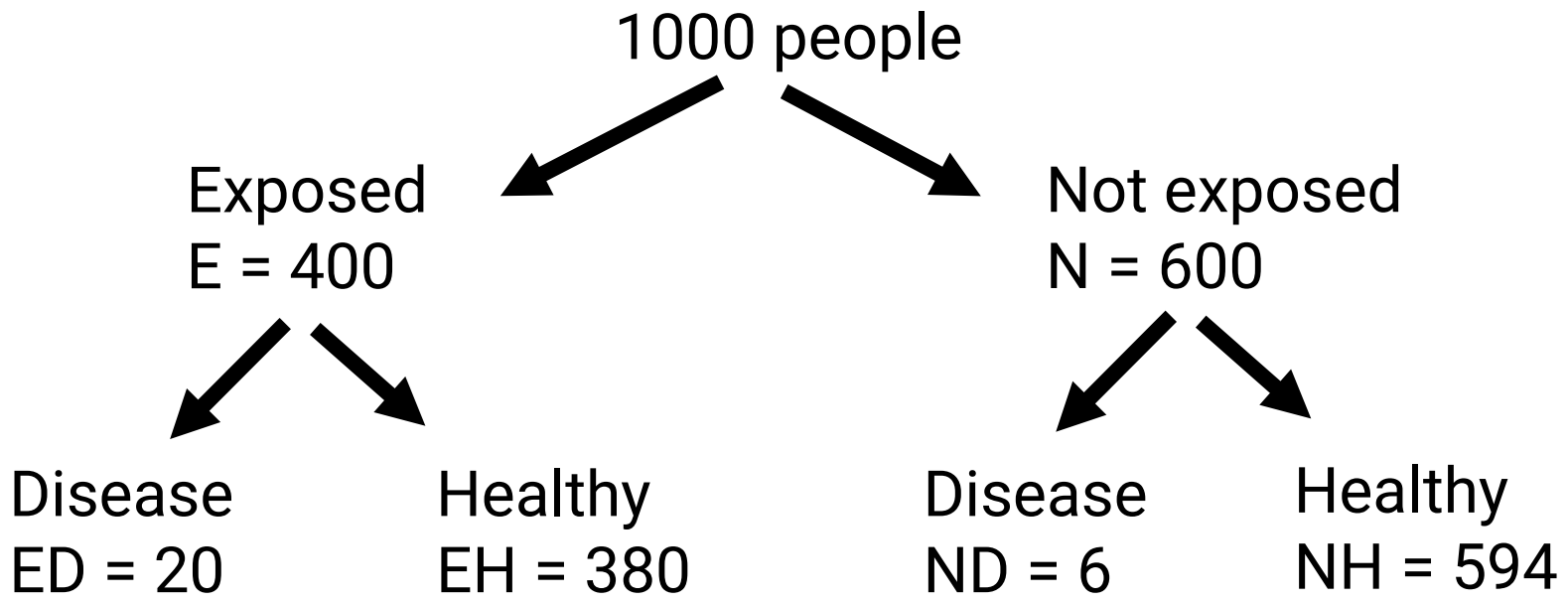


$$\text{Relative Risk} = \frac{ED / (ED+EH)}{ND/(ND+NH)} = \frac{20 / (20+380)}{6 / (6+594)} = 5$$

$$\text{Relative Risk} = \frac{ED / (ED+EH)}{ND/(ND+NH)} = \frac{20 / (20+380)}{6 / (6+594)} = 5$$

$$\text{Odds Ratio} = \frac{ED / (EH)}{ND/ (NH)} = \frac{20 / (380)}{6 / (594)} = 5.2$$

Gdy choroba jest rzadka to RR ≈ OR

Aby wyliczyć RR, chory/zdrowy muszą być wartościami losowymi.

Gdy choroba jest rzadka, to ankieta wśród 100 osób nic nie da!

Odds ratio jest trudniejsze w interpretacji, ale częściej spotykane.

