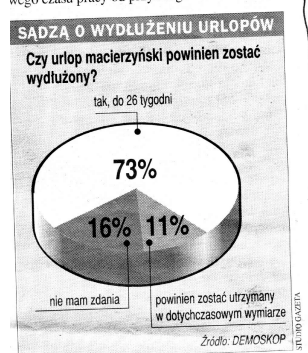


# 1 Wykład 5

**Przykład 1.1** W "Gazecie Wyborczej" z 23.10 1999 pojawiły się wyniki sondażu na temat "Czy urlop macierzyński powinien zostać wydłużony?"



Badanie z 8-11 października na 1000-osobowej próbie losowej Polaków w wieku powyżej 15 lat. Sondażu wśród menedżerów dokonano na losowej próbie 965 firm

Zapytano 1000 Polaków w wieku powyżej 15 lat. 73% respondentów odpowiedziało TAK, 11% NIE i 16% nie miało zdania w tej sprawie. W komentarzu gazety napisano, że "...3/4 Polaków jest za..."

Czy gazeta miała prawo tak napisać?

Jaki model statystyczny kryje się za tym stwierdzeniem? Wektor obserwacji  $X$  ma 1000 składowych  $x_1, x_2, \dots, x_{1000}$ , z których każda odpowiada wypowiedzi jednego z respondentów. Przyporządkujemy odpowiedzi TAK liczbę 1, pozostałym odpowiedziom liczbę 0. Odpowiedzi  $x_1, x_2, \dots, x_{1000}$  są próbą prostą o rozkładzie,  $P(X=1) = \theta$ . W "Gazecie" napisano więc, że  $\theta = 0.75$ . Wyniki sondażu oznaczają, że

$$\bar{X} = \frac{1}{1000} (x_1 + x_2 + \dots + x_{1000}) = 0.73$$

Zbudujmy przedział ufności na poziomie ufności 0.95 dla wartości  $\theta$ .

Skorzystamy tu z Centralnego Twierdzenia Granicznego, które w tym przypadku oznacza, że<sup>1</sup>

$$\frac{\bar{X} - E\bar{X}}{\text{Var}(\bar{X})} \simeq Z$$

gdzie  $Z$  ma standardowy rozkład normalny. Liczba 1000 jest dostatecznie duża, aby można było przyjąć przybliżenie rozkładem normalnym. Posłużymy się tu

<sup>1</sup>symbol  $\simeq$  oznacza asymptotyczną równość wg prawdopodobieństwa

rozwiązaniem z poprzedniego wykładu. Stąd mamy, że<sup>2</sup>

$$\begin{aligned} 0.95 &\leq P\left(-1.96 \leq \frac{\bar{X} - E\bar{X}}{\sqrt{\text{Var}(\bar{X})}} \leq 1.96\right) = \\ &P\left(-1.96 \leq \frac{\bar{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{1000}}} \leq 1.96\right) \leq P\left(-1.96 \leq \frac{\bar{X} - \theta}{\sqrt{\frac{1}{4000}}} \leq 1.96\right) = \\ &P\left(\bar{X} - \frac{1.96}{\sqrt{4000}} \leq \theta \leq \bar{X} + \frac{1.96}{\sqrt{4000}}\right) = P(0.73 - 0.031 \leq \theta \leq 0.73 + 0.031) \end{aligned}$$

Przedział ufności dla prawdopodobieństwa odpowiedzi TAK w populacji ma postać

$$[0.699, 0.761]$$

zawiera więc wielkość 0.75 podaną przez "Gazetę". ☹

W powyższym przykładzie wyznaczyliśmy dokładność sondażu, równą  $0.031 = 3.1\%$  nie znając parametru  $\theta$ . Ogólne standardy badań opinii publicznej zakładają, że należy dobrać tak dużą próbę, aby na poziomie 0.95 dokładność przedziału ufności nie przekraczała 3%.

Aby wyznaczyć na poziomie  $\gamma$  z dokładnością  $d_0$  przedział ufności dla  $\theta$  w próbie prostej z rozkładu dwupunktowego z prawdopodobieństwem wyniku 1 równym  $\theta$  wystarczy rozwiązać nierówność

$$\frac{\eta}{\sqrt{4n}} \leq d_0, \eta = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

gdzie  $\Phi$  jest dystrybucją standardowego rozkładu normalnego.

Daje to rozwiązanie

$$n \geq \left(\frac{\eta}{2d_0}\right)^2$$

W przypadku  $d_0 = 0.03, \gamma = 0.95$   $n \geq 1067.1$ .

