

Elementarna statystyka

Alexander Bendikov

28 marca 2017

Rozkłady dwumianowe

Definicja

Założmy, że wykonujemy n niezależnych eksperymentów, których rezultatem może być albo sukces z prawdopodobieństwem p albo porażka, z prawdopodobieństwem q , $p + q = 1$. Niech X będzie łączną ilością sukcesów w n eksperymentach. X nazywa się dwumianową zmienną losową.

Niech $X_k = 1$ jeżeli w k -tym eksperymencie był sukces, i $X_k = 0$ jeżeli była porażka, $k = 1, 2, \dots$. Wtedy,

- $\{X_k\}$ są niezależnymi zmiennymi losowymi Bernoulliego o jednakowym rozkładzie

X_i	0	1
P	q	p

- $X = X_1 + \dots + X_n$,
- $E(X) = \sum_k E(X_k) = np$,
- $Var(X) = \sum_k Var(X_k) = npq$,
- $Var(X) = npq$.

Przykłady: (a) Odnotowujemy płeć każdego z kolejnych 10 dzieci urodzonych w pewnym szpitalu. X oznacza ilość dziewczynek. X jest zmienną dwumianową z $n = 10$, $p = 0,51$, $q = 0,49$.

$$E(X) = 10p = 5,1, \quad Var(X) = 10pq = 5,1 \cdot 0,49 = 2,499.$$

(b) Małżeństwo postanawia starać się o kolejne dzieci, dopóki nie urodzi się pierwsza dziewczynka (a potem skutecznie przestać się starać). X oznacza ilość dzieci, jaką w rezultacie małżeństwo ma po zakończeniu eksperymentu. X nie jest zmienną dwumianową. Przyjmując, że prawdopodobieństwo urodzenia się chłopca lub dziewczynki jest takie samo równe $0,5$, mamy

X	1	2	3	...
P	$\frac{1}{2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	...

Mamy więc

$$\begin{aligned} E(X) &= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2^2} + 3 \cdot \frac{1}{2^3} + \cdots = 2, \\ \text{Var}(X) &= E(X^2) - E(X)^2 = \\ &= 1 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{2^2} + 3^2 \cdot \frac{1}{2^3} + \cdots - 4 = 6 - 4 = 2. \end{aligned}$$

Sumy, które pojawiają się w powyższych wzorach przyjmujemy jako znane (są znane).

Rozkład zmiennej X

X	0	1	2	...	n-1	n
P	q^n	npq^{n-1}	$\frac{n(n-1)}{2}p^2q^{n-1}$	...	nqp^{n-1}	p^n

- $P(X = 0) = P(X_1 = 0, X_2 = 0, \dots, X_n = 0) = q^n$,
- $P(X = 1) = P(X_1 = 1, X_2 = 0, \dots, X_n = 0) +$
 $P(X_1 = 0, X_2 = 1, \dots, X_n = 0) + \dots = npq^{n-1}$,
-
- $P(X = k) = \binom{n}{k} p^k q^{n-k}$,

gdzie

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1) \cdots (n-k+1)}{1 \cdot 2 \cdots k}.$$

Przykład: Zapomniałeś swój kod PIN ****. Wybierasz losowo 4 cyfry. Jakie jest prawdopodobieństwo, że przynajmniej 2 z nich są prawidłowe?

Rozwiązanie: X - # prawidłowych cyfr, $p = 0,1$, $q = 0,9$, $n = 4$.
Mamy

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) = \\ &= 1 - 0,9^4 - 4 \cdot 0,1 \cdot 0,9^3 = 0,3070928 \sim 0,31. \end{aligned}$$

Symetria $P(X = k)$

Łatwo zauważyć, że

- $P(X = k) \nearrow$ dla $k \leq np - q$,
- $P(X = k) \searrow$ dla $k \geq np - q$.

Dla $p = q = \frac{1}{2}$ wykres jest symetryczny wokół $k = \frac{n-1}{2}$.

Zagadnienie punktów (Pascal & Fermat, 1654)

Przeprowadzane są niezależne próby, każda z prawdopodobieństwem sukcesu p i porażki $1 - p$.

