Wstęp do rachunku prawdopodobieństwa
Wykład 12: Prawa wielkich liczb. Centralne twierdzenie graniczne

Na tym wykładzie zajmiemy się szacowaniem przybliżonych wartości prawdopodobieństwa zdarzeń losowych odwołując się do rozkładów granicznych pewnych zmiennych losowych.

Twierdzenie 1 (**Prawo Wielkich Liczb**). *Niech $X_1, X_2, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie prawdopodobieństwa, dla których $\mathbb{E}X_1 = \mu$ oraz $VarX_1 = \sigma^2 < \infty$. Wtedy dla zmiennej losowej $S_n = X_1 + \ldots + X_n$ i dowolnej stałej $\varepsilon > 0$ zachodzi*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

Zauważmy, że zmienna losowa S_n z powyższego twierdzenia ma wartość oczekiwaną równą μn, a zatem

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \mu.$$

Powyższe twierdzenie mówi nam, że w przypadku ciągu niezależnych zmiennych losowych o tym samym rozkładzie (i skończonej wariancji), średnia wartość z tych zmiennych losowych jest zmienną losową o rozkładzie „zmierzającym" do rozkładu jednopunktowego z atomem μ. Innymi słowy, dla dużych n zmienna losowa S_n/n jest dobrze skoncentrowana wokół swojej wartości oczekiwanej μ.

Przypomnijmy, że zmienna losowa X ma **rozkład normalny** z parametrami μ i σ^2, co zapisujemy jako $X \sim \mathcal{N}(\mu, \sigma^2)$, jeśli jej gęstość dana jest wzorem

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{dla} \quad x \in (-\infty, \infty).$$

W tym wypadku $\mathbb{E}X = \mu$ oraz $\mathrm{Var}X = \sigma^2$.

Jeżeli $X \sim \mathcal{N}(0,1)$, to mówimy, że X ma **standardowy rozkład normalny**. Dystrybuantę tego rozkładu będziemy oznaczać przez Φ. Funkcja Φ ma bardzo skomplikowaną postać, dlatego jej wartości będziemy odczytywać z tablic. Najczęściej interesować nas będą poniższe wartości dystrybuanty:

$$\Phi(1{,}28) \approx 0{,}9; \;\; \Phi(1{,}64) \approx 0{,}95; \;\; \Phi(1{,}96) \approx 0{,}975; \;\; \Phi(2{,}33) = 0{,}99; \;\; \Phi(2{,}58) \approx 0{,}995.$$

Tabela przedstawiająca więcej wartości dystrybuanty Φ znajduje się na końcu wykładu. Kalkulator rozkładu normalnego można znaleźć klikając w następujący link: **kalkulator**

Uwaga 1. Gęstość standardowego rozkładu normalnego jest funkcją parzystą. Oznacza to, że jej wykres jest symetryczny względem osi OY. Co więcej krzywa opisująca gęstość rozkładu normalnego przyjmuje kształt dzwonu, stąd też nazywa się ją czasem krzywą dzwonową lub krzywą Gaussa (od nazwy rozkładu, a rozkład normalny – rozkładem Gaussa). W związku z powyższym dystrybuanta tego rozkładu spełnia zależność

$$\Phi(-x) = 1 - \Phi(x)$$

dla wszystkich $x \in (-\infty, \infty)$. Dzięki temu przekształceniu, pomimo że w tablicach mamy podane wartości dystrybuanty tylko dla nieujemnych argumentów, możemy też odczytać wartości dla argumentów ujemnych.

Twierdzenie 2 (**Centralne Twierdzenie Graniczne**). *Niech $X_1, X_2, \ldots, X_n$ będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie, dla których $\mathbb{E}X_1 = \mu$, $VarX_1 = \sigma^2 < \infty$ i $\mathbb{E}(|X_1|^3) < \infty$. Wtedy dla zmiennej losowej $S_n = X_1 + \ldots + X_n$ i dowolnej stałej $x \in \mathbb{R}$ zachodzi*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - \mu n}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x).$$

Zauważmy, że wartość oczekiwana oraz wariancja zmiennej losowej S_n z powyższego twierdzenia dane są wzorami:

$$\mathbb{E}S_n = \mathbb{E}(X_1 + X_2 + \ldots + X_n) = \mu n,$$

$$\mathrm{Var}S_n = \mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \sigma^2 n.$$