Szukając przedziału ufności korzystaliśmy z pewnego "chwytu" matematycznego: szukaliśmy takiej funkcji zmiennej losowej o nieznanym parametrze  $\theta$ , że jej rozkład nie zależy od tego parametru. Nie zawsze jest jednak tak prosto, jak w dotąd rozważanych przypadkach.

**Przykład 1.2** Obserwujemy odstęp między kolejnymi przejazdami samochodów przez ustalony punkt. Chcemy dowiedzieć się jaki jest oczekiwany czas między kolejnymi przyjazdami. Przypuśćmy, że obserwacja 10 kolejnych samochodów dała średni odstęp równy 15 sekund.

Wektor obserwacji  $X$  ma 10 składowych  $x_1, x_2, \dots, x_{10}$ . Zakładamy, że jest to próba prosta z rozkładu czasu oczekiwania na zdarzenie ("Przez wybrany punkt przejedzie samochód"). W rachunku prawdopodobieństwa dowodzi się, że przy bardzo ogólnych założeniach taka zmienna losowa ma rozkład wykładniczy, którego gęstość ma postać:

$$f(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right), x \in (0, \infty)$$

---

<sup>2</sup>  $\theta(1-\theta) \leq \frac{1}{4}$  dla  $0 \leq \theta \leq 1$

gdzie  $\mu$  jest średnim czasem oczekiwania na zajście zdarzenia. Parametr  $\mu$  jest nieznanym parametrem. Będziemy szukać przedziału ufności dla tego parametru na poziomie  $\gamma = 0.95$

Zainteresujemy się zmienną losową  $S = \frac{2n}{\mu} \bar{X}$ . Znajdziemy dystrybuantę  $F_S(s)$  tej zmiennej losowej.

$$F_S(s) = P(S < s) = P\left(\sum_{i=1}^n X_i < \frac{s\mu}{2}\right) = P\left(T_n < \frac{s\mu}{2}\right) = F_{T_n}\left(\frac{s\mu}{2}\right)$$

gdzie

$$T_n \stackrel{\text{df}}{=} \sum_{i=1}^n X_i.$$

Skorzystamy tu z faktu, że gdy  $X_i$  są niezależne o rozkładzie wykładniczym ze średnią  $\mu$  to  $T_n$  ma rozkład gamma o gęstości

$$\frac{\mu^{-n} t^{n-1} \exp\left(-\frac{t}{\mu}\right)}{(n-1)!}.$$

Gęstość zmiennej losowej  $S$  będzie równa

$$f_S(s) = f_{T_n}\left(\frac{s\mu}{2}\right) \frac{\mu}{2} = \frac{\mu^{-n} \left(\frac{s\mu}{2}\right)^{n-1} \exp\left(-\frac{s}{2}\right) \frac{\mu}{2}}{(n-1)!} = 2^{-n} s^{n-1} \frac{\exp\left(-\frac{s}{2}\right)}{(n-1)!}$$

i nie zależy od  $\mu$ . Rozkład ten nazywa się rozkładem  $\chi^2$  z  $2n$  stopniami swobody. Z tablic rozkładu  $\chi^2$  z  $2n$  stopniami swobody wybieramy<sup>3</sup> taką parę liczb  $\xi$  i  $\eta$ , że  $P(\xi \leq \chi^2 \leq \eta) \geq \gamma$ . Wtedy

$$\gamma \leq P\left(\xi \leq \frac{2n}{\mu} \bar{X} \leq \eta\right) = P\left(\frac{2n}{\eta} \bar{X} \leq \mu \leq \frac{2n}{\xi} \bar{X}\right)$$

W naszym przykładzie  $n = 10$ ,  $\bar{X} = 15$ ,  $\gamma = 0.95$ ,  $\xi = 9.591$ ,  $\eta = 34.170$ . Końce przedziału ufności będą równe

$$a = \frac{2 * 10 * 15}{34.170} \simeq 8.7, b = \frac{2 * 10 * 15}{9.591} \simeq 31.3$$

Należy zwrócić uwagę, że średnia nie jest środkiem tego przedziału i na to, że jest on bardzo szeroki.✂

