

# Statystyka w analizie i planowaniu eksperymentu

Paweł Błażej

23 kwietnia 2014

- 1 Gdy badamy różnego rodzaju zjawiska (np. przyrodnicze) możemy stwierdzić, że na każde z nich ma wpływ działanie innych czynników;

# Korelacja - współczynnik korelacji

- 1 Gdy badamy różnego rodzaju rodzaju zjawiska (np. przyrodnicze) możemy stwierdzić, że na każde z nich ma wpływ działanie innych czynników;
- 2 Wydaje się oczywiste, że badanie związków pomiędzy badanymi zjawiskami jest przedmiotem badań statystycznych;

## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;

## Ważne pytania

## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;
- 2 ilość punktów z kolokwium a ilość punktów z egzaminu końcowego;

## Ważne pytania

## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;
- 2 ilość punktów z kolokwium a ilość punktów z egzaminu końcowego;
- 3 szybkość chodzenia a pozycja społeczna;

## Ważne pytania

## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;
- 2 ilość punktów z kolokwium a ilość punktów z egzaminu końcowego;
- 3 szybkość chodzenia a pozycja społeczna;
- 4 długość snu w ciągu doby a masa ciała.

## Ważne pytania

## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;
- 2 ilość punktów z kolokwium a ilość punktów z egzaminu końcowego;
- 3 szybkość chodzenia a pozycja społeczna;
- 4 długość snu w ciągu doby a masa ciała.

## Ważne pytania

- 1 Czy istnieją jakiegokolwiek zależności pomiędzy wymienionymi charakterystykami?



## Przykłady - „z życia”

- 1 Ilość wypalanych papierosów przez kobietę w trakcie ciąży a waga urodzeniowa noworodka;
- 2 ilość punktów z kolokwium a ilość punktów z egzaminu końcowego;
- 3 szybkość chodzenia a pozycja społeczna;
- 4 długość snu w ciągu doby a masa ciała.

## Ważne pytania

- 1 Czy istnieją jakiegokolwiek zależności pomiędzy wymienionymi charakterystykami?
- 2 A jeśli tak to jak mierzyć taką zależność?

## Uwaga

Zależność pomiędzy takimi charakterystykami nazywamy **związkiem korelacyjnym** albo **korelacją**.

## Zależność funkcyjna - czyli co pamiętamy ze szkoły

Tajemnicza funkcja matematyczna może opisywać zależność pomiędzy zadanymi charakterystykami. Weźmy nasze ulubione ze szkoły:

$$y = f(x), y = ax + b, y = x^2, y = a_n x^n + a_{n-1} x^{n-1} + \dots$$

Przykładów ze szkoły jest bardzo dużo.

## Zależność funkcyjna - czyli co pamiętamy ze szkoły

Tajemnicza funkcja matematyczna może opisywać zależność pomiędzy zadanymi charakterystykami. Weźmy nasze ulubione ze szkoły:

$$y = f(x), y = ax + b, y = x^2, y = a_n x^n + a_{n-1} x^{n-1} + \dots$$

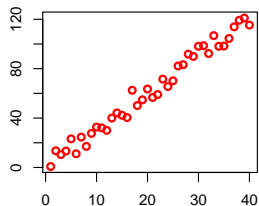
Przykładów ze szkoły jest bardzo dużo.

## Pytanie

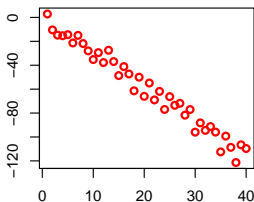
Czy nasze zjawisko da się opisać choćby w przybliżeniu przy pomocy funkcji?

# korelacja - współczynnik korelacji

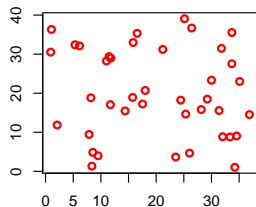
zależność liniowa dodatnia



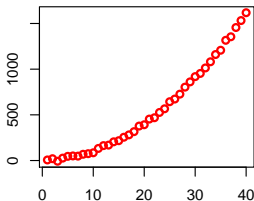
zależność liniowa ujemna



brak zależności



zależność nieliniowa



## Definicja

Współczynnikiem korelacji próbkowej nazywamy zmienną losową:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{S_{nX-1}} \right) \left( \frac{Y_i - \bar{Y}}{S_{nY-1}} \right)$$

# Własności próbkowego współczynnika korelacji

- 1 Próbkowy współczynnik korelacji jest zmienną losową ograniczoną przez liczby  $-1$  i  $1$ ;

Uwaga, Uwaga, ....