Na poprzednim wykładzie pokazaliśmy, że zmienna losowa

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}S_n}} = \frac{S_n - \mu n}{\sigma\sqrt{n}}$$

ma wartość oczekiwaną równą 0 oraz wariancję równą 1. Przypomnijmy, że przekształcenie polegające na odjęciu od zmiennej losowej X jej wartości oczekiwanej i podzielenie wyniku przez pierwiastek z jej wariancji nazywamy **standaryzacją** zmiennej losowej X. W konsekwencji nowo otrzymana zmienna losowa ma zawsze wartość oczekiwaną równą 0 i wariancję równą 1. Centralne Twierdzenie Graniczne mówi nam zatem, że po dokonaniu standaryzacji zmiennej losowej S_n (spełniającej odpowiednie założenia), rozkład tak otrzymanej nowej zmiennej losowej „zmierza" do standardowego rozkładu normalnego $\mathcal{N}(0,1)$. W związku z tym możemy dobrze przybliżać ten rozkład właśnie za pomocą standardowego rozkładu normalnego $\mathcal{N}(0,1)$.

Uwaga 2. Centralne Twierdzenie Graniczne (w skrócie CTG) odnosi sie do sytuacji, gdy n i $\mathbb{E}S_n$ są dostatecznie duże. Zwykle CTG dobrze przybliża rozkład zmiennej losowej S_n gdy $n \geqslant 30$ i $\mathbb{E}S_n \geqslant 10$. W tej sytuacji nie będziemy przejmować się błędem przybliżenia.

Uwaga 3. Błąd przybliżenia w CTG szacuje się z góry w ogólnym przypadku przez

$$\left(\frac{2\max_i |X_i|}{\sigma}\right)^3 \frac{1}{\sqrt{n}},$$

a w przypadku zmiennej losowej $S_n \sim Bin(n,p)$ przez

$$\frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}}.$$

Uwaga 4. Jeśli stosujemy CTG do przybliżania rozkładu prawdopodobieństwa zmiennej losowej S_n, która przyjmuje tylko wartości całkowite, należy pamiętać o uwzględnieniu „poprawki" wynoszącej 1/2. Na przykład, chcąc użyć CTG do oszacowania prawdopodobieństwa $\mathbb{P}(S_n = k)$, należy je zastąpić przez prawdopodobieństwo

$$\mathbb{P}(k - 1/2 < S_n \leqslant k + 1/2).$$

W przypadku zmiennej losowej $S_n = X_1 + X_2 + \ldots + X_n$ o rozkładzie dwumianowym, CTG przyjmuje poniższą postać.

Wniosek 3 (**Twierdzenie de Moivre'a–Laplace'a).** Jeżeli zmienna losowa S_n ma rozkład dwumianowy $S_n \sim Bin(n,p)$, to dla dowolnych $-\infty < a < b < \infty$ zachodzi

$$\mathbb{P}\left(a \leqslant \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}S_n}} \leqslant b\right) = \mathbb{P}\left(a \leqslant \frac{S_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \to \Phi(b) - \Phi(a).$$

Przykład 1. Niech X będzie zmienną losową o rozkładzie dwumianowym $Bin(2000, 1/4)$. A zatem $\mathbb{E}X = 500$ oraz $\mathrm{Var}X = 375$. Spróbujmy oszacować prawdopodobieństwo zdarzenia $475 < X < 525$. Możemy zrobić to stosując nierówność Czebyszewa-Bienaymé:

$$\mathbb{P}(475 < X < 525) \geqslant \mathbb{P}(|X - \mathbb{E}X| < 25) = 1 - \mathbb{P}(|X - \mathbb{E}X| \geqslant 25) \geqslant 1 - \frac{375}{625} = 0{,}4.$$

Oszacujmy teraz to prawdopodobieństwo korzystając z Centralnego Twierdzenia Granicznego:

$$\mathbb{P}\left(475 < X < 525\right) = \mathbb{P}\left(\frac{475-500}{\sqrt{375}} < \frac{X-\mathbb{E}X}{\sqrt{\mathrm{Var}X}} < \frac{525-500}{\sqrt{375}}\right) = \mathbb{P}\left(\frac{-5}{\sqrt{15}} < \frac{X-\mathbb{E}X}{\sqrt{\mathrm{Var}X}} < \frac{5}{\sqrt{15}}\right)$$
$$\approx \Phi\left(\frac{5}{\sqrt{15}}\right) - \Phi\left(-\frac{5}{\sqrt{15}}\right) = 2\cdot\Phi\left(\frac{5}{\sqrt{15}}\right) - 1 \approx 2\cdot\Phi(1{,}29) - 1 \approx 2\cdot 0{,}9 - 1 = 0{,}8.$$