Rozkład  $\chi^2$  wiąże się z rozkładem normalnym. Mamy układ współrzędnych kartezjański i losowo wybieramy punkt na płaszczyźnie. Gdy mówimy losowo to powinniśmy opisać mechanizm wyboru współrzędnych  $(x_1, x_2)$  tego punktu. Niech  $X_i$  będą niezależne i o standardowym rozkładzie normalnym. Kwadrat długości wektora, zaczepionego w początku układu współrzędnych i o końcu w losowym punkcie  $(x_1, x_2)$  jest zmienną losową, której rozkład nazywa się rozkładem  $\chi^2$  z dwoma stopniami swobody<sup>4</sup>. Uogólniając, weźmy wektor zaczepiony w początku układu współrzędnych o końcu w losowym punkcie

<sup>3</sup>rozkład  $\chi^2$  nie jest symetryczny

<sup>4</sup>nazwa pochodzi od liczby współrzędnych tego wektora

$(x_1, x_2, \dots, x_n)$ , którego współrzędne są niezależne i mają standardowy rozkład normalny. Rozkład kwadratu długości takiego wektora nazywa się rozkładem  $\chi^2$  z  $n$  stopniami swobody. Zgodnie z Centralnym Twierdzeniem Granicznym rozkład  $\chi^2$  jako suma niezależnych zmiennych losowych o tym samym rozkładzie po standaryzacji zbiega do rozkładu normalnego o średniej 0 i odchyleniu 1 gdy liczba stopni swobody zbiega do nieskończoności. Zbieżność ta jest jednak bardzo wolna.

## 1.1 Testowanie hipotez

To, o czym dotąd mówiliśmy opierało się na następującym schemacie:

Bazujemy na zebranych informacjach. Na ich podstawie mierzymy do oszacowania nieznanymi parametrów, które precyzują opis rzeczywistości (model). Oszacowanie to było albo punktowe albo przedziałowe. Można by powiedzieć, że odpowiadaliśmy na pytanie, które zaczynało się od słów: "Ile ...".

Nie zawsze jesteśmy zainteresowani odpowiedzią na tak szczegółowe pytanie. Czasami chcemy dowiedzieć się "Czy prawda jest, że...". Oto rzeczywiste przykłady:

**Przykład 1.3** *Podjeżdżając, że ludzie są zdolni do odwołania momentu ich śmierci na czas po jakimś ważnym wydarzeniu. Zbadano 1919 osób o nazwiskach żydowskich. W tygodniu poprzedzającym Święto Paschy zmarły 922 osoby, po Święcie zmarło 997 osób. Czy mimo tego można twierdzić, że śmierć następuje wśród Żydów tak samo przed Świętem Paschy jak i po tym święcie?*

**Przykład 1.4** *(badania Mendla) Mendel skrzyżował 556 gładkich, żółtych ziaren grochu z pomarszczonymi, zielonymi ziarnami. Mendel interesował się, w jaki sposób te dwie cechy (kolor i gładkość ziarna) dziedziczą się w następnym pokoleniu. Wyniki można umieścić w tabeli<sup>5</sup>:*

| typ                  | liczba obserwacji | prawdopodobieństwo teoretyczne | oczekiwana liczba ziaren |
|----------------------|-------------------|--------------------------------|--------------------------|
| gładki żółty         | 315               | 9/16                           | 312.75                   |
| gładki zielony       | 108               | 3/16                           | 104.25                   |
| pomarszczony żółty   | 102               | 3/16                           | 104.25                   |
| pomarszczony zielony | 31                | 1/16                           | 34.75                    |

*Czy te badania potwierdzają prawa dziedziczenia?*

Gdy oczekujemy odpowiedzi TAK lub NIE to mamy do czynienia z zagadnieniem, nazywanym w statystyce **testowaniem hipotez**.

W nauce takie pytania formułowane są w postaci alternatywy dwóch hipotez: hipotezy zerowej i hipotezy konkurencyjnej. Zasada konserwatyzmu, przyjęta w nauce mówi, że starą teorię będziemy odrzucać wtedy, gdy mamy dużo faktów, świadczących przeciwko niej. Tradycyjnie umieszczamy hipotezę konserwatywną jako hipotezę zerową i hipotezę nową, jako hipotezę konkurencyjną. Lekarz, który chce sprawdzić nowy lek, jako hipotezę konkurencyjną umieszcza zdanie: "Nowy lek jest lepszy od starego" a jako hipotezę zerową zdanie: "Stary lek jest lepszy od nowego"

<sup>5</sup>prawdopodobieństwo teoretyczne wyliczone na podstawie prawa Hardy'ego-Weinberga (Lista 2)