

Analiza niepewności pomiarowych

Wstęp do analizy danych

Wykłady 2 i 3

Rozkłady prawdopodobieństwa: dwumianowy, Poissona i normalny (Gaussa)

1 Uwagi wstępne

Wykład 1 zakończyliśmy definicją pomiaru jako porównania z wzorcem mierzonej wielkości. Zdefiniowaliśmy także pojęcie błędu pomiaru jako różnicy wyniku pomiaru i *wartości dokładnej*. §5 *Slajdów/transparencji do wykładu* kończy podział błędów pomiaru na: błędy grube, systematyczne oraz przypadkowe. Zaczynamy od sformułowania matematycznego modelu błędów przypadkowych. Posłużymy się **rachunkiem prawdopodobieństwa**. Aksjomaty rachunku prawdopodobieństwa przypomniane zostały w §6 *Slajdów/transparencji do wykładu*. Podczas wykładów 2 i 3 przedyskutowane zostały przykłady rozkładów prawdopodobieństwa: rozkład dwumianowy, rozkład Poissona i rozkład normalny (zwany też rozkładem Gaussa). Poniżej przytoczone są ich definicje i wyprowadzenia przydatnych wzorów.

2 Rozkład dwumianowy

Wykonujemy n losowych prób. Wynikiem każdej próby jest sukces lub porażka (trzeciej możliwości nie ma). W każdej próbie prawdopodobieństwo uzyskania sukcesu wynosi p , $0 \leq p \leq 1$, a porażki q . Oczywiście, $p + q = 1$. Próby są niezależne, tzn. wynik kolejnej próby jest niezależny od wyników wszystkich poprzednich prób.

Pytanie: Jakie jest prawdopodobieństwo $P(k; n, p)$ uzyskania dokładnie k sukcesów w serii n takich prób?

Jeśli oznaczymy sukces jako $+$, a porażkę jako $-$, to wynik kolejnych n prób zapiszemy jako ciąg n znaków $+$ i $-$:

$$+ + - + - - + \dots + - - - + + - - + ,$$

w którym wystąpi k znaków $+$ i $n - k$ znaków $-$. Prawdopodobieństwo uzyskania konkretnego ciągu uzyskamy obliczając wartość iloczynu n liczb p i q , w którym znaki $+$ zastąpimy prawdopodobieństwem p , a znaki $-$ prawdopodobieństwem q :

$$P(+ + - + - - + \dots + - - - + + - - +) = p \cdot p \cdot q \cdot p \cdot q \cdot q \cdot p \dots p \cdot q \cdot q \cdot q \cdot p \cdot p \cdot q \cdot q \cdot p = p^k \cdot q^{n-k}.$$

Prawdopodobieństwo uzyskania dokładnie k sukcesów otrzymamy sumując prawdopodobieństwa wszystkich możliwych ciągów k znaków $+$ i $n-k$ znaków $-$ różniące się położeniami tych znaków. Wiemy, że n różnych obiektów możemy przestawić na $n! := 1 \cdot 2 \cdot 3 \dots (n-1) \cdot n$ sposobów (przyjmujemy, że $0! = 1$). Jednak wszystkie przestawienia między sobą znaków $+$ w naszym ciągu opisują ten sam wynik, podobnie jest z przestawieniami znaków $-$. Liczbę różnych ciągów n znaków $+$ i $n-k$ znaków $-$ otrzymamy, gdy podzielimy $n!$ przez liczbę możliwych przestawień znaków $+$ i znaków $-$. Otrzymamy, że liczba różnych ciągów z dokładnie k znakami $+$ wynosi:

$$\frac{n!}{k!(n-k)!}.$$

Prawdopodobieństwo $P(k; n, p)$ uzyskania dokładnie k sukcesów w n niezależnych próbach, gdy prawdopodobieństwo sukcesu w każdej próbie wynosi p jest równe:

$$P(k; n, p) = \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

Sprawdzamy, że suma prawdopodobieństw otrzymania wszystkich możliwych wartości k od $k=0$ do $k=n$ wynosi 1, tak jak powinna:

$$\sum_{k=0}^{k=n} P(k; n, p) = \sum_{k=0}^{k=n} \frac{n!}{k!(n-k)!} p^k q^{n-k} = (p+q)^n = 1.$$

Otrzymany **rozkład prawdopodobieństwa dyskretnej zmiennej losowej k** nazywany jest **rozkładem dwumianowym**.