Odds ratio można wyliczyć w regresji logistycznej dla doświadczenia typu case-control. (Chorych wyszukujemy i dobieramy zdrowych)

# Uogólniony model liniowy

# Ogólny model liniowy

$$y = Xb + e$$

- Każda obserwacja ma rozkład normalny
- Każda obserwacja ma tą samą wariancję

# Dane ‘nienormalne’

- Rozkład binomialny
  np. liczba chorych krów w stadzie

- Rozkład zdarzeń rzadkich (Poissona)
  np. liczba zwierząt w stadzie z rzadkim typem raka

- Rozkład wielomianowy
  np. umaszczenie

# Uogólniony model liniowy

- pozwala zastosować model liniowy do ‘nienormalnych danych’
- **y** może być zmienną binarną (zdrowy/chory), wielowartościową (umaszczenie)
- **x** – zarówno zmienne ciągłe jak i kategoryzujące

# Uogólniony model liniowy

$$L(\text{średnia z } y) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

L to funkcja wiążąca (link)

# Funkcje wiążące

- Rozkład binomialny:
  - logit  - pozwala estymować odds ratio (OR)
  - probit
  - log
  - cloglog

- Rozkład Poissona
  - log
  - sqrt

# Rozkład dwumianowy

Określa prawdopodobieństwo y sukcesów w n próbach, jeżeli w pojedynczej próbie prawdopodobieństwo sukcesu wynosi p

Wszystkie próby są takie same i niezależne!

# Odds

$$\text{odds} = \frac{p}{1 - p} \qquad [\,0, \infty\,)$$

Np. gdy p = prawdopodobieństwo zachorowania

- **odds=5** choroba jest 5 razy bardziej prawdopodobna niż zdrowie
- **odds=1/3** szansa zachorowania jest 3 razy mniej prawdopodobna niż zdrowie
- **odds=1** szanse obu zdarzeń są równe

# log odds

$$\log\left(\frac{p}{1-p}\right)$$

$$(-\infty, \infty)$$

**log odds > 0** szansa zachorowania jest większa niż pozostania przy zdrowiu

**log odds < 0** szansa pozostania przy zdrowiu jest większa niż zachorowania

**log odds = 0** szanse obu zdarzeń są równe

# Odds ratio (OR)

$$OR = \frac{\frac{p}{1-p}}{\frac{q}{1-q}}$$

p = prawdopodobieństwo zachorowania wśród osób eksponowanych na badany czynnik ryzyka

q = prawdopodobieństwo zachorowania wśród osób NIE eksponowanych na badany czynnik ryzyka

# Logit

$$\text{logit}( x ) = \ln( x / (1-x) )$$

x – wartość z przedziału 0-1

Gdy x jest prawdopodobieństwem,
wówczas logit to **log odds**

# Regresja logistyczna

- Uogólniony model wykorzystujący funkcję logit
- Pozwala testować wpływ różnych czynników ilościowych i jakościowych na zmienną binarną (np. chory/zdrowy)
- Pozwala zmierzyć wpływ badanego czynnika (zmiana wielkości OR)

# Regresja logistyczna

$$\text{logit}(p_i) = b_0 + b_1 x_1 + b_2 x_2 \ldots$$

$p_i$ – prawd. ‘sukcesu’ dla i-tego przypadku

X – zestaw wartości zmiennych objaśniających

b – zestaw współczynników regresji

X może być zmienną kategoryzującą lub ciągłą!

# Regresja logistyczna - interpretacja

Gdy x jest zmienną binarną (0/1)

$\ln( OR ) = \ln( e^b ) = b$

$OR = \exp( b )$

Gdy x jest zmienną ciągłą

- Jeżeli x wzrasta o 1, to **log odds** wzrasta (dodawanie) o b
- Jeżeli x wzrasta o 1, to **odds** mnoży się przez $e^b$

Gdy x jest zmienną wielomianową, rzecz staje się trudniejsza. Na kursie rozszerzonym.

