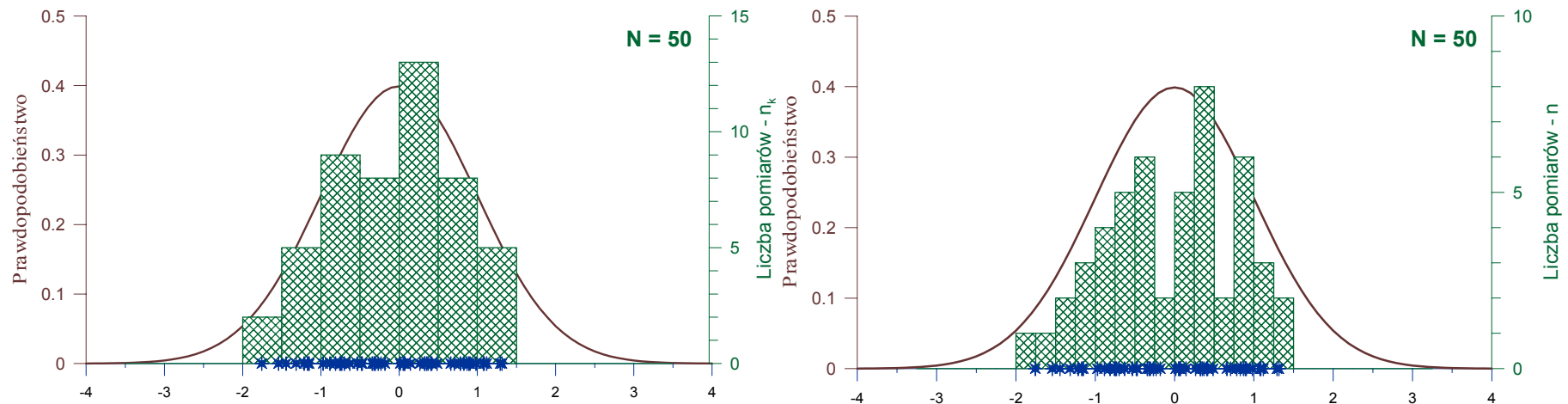


# HISTOGRAM

W pewnych przypadkach interesuje nas nie tylko określenie prawdziwej wartości mierzonej wielkości, ale także zbadanie całego rozkład prawdopodobieństwa wyników pomiarów. W takim przypadku wyniki serii  $N$  pomiarów przedstawiamy za pomocą histogramu. W tym celu obszar zmienności mierzonej wielkości  $x$  dzielimy na  $m$  przedziałów (zwanych przedziałami klasowymi), a następnie dla każdego przedziału określamy ile wartości  $x_i$  wyników pomiarów mieści się w tym przedziale. Rezultat możemy przedstawić w formie wykresu „słupkowego”. Oczywiście histogram jest tylko przybliżeniem kształtu rozkładu, a wysokość słupków (liczebność przedziałów) jest zmienną losową, co oznacza, że po powtórzeniu serii histogram będzie inny.

Wskazane jest takie dobranie wielkości i liczby przedziałów, aby na każdy z nich przypadało co najmniej kilka (5-8) pomiarów



# DYSPERSJA LICZEBNOŚCI PRZEDZIAŁÓW HISTOGRAMU

Przyjmijmy, że przeprowadzono  $N$  pomiarów wielkości  $x$ , a wyniki przedstawiono w formie histogramu. Rozważmy jeden, dowolny przedział histogramu określony jako  $(x_0, x_0 + \Delta x)$ . Zastanówmy się ile wynosi prawdopodobieństwo  $P(n, N)$  tego, że liczebność rozważanego przedziału będzie wynosiła  $n$  (tzn. z ogólnej liczby  $N$  pomiarów w serii wartości z zakresu od  $x_0$  do  $x_0 + \Delta x$  będzie miało  $n$  pomiarów).

Prawdopodobieństwo tego, że w wyniku pojedynczego pomiaru uzyska się wartość z zakresu  $(x_0, x_0 + \Delta x)$  można wyznaczyć z gęstości rozkładu prawdopodobieństwa  $f(x)$  zmiennej  $x$ . Wynosi ono:

$$p = \int_{x_0}^{x_0 + \Delta x} f(x) dx$$

Ponieważ kolejne wyniki uzyskano niezależnie od siebie prawdopodobieństwo uzyskania  $n$  wyników z przedziału  $(x_0, x_0 + \Delta x)$  oraz  $N - n$  wyników spoza tego przedziału określone jest iloczynem  $p^n \cdot (1 - p)^{N - n}$ .