**Wartość  $r = 0$  nie świadczy o braku zależności między zmiennymi.**

# Własności próbkowego współczynnika korelacji

- 1 Próbkowy współczynnik korelacji jest zmienną losową ograniczoną przez liczby  $-1$  i  $1$ ;
- 2 Znak współczynnika informuje o kierunku zależności (liniowa ujemna lub liniowa dodatnia);

Uwaga, Uwaga, ....

**Wartość  $r = 0$  nie świadczy o braku zależności między zmiennymi.**



# Własności próbkowego współczynnika korelacji

- 1 Próbkowy współczynnik korelacji jest zmienną losową ograniczoną przez liczby  $-1$  i  $1$ ;
- 2 Znak współczynnika informuje o kierunku zależności (liniowa ujemna lub liniowa dodatnia);
- 3 Wartość bezwzględna  $|r|$  informuje o sile korelacji liniowej;

Uwaga, Uwaga, ....

**Wartość  $r = 0$  nie świadczy o braku zależności między zmiennymi.**

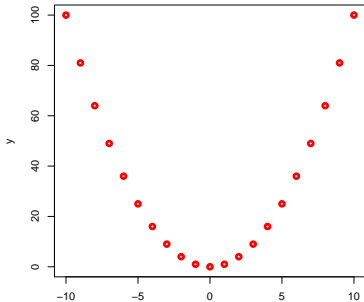
# Własności próbkowego współczynnika korelacji

- 1 Próbkowy współczynnik korelacji jest zmienną losową ograniczoną przez liczby  $-1$  i  $1$ ;
- 2 Znak współczynnika informuje o kierunku zależności (liniowa ujemna lub liniowa dodatnia);
- 3 Wartość bezwzględna  $|r|$  informuje o sile korelacji liniowej;
- 4 W szczególnym przypadku, gdy  $|r| = 1$  mamy do czynienia z dokładnie liniową zależnością.

Uwaga, Uwaga, ....

**Wartość  $r = 0$  nie świadczy o braku zależności między zmiennymi.**

# Jak oszukać współczynnik korelacji



Jak dziecko we mgle

Wartość współczynnika korelacji  
w tym przypadku wynosi  $r = 0$

Wyniki analizy korelacji liniowej dla 17 krajów europejskich (dane z 1990 roku) pomiędzy powierzchnią, liczbą mieszkańców, liczbą urodzeń a liczbą bocianów:

	powierzchnia	liczba bocianów	liczba mieszkańców	liczba urodzeń
powierzchnia	1	0.579	0.812	0.923
liczba bocianów	0.579	1	0.354	0.620
liczba mieszkańców	0.812	0.354	1	0.851
liczba urodzeń	0.923	0.620	0.851	1

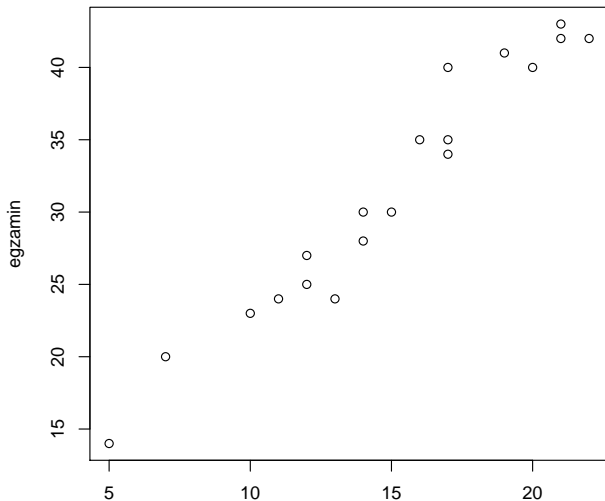
Wyniki analizy korelacji liniowej dla 17 krajów europejskich (dane z 1990 roku) pomiędzy powierzchnią, liczbą mieszkańców, liczbą urodzeń a liczbą bocianów:

	powierzchnia	liczba bocianów	liczba mieszkańców	liczba urodzeń
powierzchnia	1	0.579	0.812	0.923
liczba bocianów	0.579	1	0.354	0.620
liczba mieszkańców	0.812	0.354	1	0.851
liczba urodzeń	0.923	0.620	0.851	1

**I jak tu nie wierzyć w bociana.**

Poniżej na wykresie zostały umieszczone wyniki kolokwium zaliczeniowego i egzaminu z pewnego przedmiotu dla studentów studiujących matematykę. Zależność pomiędzy wynikiem z kolokwium a wynikiem z egzaminu można przedstawić graficznie **wykres rozproszenia**