Obliczmy ile wynosi **wartość oczekiwana $\mathcal{E}(k)$ liczby sukcesów k** , .

$$\mathcal{E}(k) = \sum_{k=0}^{k=n} k P(k; n, p) = \sum_{k=1}^{k=n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} = \sum_{k=0}^{k=n} k \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Definiujemy nowy wskaźnik sumowania, $l = k - 1$ i korzystamy z własności silni: $k! = k \cdot (k-1)! = k \cdot (l)!$. Otrzymujemy:

$$\mathcal{E}(k) = np \sum_{l=0}^{l=n-1} \frac{(n-1)!}{l!(n-1-l)!} p^l q^{n-1-l} = np(p+q)^{n-1} = np.$$

W ostatnich dwóch równościach skorzystaliśmy z rozwinięcia dwumianu Newtona i z faktu, że $p+q=1$.

Obliczmy także ile wynosi wariancja $Var(k)$ liczby sukcesów k , nazywana także średnim odchyleniem kwadratowym:

$$Var(k) := \sum_{k=0}^{k=n} (k - \mathcal{E}(k))^2 P(k; n, p).$$

W obliczaniu wariancji bardzo pomocne jest poniższe twierdzenie.

Twierdzenie: Dla dowolnej zmiennej losowej x , dla której istnieją wartości oczekiwane $\mathcal{E}(x)$ oraz $\mathcal{E}(x^2)$ i obie te wartości są skończone, zachodzi równość:

$$\mathcal{E}((x - \mathcal{E}(x))^2) = \mathcal{E}(x^2) - (\mathcal{E}(x))^2$$

Dowód przez wykonanie obliczeń. Korzystamy z faktów: (a) wartość oczekiwana jest liczbą oraz (b) wartość oczekiwana dowolnej zmiennej losowej x mnożonej przez dowolną liczbę α równa jest wartości oczekiwanej x mnożonej przez tę liczbę:

$$\mathcal{E}(\alpha x) = \alpha \mathcal{E}(x).$$

Wykonujemy obliczenia:

$$\mathcal{E}(((x - \mathcal{E}(x))^2) = \mathcal{E}(x^2 - 2x\mathcal{E}(x) + (\mathcal{E}(x))^2) = \mathcal{E}(x^2) - 2\mathcal{E}(x)\mathcal{E}(x) + (\mathcal{E}(x))^2 = \mathcal{E}(x^2) - (\mathcal{E}(x))^2.$$

Do obliczenia $Var(k)$ potrzebujemy więc już tylko obliczyć $\mathcal{E}(k^2)$. Możemy postępować podobnie jak poprzednio. Będziemy musieli dwukrotnie zmieniać wskaźnik sumowania... Spróbujmy jednak innej metody. Potraktujmy nasz wzór tak, jakby p i q były niezależnymi zmiennymi i dopiero na samym końcu obliczeń skorzystajmy z faktu, że $p + q = 1$.

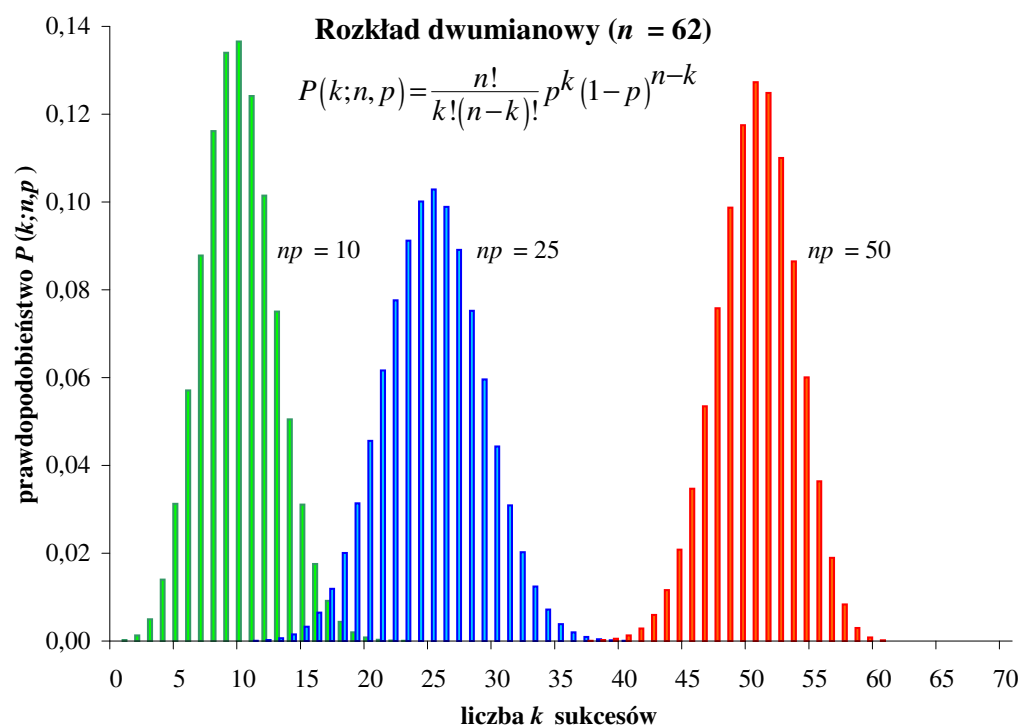
$$\begin{aligned} \mathcal{E}(k^2) &= \sum_{k=0}^{k=n} k^2 P(k; n, p) \\ &= \sum_{k=1}^{k=n} k^2 \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= p \frac{\partial}{\partial p} p \frac{\partial}{\partial p} \left(\sum_{k=0}^n \frac{n!}{k! \cdot (n-k)!} p^k q^{n-k} \right). \end{aligned}$$