Jednakże  $n$  rozważanych wyników można uzyskać w dowolnej kolejności, zatem powyższy iloczyn należy jeszcze pomnożyć przez liczbę możliwych kolejności otrzymania w serii  $N$  pomiarów  $n$  wyników z rozważanego przedziału. Ostatecznie zatem uzyskujemy wyrażenie:

$$P(n, N) = \frac{N!}{n! \cdot (N - n)!} \cdot p^n \cdot (1 - p)^{N - n}$$

**Jest to rozkład dwumianowy Bernoulliego.**

Widać zatem, że liczebność przedziału histogramu -  $n$ , jest zmienną losową opisywaną rozkładem Bernoulliego. Wartość oczekiwana jest zatem równa

$$E(n) = N \cdot p$$

a wariancja

$$V(n) = N \cdot p \cdot (1-p)$$

Pierwiastek tej ostatniej daje nam dyspersję  $D(n)$ , którą możemy traktować jako miarę rozrzutu zmiennej  $n$ , a zatem możliwych wahań liczebności przedziału histogramu (wysokości słupka).

W praktyce wartość  $n$  jest estymatorem  $E(n)$ , zatem wartość prawdopodobieństwa  $p$  możemy oszacować jako:

$$p_{est} = \frac{n}{N}$$

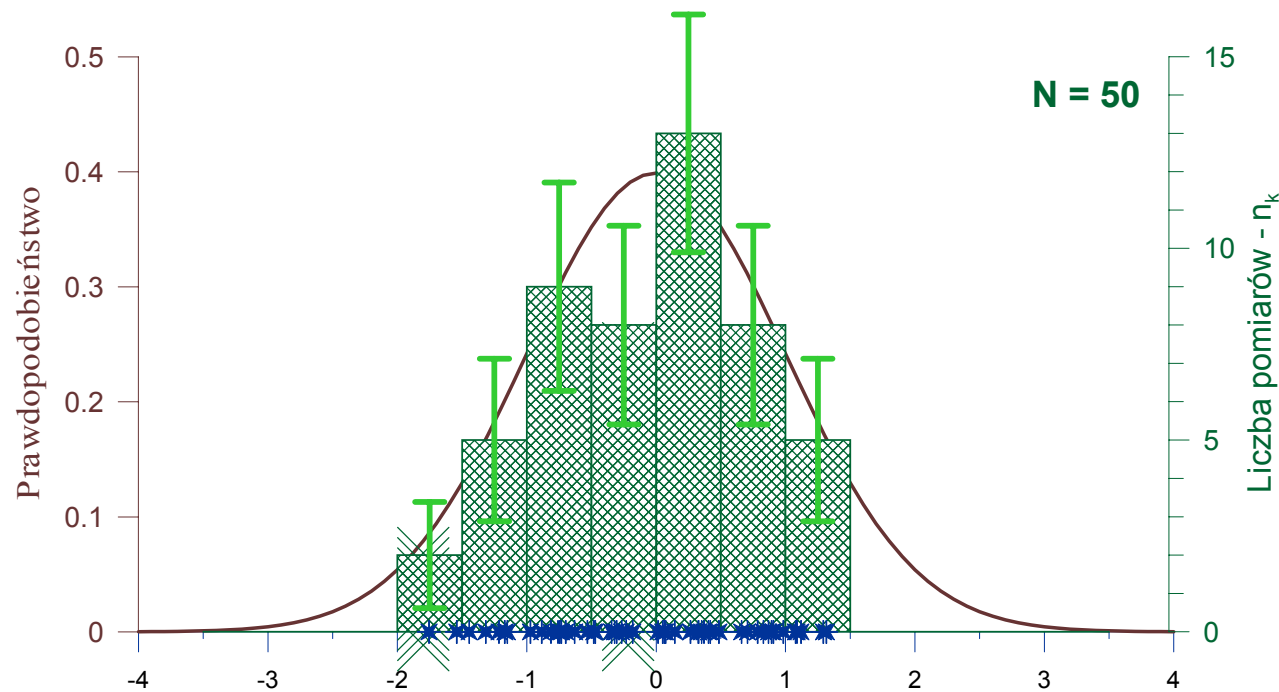
Zatem dyspersja  $D(n)$  może być oszacowana za pomocą wzoru

$$s_n = \sqrt{n \left( 1 - \frac{n}{N} \right)},$$

który może być traktowany jako wzór określający niepewność liczebności przedziału histogramu (wysokości słupka).