Pytanie: Jakie jest prawdopodobieństwo, że n sukcesów wystąpi zanim wystąpi m porażek? Sytuację można zilustrować następującym przykładem. Gracz A dostaje 1 zł w przypadku sukcesu, natomiast w przypadku porażki drugi gracz, B , dostaje 1 zł. Jakie jest prawdopodobieństwo, że A zarobi n zł zanim gracz B zarobi m zł.

Rozwiązanie: Rozważmy ciąg prób:

$$\omega = (+ - - - + + - + - \dots -).$$

Jesteśmy zainteresowani takimi ciągami ω w których n plusów wystąpi przed m minusami. Jest jasne, że „dobre” ciągi to te, w których na $n + m - 1$ pierwszych pozycjach występuje co najmniej n plusów.

Skoro prawdopodobieństwo dokładnie k ($\geq n$) plusów wynosi

$$\binom{n+m-1}{k} p^k (1-p)^{n+m-1-k},$$

więc interesujące nas prawdopodobieństwo wynosi:

$$P_{n,m} = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} p^k (1-p)^{n+m-1-k}.$$

Przybliżenie Poissona

Przykład: Przypuśćmy, że prawdopodobieństwo, że jakaś część produkowana przez pewną maszynę będzie wadliwa wynosi 0,1. Jakie jest prawdopodobieństwo, że losowa próbka kontrolna 10 części będzie zawierała nie więcej niż jedną wadliwą część?

Szukane prawdopodobieństwo p że $X = 0$ lub $X = 1$ wynosi:

$$p = p(0) + p(1) = \binom{10}{0} (0,1)^0 (0,9)^{10} + \binom{10}{1} (0,1)^1 (0,9)^9 \\ \approx 0,7361.$$

Rachunek nie jest prosty numerycznie, bo $0,9^{10}$ lub 9 jest bliskie 1.

Wzór Poissona Jeżeli liczność próbki $n \gg 1$ a prawdopodobieństwo wady $p \ll 1$, ale jednocześnie $\lambda = np$ jest umiarkowane, to

$$p(k) \approx \frac{\lambda^k}{k!} e^{-\lambda} \quad (e \approx 2,716).$$

W przykładzie powyżej $n = 10$, $p = 0,1$ $\lambda = 1$ i mamy

$$p = p(0) + p(1) \approx e^{-1} + e^{-1} \approx 0,7358.$$

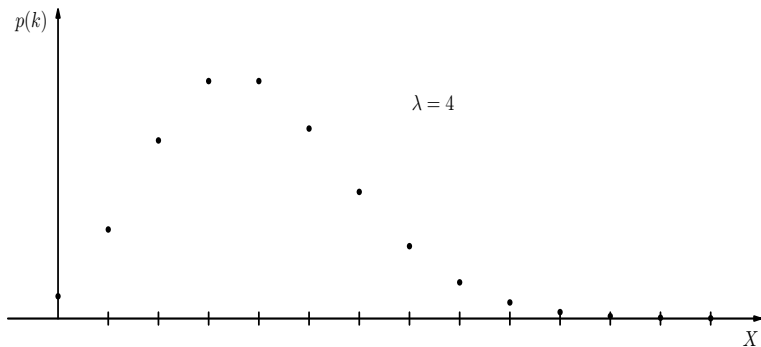
Rozkład Poissona

Definicja

Zmienna losowa X przyjmująca wartości $0, 1, 2, \dots$ nazywa się zmienną losową Poissona z parametrem λ jeżeli

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \dots$$

Skoro $p(k+1)/p(k) = \frac{\lambda}{k+1}$, to prawdopodobieństwa $p(k)$ początkowo rosną, a od $k = \lambda - 1$ maleją.



Rysunek: Rozkład zmiennej Poissona

- $EX = \lambda,$
- $VarX = EX^2 - (EX)^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$

Przykłady zmiennych losowych Poissona:

Zmienne losowe Poissona są dobrym modelem dla wielu praktycznych zdarzeń losowych. Na przykład zmiennymi losowymi Poissona są:

- Liczba błędów typograficznych na stronie pewnej książki
- Liczba ludzi w pewnym społeczeństwie w wieku powyżej 100 lat
- Liczba błędnie wybranych numerów telefonów w pewnym dniu w pewnym centrum obsługi telefonicznej
- Liczba klientów wchodzących do budynku poczty pewnego danego dnia
- Liczba nieobsadzonych miejsc w Trybunale Konstytucyjnym w danym roku
- Liczba cząsteczek α wyemitowanych w ustalonym okresie czasu przez pewien materiał radioaktywny

Prawo wielkich liczb (Bernoulli 1713)

Przypomnijmy, że zmienną losową Bernoulliego nazywamy zmienną, która przyjmuje tylko 2 wartości: 1 (tak zwany sukces) z prawdopodobieństwem p i 0 (tak zwana porażka) z prawdopodobieństwem $1 - p$.

