniej wyraźna, wówczas punkty zaczynają się rozpraszać i przesuwają, aż do kolistego tworów (chmury).

2.1. Analiza regresji

- **Korelacja dodatnia** występuje wtedy, gdy wzrostowi wartości jednej cechy odpowiada wzrost średnich wartości drugiej cechy.
- **Korelacja ujemna** występuje wtedy, gdy wzrostowi wartości jednej cechy odpowiada spadek średnich wartości drugiej cechy.
- **Współczynnik korelacji (liniowej) Pearsona** wyraża się następującym wzorem

$$r_{xy} = \frac{cov(x,y)}{s_x s_y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Oczywiście $r_{xy} = r_{yx}$ oraz $-1 \leq r_{xy} \leq 1$.

2.1. Analiza regresji

- **Współczynnik korelacji (liniowej) Pearsona jest miernikiem siły związku prostoliniowego** między dwiema cechami mierzalnymi.
- Znak współczynnika korelacji informuje nas o kierunku korelacji, natomiast jego bezwzględna wartość o sile związku.
- Wielkość współczynnika korelacji podlega wpływom wartości skrajnych i odstających.
- Do opisu i interpretacji korelacji możemy przyjąć następującą skalę określającą siłę związku (stopień zależności dwóch cech):
 - $r_{xy} = 0$ – zmienne nie są skorelowane, $0 < |r_{xy}| < 0,1$ – korelacja nikła, $0,3 \leq |r_{xy}| < 0,5$ – korelacja przeciętna, $0,5 \leq |r_{xy}| < 0,7$ – korelacja wysoka, $0,7 \leq |r_{xy}| < 0,9$ – korelacja bardzo wysoka, $0,9 \leq |r_{xy}| < 1$ – korelacja prawie pełna, $|r_{xy}| = 1$ – korelacja pełna (zależność funkcyjna: liniowa).

2.1. Analiza regresji

- Kwadrat współczynnika korelacji nazywamy **współczynnikiem determinacji** (określoności).
- Współczynnik determinacji $R^2 = r_{xy}^2 (\cdot 100\%)$ informuje o tym, jaka część zmian zmiennej objaśnianej (skutek) jest wyjaśniona przez zmiany zmiennej objaśniającej (przyczyna).
- Np. jeśli zmienna X (zmienna objaśniająca) koreluje ze zmienną Y (zmienna objaśniana) na poziomie $r_{xy} = 0,5$ to współczynnik determinacji $R^2 = r_{xy}^2 = 0,25$ oznacza, że tylko 25% zmian wartości zmiennej objaśnianej zostało wyjaśnione przez zmiany przyjętej zmiennej objaśniającej.
- Można zatem stwierdzić, że na zmienną objaśnianą oddziałują jeszcze inne zmienne objaśniające (przyczyny).

2.1. Analiza regresji

- Omówimy teraz **regresję liniową** prostą.
- Na kształtowanie się zmiennej Y (efekt, skutek, zmienna objaśniana, zmienna zależna) ma wpływ nie tylko zmienna X (przyczyna, zmienna objaśniająca, zmienna niezależna), ale również czynniki zakłócające (szum) nazywane **składnikiem losowym**.
- Najprostsze i najczęściej spotykane zależności między zmiennymi to te, które mają postać liniową.
- Prosta regresji zmiennej Y względem zmiennej X ma następującą postać
$$\hat{y} = ax + b \quad (\hat{Y} = \alpha X + \beta; Y = \alpha X + \beta + \varepsilon, \varepsilon - \text{składnik losowy})$$
- Parametry równania prostej regresji szacuje się z próby klasyczną **metodą najmniejszych kwadratów**.

2.1. Analiza regresji

- **Metoda najmniejszych kwadratów (MNK)** polega na takim oszacowaniu parametrów powyższych funkcji, aby dla danych z próby spełniony był warunek (dla funkcji $\hat{y} = ax + b$):

$$W(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - ax_i - b)^2$$

ma osiągnąć minimum, gdzie y_i oznaczają wartości empiryczne zmiennej Y , a $\hat{y}_i$ wartości „teoretyczne” wyznaczone na podstawie równania $\hat{y} = ax + b$ ($\hat{y}_i = ax_i + b$).

- Ostatecznie otrzymujemy następujące wzory na wartości szukanych współczynników:

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad b = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

2.1. Analiza regresji

- Współczynniki a i b możemy przedstawić w prostszej postaci:

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2} = r_{xy} \frac{s_y}{s_x}$$

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

lub możemy je wyliczyć z zależności:

$$\frac{\hat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}.$$

- Parametry a i b noszą nazwę parametrów prostej regresji.
- Parametr b to wyraz wolny równania prostej regresji.
- Współczynnik a określa o ile jednostek przeciętnie wzrośnie (lub zmaleje, gdy $a < 0$) wartość zmiennej zależnej, gdy wartość zmiennej niezależnej wzrośnie o jedną jednostkę.
- Ze względu na prostą i logiczną interpretację współczynnik a ma duże znaczenie praktyczne.

2.2. Badanie zależności między inteligencją a rozmiarem mózgu

- Otworzyć arkusz **IQ vs Brain**
- Zapoznać się z podstawowymi statystykami opisowymi dla dostępnych zmiennych.
- Utworzyć wykres rozrzutu zmiennej **Brain Size (pixels)** w zależności od zmiennej **Performace IQ**.
- Do wykresu dodać trend liniowy i wzór tej funkcji liniowej oraz współczynnik determinacji R^2 .
- Zapoznać się z działaniem funkcji WSP.KORELACJI, R.KWADRAT, REGLINP, REGLINX.
- Wyznaczyć rozmiar mózgu dla **Performace IQ = 110**.

2.2. Badanie zależności między inteligencją a rozmiarem mózgu

- Otworzyć arkusz **FSIQ vs MRI_Count**
- Zapoznać się z podstawowymi statystykami opisowymi dla dostępnych zmiennych.
- Zapoznać się z wykresem rozrzutu zmiennej **MRI_Count** w zależności od zmiennej **FSIQ**.
- Zapoznać się z wyznaczoną regresją liniową oraz współczynnikiem determinacji R^2 .
- Zapoznać się z działaniem funkcji WSP.KORELACJI, R.KWADRAT, REGLINP, REGLINX.
- Wyznaczyć rozmiar mózgu dla **FSIQ = 110**.

3. Uwagi o wizualizacji wyników

- **Kwartet Anscombe'a** [https://pl.wikipedia.org/wiki/Kwartet_Anscombe%E2%80%99a]
- **Same Stats, Different Graphs** [<https://www.research.autodesk.com/publications/same-stats-different-graphs/>]

- Dziękujemy za uwagę!