# ZMAD

Zespół biostatystyków ZPZSiA (WNZ)

Uniwersytet Medyczny

# TABELE KONTYNGENCJI

## Wprowadzenie

# Test $\chi^2$

## Czy to prawda?

Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

# Test $\chi^2$

## Czy to prawda?

Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

# Test $\chi^2$

## Czy to prawda?

Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

# Test $\chi^2$

## Czy to prawda?

Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

## **Eksperyment**

Cztery karty ponumerowano liczbami: 1,2,3 oraz 4. Wybieramy losowo jedną kartę. Zebrane osoby odgadują jej numer.

# Test $\chi^2$

## Czy to prawda?

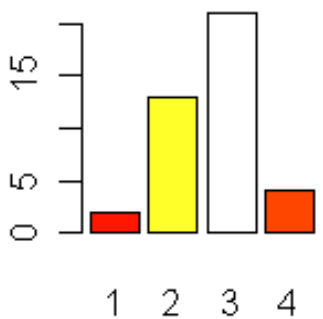
Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

## Eksperyment

Cztery karty ponumerowano liczbami: 1,2,3 oraz 4. Wybieramy losowo jedną kartę. Zebrane osoby odgadują jej numer.

## Wyniki eksperymentu

| Nr karty | Ile razy wybrana |
|---|---|
| 1 | 2 |
| 2 | 13 |
| 3 | 21 |
| 4 | 4 |



# Test $\chi^2$

## Czy to prawda?

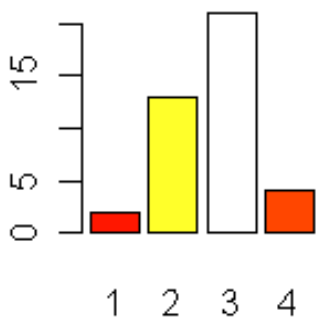
Przy odgadywaniu losowo wybranej liczby unika się wartości skrajnych, ponieważ mniema się, że wynik średni jest wynikiem najbardziej prawdopodobnym.

## Eksperyment

Cztery karty ponumerowano liczbami: 1,2,3 oraz 4. Wybieramy losowo jedną kartę. Zebrane osoby odgadują jej numer.

## Wyniki eksperymentu

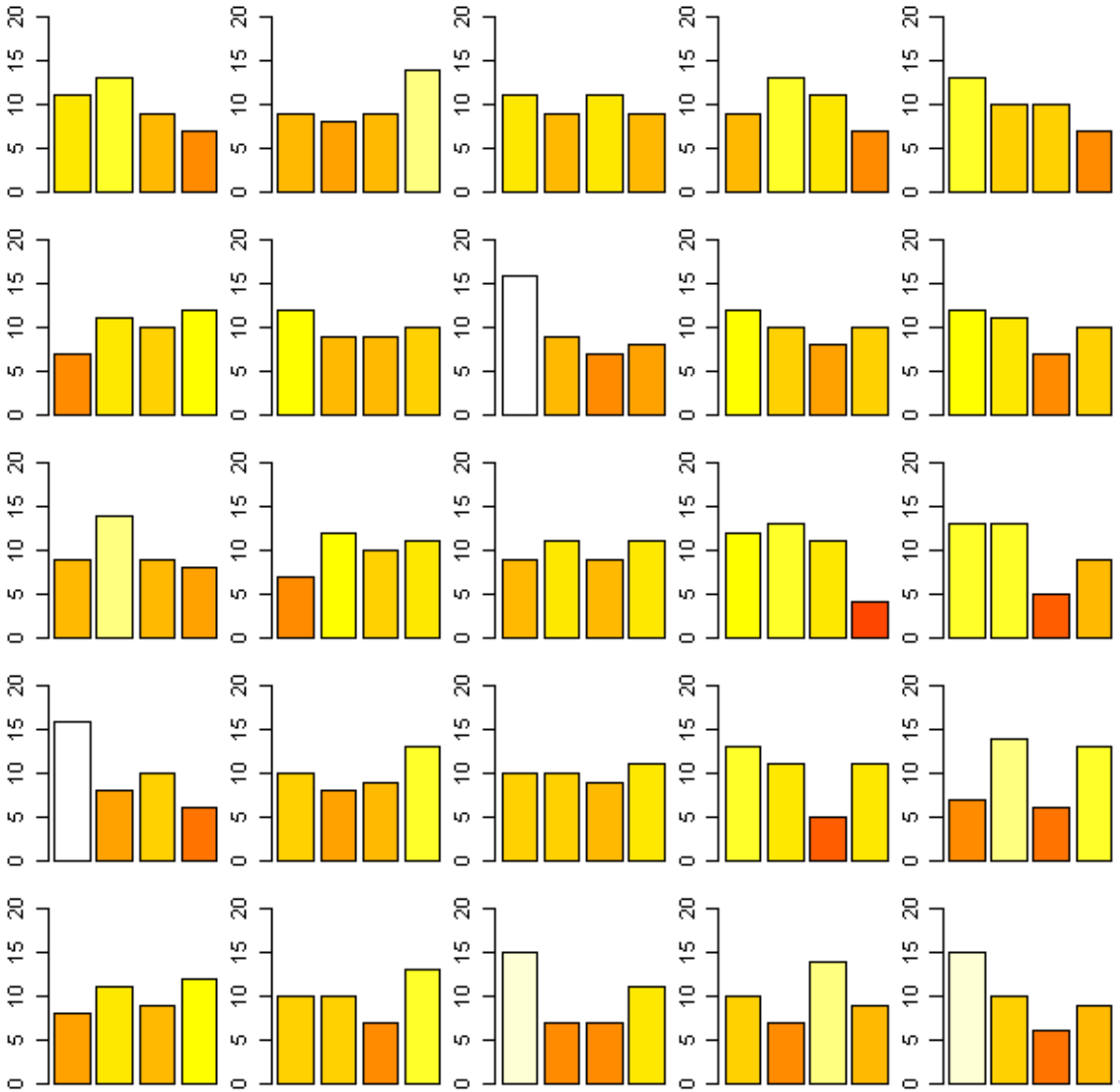
| Nr karty | Ile razy wybrana |
|---|---|
| 1 | 2 |
| 2 | 13 |
| 3 | 21 |
| 4 | 4 |




## O czym świadczą uzyskane wyniki?

# Gdy odgadujemy losowo...

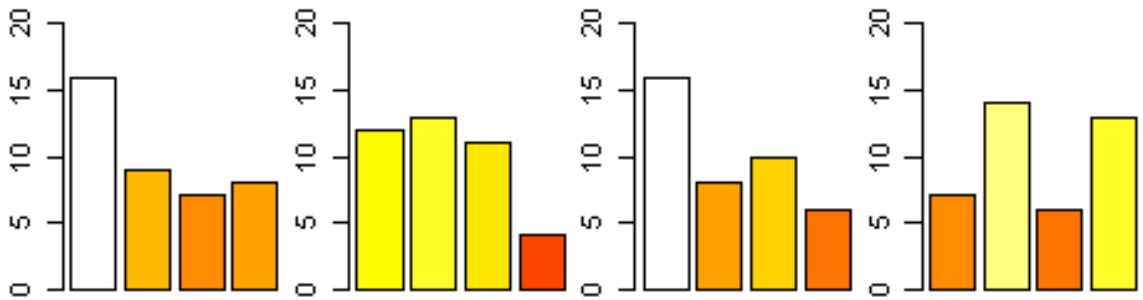
## UKŁADY LOSOWE



| Realizacje statystyki $\chi^2$ | | | | |
|---|---|---|---|---|
| 2.0 | 2.2 | 0.4 | 2.0 | 1.8 |
| 1.4 | 0.6 | 5.0 | 0.8 | 1.4 |
| 2.2 | 1.4 | 0.4 | 5.0 | 4.4 |
| 5.6 | 1.4 | 0.2 | 3.6 | 5.0 |
| 1.0 | 1.8 | 4.4 | 2.6 | 4.2 |

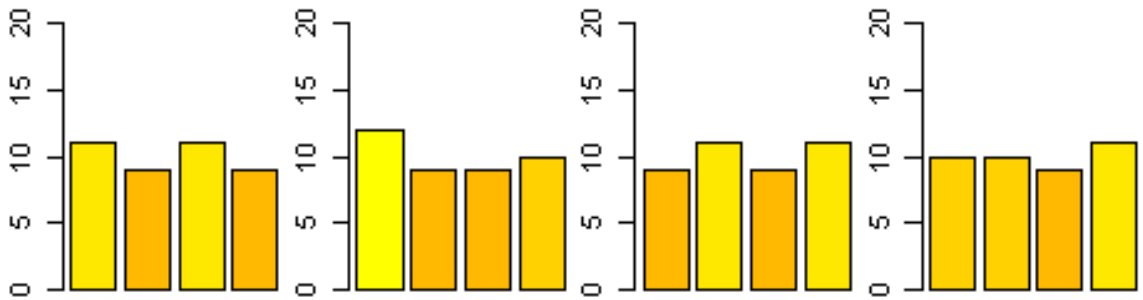
| Realizacje statystyki $\chi^2$ | | | | |
|---|---|---|---|---|
| 2.0 | 2.2 | 0.4 | 2.0 | 1.8 |
| 1.4 | 0.6 | 5.0 | 0.8 | 1.4 |
| 2.2 | 1.4 | 0.4 | 5.0 | 4.4 |
| 5.6 | 1.4 | 0.2 | 3.6 | 5.0 |
| 1.0 | 1.8 | 4.4 | 2.6 | 4.2 |

Cztery największe wartości



| Realizacje statystyki $\chi^2$ | | | | |
|---|---|---|---|---|
| 2.0 | 2.2 | 0.4 | 2.0 | 1.8 |
| 1.4 | 0.6 | 5.0 | 0.8 | 1.4 |
| 2.2 | 1.4 | 0.4 | 5.0 | 4.4 |
| 5.6 | 1.4 | 0.2 | 3.6 | 5.0 |
| 1.0 | 1.8 | 4.4 | 2.6 | 4.2 |

Cztery najmniejsze wartości



Hipoteza: Odgadywano przypadkowo

| Karta $(i)$ | $n_i$ | $n_i^t$ | $\frac{(n_i - n_i^t)^2}{n_i^t}$ |
|---|---|---|---|
| 1 | 2 | 10 | 6.4 |
| 2 | 13 | 10 | 0.9 |
| 3 | 21 | 10 | 12.1 |
| 4 | 4 | 10 | 3.6 |
| Suma | 40 | 40 | 23.0 |

$$\chi^2 = \sum_i \frac{(n_i - n_i^t)^2}{n_i^t} = 23.0$$

$n_i$ – liczebność zaobserwowana

$n_i^t$ – oczekiwana/spodziewana liczebność przy założeniu, że odgadywano całkowicie przypadkowo nie kierując się żadnymi uświadomionymi lub nieuświadomionymi regułami

Hipoteza: Odgadywano przypadkowo

| Karta $(i)$ | $n_i$ | $n_i^t$ | $\frac{(n_i - n_i^t)^2}{n_i^t}$ |
|---|---|---|---|
| 1 | 2 | 10 | 6.4 |
| 2 | 13 | 10 | 0.9 |
| 3 | 21 | 10 | 12.1 |
| 4 | 4 | 10 | 3.6 |
| Suma | 40 | 40 | 23.0 |

$$\chi^2 = \sum_i \frac{(n_i - n_i^t)^2}{n_i^t} = 23.0$$

$n_i$ – liczebność zaobserwowana

$n_i^t$ – oczekiwana/spodziewana liczebność przy założeniu, że odgadywano całkowicie przypadkowo nie kierując się żadnymi uświadomionymi lub nieuświadomionymi regułami

Pytanie: **Czy otrzymana wartość statystyki jest na tyle duża, aby można było uznać, że w odgadując losowo numer karty kierujemy się jakimiś regułami?**

Sposób 1: Dla 2000 losowych układów sprawdzamy, ile razy empiryczna wartość statystyki Chi-kwadrat przekroczy 23. Następnie wyznaczamy procent takich układów:

$$0.04998\%$$

Sposób 2: Wyznaczamy prawdopodobieństwo zdarzenia losowego polegającego na tym, że zmienna losowa o rozkładzie $\chi^2(3)$ zrealizuje się powyżej wartości 23. W ujęciu procentowym otrzymujemy:

$$0.0040383\%$$

## Czy to prawda?

Niezdecydowani wyborcy głosują na pierwsze osoby z listy,których na przykład imiona brzmią ładnie lub znajomo. W przypadku alfabetycznego porządku na listach, kandydaci o nazwiskach zaczynających się na pierwsze litery alfabetu mają większe szanse na wybór niż pozostali kandydaci.

## Czy to prawda?

Niezdecydowani wyborcy głosują na pierwsze osoby z listy,których na przykład imiona brzmią ładnie lub znajomo. W przypadku alfabetycznego porządku na listach, kandydaci o nazwiskach zaczynających się na pierwsze litery alfabetu mają większe szanse na wybór niż pozostali kandydaci.

## Eksperyment

| Pierwsza litera nazwiska | Frakcja wyborców | Liczba kandydatów | Liczba wybranych |
|---|---|---|---|
| A-C | 0.203 | 91 | 47 |
| D-G | 0.179 | 63 | 32 |
| H-L | 0.172 | 59 | 23 |
| M-O | 0.253 | 75 | 22 |
| P-Z | 0.194 | 47 | 20 |
| Suma | 1 | 335 | 144 |

## Czy to prawda?

Niezdecydowani wyborcy głosują na pierwsze osoby z listy,których na przykład imiona brzmią ładnie lub znajomo. W przypadku alfabetycznego porządku na listach, kandydaci o nazwiskach zaczynających się na pierwsze litery alfabetu mają większe szanse na wybór niż pozostali kandydaci.

## Eksperyment

| Pierwsza litera nazwiska | Frakcja wyborców | Liczba kandydatów | Liczba wybranych |
|---|---|---|---|
| A-C | 0.203 | 91 | 47 |
| D-G | 0.179 | 63 | 32 |
| H-L | 0.172 | 59 | 23 |
| M-O | 0.253 | 75 | 22 |
| P-Z | 0.194 | 47 | 20 |
| Suma | 1 | 335 | 144 |

## Jakie hipotezy można postawić i zweryfikować?

Hipoteza: Kandydaci zostali nominowani przez partię bez względu na nazwisko

Hipoteza: Porządek alfabetyczny na listach nie wpływa na wyniki wyborów

| Pierwsza litera nazwiska | Liczba kandydatów | Oczekiwana liczba kandydatów |
|---|---|---|
| A-C | 91 | 0.203(335)=68.0 |
| D-G | 63 | 0.179(335)=59.9 |
| H-L | 59 | 0.172(335)=57.6 |
| M-O | 75 | 0.253(335)=84.7 |
| P-Z | 47 | 0.194(335)=65.0 |
| Suma | 335 | 335.2 [335 bez zaokrągleń] |

$$\chi^2 = 14.0785,\ df = 4,\ p - value = 0.007049$$

Hipoteza: Kandydaci zostali nominowani przez partię bez względu na nazwisko

Hipoteza: Porządek alfabetyczny na listach nie wpływa na wyniki wyborów

| Pierwsza litera nazwiska | Liczba wybranych | Oczekiwana liczba wybrnych |
|---|---|---|
| A-C | 47 | (91/335)144=39.1 |
| D-G | 32 | (63/335)144=27.1 |
| H-L | 23 | (59/335)144=25.4 |
| M-O | 22 | (75/335)144=32.2 |
| P-Z | 20 | (47/335)144=20.2 |
| Suma | 144 | 144 |

$$\chi^2 = 5.9562,\ df = 4,\ p - value = 0.2024$$

## Czy to prawda?

Starsze rodzeństwo jest bardziej prawomyślne niż młodsze.