Twierdzenie

Niech $\{X_k\}$ będą niezależnymi zmiennymi losowymi Bernoulliego, z tym samym parametrem p . Obliczamy średnią

$$\bar{x} = \frac{1}{n} (X_1 + \cdots + X_n),$$

i wtedy

$$\bar{x} \longrightarrow p \quad \text{według prawdopodobieństwa,}$$

to znaczy dla dowolnego $\epsilon > 0$ $P(|\bar{x} - p| > \epsilon) \rightarrow 0$ gdy $n \rightarrow \infty$.

Dowód.

Wiemy, że $X_1 + \dots + X_n$ jest zmienną dwumianową, a więc wiemy, że $E(\bar{x}) = \frac{1}{n}np = p$, oraz $Var(\bar{x}) = \frac{1}{n^2}npq = \frac{pq}{n}$. Następnie korzystamy z tak zwanej nierówności Czebyszewa:

$$P(|X - E(X)| > \epsilon) \leq \frac{1}{\epsilon^2} Var(X), \quad (\text{NC})$$

prawdziwej dla dowolnej zmiennej X . Widzimy, korzystając z (NC), że

$$P(|\bar{x} - p| > \epsilon) \leq \frac{pq}{n\epsilon^2} \longrightarrow 0 \text{ gdy } n \rightarrow \infty.$$



Względna częstość i prawdopodobieństwo

Przypomnijmy sposób, w jaki wprowadziliśmy pojęcie prawdopodobieństwa. Ustalmy eksperyment losowy Ω i zdarzenie $A \subset \Omega$. Przeprowadzamy n niezależnych eksperymentów Ω_k , i niech dla $k \leq n$

$$X_k = \begin{cases} 1 & \text{jeżeli } A \text{ zaszło w } k\text{-tej próbie,} \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

Zauważmy, że

- $\{X_k\}$ są zmiennymi losowymi Bernoulliego z $p = P(A)$,
- Względna częstość występowania zdarzenia A wynosi

$$W(A) = \frac{1}{n} (X_1 + X_2 + \cdots + X_n).$$

Korzystając z prawa wielkich liczb Bernoulliego widzimy, że

$$W(A) \longrightarrow P(A), \quad \text{gdy } n \rightarrow \infty \quad (\text{według prawdopodobieństwa}),$$

a więc nasza intuicyjna definicja prawdopodobieństwa zdarzenia, jako „granicy” względnych częstości danego zdarzenia przy wielu niezależnych powtórzeniach tego samego eksperymentu jest prawidłowa. Ogólnie, gdy X_k jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, to prawdziwe jest tak zwane „mocne prawo wielkich liczb” („Mocne LLN”), które mówi, że

$$W(A) \longrightarrow P(A) \quad \text{prawie na pewno,}$$

gdzie „prawie na pewno” znaczy:

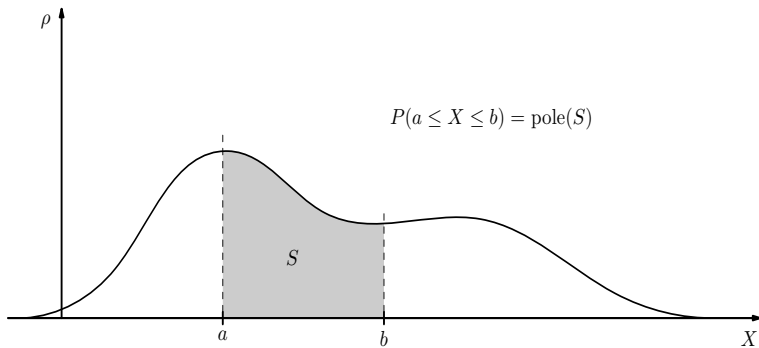
$$P(\omega : W(A)(\omega) \not\rightarrow P(A)) = 0.$$

Ciągłe zmienne losowe

Przypomnijmy, że zmienną losową nazwaliśmy przyporządkowanie liczb wynikom eksperymentu, czyli funkcję na Ω , która przyjmuje wartości rzeczywiste.