Dla dużej liczby przeprowadzonych pomiarów tzn.  $N \rightarrow \infty$  możliwe jest podzielenie obszaru zmienności wielkości  $x$  na wiele przedziałów klasowych, a co za tym idzie prawdopodobieństwo  $p$  maleje ( $p \rightarrow 0$ ). W takim przypadku rozkład prawdopodobieństwa zmiennej  $n$  przechodzi w rozkład Poissona, a następnie w rozkład Gaussa.

Jeżeli przedstawimy niepewności liczebności poszczególnych przedziałów klasowych histogramu na wykresie zauważymy, że możliwe wahania wysokości słupka są dość znaczne.



Optymalny efekt otrzymuje się zmierzając do minimalizacji niepewności względnej wysokości słupka histogramu  $s_n/n$ .

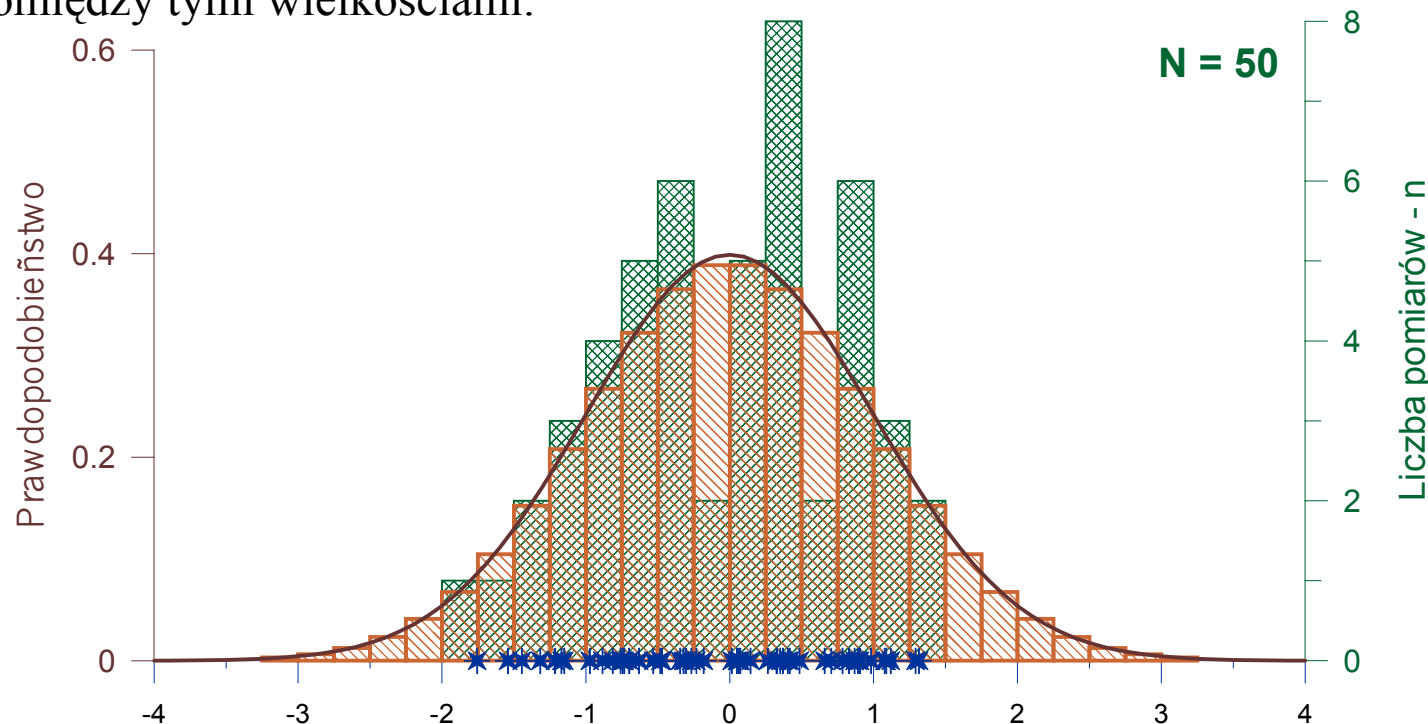
$$\frac{s_n}{n} = \frac{\sqrt{n\left(1 - \frac{n}{N}\right)}}{n} = \sqrt{\frac{n\left(1 - \frac{n}{N}\right)}{n^2}} = \sqrt{\frac{1 - \frac{n}{N}}{n}} = \sqrt{\frac{1}{n} - \frac{1}{N}}$$

Łatwo zauważyć, że  $s_n/n$  maleje ze wzrostem  $n$ .