A zatem widzimy, że oszacowanie z nierówności Czebyszewa-Bienaymé daje mniej więcej 50–procentowy błąd. Co więcej, tutaj skorzystaliśmy z tego, że szacowaliśmy prawdopodobieństwo przedziału o środku w $\mathbb{E}X$. Gdybyśmy natomiast chcieli oszacować prawdopodobieństwo zdarzenia $500 \leqslant X < 525$, nierówność Czebyszewa-Bienaymé za bardzo by się nam nie przydała, bo musielibyśmy oszacować prawdopodobieństwo tego zdarzenia z góry przez prawdopodobieństwo zdarzenia $475 < X < 525$, a z kolei korzystając z nierówności Czebyszewa-Bienaymé to prawdopodobieństwo moglibyśmy oszacować jedynie z dołu. Natomiast możemy jak najbardziej zastosować CTG:

$$\mathbb{P}\left(500 \leqslant X < 525\right) = \mathbb{P}\left(0 \leqslant \frac{X-\mathbb{E}X}{\sqrt{\mathrm{Var}X}} < \frac{525-500}{\sqrt{375}}\right) = \mathbb{P}\left(0 \leqslant \frac{X-\mathbb{E}X}{\sqrt{\mathrm{Var}X}} < \frac{5}{\sqrt{15}}\right)$$
$$\approx \Phi\left(\frac{5}{\sqrt{15}}\right) - \Phi(0) \approx \Phi(1{,}29) - \Phi(0) \approx 0{,}9 - 0{,}5 = 0{,}4.$$

Przykład 2. Oszacujmy prawdopodobieństwo, że zmienna losowa Z o rozkładzie dwumianowym $Bin(100, 1/2)$ przyjmuje wartość 60. Ponieważ zmienna ta przyjmuje tylko wartości naturalne, możemy oszacować szukane prawdopodobieństwo przez

$$\mathbb{P}\left(59{,}5 < Z \leqslant 60{,}5\right).$$

Stosując Centralne Twierdzenie Graniczne mamy:

$$\mathbb{P}\left(59{,}5 < Z \leqslant 60{,}5\right) = \mathbb{P}\left(\frac{59{,}5-50}{\sqrt{25}} < \frac{Z-\mathbb{E}Z}{\sqrt{VarZ}} \leqslant \frac{60{,}5-50}{\sqrt{25}}\right) = \mathbb{P}\left(\frac{9{,}5}{5} < \frac{Z-\mathbb{E}Z}{\sqrt{VarZ}} \leqslant \frac{10{,}5}{5}\right)$$
$$\approx \Phi(2{,}1) - \Phi(1{,}9) \approx 0{,}98214 - 0{,}97128 = 0{,}01086.$$

Dokładna wartość tego prawdopodobieństwa wynosi:

$$\mathbb{P}\left(Z = 60\right) = \binom{100}{60}\left(\frac{1}{2}\right)^{100} \approx 0{,}01084$$

a zatem widzimy, że oszacowanie z CTG daje nam bardzo dobry wynik.

Przykład 3. Profesor wie z doświadczenia, że wynik studenta na egzaminie jest zmienną losową X o $\mathbb{E}X = 75$ i $\mathrm{Var}X = 50$. Zakładamy idealistycznie, że wyniki studentów są niezależne. Oszacujmy na początek, z jakim prawdopodobieństwem średnia z wyników 50 studentów zmieści się w przedziale $(72, 78)$.

W tym celu wprowadźmy pomocnicze zmienne losowe $X_1, X_2, \ldots, X_{50}$ oznaczające wyniki poszczególnych studentów oraz zmienną losową $S_{50} = X_1 + X_2 + \ldots + X_{50}$. Naszym celem jest oszacować prawdopodobieństwo zdarzenia $S_{50}/50 \in (72, 78)$. Zauważmy, że $\mathbb{E}S_{50} = \mathbb{E}(X_1 + X_2 + \ldots + X_{50}) = 3750$ oraz, ponieważ zmienne losowe X_i są niezależne, $\mathrm{Var}S_{50} = \mathrm{Var}(X_1 + X_2 + \ldots + X_{50}) = 2500$. Posłużymy się Centralnym Twierdzeniem Granicznym:

$$\mathbb{P}\left(72 < \frac{S_{50}}{50} < 78\right) = \mathbb{P}\left(3600 < S_{50} < 3900\right) = \mathbb{P}\left(\frac{3600-3750}{\sqrt{2500}} < \frac{S_{50}-\mathbb{E}S_{50}}{\sqrt{\mathrm{Var}S_{50}}} < \frac{3900-3750}{\sqrt{2500}}\right)$$
$$= \mathbb{P}\left(-3 < \frac{S_{50}-\mathbb{E}S_{50}}{\sqrt{\mathrm{Var}S_{50}}} < 3\right) \approx \Phi(3) - \Phi(-3) = 2\cdot\Phi(3) - 1 \approx 0{,}9973.$$

Ilu studentów potrzeba, aby z prawdopodobieństwem co najmniej 0,99 średni wynik różnił się od 75 o mniej niż 1?