Definicja

Zmienna losowa X nazywa się ciągłą, jeżeli istnieje funkcja ρ taka, że prawdopodobieństwo, że zmienna X przyjmie wartości pomiędzy liczbami a, b , czyli $P(a \leq X \leq b)$ jest równe polu pod wykresem funkcji ρ nad przedziałem $[a, b]$. Funkcja ρ nazywa się gęstością prawdopodobieństwa, a jej wykres krzywą gęstości.



Rysunek: Krzywa gęstości prawdopodobieństwa

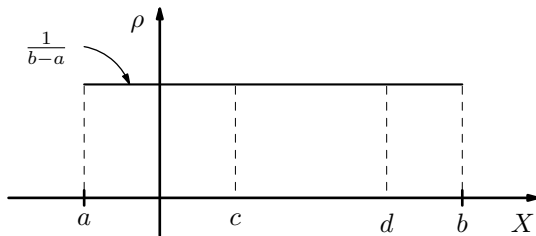
Przykłady: (a) Rozkład jednostajny na $[a, b]$.

$$P(X \in [c, d]) = \rho \cdot (d - c),$$

$$\rho(x) = \rho = \frac{1}{b - a} \quad \text{dla } x \in [a, b],$$

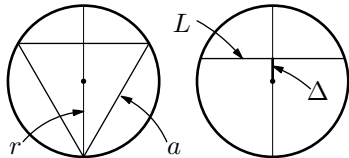
$$\rho(x) = 0 \quad \text{dla } x \notin [a, b],$$

$$P(X \in [c, d]) = \frac{d - c}{b - a}.$$



(b) L. F. Bertrand (1889). Wyobraźmy sobie „losową cięciwę” wewnątrz ustalonego okręgu. Jakie jest prawdopodobieństwo, że taka losowa cięciwa będzie dłuższa niż bok trójkąta równobocznego wpisanego w ten sam okrąg?

Rozwiązanie: Długość wylosowanej cięciwy możemy ustalić posługując się odległością Δ cięciwy od środka okręgu.



Odległość ta może wynieść od 0 do r (r – promień okręgu). Załóżmy, że Δ ma rozkład jednostajny na $[0, r]$. Nie jest trudno sprawdzić, używając argumentów geometrycznych, że długość cięciwy L jest większa niż

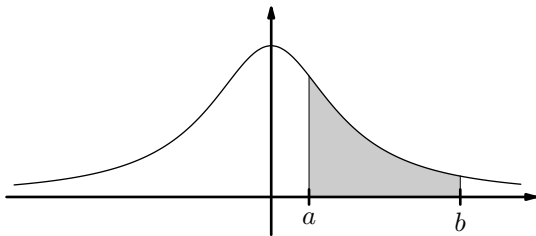
długość boku trójkąta równobocznego wpisanego (długość tego boku to $a = \sqrt{3}r$) dokładnie wtedy, gdy $\Delta \leq \frac{r}{2}$. Otrzymujemy

$$P(L \geq a) = P\left(\Delta \leq \frac{r}{2}\right) = \frac{r/2}{r} = \frac{1}{2}.$$

(c) Rozkład Cauchy'ego.

$$p(x) = \frac{1}{\pi} \frac{1}{1+x^2},$$

$$P(a \leq X \leq b) = \frac{1}{\pi} \arctan x \Big|_a^b.$$



Przykład: rozkład normalny (Abraham de Moivre 1773)

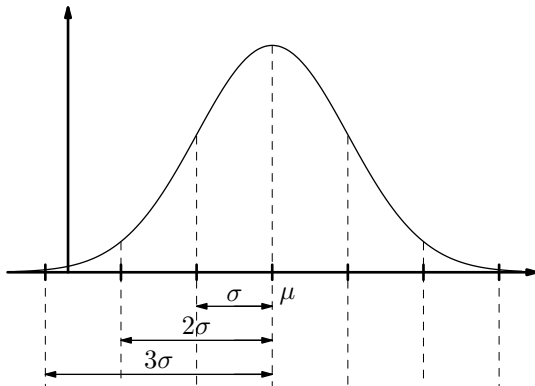
Definicja

Zmienna losowa X nazywa się zmienną normalną, $X \in N(\mu, \sigma)$, jeżeli jej gęstość prawdopodobieństwa jest dana przez funkcję:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