# Przykład

Badamy wpływ wieku i palenia na nadciśnienie.

Dane:

y - nadciśnienie (TAK=1 / NIE=0)

$x_1$ - wiek (liczba lat – zmienna ciągła)

$x_2$ - palenie (pali=1 / nie pali = 0)

**0 30 0**
**0 30 1**
**0 30 0**
**1 40 1**
**1 40 0**
**1 60 1**
**1 50 0**
**0 30 1**
**1 30 1**
**0 40 1**
**1 60 1**
**1 50 0**
**1 50 1**
**0 60 0**

# Przykład – rozwiązanie

$b_1 = 0,10$ czyli każdy rok zwiększa ryzyko nadciśnienia o $e^{0,10} = 1,10$, czyli o 10%

$b_2 = 0,81$ a więc palenie zwiększa ryzyko nadciśnienia o OR = $e^{0,81} = 2,2$, czyli o 120% (te słowa są trochę mylące bo to nie jest RR)

$b_0 = -4.37$ (intercept) brak interpretacji

# 95% przedział ufności

b = 0.81  SE = 0.40

OR = 2.2

$e^{0.81 \pm 1.96 * 0.40} = [\, e^{0.03}; e^{1.59} \,] = [\, 1.03 ; 4.9 \,]$

Wartość 1 jest poza przedziałem. Asocjacja jest statystycznie istotna.

# Przygotowanie dane w formie surowej (w R)

Dane mogą być wpisane w formie surowej, przy czym p = P(y=1) czyli (sukces=1)

```
y <- c( 0, 0, 0, 1, 1, 1 ... )
wiek <- c( 30, 30, 30, 40, 40, 40 ...)
palenie <- c( 0, 1, 0, 1, 0, 1 ... )
mojedane <- data.frame( y, wiek, palenie)
```

Lub

surowe dane mogą być wczytane z pliku razem z nagłówkami

```
mojedane <- read.table( „nazwapliku.txt”, header=TRUE )
```

# Przygotowanie danych w formie zbiorczej

| wiek | palenie | liczba chorych | liczba osobników zdrowych |
|---|---|---|---|
| 30 | 0 | 0 | 2 |
| 30 | 1 | 1 | 2 |
| 40 | 0 | 1 | 0 |
| 40 | 1 | 1 | 1 |
| itd. | | | |

```
yc <- c( 0, 1, 1, 1, ... )
yz <- c( 2, 2, 0, 1, ... )
y  <- cbind( yc, yz )     #Uwaga, y to macierz!

wiek         <- c( 30, 30, 40, 40, ...)
palenie      <- c( 0, 1, 0, 1, ... )

mojedane <- data.frame(y, wiek, palenie)
```

```
wynik <- glm( y ~ wiek + palenie,

family=binomial(link=logit),
        data=mojedane )

summary( wynik )
confint( wynik )
```

Uwaga:
Dla danych surowych y jest wektorem zer i jedynek, natomiast dla danych zbiorczych y jest macierzą z 2 kolumnami (chore, zdrowe)

# Zadanie 1

Zbadaj wpływ wieku na występowanie zaćmy u psów.

| wiek | 10 | 13 | 14 | 15 | 17 |
|---|---|---|---|---|---|
| liczba badanych | 50 | 45 | 48 | 44 | 37 |
| liczba chorych | 1 | 3 | 8 | 9 | 9 |

# Zadanie 2

Zbadaj wpływ obecności w genotypie zmutowanego genu *CNEP1R1* na otyłość psów, z uwzględnieniem płci i wieku.

Dane „otylosc-psow.txt” pobierz ze strony
http://merlin.up.poznan.pl/~mcszyd/dyda/Biostatystyka/

Opis danych

obese = 0 lub 1 (pies otyły)
sex = Male lub Female
age = wiek w miesiącach
del = 0 lub 1 (delecja w co najmniej jednej kopii genu *CNEP1R1*)