## Czy to prawda?

Starsze rodzeństwo jest bardziej prawomyślne niż młodsze.

## Eksperyment

1137 chłopców sklasyfikowano na dwie kategorie: bardzie i mniej niegrzecznych. Następnie sprawdzono, w jakiej kolejności się rodzili.

## Czy to prawda?

Starsze rodzeństwo jest bardziej prawomyślne niż młodsze.

## Eksperyment

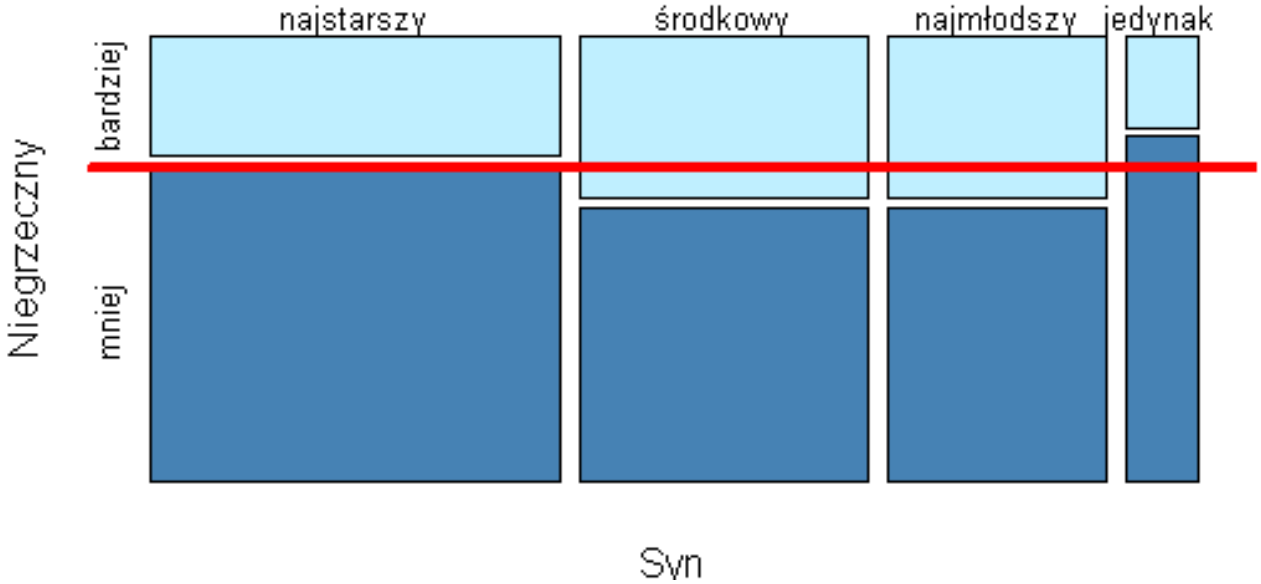
1137 chłopców sklasyfikowano na dwie kategorie: bardzie i mniej niegrzecznych. Następnie sprawdzono, w jakiej kolejności się rodzili.

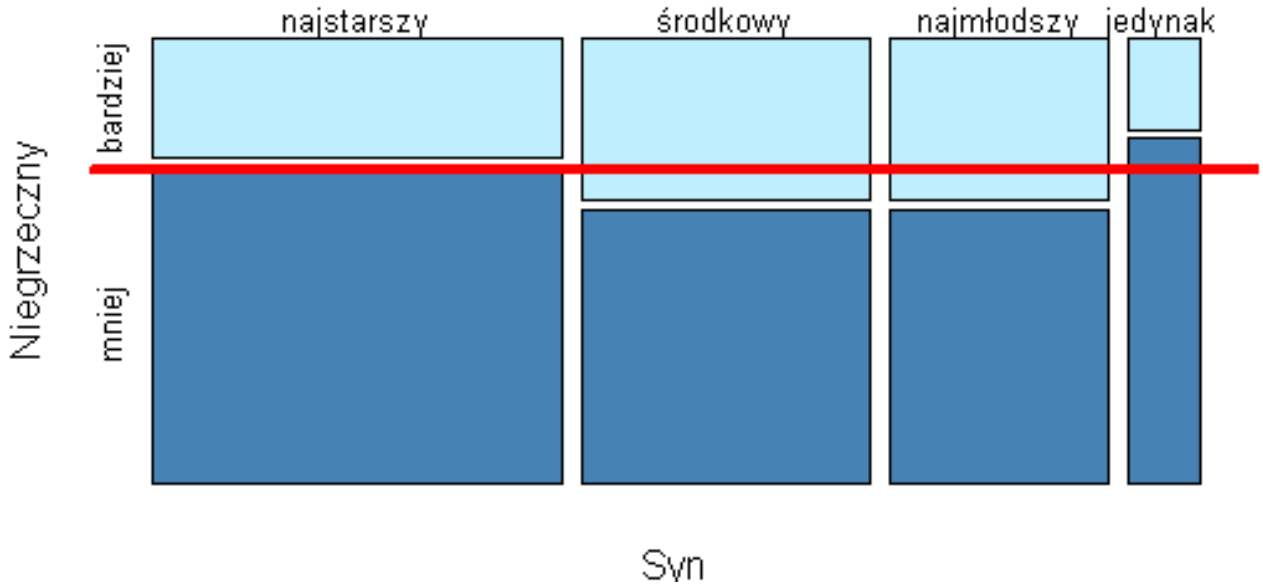
| Niegrzeczny | Najstarszy | Środkowy | Najmłodszy | Jedynak | Suma |
|---|---|---|---|---|---|
| bardziej | 127 | 123 | 93 | 17 | 360 |
| mniej | 345 | 209 | 158 | 65 | 777 |
| Suma | 472 | 332 | 251 | 82 | 1137 |

Liczebności

| Niegrzeczny | Najstarszy | Środkowy | Najmłodszy | Jedynak | Suma |
|---|---|---|---|---|---|
| bardziej | 26.9% | 37.0% | 37.1% | 20.7% | 31.7% |
| mniej | 73.1% | 63.0% | 62.9% | 79.3% | 68.3% |
| Suma | 100% | 100% | 100% | 100% | 100% |

Kolumnowy rozkład procentowy


najstarszy
środkowy
najmłodszy
jedynak
bardziej
mniej
Niegrzeczny
Syn




| Niegrzeczny | Najstarszy | Środkowy | Najmłodszy | Jedynak | Suma |
|---|---|---|---|---|---|
| bardziej | 149.4 | 105.1 | 79.5 | 26.0 | 360 |
| mniej | 322.6 | 226.9 | 171.5 | 56.0 | 777 |
| Suma | 472 | 332 | 251 | 82 | 1137 |

Liczebności oczekiwane

$$149.4 = 0.371 \cdot 472$$
$$105.1 = 0.371 \cdot 332$$
$$56.0 = 0.683 \cdot 82$$

$$\chi^2 = \sum \frac{(Zaobserwowana - Oczekiwana)^2}{Oczekiwana}$$

$$\chi^2 = \frac{(127-149.5)^2}{149.5} + \frac{(123-105.1)^2}{105.1} + \ldots + \frac{(82-56.0)^2}{56.0}$$

$$\chi^2 = 17.2816, \quad df = 3, \quad p - value = 0.0006185$$

# TABELE $2\ x\ 2$

Różne sposoby zbierania danych

Forma przedstawiania danych

| | X | | |
|---|---|---|---|
| $Y$ | 0 | 1 | Suma |
| 0 | $d$ | $c$ | $r$ |
| 1 | $b$ | $a$ | $s$ |
| Suma | $m$ | $n$ | $N$ |

Literki $a,\ b,\ c,\ d$ oznaczają liczebności.

POSTAĆ TESTU $\chi^2$ zależy od sposobu zebrania danych

$$\chi^2_1 = \frac{[a(m-b) - b(n-a)]^2 N}{mnsr}, \quad \chi^2_2 = \frac{(ad-bc)^2(c+d)}{N(a+b)cd},$$

$$\chi^2_3 = u^2, \quad u = \frac{b-c}{\sqrt{b+c}}$$

# Eksperyment 1, $\chi^2_1$

Jeden z marginesów tabelki jest ustalony.

## Przykład

Chcemy porównać skuteczność dwóch terapii w leczeniu choroby oczu. Do każdej terapii przydzielamy losowo po 28 pacjentów. Po zakończonych terapiach zliczamy zdrowych i chorych pacjentów.

| Choroba | Terapia $1$ | Terapia $2$ |
|---|---|---|
| Nie ustąpiła | 14 | 15 |
| Ustąpiła | 14 | 13 |
| Ogółem | 28 | 28 |

# Eksperyment 2, $\chi^2_1$

Ustalona jest całkowita liczba obserwacji

## Przykład

Chcemy zbadać, czy poziom hormonów u pacjenta ma wpływ na wynik leczenia. W tym celu wybieramy losowo stu pacjentów, z których część ma niski poziom hormonów, a część wysoki. Po zakończeniu procedury leczenia, tworzymy tabelkę:

| Poziom hormonu | Wynik leczenia | | Suma |
|---|---|---|---|
| | Negatywny | Pozytywny | |
| Niski | 18 | 65 | |
| Wysoki | 10 | 7 | |
| | | | 100 |

# Eksperyment 3, $\chi^2_1$

Wszystkie wartości w tabelce uzyskane są drogą losową

## Przykład 1

Badamy potomstwo pewnego wodnego organizmu. Część potomstwa trafia losowo do grupy kontrolnej, a część do grupy ryzyka. W obu grupach po zakończeniu eksperymentu notujemy liczbę żeńskich i męskich osobników, które przetrwały osiągając dojrzałość. Początkowa liczba osobników nie jest znana.

| | Osobniki | |
|---|---|---|
| Grupa | żeńskie | męskie |
| kontrolna | 363 | 103 |
| ryzyka | 376 | 182 |

W niektórych badaniach notowane liczebności są związane z

- czasem obserwacji
- rozmiarem obszaru, z którego pochodzą obserwacje

### Przykład 2

Badamy czy praca przy pewnych chemikaliach ma związek z ryzykiem zachorowania na raka i czy ryzyko to zależy od płci. W tym celu przez określony czas obserwujemy grupę kontrolną i grupę narażoną na działanie chemikaliów. Otrzymujemy wyniki

| | Płeć | |
|---|---|---|
| Chemikalia | żeńska | męska |
| Tak | (482; 8773) | (85; 1974) |
| Nie | (549; 13634) | (13; 303) |

(liczba zachorowań; łączny czas obserwacji)

# Eksperyment 4, $\chi^2_1$

Badania retrospektywne

## Przykład

Podejrzewamy, że wykonywanie pewnej pracy może być przyczyną choroby, która ujawnia się dopiero po wielu latach. Aby zweryfikować te podejrzenia, wybieramy do badań dwie grupy osób. Pierwsza grupa składa się z 50 zdrowych osób, a druga z 20 chorych. Sprawdzamy, które osoby wykonywały tę pracę.

| | Praca | | Suma |
|---|---|---|---|
| Choroba | Nie | Tak | |
| Nie | 12 | 38 | 50 |
| Tak | 2 | 18 | 20 |

# Eksperyment 5, $\chi^2_2$

## Przykład

W leczeniu pewnej choroby stosuje się zamiennie dwa preparaty. Wiadomo, że mogą one powodować uszkodzenie wątroby. Chcąc porównać szkodliwość obu preparatów wykonujemy następujący eksperyment. Z grupy pacjentów, którym podano pierwszy preparat losujemy osoby do momentu uzyskania pięciu przypadków uszkodzenia wątroby. To samo robimy w stosunku do pacjentów, którzy otrzymali drugi preparat. Wyniki losowania przedstawiamy w tabelce.

| | Preparat | |
|---|---|---|
| Uszkodzenie | pierwszy | drugi |
| Tak | 5 | 5 |
| Nie | 53 | 312 |

## Zadanie

Porównywano skuteczność dwóch terapii w leczeniu pewnej choroby. Do każdej terapii przydzielono losowo po 36 pacjentów. Po zakończonych terapiach zliczono zdrowych i chorych pacjentów.

| Choroba | Terapia 1 | Terapia 2 |
|---|---|---|
| nie ustąpiła | 18 | 25 |
| ustąpiła | 18 | 11 |
| Ogółem | 36 | 36 |

1. Które wartości w tabelce są realizacjami zmiennych losowych, a które są ustalone przez prowadzącego eksperyment?
2. Czy na podstawie tabelki można wyznaczyć odsetek osób biorących udział w terapii pierwszej? Czy można uznać, że odsetek ten jest rozsądnym oszacowaniem odsetka osób, którym lekarze przydzielają terapię pierwszą w całej populacji pacjentów, z której pochodzi badana próba osób?
3. Czy można uznać, że liczba wyleczonych pacjentów za pomocą terapii pierwszej jest realizacją zmiennej losowej z rozkładu dwumianowego?
4. W jaki sposób można oszacować odsetek pacjentów leczonych za pomocą terapii pierwszej, którzy przychodzą do zdrowia?
5. Ile razy skuteczniejsza jest terapia pierwsza od drugiej?
6. Stosunek liczby pacjentów leczonych skutecznie do liczby pacjentów leczonych nieskutecznie zależy od terapii. Ile razy większy jest ten stosunek przy terapii pierwszej niż przy drugiej?

## Zadanie

Chcemy zbadać, czy poziom hormonów u pacjenta ma wpływ na wynik leczenia. W tym celu wybieramy losowo stu pacjentów, z których część ma niski poziom hormonów, a część wysoki. Po zakończeniu procedury leczenia, tworzymy tabelkę:

| Poziom hormonu | Wynik leczenia | | Suma |
|---|---|---|---|
| | negatywny | pozytywny | |
| niski | 18 | 65 | 83 |
| wysoki | 10 | 7 | 17 |
| Suma | 28 | 72 | 100 |

1. Które wartości w tabelce są realizacjami zmiennych losowych, a które są ustalone przez prowadzącego eksperyment?
2. Czy na podstawie przeprowadzonego eksperymentu można oszacować odsetek pacjentów o wysokim poziomie hormonu?
3. Czy na podstawie przeprowadzonego eksperymentu można oszacować skuteczność leczenia?
4. Czy na podstawie przeprowadzonego eksperymentu można ocenić, jaki odsetek pacjentów o wysokim poziomie hormonu kończy leczenie z wynikiem pozytywnym?
5. Czy na podstawie przeprowadzonego eksperymentu można ocenić, jaki odsetek pacjentów, którzy kończą leczenie z wynikiem pozytywnym, charakteryzuje się niskim poziomem hormonu?
6. Stosunek liczby pacjentów kończących leczenie z wynikiem pozytywnym do liczby pacjentów kończących leczenie z wynikiem negatywnym zależy od poziomu hormonu. Ile razy większy jest ten stosunek w przypadku niskiego poziomu hormonu, niż w przypadku, gdy poziomu tego hormonu jet wysoki?

## Zadanie

Pobrano próbkę pewnego płynu i rozdzielono ją na dwie części. Następnie jedną część poddano działaniu pierwszego preparatu, a drugą drugiego. Po zakończeniu eksperymentu w obu częściach zliczono liczbę osobników żeńskich i męskich pewnego organizmu. Liczono osobniki, które przetrwały działanie preparatów. Uzyskano wyniki:

| | Osobniki | | |
|---|---|---|---|
| Preparat | żeńskie | męskie | Suma |
| pierwszy | 363 | 103 | 466 |
| drugi | 376 | 182 | 558 |
| Suma | 736 | 285 | 1024 |

1. Które wartości w tabelce są realizacjami zmiennych losowych, a które są ustalone przez prowadzącego eksperyment?
2. Czy na podstawie wyników w tabelce można oszacować odsetek żeńskich osobników w płynie, z którego pochodzi próbka?
3. Czy można uznać, że odsetek osobników męskich jest mniejszy od odsetka osobników żeńskich, jeżeli dotyczy on osobników, które są w stanie przeżyć pod działaniem preparatu pierwszego?
4. Ile razy szansa na przetrwanie osobników męskich jest większa przy preparacie drugim niż przy preparacie pierwszym?