Suma w nawiasie jest rozwinięciem dwumianu $(p + q)^n$. Skorzystajmy z tego:

$$\begin{aligned} \mathcal{E}(k^2) &= p \frac{\partial}{\partial p} p \frac{\partial}{\partial p} \left(\sum_{k=0}^n \frac{n!}{k! \cdot (n-k)!} p^k q^{n-k} \right) \\ &= p \frac{\partial}{\partial p} p \frac{\partial}{\partial p} (p + q)^n \\ &= p \frac{\partial}{\partial p} p(n(p + q)^{n-1}) \\ &= p(n(p + q)^{n-1} + n(n-1)p(p + q)^{n-2}) \\ &= np + n(n-1)p^2 \\ &= np + n^2p^2 - np^2 \end{aligned}$$

W ostatnich dwóch liniijkach skorzystaliśmy z faktu że $p + q = 1$. Możemy teraz skorzystać z udowodnionego przed chwilą twierdzenia i otrzymać:

$$\mathcal{E}((k - \mathcal{E}(k))^2) = np - np^2 = np(1 - p) = npq.$$



Histogramy prawdopodobieństw rozkładu dwumianowego dla $n = 62$ i różnych wartości p .

3 Rozkład Poissona

Zbadajmy graniczną postać rozkładu dwumianowego, gdy liczba prób $n \rightarrow \infty, p \rightarrow 0$ i jednocześnie wartość oczekiwana liczby sukcesów pozostaje stała, $np = \lambda$. Podstawmy wartość $p = \lambda/n$ do wzoru na $P(k; n, p)$:

$$\begin{aligned} P(k; n, p) &= \frac{n!}{k! \cdot (n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot 1 \cdot \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

Przechodząc do granicy z $n \rightarrow \infty$ skorzystaliśmy z faktu, że iloczyn skończonej liczby (k jest znane i skończone) wyrazów mających skończone granice dąży do iloczynu tych granic oraz tego, że, gdy $n \rightarrow \infty$:

$$\left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

Otrzymaliśmy rozkład prawdopodobieństwa dyskretnej zmiennej losowej k zwany **rozkładem Poissona**:

$$P(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Podobnie jak w przypadku rozkładu dwumianowego sprawdzamy, że podany wyżej rozkład prawdopodobieństwa liczby sukcesów „sumuje się” do 1 oraz obliczymy wartości oczekiwane $\mathcal{E}(k)$, $\mathcal{E}(k^2)$ oraz $\mathcal{E}((k - \mathcal{E}(k))^2)$:

$$\sum_{k=0}^{\infty} P(k; \lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Skorzystaliśmy z rozwinięcia funkcji e^{λ} w szereg potęgowy. Przekonaliśmy się, że tak, jak być powinno suma prawdopodobieństw jest równa 1.