Zasada 3 σ :

- $P(|X - \mu| \geq \sigma) \approx 0,32,$
- $P(|X - \mu| \geq 2\sigma) \approx 0,05,$
- $P(|X - \mu| \geq 3\sigma) \approx 0,003.$



Jak obliczyć prawdopodobieństwo dla zmiennej normalnej:

- 1 Unormowanie: $Z = \frac{X - \mu}{\sigma} \in N(0, 1)$,
- 2 $P(a < X < b) = P(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma})$,
- 3 Dystrybuanta: $\Phi(x) = P(Z \leq x)$,
- 4 $P(a < X < b) = \Phi(\frac{b - \mu}{\sigma}) - \Phi(\frac{a - \mu}{\sigma})$.

Przykład: Przyjmijmy, że iloraz inteligencji (IQ) Amerykanów jest zmienną normalną, o rozkładzie $N(100, 25)$. Jaki procent Amerykanów ma iloraz $85 < IQ < 120$? A jaki $IQ > 120$?
Rozwiązanie: Możemy użyć Tabeli A w książce D. Moore'a (albo Google'a):

- 1 $P(85 < IQ < 120) = P\left(\frac{85-100}{25} < Z < \frac{120-100}{25}\right) =$
 $= P(0,6 < Z < 0,8) = \Phi(0,8) - \Phi(0,6) \approx$
 $\approx 0,7881 - 0,2743 = 0,5138.$
- 2 $P(IQ > 120) = P\left(Z > \frac{120-100}{25}\right) = P(Z > 0,8) =$
 $= 1 - P(Z \leq 0,8) \approx 1 - 0,7881 = 0,2119$

Ciągłe zmienne losowe: wartość oczekiwana i wariancja

Definicja

Niech X będzie zmienna losowa z zakresem wartości $[a, b]$. Weźmy dowolny podział przedziału $[a, b]$ punktami

$a = x_1 < x_2 < \dots < x_n = b$ i rozważmy dyskretną zmienną losową $\tilde{X}$:

$$\tilde{X} = x_i, \quad \text{jeżeli } x_i \leq X < x_{i+1}, \quad i = 1, 2, \dots, n-1.$$

Wartość oczekiwaną $E(X)$ definiujemy jako granicę $E(\tilde{X})$ gdy rozmiar rozkładu $\max(x_{i+1} - x_i) \rightarrow 0$. Tak jak poprzednio definiujemy też:

- $\text{Var}(X) = E(X - E(X))^2$,
- Odchylenie standardowe: $SD(X) = \sqrt{\text{Var}(X)}$.

Fakt

Jeżeli X jest ciągłą zmienną losową z gęstością $p(x)$, to

- $E(X) = \int_{-\infty}^{\infty} xp(x) dx$,
- $Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 p(x) dx$.

Przykłady:

- 1 Rozkład jednostajny na $[a, b]$: $E(X) = \frac{1}{2}(a + b)$,
 $Var(X) = \frac{1}{12}(b - a)^2$,
- 2 Rozkład Cauchy'ego: $E(X) = Var(X) = +\infty$,
- 3 Rozkład normalny $N(\mu, \sigma)$: $E(X) = \mu$, $Var(X) = \sigma^2$.

$E(X)$ oraz $Var(X)$ mają takie same własności jak w przypadku dyskretnym, czyli:

- $E(X + Y) = EX + EY$, $E(cX) = cEX$,
- $E(XY) = EX \cdot EY$, jeżeli X i Y są niezależne,
- $Var(X + Y) = VarX + VarY$, jeżeli X i Y są niezależne,
- $Var(cX) = c^2 VarX$, $Var(c) = 0$.