## Zadanie

Podejrzewano, że wykonywanie pewnej pracy może być przyczyną choroby, która ujawnia się dopiero po wielu latach. Aby zweryfikować te podejrzenia, wybrano do badań dwie grupy osób. Pierwsza grupa składała się z 50 zdrowych osób, a druga z 20 chorych. Sprawdzono, które z tych osób wykonywały tę pracę.

| | Praca | | |
|---|---|---|---|
| Choroba | Nie | Tak | Suma |
| Nie | 12 | 38 | 50 |
| Tak | 2 | 18 | 20 |

1. Które wartości w tabelce są realizacjami zmiennych losowych, a które są ustalone przez prowadzącego eksperyment?
2. Czy na podstawie tabelki można oszacować odsetek osób cierpiących na tę chorobę?
3. Czy można oszacować (w odniesieniu do populacji) odsetek cierpiących na tę chorobę wśród osób wykonujących tę pracę?
4. Stosunek liczby osób wykonujących tę pracę do liczby osób niewykonujących tej pracy zależy od tego, czy ktoś cierpi na badaną chorobę. Ile razy większy jest ten stosunek dla chorych w odniesieniu do zdrowych?
5. Stosunek liczby cierpiących na tę chorobę do liczby osób zdrowych zależy od tego, czy ktoś wykonywał daną pracę. Ile razy ten stosunek jest większy dla wykonujących tę pracę w odniesieniu do niewykonujących tej pracy?

## Zadanie

W leczeniu pewnej choroby stosuje się zamiennie dwa preparaty. Wiadomo, że mogą one powodować uszkodzenie wątroby. Chcąc porównać szkodliwość obu preparatów wykonano następujący eksperyment. Z grupy pacjentów, którym podano pierwszy preparat losowano osoby do momentu uzyskania pięciu przypadków uszkodzenia wątroby. To samo zrobiono w stosunku do pacjentów, którzy otrzymali drugi preparat. Uzyskano wyniki:

| | Preparat | |
|---|---|---|
| Uszkodzenie | pierwszy | drugi |
| Tak | 5 | 5 |
| Nie | 53 | 312 |

1. Które wartości w tabelce są realizacjami zmiennych losowych, a które są ustalone przez prowadzącego eksperyment?
2. Czy w tym przypadku można zastosować klasyczny test $\chi^2$ na niezależność, aby zbadać zależność uszkodzenia od podawanego preparatu?

## Zadanie

Przebadano 195 pacjentów pod kątem obecności pewnych bakterii. Wykryto je u 103 osób. Po upływie 6 miesięcy leczenia przebadano tych pacjentów ponownie. Bakterie wykryto u 47 osób, w tym u 39, u których wynik wcześniejszego badania był pozytywny.

| | Po leczeniu | |
|---|---|---|
| Przed leczeniem | wynik pozytywny | wynik negatywny |
| wynik pozytywny | 39 | 64 |
| wynik negatywny | 8 | 84 |

1. Zgodnie z treścią zadania, leczono osoby, u których nie wykryto bakterii. Co można o tym sądzić?
2. Czy o wpływie leczenia na obecność bakterii może nam coś powiedzieć porównanie wartości 39 z 84?
3. Czy o wpływie leczenia na obecność bakterii może nam coś powiedzieć porównanie wartości 8 z 64?
4. Czy dane z tabelki sugerują nam, że leczenie ma wpływ na zmniejszenie się liczby osób zakażonych?

## Zadanie

Przeprowadzono ankietę, w której pytano się, czy wirusem HIV można się zakazić przez podanie ręki. Po zastosowanej ankiecie przeprowadzono wśród uczestników wykład na temat choroby AIDS i metodach przenoszenia wirusa HIV. Po wykładzie ponowiono pytanie. Uzyskano wyniki:

| | Po wykładzie | | |
|---|---|---|---|
| Przed wykładem | Tak | Nie | Suma |
| Tak | 20 | 47 | 67 |
| Nie | 0 | 33 | 33 |
| Suma | 20 | 80 | 100 |

Czy przeprowadzony wykład wpłynął na rodzaj udzielanych odpowiedzi?

## Zadanie

Porównywano dwie metody leczenia. Aby zredukować zmienność osobniczą, dobrano osoby w pary, porównywalne ze względu na płeć, wiek oraz status ekonomiczny. Jedna osoba z pary była leczona jedną metodą, a druga drugą. Metoda leczenia była przydzielana do osoby w parze losowo. Otrzymano wyniki:

| | Osoba leczona metodą 2 | | |
|---|---|---|---|
| Osoba leczona metodą 1 | wyleczona | niewyleczona | Suma |
| wyleczona | 33 | 47 | 80 |
| niewyleczona | 27 | 33 | 60 |
| Suma | 60 | 80 | 140 |

Czy można uznać, że jedna z metod jest lepsza? Czy w tym przypadku możemy zastosować test postaci $\chi^2_3$ ?

# ANALIZA WARIANCJI

## Wprowadzenie

# Jednoczynnikowa analiza wariancji

**Model**
$Y_{ij} \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$, są niezależnymi zmiennymi losowymi

**Porównanie wartości średnich**

$$H_0 :\ \mu_1 = \ldots = \mu_k$$

**Test** $F$ (poziom istotności $\alpha$)
**Statystyka testowa**

$$F_{\text{emp}} = \frac{S_a^2}{S_e^2}$$

Jeżeli $F_{\text{emp}} > F(\alpha; k-1, N-k)$, to hipotezę $H_0 : \mu_1 = \cdots = \mu_k$ odrzucamy.

**Wniosek praktyczny:** przynajmniej jedna ze średnich $\mu_1, \ldots, \mu_k$ jest inna od pozostałych

# Podział całkowitej sumy kwadratów

$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{zmienność całkowita}} = \underbrace{\bar{Y}_{i.} - \bar{Y}_{..}}_{\text{zmienność średnich}} + \underbrace{Y_{ij} - \bar{Y}_{i.}}_{\text{zmienność wokół średniej}}$$

$$\underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2}_{SST} = \underbrace{\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{SSTR} + \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2}_{SSE}$$

$SST$ – całkowita suma kwadratów

$SSTR$ – suma kwadratów dla czynnika

$SSE$ – suma kwadratów reszt

$S_a^2 = SSTR/(k-1), \quad S_e^2 = SSE/(N-k), \quad N = \sum_i n_i$

**Grupy jednorodne** — podzbiory średnich, które można uznać za takie same

**Procedury porównań wielokrotnych** — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheffégo, Bonfferroniego, Duncana, Newmana–Kuelsa i inne.

Ogólna idea procedur porównań wielokrotnych
$(n_1 = \cdots = n_k)$

$NIR$ — najmniejsza istotna różnica

Jeżeli $|\bar{Y}_i - \bar{Y}_j| < NIR$, to uznajemy, że $\mu_i = \mu_j$. Jeżeli

$|\bar{Y}_i - \bar{Y}_j| < NIR$
$|\bar{Y}_i - \bar{Y}_l| < NIR$
$|\bar{Y}_l - \bar{Y}_j| < NIR$,
to uznajemy, że $\mu_i = \mu_j = \mu_l$.

Badając w ten sposób wszystkie pary średnich próbkowych otrzymujemy podział zbioru średnich na grupy jednorodne.

# Procedura Tukeya

Założenie: $n_1 = \cdots = n_k = n$

$$NIR = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{n}}$$

$t(\alpha; k, N-k)$ — wartość krytyczna studentyzowanego rozstępu
Przypadek nierównolicznych prób
Jedna z modyfikacji procedury Tukeya

$$NIR_{ij} = t(\alpha; k, N-k) S_e \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

# Dwuczynnikowa analiza wariancji

## Przykład

Obserwujemy czas przetrwania zwierząt losowo rozdzielonych do dwunastu grup.

| | KURACJA | | | |
|---|---|---|---|---|
| SUBSTANCJA | A | B | C | D |
| I | 0.31 | 0.82 | 0.43 | 0.45 |
| | 0.45 | 1.10 | 0.45 | 0.71 |
| | 0.46 | 0.88 | 0.63 | 0.66 |
| | 0.43 | 0.72 | 0.76 | 0.62 |
| II | 0.36 | 0.92 | 0.44 | 0.56 |
| | 0.29 | 0.61 | 0.35 | 1.02 |
| | 0.40 | 0.49 | 0.31 | 0.71 |
| | 0.23 | 1.24 | 0.4 | 0.38 |
| III | 0.22 | 0.30 | 0.23 | 0.30 |
| | 0.21 | 0.37 | 0.25 | 0.36 |
| | 0.18 | 0.38 | 0.24 | 0.31 |
| | 0.23 | 0.29 | 0.22 | 0.33 |

## Tabelka analizy wariancji

| źródło zmienności | suma kwadratów | stopnie swobody | średnia suma kwadratów | Femp |
|---|---|---|---|---|
| KURACJA | 1.0333 | 2 | 0.5165 | 23.2 |
| SUBSTANCJA | 0.9224 | 3 | 0.3075 | 13.8 |
| WSPÓŁDZIAŁANIE | 0.2501 | 6 | 0.0417 | 1.9 |
| BŁĄD | 0.8007 | 36 | 0.0222 | |
| OGÓŁEM | 3.0062 | 47 | | |

Wartości krytyczne: $3.259446, 2.866266, 2.363751$

# Porównania szczegółowe

| | Różnica | Lewy | Prawy | p-wartość |
|---|---|---|---|---|
| B-A | 0.363 | 0.199 | 0.526 | 0.000 |
| C-A | 0.078 | -0.086 | 0.242 | 0.577 |
| D-A | 0.220 | 0.056 | 0.384 | 0.005 |
| C-B | -0.284 | -0.448 | -0.120 | 0.000 |
| D-B | -0.142 | -0.306 | 0.021 | 0.108 |
| D-C | 0.142 | -0.022 | 0.306 | 0.111 |

| | Różnica | Lewy | Prawy | p-wartość |
|---|---|---|---|---|
| II-I | -0.073 | -0.202 | 0.056 | 0.358 |
| III-I | -0.341 | -0.470 | -0.212 | 0.000 |
| III-II | -0.268 | -0.397 | -0.139 | 0.000 |

## Zadanie

Zmierzono poziom mleczanów u pacjentów wylosowanych z pięciu grup: A, B, C, D, E. Zbadać, czy grupy te różnią się ze względu na średni poziom mleczanów. Jeżeli tak, przeprowadzić porównania szczegółowe.

| A | B | C | D | E |
|---|---|---|---|---|
| 5.24 | 6.51 | 9.49 | 9.01 | 7.29 |
| 3.48 | 8.48 | 9.79 | 10.78 | 8.26 |
| 5.05 | 5.86 | 9.11 | 10.91 | 9.23 |
| 2.42 | 7.58 | 6.57 | 11.21 | 10.2 |
| 3.06 | 5.40 | 8.15 | 11.41 | 7.57 |
| 2.01 | 5.39 | 11.39 | 12.94 | 7.29 |
| 3.49 | 7.59 | 10.49 | 9.01 | 7.29 |
| 5.06 | 5.70 | 9.78 | 9.78 | 8.26 |
| 2.41 | 6.53 | 10.11 | 9.93 | 7.29 |
| 2.03 | 8.47 | 6.57 | 11.2 | 9.21 |
| 5.25 | 5.4 | 8.19 | 10.41 | 7.57 |
| 3.07 | 5.38 | 11.3 | 12.74 | 8.00 |

## Zadanie

Zmierzono poziom cytokin (pg/ml) u chorych z czterech grup. Czy grupy te różnią się ze względu na średni poziom cytokin. Jeżeli tak, przeprowadzić porównania szczegółowe.

| A | B | C | D |
|---|---|---|---|
| 29 | 34 | 38 | 43 |
| 30 | 44 | 37 | 37 |
| 37 | 41 | 31 | 50 |
| 32 | 35 | 33 | 36 |
| 27 | 34 | 38 | 44 |
| 34 | 39 | 33 | 43 |
| 34 | 40 | 40 | 40 |
| 29 | 42 | 32 | 38 |
| 35 | 41 | 41 | 36 |
| 37 | 38 | 38 | 47 |
| 41 | 33 | 33 | 39 |

## Zadanie

Badano poziom leukocytów ($\times 10^3/\mu L$) we krwi w zależności od płci i wieku w grupie pacjentów. Uzyskane wyniki zebrano w tabeli. Przeprowadzić analizę wariancji.

| Wiek | męska | żeńska |
|---|---|---|
| młody | 12.45 | 6.61 |
| | 8.40 | 5.77 |
| | 10.32 | 8.81 |
| | 11.76 | 8.11 |
| | 12.44 | 2.25 |
| stary | 4.78 | 4.28 |
| | 8.03 | 5.37 |
| | 5.27 | 3.15 |
| | 9.90 | 4.44 |
| | 7.31 | 7.30 |
| średni | 7.05 | 8.04 |
| | 11.18 | 9.78 |
| | 8.78 | 3.96 |
| | 5.92 | 3.75 |
| | 7.39 | 7.88 |

# REGRESJA

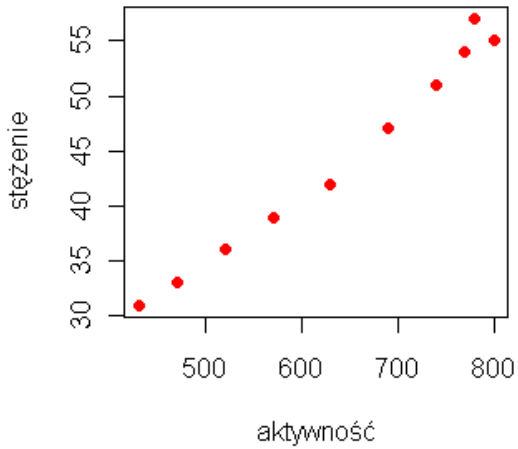
## Wprowadzenie

## Przykład

W pewnej klinice badano związek między aktywnością enzymów aminotransferazy a stężeniem amoniaku we krwi u chorych z ostrą niewydolnością wątroby. Pobrano losową próbę 10 pacjentów i otrzymano następujące wyniki:

| aktywność | 430 | 470 | 520 | 570 | 630 | 690 | 740 | 770 | 800 | 780 |
|---|---|---|---|---|---|---|---|---|---|---|
| stężenie | 31 | 33 | 36 | 39 | 42 | 47 | 51 | 54 | 55 | 57 |

1. Czy stężenie amoniaku we krwi można rozpoznać po poziomie aktywności enzymów?
2. Czy w przypadku istnienia takiej zależności, można ją przedstawić w zwartej formie, za pomocą funkcji?