Obliczmy **wartość oczekiwaną liczby sukcesów**, k , podlegającej rozkładowi Poissona:

$$\mathcal{E}(k) = \sum_{k=0}^{\infty} k P(k; \lambda) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{l=0}^{\infty} \frac{\lambda^l}{l!} e^{-\lambda} = \lambda.$$

Obliczmy także wartość oczekiwaną k^2 :

$$\mathcal{E}(k^2) = \sum_{k=0}^{\infty} k^2 P(k; \lambda) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \sum_{l=0}^{\infty} (l+1) \frac{\lambda^{l+1}}{l!} e^{-\lambda}.$$

Ostatnią z równości otrzymaliśmy zmieniając wskaźnik sumowania na $l = k - 1$. Zauważamy, że otrzymaliśmy sumę poprzedniego wyniku mnożonego przez λ oraz λ :

$$\sum_{l=0}^{\infty} (l+1) \frac{\lambda^{l+1}}{l!} e^{-\lambda} = \lambda \sum_{l=0}^{\infty} l \frac{\lambda^l}{l!} e^{-\lambda} + \lambda = \lambda^2 + \lambda.$$

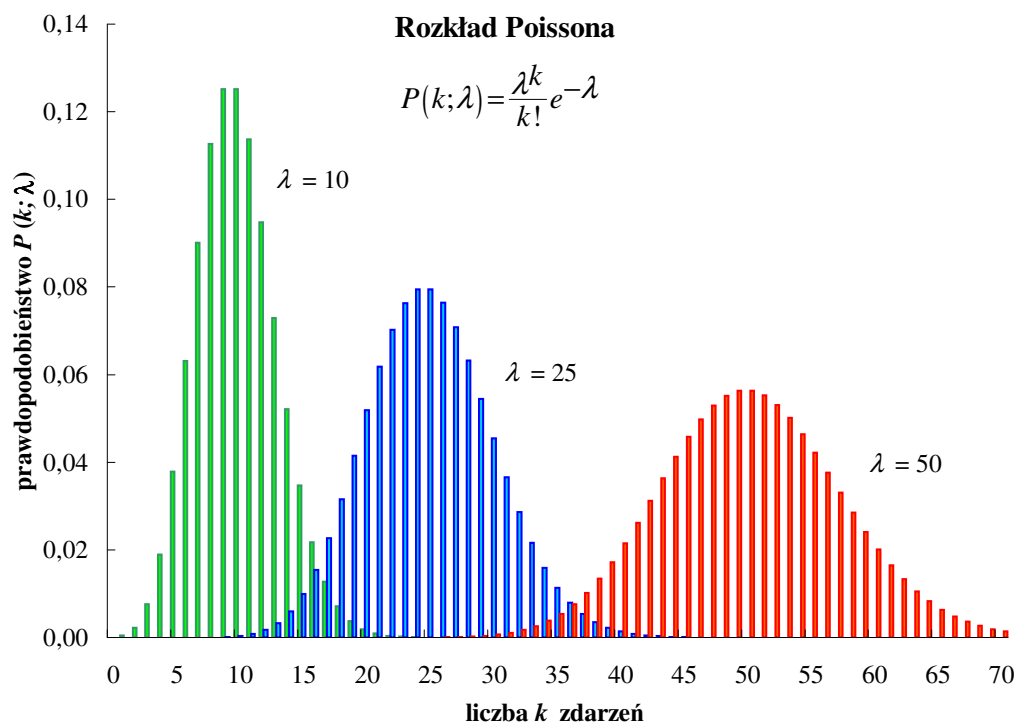
Korzystając z poprzednio idowodnionego twierdzenia otrzymujemy wartość **wariancji dyskretnej zmiennej losowej podlegającej rozkładowi Poissona**:

$$Var(k) = \mathcal{E}((k - \mathcal{E}(k))^2) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Warto zapamiętać, że **wartość oczekiwana i wariancja zmiennej losowej podlegającej rozkładowi Poissona są sobie równe**.

Rozkład Poissona jest przydatny do analizowania zjawisk, w których prawdopodobieństwo „sukcesu” jest małe, a wartość oczekiwana liczby sukcesów jest znana. Na przykład liczba goli strzelonych przez jednego zawodnika w jednym meczu. Jeśli na podstawie wyników wielu meczów ustalimy, że zawodnik ten strzela średnio $\lambda = 1,2$ gola w jednym meczu, to rozkład Poissona pozwoli nam obliczyć prawdopodobieństwa strzelenia przez niego $k = 0, 1, 2, \dots$ bramek w kolejnym meczu (takie analizy są rzeczywiście wykonywane przez zawodowych komentatorów sportowych). Podobną analizę możemy przeprowadzić w odniesieniu do liczby samochodów przejeżdżających w ciągu godziny mało uczęszczaną uliczką.

W fizyce rozkład Poissona używany jest między innymi do analizowania procesów rozpadu promieniotwórczego.



Histogramy prawdopodobieństw rozkładu Poissona dla różnych wartości λ .