Normalne przybliżenia rozkładu dwumianowego

Twierdzenie graniczne De Moivre'a-Laplace'a (1733, $p = \frac{1}{2}$, 1812 $0 < p < 1$)

Niech X oznacza ilość sukcesów w n niezależnych próbach. Wtedy, dla dowolnych $a < b$

$$P\left(a \leq \frac{X - np}{\sqrt{npq}} \leq b\right) \rightarrow P(a \leq Z \leq b) \quad \text{gdy } n \rightarrow \infty.$$

Uwaga: Mamy dwa niezależne przybliżenia rozkładu dwumianowego:

- 1 Przybliżenie rozkładem Poissona, które jest dobre dla $n \gg 1$, $\lambda = np$ średnie,
- 2 Przybliżenie rozkładem normalnym, które jest dobre dla $n \gg 1$, oraz $npq \geq 10$ (czyli $\lambda \geq \frac{10}{q} \gg 10$).

Przykład: $p = 0,0002$, $q = 0,9998$, $n = 5000$. Żeby obliczyć $P(X = 3)$ stosujemy:

- ① Przybliżenie rozkładem Poissona: $\lambda = n \cdot p = 1$ (!),

$$P(X = 3) \approx \frac{1}{3!} e^{-1} = \frac{1}{6e} \approx 0,006,$$

- ② przybliżenie rozkładem normalnym: (?),

$$\begin{aligned} P(X = 3) &= P(2,5 < X < 3,5) \approx P(1,5 < Z < 2,5) = \\ &= \Phi(2,5) - \Phi(1,5) \approx 0,9938 - 0,9332 = 0,0606, \quad !!! \end{aligned}$$

$$\begin{aligned} P(X = 3) &= P(2,9 < X < 3,1) \approx P(1,9 < Z < 2,1) = \\ &= \Phi(2,1) - \Phi(1,9) \approx 0,9821 - 0,9731 = 0,0108, \quad ?! \end{aligned}$$

Uwaga: Przybliżenia rozkładem normalnym używamy gdy $n \cdot p \cdot q \geq 10$!

Przykład: $p = q = 0,5$, $n = 100$. Mamy więc $n \cdot p \cdot q = 25 \gg 10$!

$$\begin{aligned} P(40 \leq X \leq 60) &= P\left(\frac{40 - 50}{5} \leq Z \leq \frac{60 - 50}{5}\right) = \\ &= P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) \approx 0,9544, \\ P(X < 40 \text{ lub } X > 60) &\approx 0,0456. \end{aligned}$$

Średnia próbki (zastosowanie tw. DeMoivre'a-Laplace'a)

Przykład: (Równouprawnienie kobiet) Czy wysiłki w celu zapewnienia kobietom równego statusu w USA osiągnęły swój cel? Przepytano losową próbkę 1019 osób: 550 z nich (ok. 54%) powiedziało "Tak" (czyli cel został osiągnięty).

Na podstawie danych z tej losowej próbki, co możemy powiedzieć o odsetku wszystkich obywateli, którzy odpowiedzieliby "Tak"?

Innymi słowy, chcemy oszacować proporcje w populacji. Interesuje nas nieznaną odsetek p populacji, dla której mamy, powiedzmy, "sukces". Parametr statystyczny, który służy do oszacowania p nazywa się "średnią próbki" $\hat{p}$:

$$\hat{p} = \frac{\text{ilość sukcesów w próbce}}{\text{ilość wszystkich elementów próbki}}.$$

Uwagi: Skoro próbka była pobrana losowo, to wynik $\hat{p}$ jest zmienną losową. Mamy następujące własności:

- 1 $E(\hat{p}) = p$, $Var(\hat{p}) = \frac{p \cdot q}{n}$, gdzie $q = 1 - p$,
- 2 $\hat{p} \rightarrow p$ (według prawdopodobieństwa), gdy $n \rightarrow \infty$,
- 3 Rozkład zmiennej $\hat{p}$ można przybliżyć przez rozkład normalny $N\left(p, \sqrt{\frac{pq}{n}}\right)$:

$$P\left(a < \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} < b\right) \approx P(a < Z < b).$$

Model matematyczny (prawie) Rozkład dwumianowy:

- Eksperyment: wybór losowy, niech $X = 0$ lub 1 (sukces lub porażka) będzie wynikiem eksperymentu. Oczywiście, X jest zmienną losową Bernoulliego, z parametrami p i $q = 1 - p$.
- Powtarzamy eksperyment n razy, i otrzymujemy wyniki $X_1, X_2, \dots, X_n$.
- Ponieważ po k -tym losowaniu nie zwracamy wylosowanego osobnika do populacji, to każde kolejne losowanie ma inny parametr p (!)
- Od tej pory będziemy zakładali, że liczebność populacji $N \gg 1$ oraz rozmiar próbki $n \gg 1$ przy czym $n \ll N$.