Ocena wykresu daje pewne podstawy do przyjęcia następującej zależności:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n,$$

gdzie
$y_i$ – stężenie amoniaku we krwi $i$ – tego pacjenta
$x_i$ – aktywność enzymu w organiźmie $i$ – tego pacjenta
$e_i$ – błąd dopasowania

# Metoda najmniejszych kwadratów

Parametry $\beta_0$ oraz $\beta_1$ dobieramy tak, aby średniokwadratowy błąd dopasowania, mianowicie $\sum_i e_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$, był minimalny. W ten sposób dobrane parametry oznaczamy przez $\hat{\beta}_0$ oraz $\hat{\beta}_1$. Wyrażają się one wzorami

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x}_i)^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

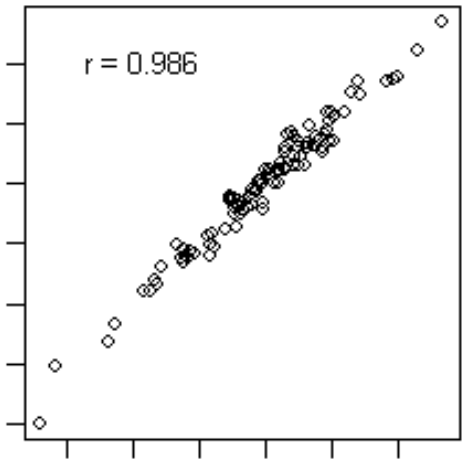
Wówczas

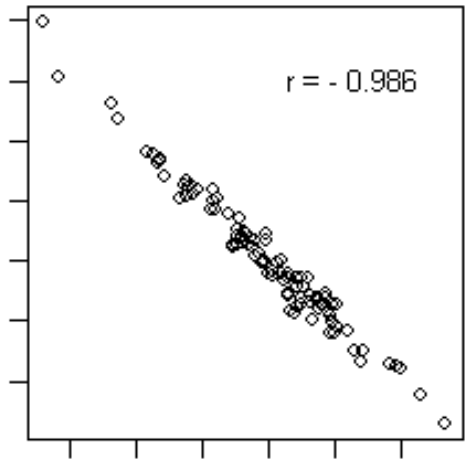
$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (1 - r^2) \sum_i (y_i - \bar{y})^2,$$

gdzie

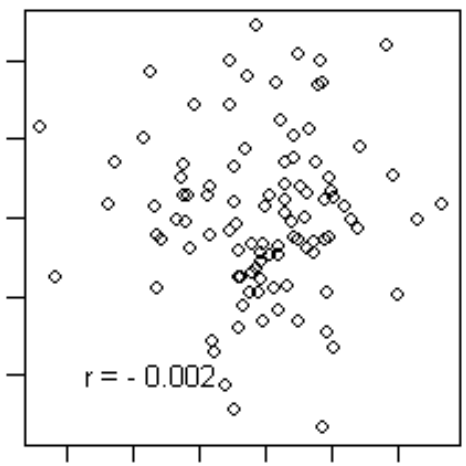
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i (x_i - \bar{x})^2}}$$

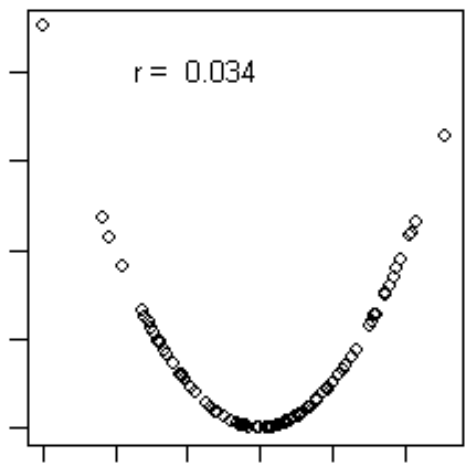
Współczynnik $r$ jest miernikiem zależności liniowej.




**Wartość r jest zawsze z przedziału <-1,1>**



## Przykład, ciąg dalszy

1. Czy stężenie amoniaku we krwi można rozpoznawać po aktywności enzymów?
2. W jaki sposób wyznaczyć ewentualną zależność liniową między stężeniem amoniaku a aktywnością enzymów? W jaki sposób sprawdzić wiarygodność tej zależności?

Aby odpowiedzieć na postawione pytania, musimy przyjąć model statystyczny, który pozwoli nam na uogólnienie wniosków z próby 10 pacjentów na całą populację pacjentów.

Przyjmujemy następujący model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Przyjmujemy następujący model:

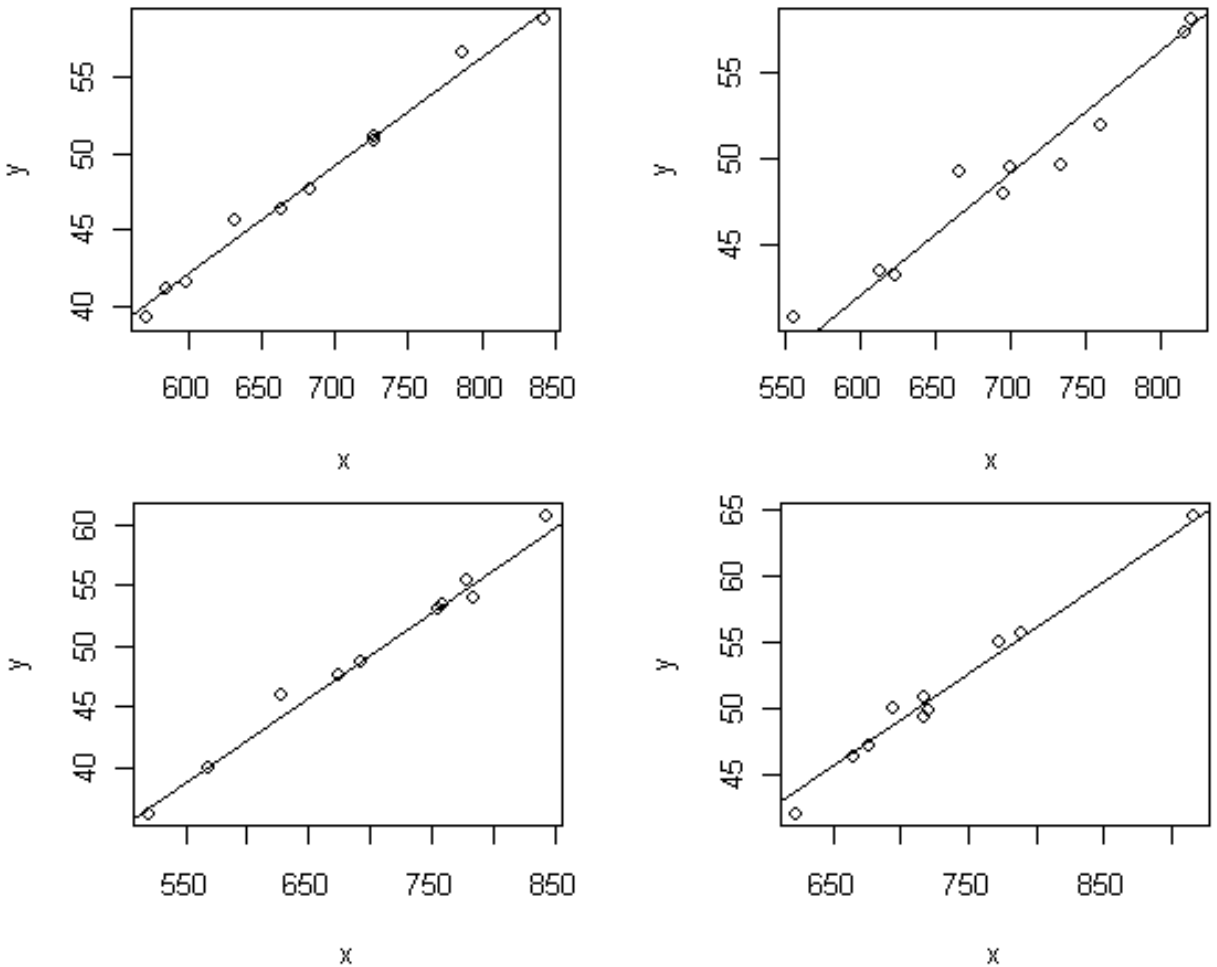
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

gdzie $\varepsilon_i, \quad i = 1, 2, \ldots, n$, są niezależnymi zmiennymi losowymi o tym samym rozkładzie $N(0, \sigma^2)$.

Uwagi

- $Y_1, Y_2, \ldots, Y_n$, są zmiennymi losowymi, a $y_1, y_2, \ldots, y_n$ są ich realizacjami.
- Model dotyczy rozkładu warunkowego $Y|X = x$.

Cztery hipotetyczne realizacje doświadczenia według modelu. Takich realizacji jest nieskończenie wiele. Wśród nich znajduje się rzeczywiste doświadczenie.



# Estymacja

- $\hat{\beta}_0,\ \hat{\beta}_1$ są oszacowaniami punktowymi parametrów $\beta_0$ oraz $\beta_1$.

# Estymacja

- Oszacowania przedziałowe dla $\beta_0$ oraz $\beta_1$ są postaci

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n-2)S_{\beta_1}, \hat{\beta}_1 + t(\alpha; n-2)S_{\beta_1})$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n-2)S_{\beta_0}, \hat{\beta}_0 + t(\alpha; n-2)S_{\beta_0})$$

gdzie

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}x}, \quad S^2_{\beta_0} = \frac{S^2}{\mathrm{var}x}\left(\frac{\mathrm{var}x}{n} + \bar{x}^2\right)$$

$$S^2 = \frac{\mathrm{var}y - \hat{\beta}_1 \mathrm{cov}(x,y)}{n-2} = \frac{\mathrm{var}y(1-r^2)}{n-2}$$

# Weryfikacja hipotez

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Statystyka testowa

$$F_{\text{emp}} = \frac{\hat{\beta}_1^2}{S_{\beta_1}^2} = \frac{\hat{\beta}_1 \text{cov}(x, y)}{S^2}$$

Hipotezę odrzucamy, jeżeli $F_{\text{emp}} > F(\alpha; 1, n-2)$.

$F(\alpha; 1, n-2)$ jest wartością krytyczną rozkładu $F$.

# Weryfikacja hipotez

$$H_0 : \beta_1 = a$$
$$H_1 : \beta_1 \neq a$$

Statystyka testowa

$$t_{\text{emp}} = \frac{\hat{\beta}_1 - a}{S_{\beta_1}}$$

Hipotezę odrzucamy, jeżeli $|t_{\text{emp}}| > t(\alpha; n-2)$.

$t(\alpha; n-2)$ jest wartością krytyczną rozkładu $t-$Studenta.

## Przykład, ciąg dalszy

Obliczenia:

$$r^2 = 0.9826,\ F_{\text{emp}} = 452.5,\ t_{\text{emp}} = 21.271\ (\text{przy } a = 0)$$

Przedział ufności dla $\beta_1$ (na poziomie ufności 0.95):

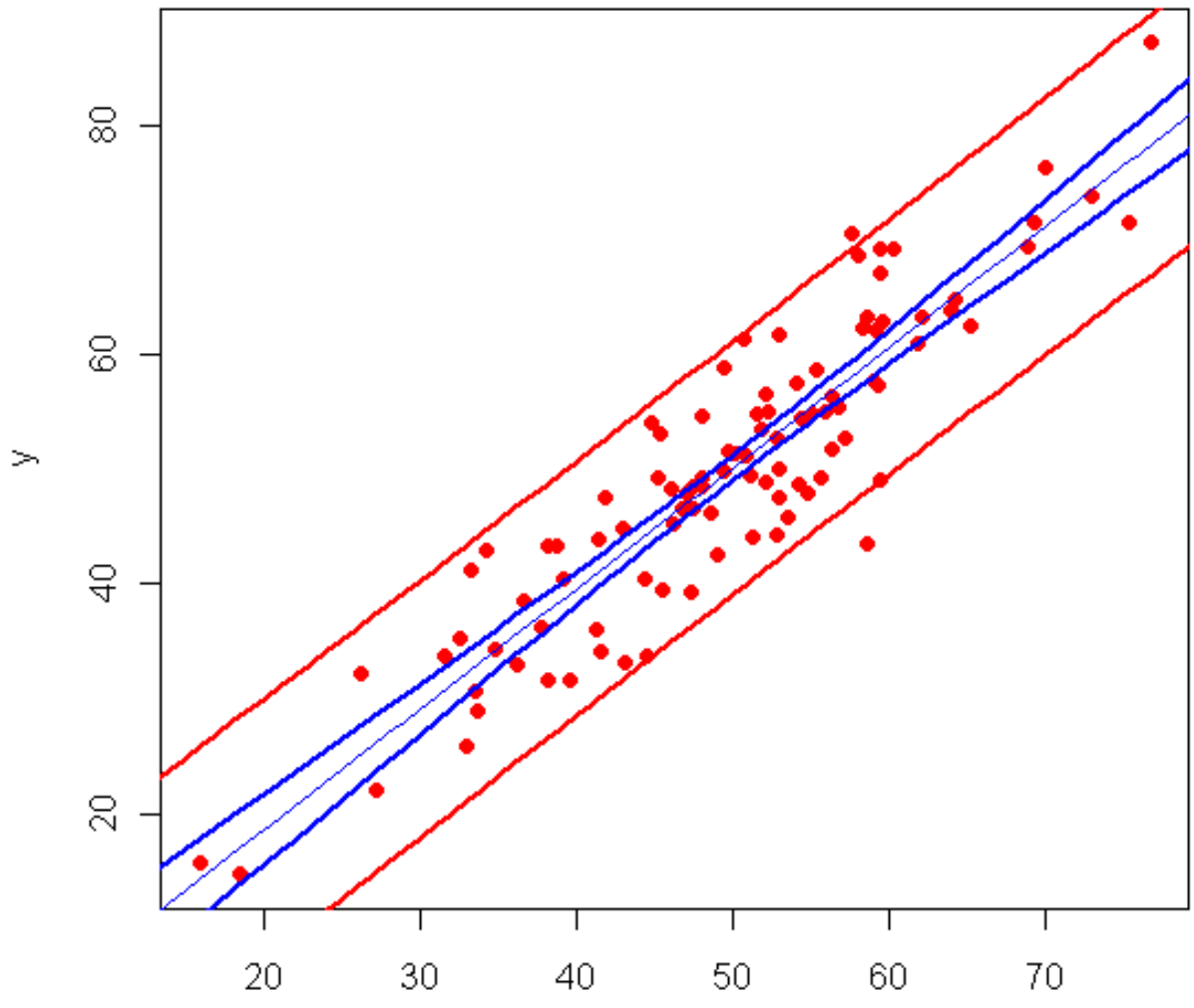
$$(0.0622,\ 0.0774)$$

Przedział ufności dla $\beta_0$ (na poziomie ufności 0.95):

$$(-5.1272, 4.7571)$$

.