4 Rozkład Gaussa – Graniczna postać rozkładu dwumianowego ($n \rightarrow \infty$).

Obliczenie prawdopodobieństw zadanych rozkładem dwumianowym dla dużej liczby prób n jest kłopotliwe ze względu na konieczność obliczania silni dużych liczb. Zbadajmy jak zachowuje się $P(k; p, n)$ dla dużych n . Posłużymy się przybliżeniem silni za pomocą wzoru Stirlinga:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} = n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}.$$

do wyrażenia przybliżonej wartości logarytmu naturalnego $P(k; n, p)$. Dla uproszczenia notacji zapisujemy dalej $\ln(P(k; n, p))$ jako $\ln(P)$:

$$\begin{aligned} \ln(P) &= \ln\left(\frac{n!}{k!(n-k)!} p^k q^{n-k}\right) \\ &\approx \left(n + \frac{1}{2}\right) \ln(n) - \left(k + \frac{1}{2}\right) \ln(k) - \left(n - k + \frac{1}{2}\right) \ln(n - k) + k \ln(p) - (n - k) \ln(q) - \frac{1}{2} \ln(2\pi) \\ &= -\left(k + \frac{1}{2}\right) \ln\left(\frac{k}{np}\right) - \left(n - k + \frac{1}{2}\right) \ln\left(\frac{n - k}{nq}\right) - \frac{1}{2} \ln(2\pi npq). \end{aligned}$$

Znak przybliżonej równości związany jest z podstawieniem przybliżenia Stirlinga zamiast $n!, k!, (n - k)!$. Dalej dokonywaliśmy jedynie przegrupowania wyrazów i korzystaliśmy z właściwości funkcji logarytm. Otrzymany wzór nie jest jeszcze dość wygodny do badania zachowania $P(k; n, p)$ dla dużych n . Przyjrzyjmy się histogramom $P(k; n, p)$. Dosyć łatwo zauważyć, że maksimum osiągnięte jest dla $k \approx np$, a wartości istotnie różne od zera skupiają się w pobliżu tego maksimum, w obszarze wartości: $np - \sqrt{npq} \leq k \leq np + \sqrt{npq}$. Zdefiniujmy więc nową zmienną:

$$x := \frac{k - np}{\sqrt{npq}},$$

Mamy wówczas:

$$k = np + x\sqrt{npq}; \quad n - k = nq - x\sqrt{npq}$$

Podstawiamy oba wyrażenia do otrzymanego wyżej wzoru na przybliżenie $\ln(P)$. Dostajemy:

$$\ln(P) \approx -\frac{1}{2} \ln(2\pi npq) - (np + x\sqrt{npq} + \frac{1}{2}) \ln\left(1 + x\sqrt{\frac{q}{np}}\right) - (nq - x\sqrt{npq} + \frac{1}{2}) \ln\left(1 - x\sqrt{\frac{p}{nq}}\right).$$

Do znalezienia przydatnego rozwinięcia $\ln(P)$ potrzebne będzie nam jeszcze rozwinięcie logarytmu naturalnego. Zaczniemy od przedstawienia logarytmu za pomocą całki:

$$\ln(1 + z) = \int_0^z \frac{dt}{1 + t}.$$

Interesuje nas rozwinięcie dla małych wartości z . Przyjmijmy $|z| < 1$, wówczas także $|t| < 1$ w wyrażeniu podcałkowym i możemy to wyrażenie przedstawić jako sumę szeregu geometrycznego:

$$\ln(1 + z) = \int_0^z dz \left(\sum_{l=0}^{\infty} (-t)^l \right) = \sum_{l=0}^{\infty} (-1)^l \frac{z^{l+1}}{l+1} = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots$$

Dla $\ln(1 - z)$, gdy $|z| < 1$ otrzymujemy rozwinięcie (wystarczy w powyższym wzorze podstawić $-z$ w miejsce z):

$$\ln(1 - z) = -z - \frac{z^2}{2} - \frac{z^3}{3} - \frac{z^4}{4} - \dots$$

Po podstawieniu do przybliżenia $\ln(P)$, wykonaniu mnożeń i uporządkowaniu wyrazów według potęg x i n otrzymujemy przybliżenie:

$$\ln(P) \approx -\frac{1}{2} \ln(2\pi npq) - \frac{x^2}{2} + \frac{x(p-q)}{2\sqrt{npq}} - \frac{x^2(p^2+q^2)}{npq} + \frac{x^3(q^2-p^2)}{6\sqrt{npq}} + \dots$$