Przy tych założeniach,

$$\hat{p} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

gdzie $\{X_i\}$ są jednakowo rozłożonymi, niezależnymi zmiennymi losowymi Bernoulliego, z parametrami p i $q = 1 - p$. W takim razie, możemy zastosować prawo wielkich liczb Bernoulliego i twierdzenie graniczne DeMoivre'a-Laplace'a do uzasadnienia Uwag 1-3.

Przykład: Wybranej losowo próbce $n = 1500$ studentów pierwszego roku zadamy pytanie, czy starali się również o przyjęcie na inną uczelnię. Wiadomo, że 35% wszystkich studentów pierwszego roku w rzeczywistości starało się o przyjęcie na jakąś inną uczelnię oprócz tej, na której studiują. Jakie jest prawdopodobieństwo, że nasza losowa próbka da wynik różny od prawdziwego dla całej populacji o mniej niż 2 punkty procentowe?

Rozwiązanie: Jest jasne, że spełnione są nasze założenia, gdyż ilość wszystkich studentów pierwszego roku na wszystkich uczelniach (w USA) jest większa niż 100 000.

- Próbkę losową populacji, w której prawdziwa proporcja wynosi $p = 0,35$ ma licznosc $n = 1500$.
- $E(\hat{p}) = p = 0,35$,
- $Var(\hat{p}) = \frac{p(1-p)}{n} = \frac{0,35 \cdot 0,65}{1500} \approx 0,0001517 \approx 0,0123^2$.
- Interesuje nas prawdopodobieństwo

$$\begin{aligned} P(0,35 - 0,02 < \hat{p} < 0,35 + 0,02) &= P(0,33 < \hat{p} < 0,37) = \\ &= P\left(\frac{0,33 - 0,35}{0,0123} < \frac{\hat{p} - 0,35}{0,0123} < \frac{0,37 - 0,35}{0,0123}\right) \approx \\ &\approx P(-1,63 < Z < 1,63) \approx 0,9484 - 0,0516 = 0,8968. \end{aligned}$$

Prawo Wielkich Liczb i Centralne Twierdzenie Graniczne

Niech $\{X_i\}$ będą niezależnymi, jednakowo rozłożonymi zmiennymi losowymi, z $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Niech

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

będzie średnią arytmetyczną. Mamy:

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}, \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Twierdzenie (Prawo Wielkich Liczb, Chińczyn, Kołmogorow)

$$\bar{X} \longrightarrow \mu \quad \text{wg prawd. gdy } n \rightarrow \infty,$$

czyli

$$\forall \epsilon > 0 \quad P(|\bar{X} - \mu| > \epsilon) \longrightarrow 0 \quad \text{gdy } n \rightarrow \infty.$$

Twierdzenie (Centralne Twierdzenie Graniczne, Liapunow)

Dla dowolnych $A, B \in \mathbb{R}$ mamy

$$P\left(A < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < B\right) \longrightarrow P(A < Z < B),$$

gdy $n \rightarrow \infty$, jednostajnie względem A, B .

Uwaga: Powołując się na powyższe twierdzenie, dla a, b bliskich μ i $n \gg 1$ mamy

$$\begin{aligned} P(a < \bar{X} < b) &\approx P\left(\frac{a - \mu}{\sigma/\sqrt{n}} < Z < \frac{b - \mu}{\sigma/\sqrt{n}}\right) = \\ &= P\left(\sqrt{n} \frac{a - \mu}{\sigma} < Z < \sqrt{n} \frac{b - \mu}{\sigma}\right). \end{aligned}$$

Przykłady: (a) Ilość wypadków X na pewnym skrzyżowaniu w ciągu tygodnia jest zmienną z $E(X) = 2,2$ i $Var(X) = 1,4$. Niech $\bar{X}$ będzie średnią tygodniową ilością wypadków w ciągu roku (52 tygodnie), czyli

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{52}}{52}.$$

- 1 Jaki jest (przybliżony) rozkład $\bar{X}$?
- 2 Znajdź przybliżoną wartość $P(\bar{X} < 2)$.
- 3 Znajdź przybliżone prawdopodobieństwo, że wydarzy się najwyżej 100 wypadków w ciągu roku.