# Obszar ufności dla prostej regresji. Obszar predykcji



# Obszar ufności dla prostej regresji

**Obszar ufności dla prostej regresji** umożliwia nam wnioskowanie o wartościach średnich zmiennej $Y$ jednocześnie dla wielu wybranych wartości zmiennej $X$.

$$f(x) \in (\hat{f}(x) - t(\alpha; n-2)S_Y; \hat{f}(x) + t(\alpha; n-2)S_Y)$$

gdzie

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S_Y^2 = S^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\mathrm{var}x} \right)$$

# Obszar predykcji

**Obszar predykcji** umożliwia nam wnioskowanie o wartościach zmiennej $Y$ jednocześnie dla wielu wybranych wartości zmiennej $X$.

$$Y(x) \in (\hat{f}(x) - t(\alpha; n-2)S_{Y(x)}; \hat{f}(x) + t(\alpha; n-2)S_{Y(x)})$$

gdzie $Y(x)$ oznacza wartość zmiennej $Y$ dla wybranej wartości $x$ zmiennej $X$ oraz

$$S^2_{Y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\text{var}x}\right)$$

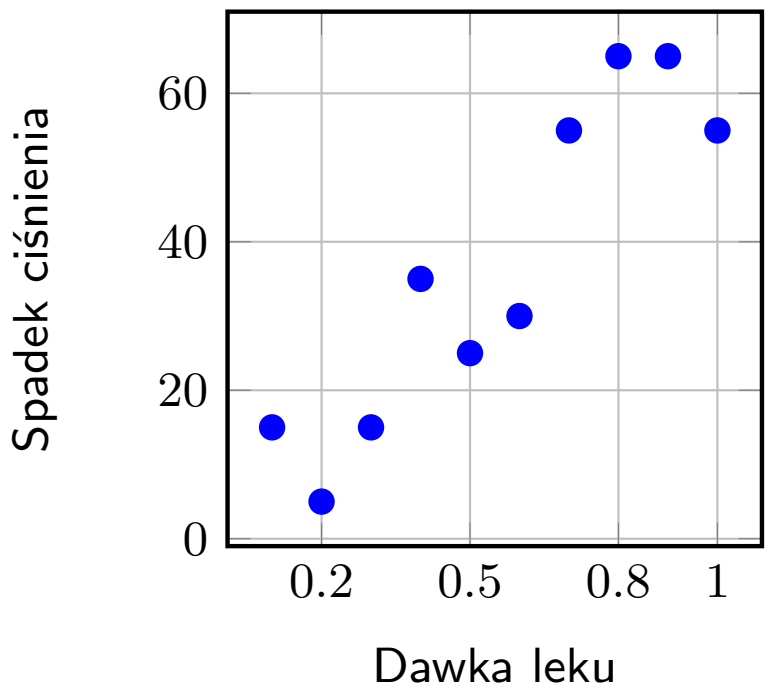
## Przykład, ciąg dalszy

Obliczenia (na poziomie ufności 0.95):

Przedział ufności dla $f(635)$: (43.17199, 45.1298)
Przedział ufności dla $Y(635)$: (40.90645, 47.39535)

## Zadanie

W pewnym doświadczeniu farmakologicznym bada się wpływ leku hipotensyjnego na ciśnienie tętnicze krwi zwierząt doświadczalnych. Podano 10 różnej wielkości dawek (w mg/kg wagi ciała) tego leku i otrzymano następujące spadki ciśnienia tętniczego krwi (w mm Hg):




| dawka | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| spadek | 15 | 5 | 15 | 35 | 25 | 30 | 55 | 65 | 65 | 55 |

1. Czy zależność spadku ciśnienia od dawki leku jest statystycznie istotna?
2. Jakiego spadku ciśnienia należy oczekiwać przy dawce 0.35?
3. Jaki jest przeciętny spadek ciśnienia przy dawce 0.45?
4. O ile zmienia się średnio spadek ciśnienia przy zwiększaniu dawki o 0.1?

## Zadanie

Badano, czy aktywność enzymu zależy od czasu leczenia (w dniach). Zebrano wyniki dla 12 osób:

| Czas leczenia | 1 | 2 | 3 | 4 | 5 | 7 | 10 | 14 | 18 | 20 | 24 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aktywność enzymu | 42 | 40 | 35 | 44 | 36 | 35 | 30 | 33 | 22 | 20 | 16 | 18 |

1. Czy można uznać, że czas leczenia obniża aktywność enzymu?
2. Jakiej aktywności enzymu należy oczekiwać u osoby, która leczyła się 8 dni?

## Zadanie

Badano zależność między wzrostem a obwodem klatki piersiowej w populacji osób chorych na choroby reumatycne kręgosłupa. Otrzymano wyniki:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| wzrost | 153 | 158 | 160 | 163 | 166 | 170 | 175 | 178 |
| obwód | 74 | 76 | 77 | 78 | 80 | 83 | 85 | 88 |

Przeprowadzić analizę regresji.

## Zadanie

Wskaźnik szczepień i zapadalności na błonnicę w 12 wylosowanych miastach przedstawia tabela:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wskaźnik szczepień | 3.32 | 3.42 | 3.45 | 3.85 | 4.25 | 5.20 | 5.84 | 6.25 | 6.50 | 6.88 | 7.60 | 7.60 |
| Wskaźnik zapadalności | 3.20 | 3.00 | 3.50 | 2.25 | 2.50 | 1.80 | 1.00 | 0.79 | 0.52 | 1.00 | 0.30 | 0.06 |

1. Stosując model regresji liniowej, zbadać istotność regresji.
2. Czy można uznać, że im większy wskaźnik szczepień, tym mniejsza zapadalność?
3. O ile powinna zmniejszyć się zapadalność na błonnicę, gdy wskaźnik szczepień zostanie zwiększony o jednostkę?
4. Ile średnio wynosi zapadalność w miastach, w których wskaźnik szczepień jest na poziomie dwóch jednostek?
5. Jaka może być zapadalność w mieście, w którym wskaźnik szczepień wynosi 2.3?

# MODELE LOGITOWE

## Wprowadzenie

Rozważmy dwuwartościowy wynik obserwacji cechy $D$ (Diseased)

| | |
|---|---|
| $D$ | 0 – zdrowy |
| | 1 – chory (choroba wieńcowa) |

Jesteśmy zainteresowani czynnikiem ryzyka $E$ (Exposure)

| | |
|---|---|
| $E$ | 0 – nie pali papierosów |
| | 1 – pali papierosy |

- Jak określić związek choroby wieńcowej z paleniem papierosów?
- Jak związek ten wyrazić w sposób ilościowy?
- Jakie zmienne kontrolne musimy wziąć pod uwagę?
  1. $C_1$ – wiek
  2. $C_2$ – rasa
  3. $C_3$ – płeć

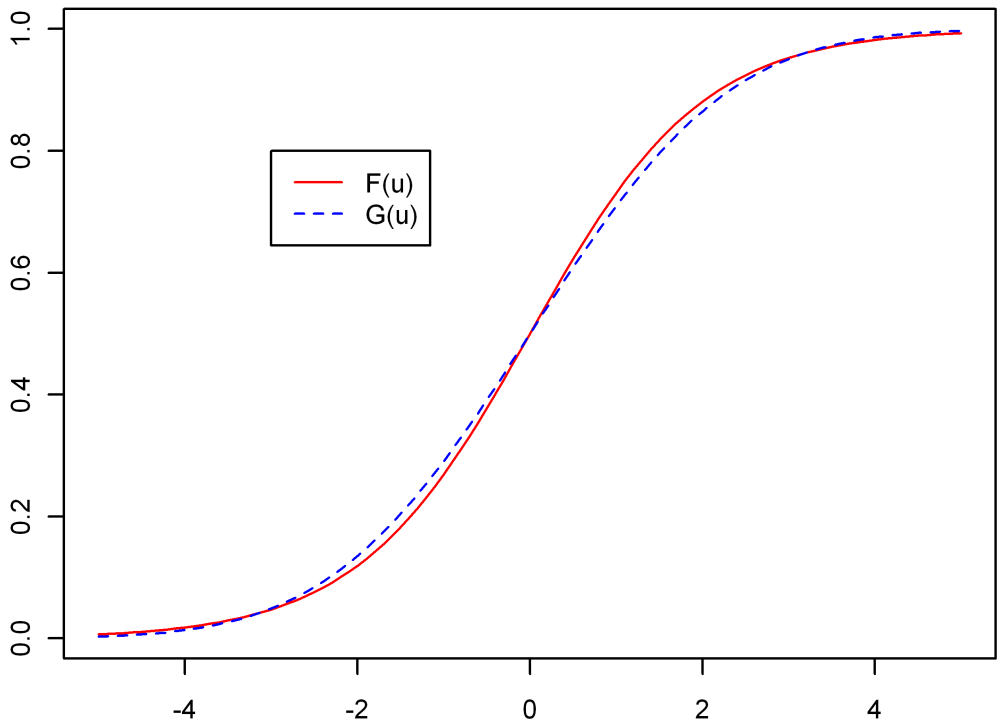
Zmienne $E$, $C_1$, $C_2$, $C_3$ są to tzw. zmienne niezależne, za pomocą których chcemy objaśnić zmienną zależną $D$.

Ogólnie, chcemy objaśnić zależność $D$ od $X_1,\ X_2,\ X_3,\ \ldots$

| Efekty główne | Efekty współdziałań | Efekty nieliniowe |
|---|---|---|
| $X_1 = E$ | $X_4 = E \times C_1$ | $X_6 = C_1^2$ |
| $X_2 = C_1$ | $X_5 = C_1 \times C_2$ | |
| $X_3 = C_2$ | | |

Do opisu tej zależności stosujemy regresję logitową lub probitową

Funkcja logistyczna $F(u) = \frac{1}{1+\exp(-u)} \in (0,1)$
Funkcja probitowa $G(u) = \Phi(\frac{u}{\pi/\sqrt{3}})$, gdzie $\Phi$–dystrybuanta $N(0,1)$.




Ze względu na to, ze funkcje wiążące $F$, $G$ są rosnące, duże ryzyko zachorowania jest osiągane dla relatywnie dużych wartości $u$. Przyjmujemy $u = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$.

Nieznane parametry $\alpha, \beta_1, \ldots, \beta_k$ należy oszacować na podstawie danych (eksperymentu). Postać modelu

$$P(D = 1|X_1, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum_i \beta_i x_i)}}$$

Oznaczenia

$$\mathbb{X} = (X_1, X_2, \ldots, X_n)$$

$$P(\mathbb{X}) = P(D = 1|X_1, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum_i \beta_i x_i)}}$$

## Przykład

- D – choroba wieńcowa, $\{0$ - zdrowy, $1$ - chory$\}$
- $X_1$ – poziom katecholamin, $\{0$ - niski, $1$ - wysoki$\}$
- $X_2$ – wiek; cecha ciągła
- $X_3$ – EKG, $\{0$ - normalne, $1$ - zaburzone$\}$

Próba liczy $n = 609$ mężczyzn. Eksperyment prospektywny trwający 9 lat.

$$P(X_1, X_2, X_3) = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)}}$$

Oszacowane na podstawie próby wartości parametrów:

- $\hat{\alpha} = -3.911$
- $\hat{\beta}_1 = 0.652$: im wyższy poziom katecholamin, tym większe ryzyko choroby
- $\hat{\beta}_2 = 0.029$: im osoba starsza, tym większe ryzyko choroby
- $\hat{\beta}_3 = 0.342$: zburzone EKG oznacza większe ryzyko choroby

Ryzyko choroby (prawdopodobieństwo zachorowania). Chcemy wyznaczyć ryzyko choroby wieńcowej u wybranej osoby, która charakteryzuje się wysokim poziomem katecholamin, ma 40 lat oraz ma EKG w normie:

$$\hat{p}(1, 40, 0) = \frac{1}{1 + e^{-(-3.911 + 0.652 \cdot 1 + 0.029 \cdot 40 + 0.342 \cdot 0)}} = 0.109$$

Ryzyko zachorowania wynosi ok. 11%.

Podobnie możemy wyznaczyć ryzyko choroby u osoby 40-letniej, normalnym EKG, ale o niskim poziomie katecholamin:

$$\hat{p}(0, 40, 0) = 0.06$$

Zauważmy, że dzieląc te wartości przez siebie otrzymamy ryzyko względne:

$$RR = \frac{\hat{p}(1, 40, 0)}{\hat{p}(0, 40, 0)} = 1,82$$

Interpretacja: Czterdziestolatkowie, którzy mają EKG w normie, mają o 82% większe ryzyko choroby wieńcowej, jeżeli ich poziom katecholamin jest podwyższony.

## Do domu

a) Opisać i porównać grupy (I vs II)

| I | | | II | | |
|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ |
| 1 | 30 | 0 | 0 | 30 | 0 |
| 1 | 35 | 1 | 1 | 35 | 0 |
| 0 | 20 | 0 | 0 | 40 | 0 |

b) Narysować wykresy funkcji

1. na jednym wykresie

$$f(t) = \hat{p}(1, t, 1) \text{ dla } t \in (20, 50)$$
$$g(t) = \hat{p}(1, t, 0) \text{ dla } t \in (20, 50)$$

2. na osobnym wykresie

$$RR(t) = \frac{f(t)}{g(t)} \text{ dla } t \in (20, 50)$$

Zinterpretować otrzymane wykresy.

# Iloraz szans

## model logitowy

$$\text{logit}(P(\mathbb{X})) = \ln \frac{P(\mathbb{X})}{1 - P(\mathbb{X})} = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

Niech $k = 3$. Porównujemy grupę o zaburzonym EKG ($X_3 = 1$) z grupą o prawidłowym EKG ($X_3 = 0$)

$$\ln \Big[ OR(X3 = 1 | X_3 = 0) \Big] = \ln \left[ \frac{P(X_1, X_2, 1)}{1 - P(X_1, X_2, 1)} \Big/ \frac{P(X_1, X_2, 0)}{1 - P(X_1, X_2, 0)} \right] = \beta_3$$

Porównujemy grupę wiekową $X1 = Wiek + k$ z grupą wiekową $X_1 = Wiek$

$$\ln \Big[ OR(Wiek + k | Wiek) \Big] = \ln \left[ \frac{\frac{P(Wiek + k, X_2, X_3)}{1 - P(Wiek + k, X_2, X_3)}}{\frac{P(Wiek, X_2, X_3)}{1 - P(Wiek, X_2, X_3)}} \right] = k \cdot \beta_1$$

Ogólnie

$$\mathbb{X}_1 = (X_{11}, X_{12}, \ldots, X_{1k}), \quad \mathbb{X}_0 = (X_{01}, X_{02}, \ldots, X_{0k})$$

$$\ln[OR(\mathbb{X}_1|\mathbb{X}_0)] = \sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})$$

Współdziałanie

$X_1 = A,\ A \in \{0,1\};\ X_2 = B,\ B \in \{0,1\};\ X_3 = A \times B$

$$A = 1, B = 1 \rightarrow \mathbb{X} = (A, B, A \times B) = (1, 1, 1)$$
$$A = 0, B = 1 \rightarrow \mathbb{X} = (A, B, A \times B) = (0, 1, 0)$$

$\ln[OR(A = 1, B = 1|A = 0, B = 1)] = \beta_1(1-0)+\beta_2(1-1)+\beta_3(1-0) = \beta_1+\beta_2$

Analogicznie

$\ln[OR(A = 1, B = 0|A = 0, B = 0)] = \beta_1(1-0) + \beta_2(0-0) + \beta_3(0-0) = \beta_1$

Przy współdziałaniu $OR$ dla $A = 1$ względem $A = 0$ zależy od poziomu $B$.

## Przykład dla $OR(E = 1|E = 0)$

1. $E$–poziom katecholamin
2. $C_1$–wiek
3. $C_2$–cholesterol, 0-niski, 1-wysoki
4. $C_3$–palenie papierosów, 0-nie pali, 1-pali
5. $C_4$– EKG, 0-w normie, 1-zaburzone
6. $C_5$– ciśnienie krwi, 0-w normie, 1-zaburzone
7. $E \times C_2$
8. $E \times C_5$

$$\ln[OR(E = 1|E = 0)] = \beta_1 + \beta_7 C_2 + \beta_8 C_5$$

# Krzywa ROC

Krzywa ROC (Receiver operating characteristic curve). Każdego pacjenta klasyfikujemy do kategorii pierwszej, jeżeli oszacowane prawdopodobieństwo przynależności do tej kategorii przekracza umowną granicę $\gamma \in (0,1)$, na przykład $\gamma = 1/2$. W przeciwnym wypadku klasyfikujemy pacjenta do drugiej kategorii.