Proszę zwrócić uwagę, że im większa liczba studentów n, tym bardziej średni wynik będzie skoncentrowany wokół wartości oczekiwanej $\mathbb{E}X = 75$. Oszacujemy szukaną liczbę studentów n postępując podobnie jak powyżej. W tym celu znów wprowadzamy pomocnicze zmienne losowe $X_1, X_2, \ldots, X_n$ będące wynikami poszczególnych studentów, oraz zmienną losową $S_n = X_1 + X_2 + \ldots + X_n$. W szczególności $\mathbb{E}S_n = 75n$ i $\mathrm{Var}S_n = 50n$. Chcemy tak dobrać parametr n aby $\mathbb{P}\left(S_n/n \in (74, 76)\right) > 0{,}99$. A zatem oszacujmy to prawdopodobieństwo, korzystając z CTG:

$$\mathbb{P}\left(74 < \frac{S_n}{n} < 76\right) = \mathbb{P}\left(74n < S_n < 76n\right) = \mathbb{P}\left(\frac{74n - 75n}{\sqrt{50n}} < \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}S_n}} < \frac{76n - 75n}{\sqrt{50n}}\right)$$
$$= \mathbb{P}\left(-\frac{\sqrt{n}}{5\sqrt{2}} < \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}S_n}} < \frac{\sqrt{n}}{5\sqrt{2}}\right) \approx \Phi\left(\frac{\sqrt{n}}{5\sqrt{2}}\right) - \Phi\left(-\frac{\sqrt{n}}{5\sqrt{2}}\right) = 2 \cdot \Phi\left(\frac{\sqrt{n}}{5\sqrt{2}}\right) - 1.$$

Teraz wystarczy tak dobrać n, aby

$$2 \cdot \Phi\left(\frac{\sqrt{n}}{5\sqrt{2}}\right) - 1 > 0{,}99.$$

Równoważnie

$$\Phi\left(\frac{\sqrt{n}}{5\sqrt{2}}\right) > 0{,}995.$$

Następnie odczytujemy z tablic rozkładu normalnego dla jakiej wartości x mamy $\Phi(x) > 0{,}995$. Tą wartością jest $x = 2{,}58$, a ponieważ dystrybuanta jest funkcją niemalejącą, wystarczy wziąć n spełniające nierówność

$$\frac{\sqrt{n}}{5\sqrt{2}} \geqslant 2{,}58,$$

skąd otrzymujemy

$$n \geqslant (2{,}58 \cdot 5\sqrt{2})^2 \approx 332{,}82.$$

Ostatecznie wystarczy 333 studentów.

TABELA 1. Dystrybuanta $\Phi(x)$ standardowego rozkładu normalnego $\mathcal{N}(0,1)$

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,5279	0,53188	0,53586
0,1	0,53983	0,5438	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,6293	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,6591	0,66276	0,6664	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,7054	0,70884	0,71226	0,71566	0,71904	0,7224
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,7549
0,7	0,75804	0,76115	0,76424	0,7673	0,77035	0,77337	0,77637	0,77935	0,7823	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,8665	0,86864	0,87076	0,87286	0,87493	0,87698	0,879	0,881	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,9032	0,9049	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,9222	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,9452	0,9463	0,94738	0,94845	0,9495	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,9608	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,9732	0,97381	0,97441	0,975	0,97558	0,97615	0,9767
2	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,9803	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,983	0,98341	0,98382	0,98422	0,98461	0,985	0,98537	0,98574
2,2	0,9861	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,9884	0,9887	0,98899
2,3	0,98928	0,98956	0,98983	0,9901	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,9918	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,9943	0,99446	0,99461	0,99477	0,99492	0,99506	0,9952
2,6	0,99534	0,99547	0,9956	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,9972	0,99728	0,99736
2,8	0,99744	0,99752	0,9976	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,999
3,1	0,99903	0,99906	0,9991	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,9994	0,99942	0,99944	0,99946	0,99948	0,9995
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,9996	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,9997	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,9998	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,9999	0,9999	0,9999	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998