Dalsze wyrazy, pominięte w powyższym rozwinięciu, mają w liczniku x w potęgach 4 lub większej, a w mianowniku npq w potęgach $3/2$ lub większej. Ze wzrostem n pozostają dwa dominujące wyrazy:

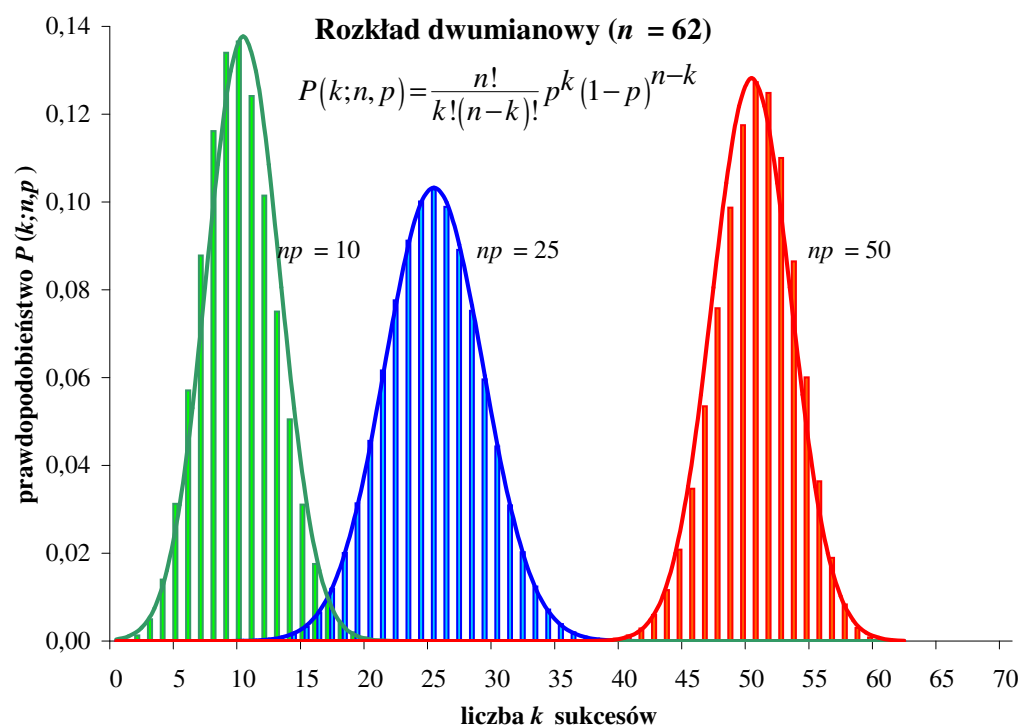
$$\ln(P) \approx -\frac{1}{2} \ln(2\pi npq) - \frac{x^2}{2}.$$

Ostatecznie, dla dużych n otrzymaliśmy przybliżenie:

$$P(k; n, p) \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{x^2}{2}\right) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k - np)^2}{2npq}\right).$$

Rysunek na następnej stronie pokazuje, jak dobrze, otrzymane przybliżenie (linie ciągłe) „pasuje” do histogramów $P(k; n, p)$ dla $n = 62$ i różnych wartości p . Niebieska linia ciągła, dla $p = 25/62$, przebiega niemal dokładnie przez wierzchołki słupków niebieskiego histogramu. Dla $p = 10/62$ – histogram i linia zielona i dla $p = 50/62$ – histogram i linia czerwona, zgodność jest nieco gorsza, ale ciągle niezła. Zielony histogram przewyższa odpowiednią ciągłą krzywą na lewo od swego maksimum, a czerwony na prawo – zgodnie z „wpływem” pierwszego odrzuconego wyrazu, liniowego w x , w przybliżeniu $\ln(P)$.

Otrzymany wzór przybliżony ułatwia posługiwanie się rozkładem dwumianowym, gdy liczba prób jest duża, a prawdopodobieństwo sukcesu jest bliskie $p = 1/2$. W dalszej części wykładu – §11 i następne *Slajdów/transparencji do wykładu* – otrzymany tu wzór posłuży do sformułowania modelu rozkładu prawdopodobieństwa dla wartości przyjmowanych przez błędy przypadkowe.



Porównanie dokładnych wartości prawdopodobieństw rozkładu dwumianowego (histogramy) z obliczonymi za pomocą wyprowadzonego wyżej wzoru przybliżonego (linie ciągłe).