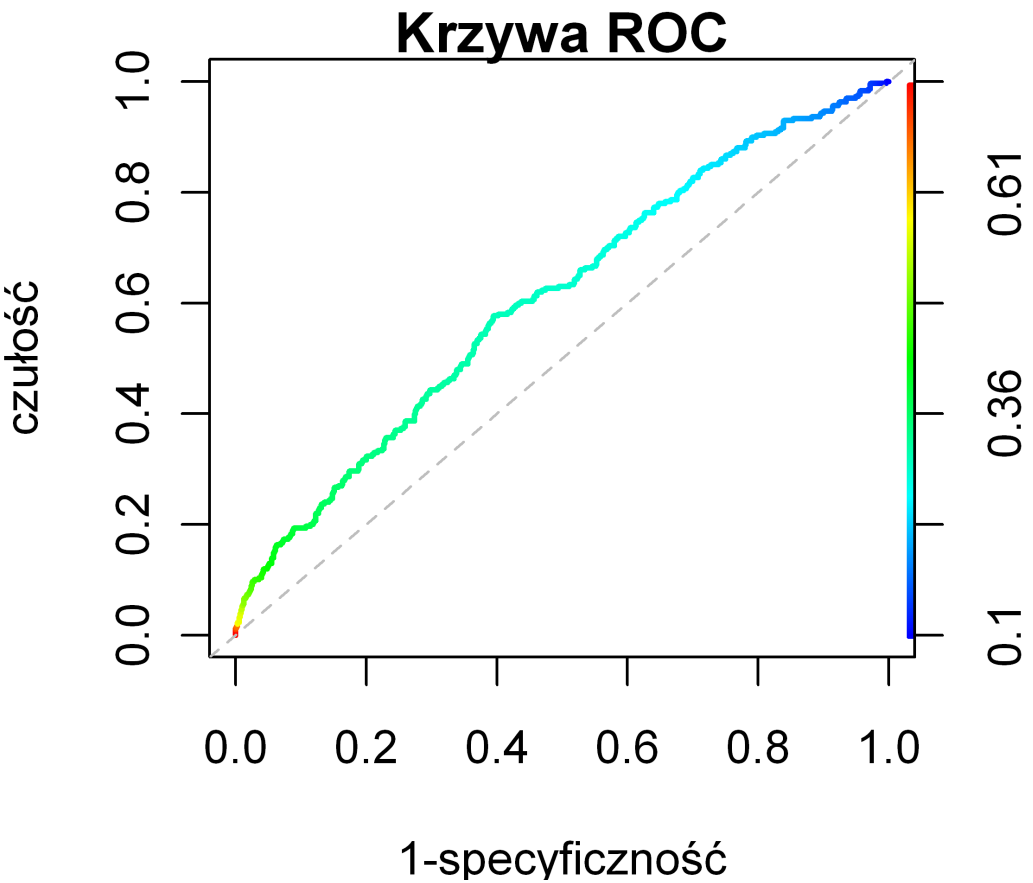
| | Klasyfikowany do | |
|---|---|---|
| Należy do | pierwszej | drugiej |
| pierwszej | TN | FP |
| drugiej | FN | TP |

$\text{Czułość} = \frac{\text{TP}}{\text{TP+FN}}$
$\text{Specyficzność} = \frac{\text{TN}}{\text{TN+FP}} = 1 - \frac{\text{FP}}{\text{TN+FP}}$
Ponieważ tabelkę można konstruować dla różnych wartości $\gamma$, zdefinowane pojęcia są funkcją $\gamma$: Czułość=Czułość($\gamma$), Specyficzność=Specyficzność($\gamma$). Krzywa ROC to wykres zbioru

$$\{(1 - \text{Specyficzność}(\gamma), \text{Czułość}(\gamma)) : \ \gamma \in (0,1)\}$$

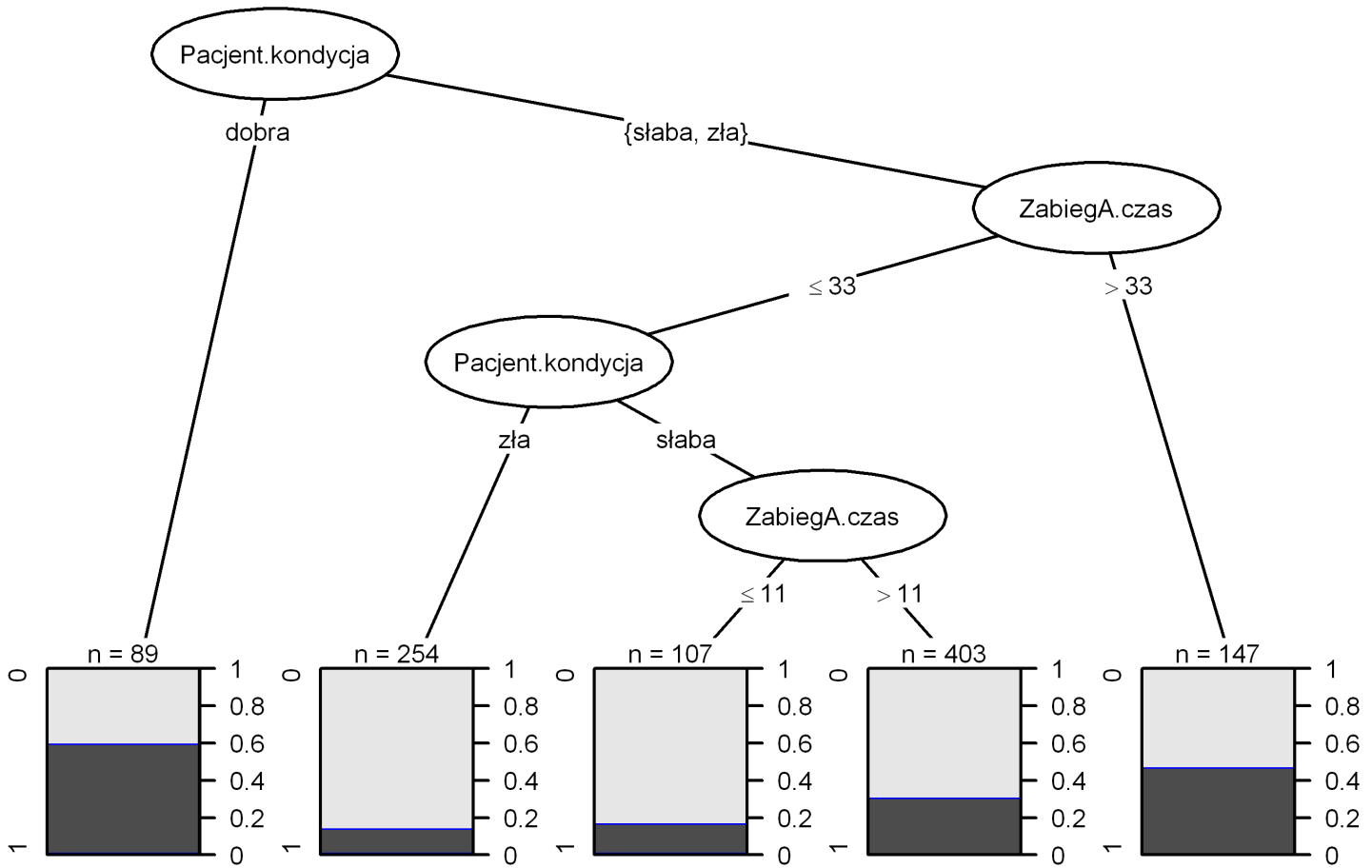
## Krzywa ROC do modelu

```
Pacjent.kategoria ~ Pacjent.wiek + Pacjent.wiek:Cecha.pomiar
```

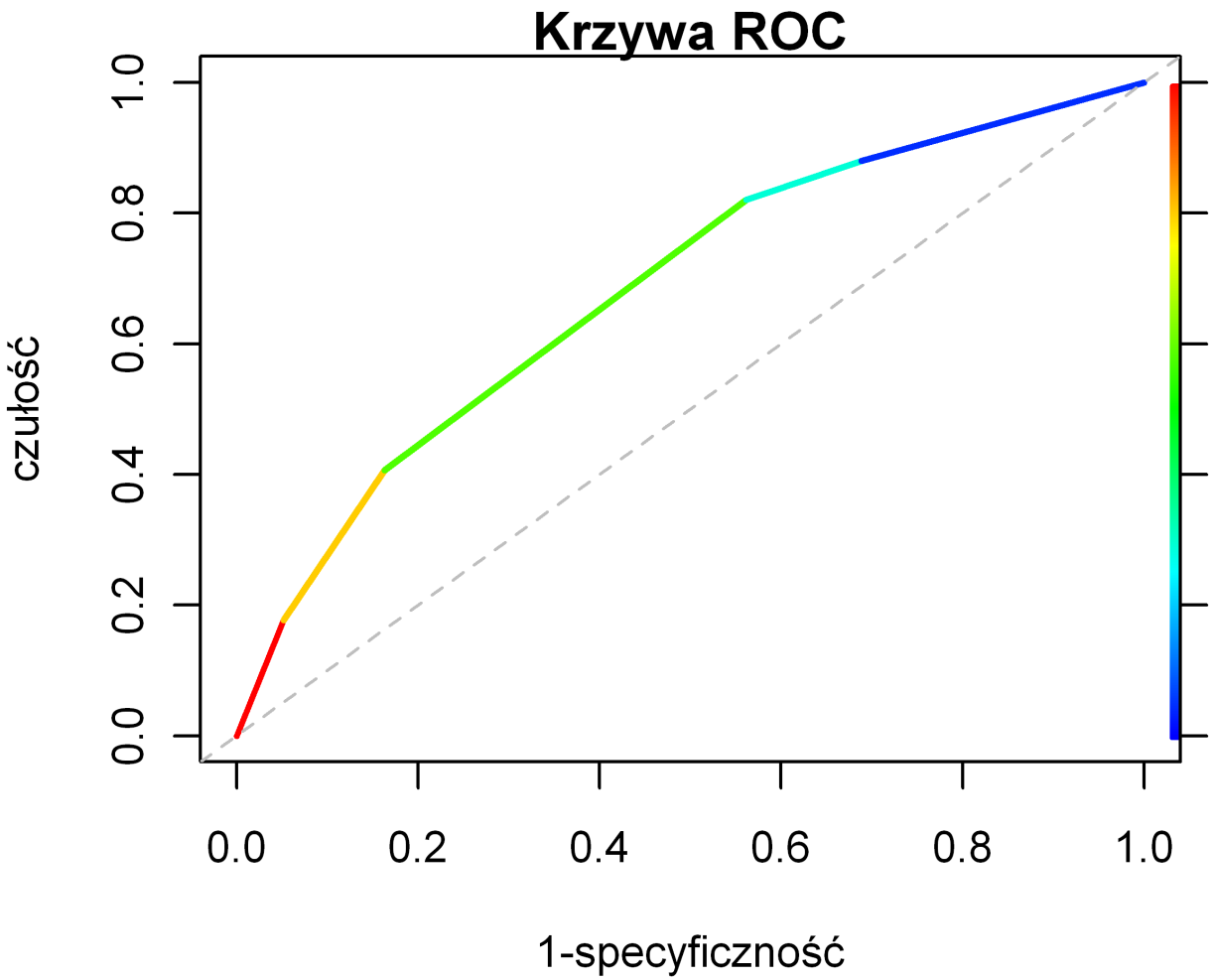


# Drzewa decyzyjne

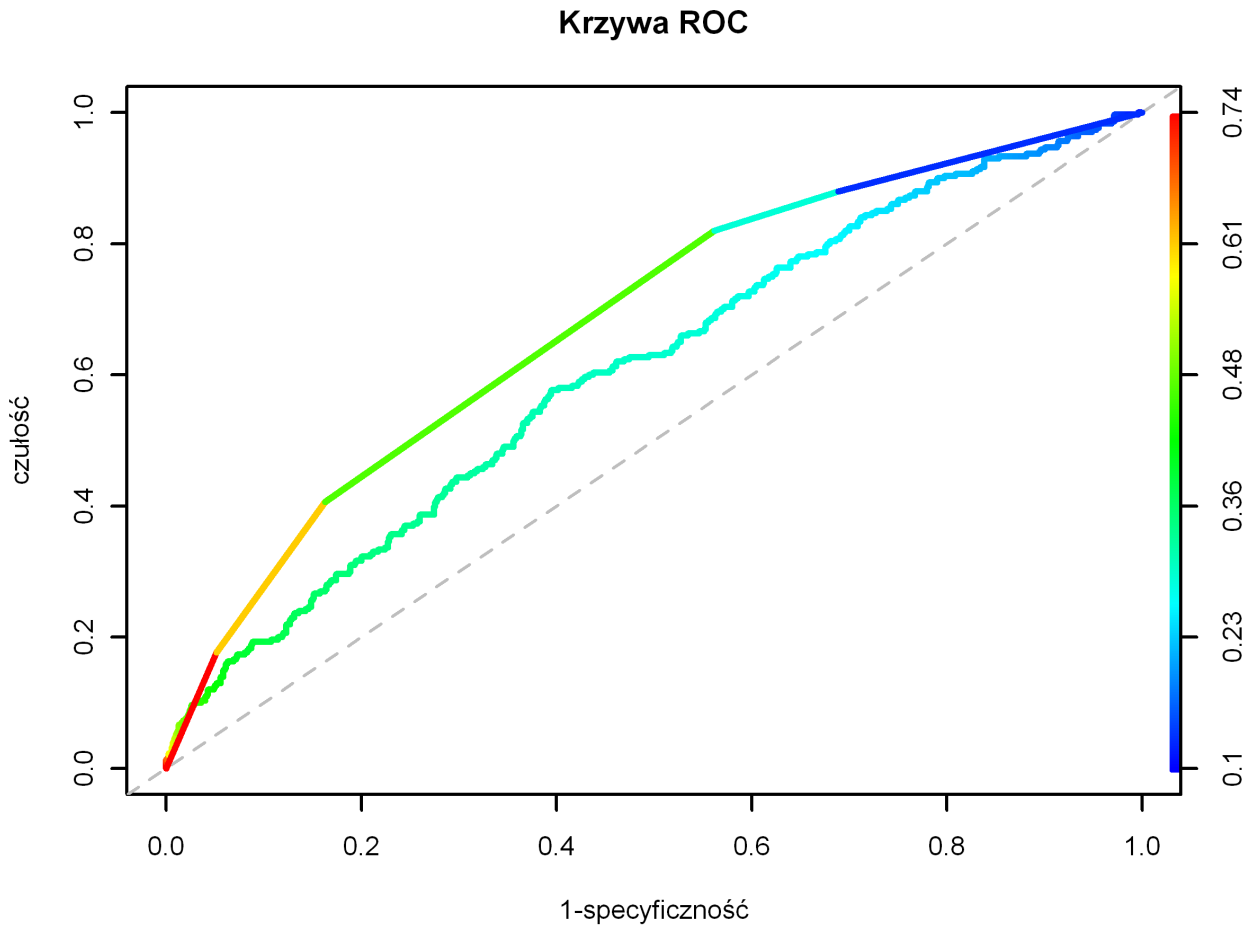
```
Kategoria~ZabiegA.czas+Cecha.pomiar+ZabiegB+Pacjent.wiek+
          +Pacjent.kondycja+ZabiegC+Lek+ZabiegD
```



## Krzywa ROC na podstawie drzewa



## Porównanie krzywych



## Zadanie (a)

Sprawdzano, czy śmiertelność z powodu choroby serca zależy od statusu socjalnego $SOC$ (0-niski status, 1-wysoki status). W tym celu przeprowadzono dwunastoletni eksperyment prospektywny. Obserwowano dwustu mężczyzn w wieku co najmniej 60 lat. Jako zmienne kontrolne przyjęto $SMK$ – status palenia (0-nie pali, 1-pali) oraz $SBP$ – ciśnienie skurczowe serca. Dopasowano dwa modele logitowe:

## Zadanie (a)

Model 1

| Zmienna | Współczynnik |
|---|---|
| stała | -1.180 |
| $SOC$ | -0.520 |
| $SBP$ | 0.040 |
| $SMK$ | -0.560 |
| $SOC \times SBP$ | -0.033 |
| $SOC \times SMK$ | 0.175 |

Model 2

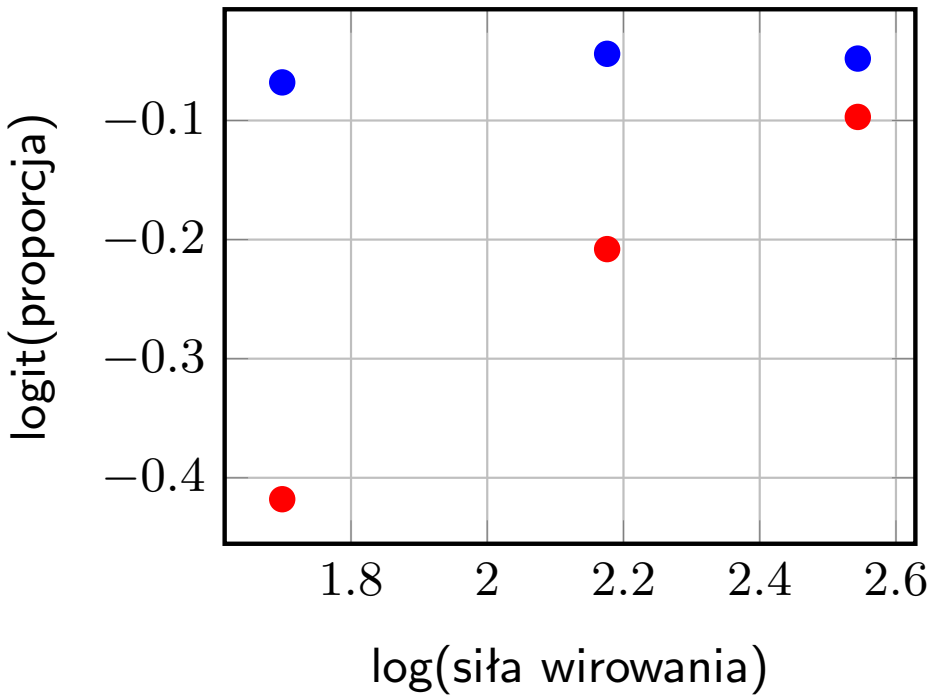
| Zmienna | Współczynnik |
|---|---|
| stała | -1.19 |
| $SOC$ | -0.50 |
| $SBP$ | 0.01 |
| $SMK$ | -0.42 |

1. Za pomocą modelu 1. oszacuj ryzyko śmierci osoby o wysokim statusie społecznym, która pali paierosy oraz ma wysokie ciśnienie.
2. Za pomocą modelu 2. wyznacz ryzyko śmierci dla dwóch osób: Osoba 1 ($SOC = 1$, $SMK = 1$, $SBP = 155$), Osoba 2 ($SOC = 0$, $SMK = 1$, $SBP = 155$).
3. W modelu 2. wyznacz iloraz szans dla grupy o wysokim statusie społecznym względem grupy o niskim statusie społecznym. Uwzględnij warianty dla różnych wartości zmiennych kontrolnych.

## Zadanie (b)

Obserwujemy liczbę zarodkowych pylników roślin z gatunku Datura innoxia, uzyskanych przy różnych warunkach, określonych przez dwa czynniki. Jeden z czynników jest jakościowy na dwóch poziomach. Określa sposób przechowywania pylników (kontrolny/specyficzny). Drugi czynnik jest ilościowy i określa siłę wirowania (40, 150, 350).

## Zadanie (b)




$$\textbf{proporcja} = \frac{\textbf{liczba pylników zarodkowych}}{\textbf{liczba wszystkich pylników}}$$

## Zadanie (b)

| Warunki przechowywania | | Siła wirowania 40 | 150 | 350 |
|---|---|---|---|---|
| kontrolne | $y_{1k}$ | 55 | 52 | 57 |
| | $n_{1k}$ | 102 | 99 | 108 |
| specyficzne | $y_{2k}$ | 55 | 50 | 50 |
| | $n_{2k}$ | 76 | 81 | 90 |

## Zadanie (b)

Sposób wprowadzenia danych w programie STATISTICA:

| Przechowywanie | Wirowanie | Liczba | Zarodkowe |
|---|---|---|---|
| kontrolne | 40 | 55 | 1 |
| kontrolne | 150 | 52 | 1 |
| kontrolne | 350 | 57 | 1 |
| kontrolne | 40 | 47 | 0 |
| kontrolne | 150 | 47 | 0 |
| kontrolne | 350 | 51 | 0 |
| specyficzne | 40 | 55 | 1 |
| specyficzne | 150 | 50 | 1 |
| specyficzne | 350 | 50 | 1 |
| specyficzne | 40 | 21 | 0 |
| specyficzne | 150 | 31 | 0 |
| specyficzne | 350 | 40 | 0 |

Zauważmy, jak są wprowadzone wyniki dla grupy pylników otrzymanych przy sile wirowania 40 oraz przechowywanych w warunkach określonych, jako kontrolne. Wszystkich pylników jest 102, zarodnikowych jest 55, a pozostałych niezarodnikowych jest 102-55=47. Stąd mamy wiersz (dla Zarodnikowe=1): (kontrolne, 40, 55, 1) oraz wiersz (dla Zarodnkowe=0): (kontrolne, 40, 47, 0).

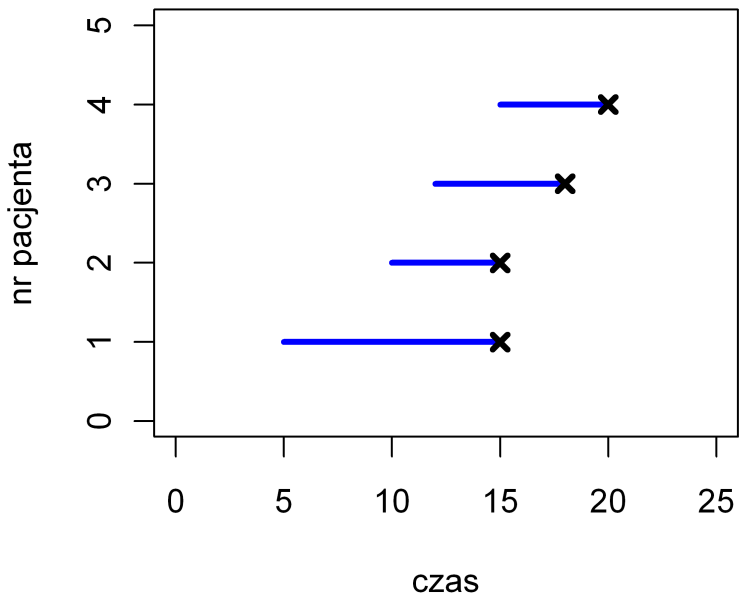
# Analiza przeżycia

W analizie przeżycia zmienna zależna mierzy czas do momentu interesującego nas zdarzenia.

- Kiedy pacjent wyzdrowieje?
- Kiedy pacjent umrze?
- Kiedy małżeńswo się skończy?
- Kiedy recydywista wróci do więzienia?
- Jak długo liderzy polityczni będą cieszyć się wysokim poparciem?

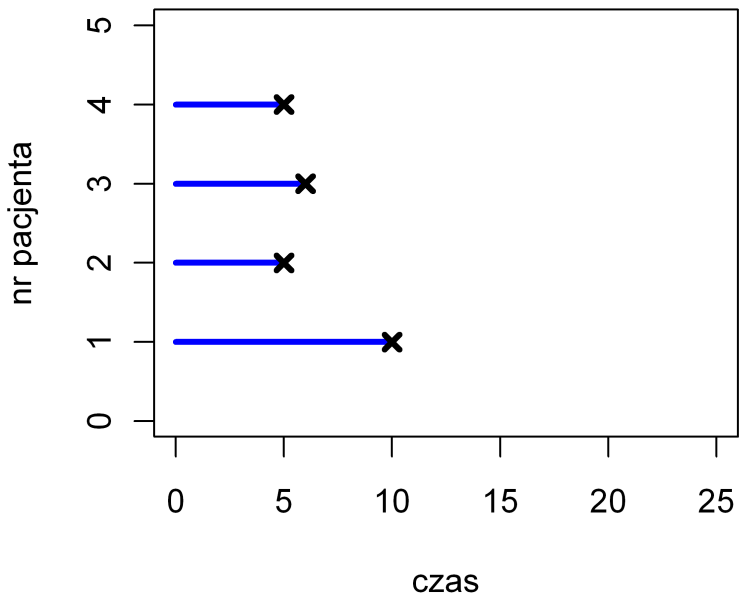
# Wyrównywanie danych

## Przed wyrównaniem



# Wyrównywanie danych

## Po wyrównaniu



# Cenzurowanie/ Kontrolowanie/ Ucinanie

- Pierwszego typu: Obserwujemy $n$ obiektów do momentu $t_0$. W tym czasie „zawiodło" $n_u$ obiektów. Pozostałe $n_c = n - n_u$ nadal „pracują". Nazywamy je obiektami cenzurowanymi.
- Drugiego typu: Obserwujemy $n$ obiektów do momentu aż $k < n$ spośród nich „zawiedzie".
- Cenzurowanie losowe (prawostronne): całkowity „czas życia" nie został zaobserwowany z powodów, które są poza kontrolą prowadzącego eksperyment
  - pacjent wycjechał i nie ma z nim kontaktu
  - trzeba przerwać terapię u pacjenta z powodu niebezpiecznych efektów ubocznych
  - nie zaszło interesujące nas zdarzenie przed końcem eksperymentu

# Funkcje charakteryzujące czas życia $T$

- funkcja przeżycia (niezawodności)

$$S(t) = 1 - F(t) = P(T > t)$$

- gęstość

$$f(t) = \frac{d}{dt} F(t)$$

- funkcja hazardu

$$h(t) = \lim_{0<h\to 0} \frac{1}{h} P[t < T \leq t + h | T > t] = \frac{f(t)}{S(t)}$$

- Skumulowana funkcja hazardu

$$H(t) = \int_0^t h(t)\, dt$$

Stąd wynika, że $S(t) = \exp[-H(t)]$

# Funkcja hazardu

**Własności.**

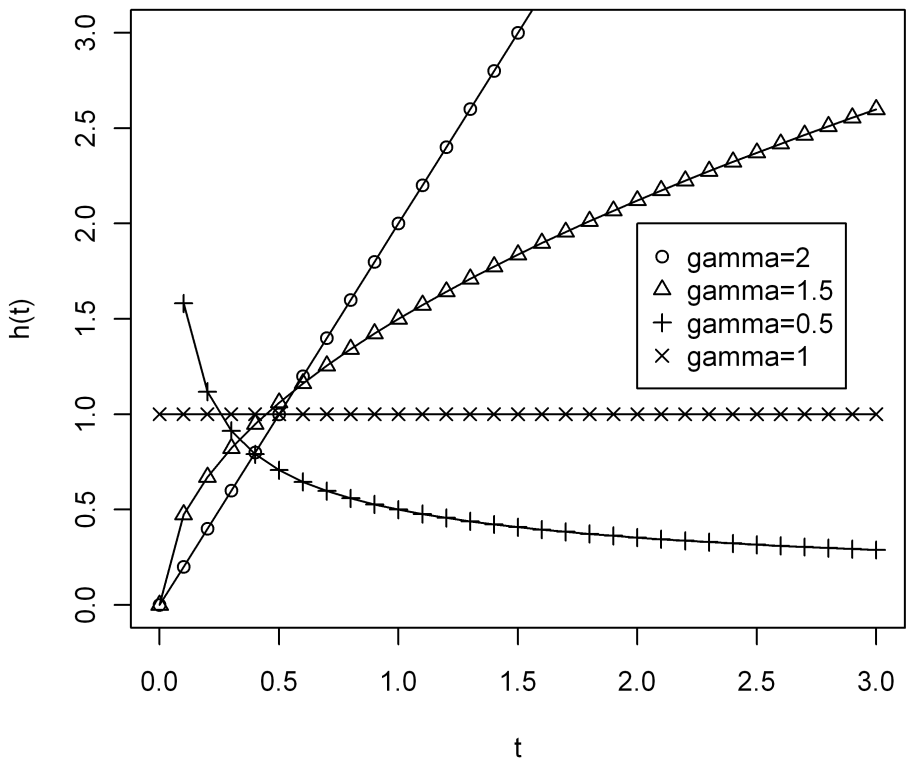
- Funkcja hazardu jest rosnąca $\iff$ Obiekt w starszym wieku ma mniejsze szanse na dalsze życie niż obiekt w młodszym wieku
- Funkcja hazardu jest malejąca $\iff$ Obiekt w starszym wieku ma większe szanse na dalsze życie niż obiekt w młodszym wieku.
- Funkcja hazardu jest stała $\iff$ Obiekt w starszym wieku ma identyczne szanse na dalsze życie jak obiekt w młodszym wieku.

# Rozkłady prawdopodobieństwa charakteryzujące czas życia $T$

## Rozkład Weibulla

$$F(t) = 1 - e^{-\lambda t^{\gamma}}, \quad t > 0,\ \lambda > 0,\ \gamma > 0$$

Funkcja hazardu

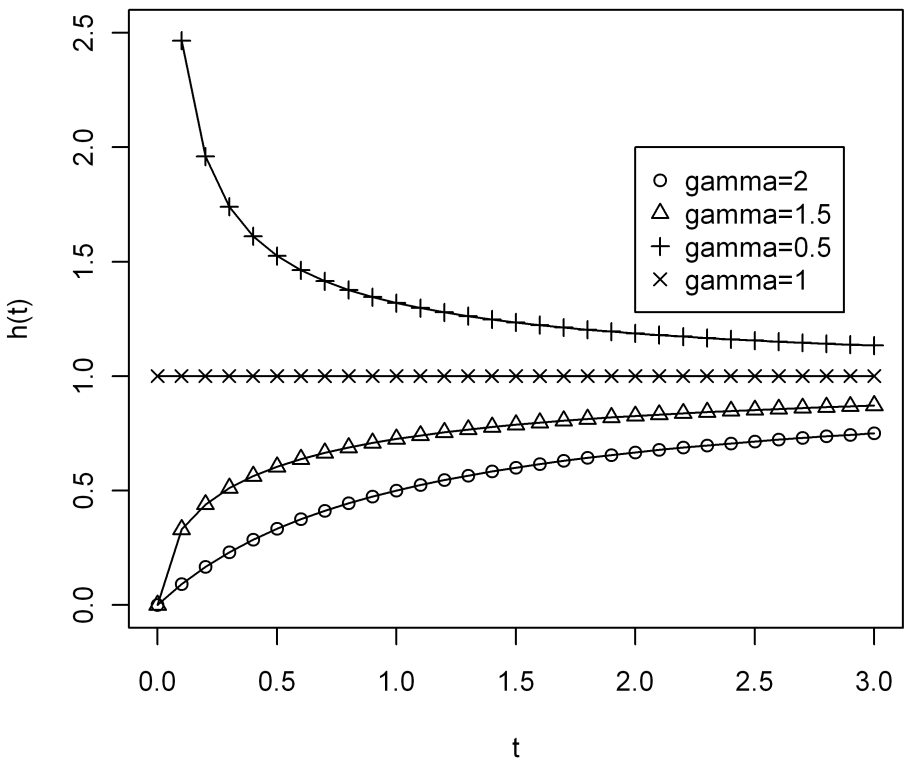


# Rozkłady prawdopodobieństwa charakteryzujące czas życia $T$

## Rozkład Gamma

$$f(t) = \frac{\lambda^{\gamma} t^{\gamma-1} e^{-\lambda t}}{\Gamma(\gamma)}, \quad t > 0,\ \lambda > 0,\ \gamma > 0$$