# TEST $\chi^2$ ZGODNOŚCI ROZKŁADÓW

W wielu przypadkach interesować nas będzie odpowiedź na pytanie, czy otrzymany rozkład empiryczny (histogram) badanej wielkości jest zgodny rozkładem wynikającym z teorii lub założonym na innej podstawie. Przykładowo - niezgodność rozkładu wyników pomiarów z rozkładem Gaussa może świadczyć o tym, że w trakcie trwania serii pomiarów nastąpiła systematyczna zmiana warunków, w jakich były one dokonywane.

W celu sprawdzenia zgodności rozkładów porównujemy liczebność poszczególnych przedziałów histogramu otrzymanego w oparciu o wyniki pomiarów z wartością oczekiwaną liczebności jaka wynika z założonej postaci rozkładu poprzez wyznaczenie sumy kwadratów różnic pomiędzy tymi wielkościami.



Założmy, że na podstawie wyników  $N$  pomiarów konstruujemy histogram składający się z  $m$  przedziałów klasowych, których granice mają wartości  $X_0, X_1, \dots, X_m$ . Wyznaczone doświadczalnie liczebności poszczególnych przedziałów wynoszą  $n_i$ , gdzie  $i = 1, 2, \dots, m$ , a

$$\sum_{i=1}^m n_i = N$$

Wartość oczekiwana liczebności przedziału  $i$  wynosi

$$E(n_i) = N \cdot p_i$$

gdzie prawdopodobieństwo  $p_i$  wyznacza się w oparciu o zakładaną postać rozkładu gęstości prawdopodobieństwa  $f(x)$  na podstawie zależności:

$$p_i = \int_{X_{i-1}}^{X_i} f(x) dx$$

W celu przeprowadzenia testu zgodności rozkładów wyznaczamy wartość wyrażenia:

$$z = \sum_{i=1}^m \frac{(n_i - Np_i)^2}{Np_i}$$

Wielkość  $z$  jako funkcja zmiennych losowych  $n_i$  również jest zmienną losową.

Jak pokazano wcześniej zmienne  $n_i$  mają rozkład dwumianowy, ale dla dużych  $N$  rozkłady te zbiegają do rozkładów Gaussa  $N(Np_i, (Np_i)^{1/2})$ . Zatem każdy z elementów sumy będzie wówczas charakteryzował rozkładem normalnym  $N(0,1)$ .

Rozkład prawdopodobieństwa zmiennej wyznaczonej jako suma kwadratów zmiennych losowych o rozkładzie normalnym  $N(0,1)$  jest w statystyce matematycznej nazywany **rozkładem  $\chi^2$** .