Rozwiązanie:

- 1 X przyjmuje wartości całkowite, w takim razie nie ma rozkładu normalnego. To samo odnosi się do $\bar{X}$. Z CTG rozkład $\bar{X}$ jest w przybliżeniu $N(\mu, \sigma)$, dla $\mu = 2,2$ i $\sigma = \frac{1,4}{\sqrt{52}}$.
- 2 Korzystając z powyższego, mamy

$$P(\bar{X} < 2) = P\left(\frac{\bar{X} - 2,2}{1,4/\sqrt{52}} < \frac{2 - 2,2}{1,4/\sqrt{52}}\right) \approx P(Z < -1,03) \approx 0,1515$$

- 3 Argumentując podobnie,

$$\begin{aligned} P(\# \leq 100) &= P\left(\bar{X} \leq \frac{100}{52}\right) \approx P\left(Z \leq \frac{1,923 - 2,2}{1,4/\sqrt{52}}\right) \approx \\ &\approx P(Z \leq -1,4264) \approx 0,0708. \end{aligned}$$

(b) Ilość NOX (mieszanka różnych tlenków azotu) wyemitowanego do atmosfery w spalinach pewnego modelu samochodu jest zmienną losową ze średnią $1,4\text{g/mi}$ (gramów na przejechaną milę) i odchyleniem standardowym $0,3\text{g/mi}$. Firma ma 125 takich samochodów w swojej flocie. Niech $\bar{X}$ oznacza średnią emisję NOX przez samochód floty.

- 1 Jaki jest przybliżony rozkład $\bar{X}$?
- 2 Oszacuj prawdopodobieństwo, że $1,4 < \bar{X} < 2,0\text{g/mi}$.
- 3 Jaki poziom emisji L ma tę własność, że $P(\bar{X} > L) = 0,05$

Rozwiązanie: $E(\bar{X}) = 1,4$, $SD(\bar{X}) = 0,3/\sqrt{125} \approx 0,0268$.

① Korzystając z CTG, $\bar{X} \sim N(1,4, 0,0268)$,

② $P(1,4 < \bar{X} < 2,0) \approx P\left(0 < Z < \frac{2,0-1,4}{0,0268}\right) = P(0 < Z < 22) \approx 0,5$,

③

$$\begin{aligned} 0,05 &= P(\bar{X} > L) = P\left(\frac{\bar{X} - 1,4}{0,0268} > \frac{L - 1,4}{0,0268}\right) \approx \\ &\approx P\left(Z > \frac{L - 1,4}{0,0268}\right) = 1 - \Phi\left(\frac{L - 1,4}{0,0268}\right). \end{aligned}$$

Korzystając z tabel, mamy

$$\frac{L - 1,4}{0,0268} = 1,645 \Rightarrow L = 1,444.$$

(c) Ilość studentów, która wybiera kurs Statystyki Elementarnej jest zmienną losową o rozkładzie Poissona ze średnią $\lambda = 100$ (taki był plan). Prowadzący zajęcia postanowił, że jeżeli ilość zapisanych studentów przekroczy 120 powstaną 4 grupy ćwiczeniowe, a jeżeli nie przekroczy, to będą 3 grupy. Jakie jest prawdopodobieństwo, że powstaną 4 grupy?

Rozwiązanie: Dokładna wartość tego prawdopodobieństwa to

$$e^{-100} \sum_{i \geq 120} \frac{100^i}{i!},$$

która nie jest łatwa do oszacowania. Ale zauważmy, że zmienna Poissona ze średnią $\lambda = 100$ jest sumą 100 jednakowo rozłożonych niezależnych zmiennych Poissona ze średnią 1. Możemy więc użyć CTG.

$$\begin{aligned} P(X \geq 120) &= P\left(\frac{X - 100}{\sqrt{100}} \geq \frac{120 - 100}{\sqrt{100}}\right) \approx \\ &\approx 1 - \Phi(2) = 0,0228. \end{aligned}$$