Funkcja hazardu



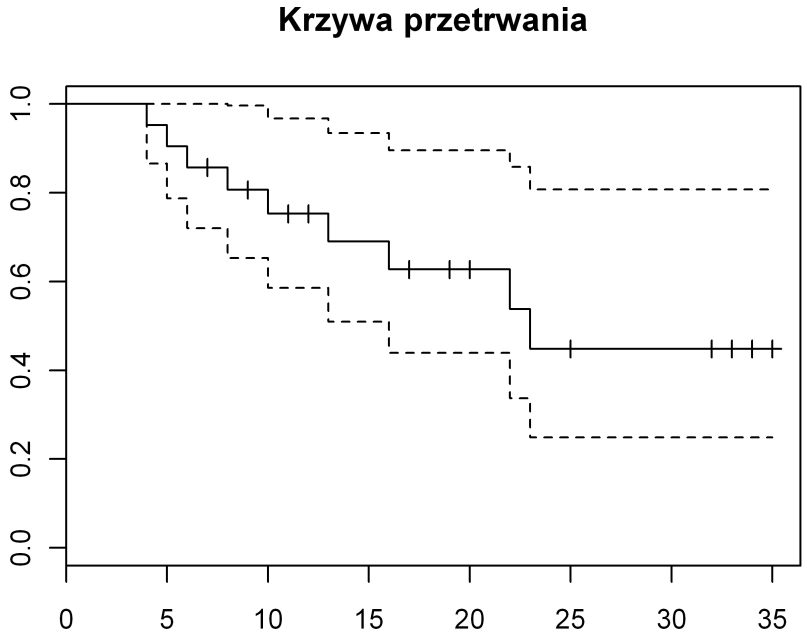
## Wpływ leku na remisję raka

Przeprowadzono eksperyment w celu określenia efektu leku na czas remisji raka. W eksperymencie wzięło udział dwudziestu jeden pacjentów. U sześciu pacjentów zaobserwowano czas remisji. Pozostałych dwunastu dostarczyło dane cenzurowane losowo (prawostronnie).

```
-------------------------------------------------------------
R: Wprowadzenie danych
> czas=c(4,5,6,7,8,9,10,11,12,13,16,17,19,20,22,23,25,32,33,34,35)
> status=c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0)
> Surv(czas,status)
4   5   6   7+  8   9+ 10  11+ 12+ 13  16
         17+ 19+ 20+ 22  23  25+ 32+ 33+ 34+ 35+
-------------------------------------------------------------
```

## Oszacowanie nieparametryczne

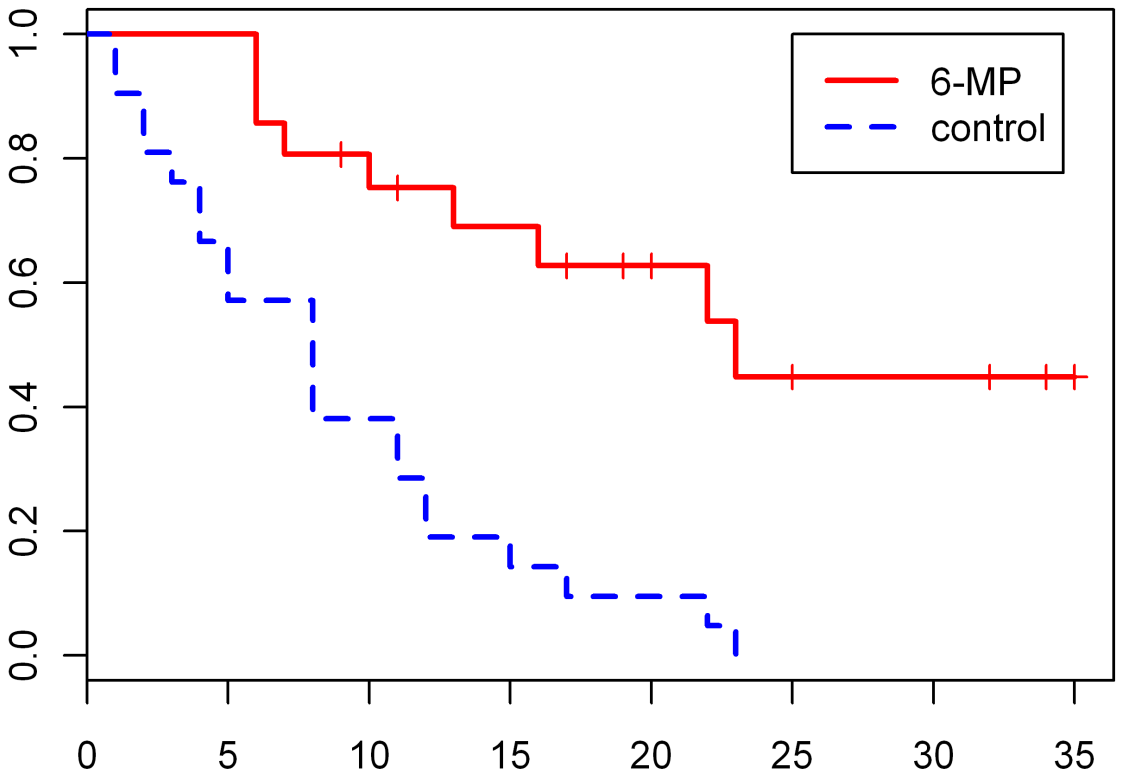
```
-----------------------------------------------------------
remisja=survfit(Surv(czas,status)~1)
plot(remisja,conf.int=T,main="Krzywa przetrwania")
-----------------------------------------------------------
```

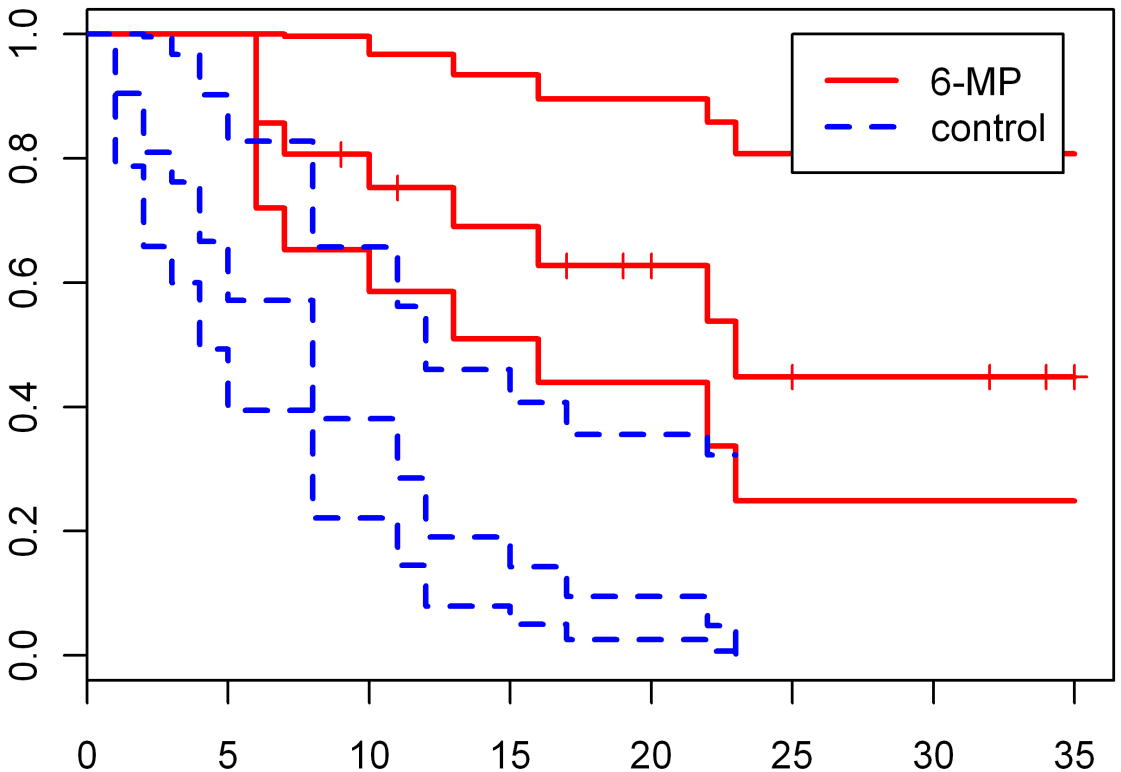


## Czy lek ma istotny wpływ na remisję raka

Obserwujemy czas remisji raka (w tygodniach) dwóch grup pacjentów. Jedna z grup otrzymuje placebo.

```
>library(MASS);library(survival)
>data(gehan)
> attach(gehan)
> Surv(time,cens)
 [1]  1  10  22   7   3  32+ 12  23   8  22  17
[12]  6   2  16  11  34+  8  32+ 12  25+  2  11+
[23]  5  20+  4  19+ 15   6   8  17+ 23  35+  5
[34]  6  11  13   4   9+  1   6+  8  10+
> table(treat)
treat
   6-MP control
     21      21
```


6-MP
control
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35


6-MP
control
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35

## Sprawdzenie, czy są różnice między grupami

```
> survdiff(Surv(time, cens) ~ treat, data = gehan)
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = gehan)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP     21        9     19.3      5.46      16.8
treat=control  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Jest istotna różnica

# Modele parametryczne

## Model parametryczny (przykłady)

$$\ln(T) = \beta^T x + \sigma \ln(E)$$

$\sigma-$ parametr skali
$x-$ wektor cech objaśniających
$\beta-$ wektor parametrów
$E-$ zmienna losowa z rozkładu wykładniczego

W zależności od rozkładu zmiennej losowej $E$ rozróżniamy modele, np: Weibulla, Wykładniczy (wtedy $\sigma = 1$)

# Modele parametryczne

## Model Weibulla dla poprzedniego przykładu

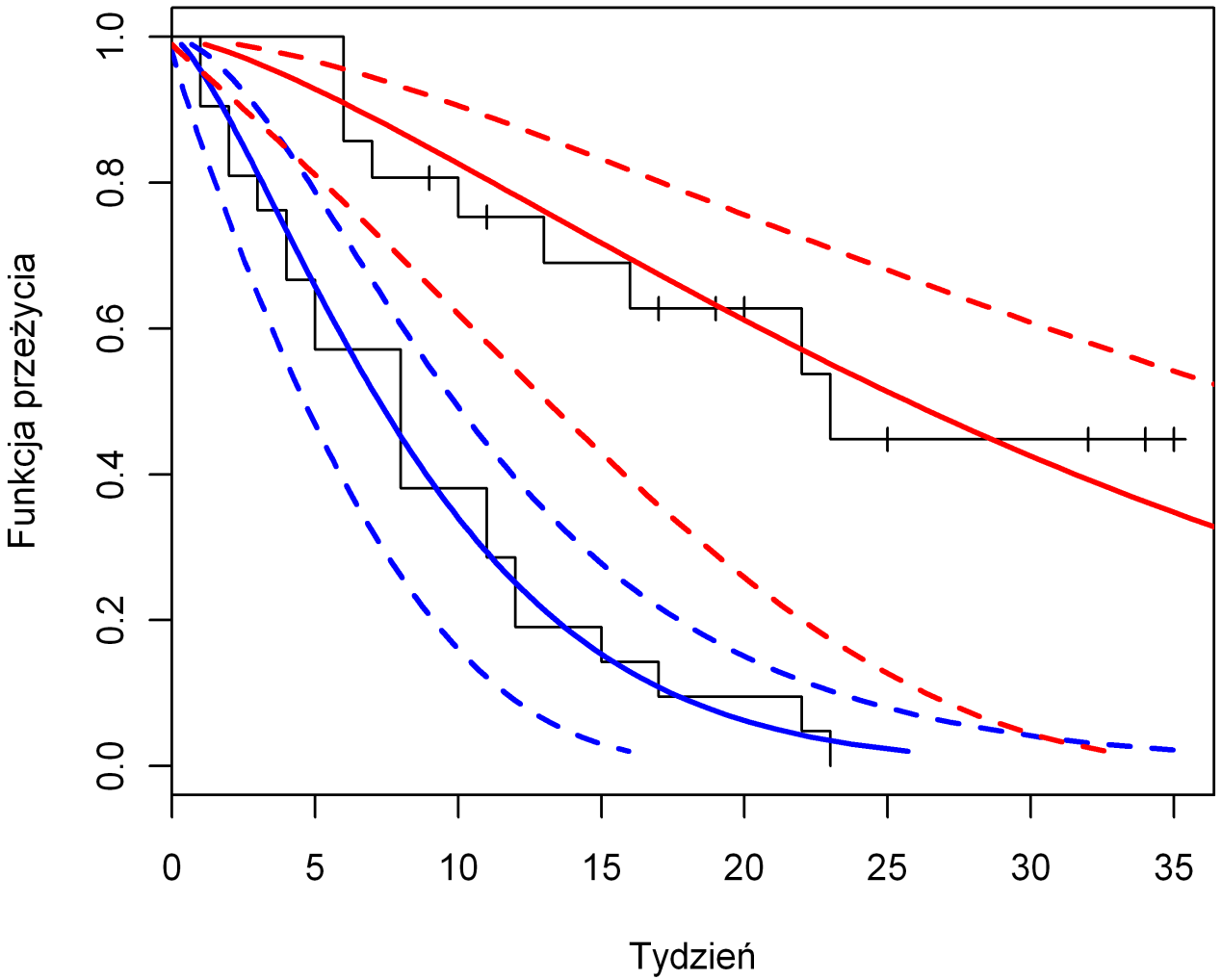
```
> remisja=survreg(Surv(time, cens) ~ treat, data = gehan)
> summary(remisja)

Call:
survreg(formula = Surv(time, cens) ~ treat, data = gehan)
              Value Std. Error     z        p
(Intercept)   3.516      0.252 13.96 2.61e-44
treatcontrol -1.267      0.311 -4.08 4.51e-05
Log(scale)   -0.312      0.147 -2.12 3.43e-02

Scale= 0.732

Weibull distribution
Loglik(model)= -106.6   Loglik(intercept only)= -116.4
        Chisq= 19.65 on 1 degrees of freedom, p= 9.3e-06
Number of Newton-Raphson Iterations: 5
n= 42
```

## Porównanie modelu Weibulla z modelem nieparametrycznym



# Model proporcjonalnego hazardu Cox'a

## Model

$$h(t) = h_0(t) \exp(\beta^T x)$$

$x-$ wektor cech objaśniających
$\beta-$ wektor parametrów
$h_0(t)-$ bazowa funkcja hazardu (nieznana); dla $x = 0$

## Model proporcjonalnego hazardu Cox'a dla poprzedniego przykładu

```
> remisja=coxph(Surv(time, cens) ~ treat, data = gehan,method="exact")
> summary(remisja)

  n= 42, number of events= 30

               coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.6282    5.0949   0.4331 3.759  0.00017 ***
---
             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol     5.095     0.1963      2.18     11.91

Rsquare= 0.321   (max possible= 0.98 )
Likelihood ratio test= 16.25  on 1 df,   p=5.544e-05
Wald test            = 14.13  on 1 df,   p=0.0001704
Score (logrank) test = 16.79  on 1 df,   p=4.169e-05
```