## Rozkład $\chi^2$

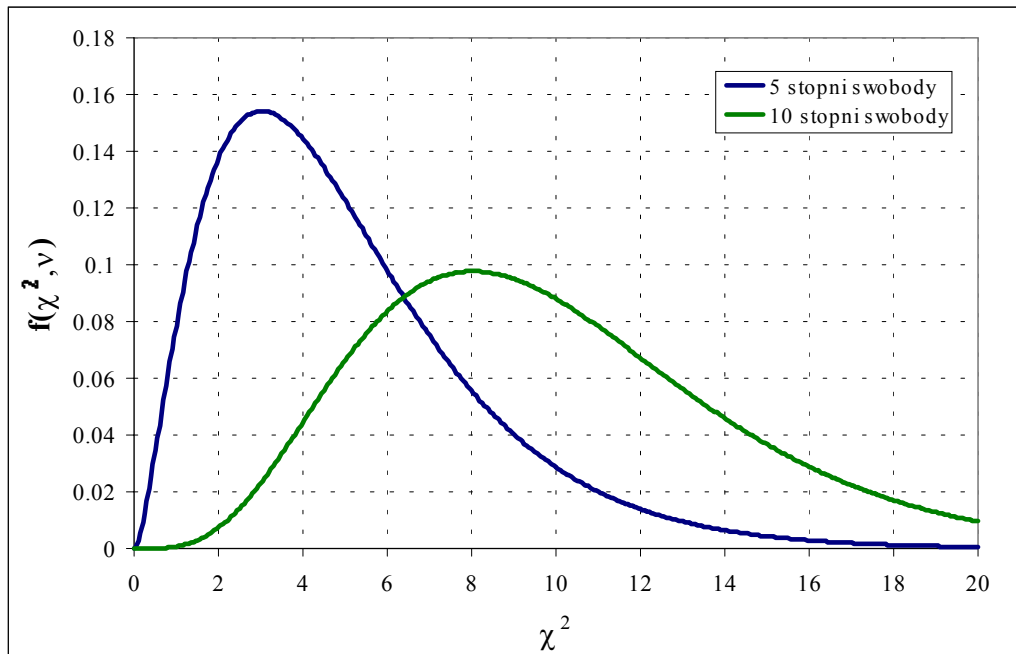
Zmienna losowa  $\chi^2$  otrzymana poprzez zsumowanie  $\nu$  kwadratów niezależnych zmiennych  $x_i$  o rozkładach normalnych  $N(0,1)$

$$\chi^2 = \sum_{i=1}^{\nu} x_i^2$$

ma rozkład zwany , którego funkcja gęstości prawdopodobieństwa opisana jest wzorem:

$$f(\chi^2, \nu) = \frac{(\chi^2)^{\nu/2-1} \cdot e^{-(\chi^2)/2}}{\Gamma(\nu/2) \cdot 2^{\nu/2}}$$

Wielkość  $\nu$  nosi nazwę ilości stopni swobody, a funkcja  $\Gamma(x)$  jest uogólnieniem silni na zbiór liczb rzeczywistych tzn.  $\Gamma(x) = (x-1) \cdot \Gamma(x-1)$ .



Wartość oczekiwana zmiennej  $\chi^2$  jest równa liczbie stopni swobody

$$E(\chi^2) = \nu$$

a wariancja wynosi

$$V(\chi^2) = 2\nu$$

**Procedura testowania hipotezy, że rozkład (histogram) otrzymany doświadczalnie jest zgodny z założoną postacią rozkładu przebiega następująco:**

1. W oparciu o histogram z danych pomiarowych wyznaczamy wartość  $z = \sum_{i=1}^m \frac{(n_i - Np_i)^2}{Np_i}$

Jeżeli testowane rozkłady ze sobą zgodne zmienna  $z$  ma rozkład  $\chi^2$ . Jednakże ilość stopni swobody tego rozkładu jest mniejsza od  $m$  i wynosi

$$v = m - b - 1$$

gdzie  $b$  oznacza ilość parametrów rozkładu (np.  $\mu$  i  $\sigma$  dla rozkładu Gaussa), które wyznaczyliśmy na podstawie danych pomiarowych. (Zmienne  $n_i$  nie są od siebie niezależne, gdyż ich suma jest ściśle określona i równa  $N$ ).

2. Przyjmujemy wartość poziomu istotności  $\alpha$  (określającą prawdopodobieństwo tego, że w wyniku testu odrzucimy hipotezę mimo, iż jest ona prawdziwa).

3. Wyznaczamy wartość krytyczną  $\chi^2_{kr}$  odpowiadającą przyjętemu poziomowi istotności  $\alpha$ . Jeżeli testowane rozkłady nie są ze sobą zgodne to zmienna  $z$  powinna mieć dość dużą wartość, gdyż liczebność każdego z przedziałów klasowych (wysokość każdego ze słupków histogramu)  $n_i$  istotnie różni się od wartości oczekiwanej  $N \cdot p_i$  wynikającej z założonej postaci rozkładu. Zatem obszar krytyczny jest w tym przypadku obszarem wartości zmiennej  $\chi^2$  większych od  $\chi^2_{kr}$  takim, że  $P(\chi^2 > \chi^2_{kr}) = \alpha$ .

4. Porównujemy wartość  $z$  z wartością  $\chi^2_{kr}$ . **Jeżeli  $z > \chi^2_{kr}$  to z prawdopodobieństwem pomyłki równym poziomowi istotności możemy uznać, że testowane rozkłady nie są ze sobą zgodne.**