# Regresja liniowa - Wykres rozproszenia



kolokwium

Paweł Błażej

Statystyka w analizie i planowaniu eksperymentu



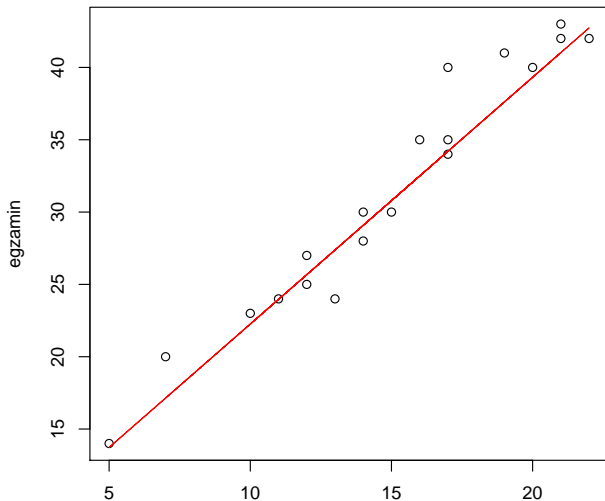
Zauważmy, że w naszym przykładzie dużym (odpowiednio małym) wartościom wyniku kolokwium odpowiadają z reguły duże (małe) wartości egzaminu końcowego. Mówimy w takim przypadku o zależności dodatniej między zmiennymi, w odróżnieniu od zależności ujemnej, gdy duże wartości jednej zmiennej odpowiadają w większości przypadków małym wartościom drugiej zmiennej (zależność ujemna).



Zmienną  $Y$  będącą wynikiem doświadczenia będziemy nazywali **zmienną objaśnianą** w odróżnieniu od  $X$  **zmiennnej objaśniającej** ( za pomocą zmian  $X$  chcemy wyjaśnić zmiany zmiennej  $Y$ ).

Powstaje pytanie - Czy przedstawiona na tym wykresie chmura punktów ma choćby w przybliżeniu zależność funkcyjną ? To znaczy czy punkty z wykresu układają się wzdłuż wykresu pewnej funkcji?

# Regresja liniowa - zależność funkcyjna



kolokwium

Paweł Błażej

Statystyka w analizie i planowaniu eksperymentu



# Współczynnik korelacji

W naszym przypadku (kolokwia, egzamin) współczynnik korelacji wynosi 0.973430184460792 co świadczy o silnej dodatniej zależności pomiędzy tymi zmiennymi.

## Zadanie

Wyznaczenie prostej reprezentującej w sposób adekwatny chmurę punktów z wykresy rozproszenia.

## Zadanie

Wyznaczenie prostej reprezentującej w sposób adekwatny chmurę punktów z wykresy rozproszenia.

## Rozwiązanie

Jeżeli przyjmiemy, że funkcja liniowa ma przybliżać chmurę punktów, to wartość:

$$y_i^* = ax_i + b$$

można interpretować jako wartość przewidywaną  $y_i$  na podstawie rozpatrywanej prostej. Błąd oszacowania wynosi wtedy  $y_i^* - y_i$  (wartość resztowa rezyduum).

Niech

$$S(a, b) = \sum_{i=1}^n (y_i - y_i^*)^2.$$

## Definicja

Prostą regresji opartą na metodzie najmniejszych kwadratów nazywamy prostą  $y = ax + b$ , dla której wartość  $S(a, b)$  jest minimalna.

①  $b = \bar{y} - a\bar{x};$

## Pytanie

Jak te wzory uzyskano?



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

2  $a = r \frac{S_{nX}}{S_{nY}}.$

## Pytanie

Jak te wzory uzyskano?

## Całkowita suma kwadratów

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

## Suma kwadratów błędów

$$SSE = \sum_{i=1}^n (y_i - y_i^*)^2$$

## Regresyjna suma kwadratów

$$SSR = \sum_{i=1}^n (y_i^* - \bar{y})^2$$

iloraz

$$R^2 = \frac{SSR}{SST}$$

nazywamy współczynnikiem determinacji.

## Własności $R^2$

- 1  $R^2$  jest miarą stopnia dopasowania funkcji regresji do danych empirycznych;

iloraz

$$R^2 = \frac{SSR}{SST}$$

nazywamy współczynnikiem determinacji.

## Własności $R^2$

- 1  $R^2$  jest miarą stopnia dopasowania funkcji regresji do danych empirycznych;
- 2 W naszym przypadku  $R^2 = r^2$ .

# Współczynniki prostej regresji

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
5.20	1.76

# Bardziej szczegółowy wydruk

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.0855	-1.2261	-0.1272	0.7530	4.8728

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	5.1999	1.5660	3.321	0.00405	**
x	1.7604	0.1004	17.528	2.56e-12	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 17 degrees of freedom

Multiple R-squared: 0.9476, Adjusted R-squared: 0.9445

F-statistic: 307.2 on 1 and 17 DF, p-value: 2.555e-12