STANDARYZACJA

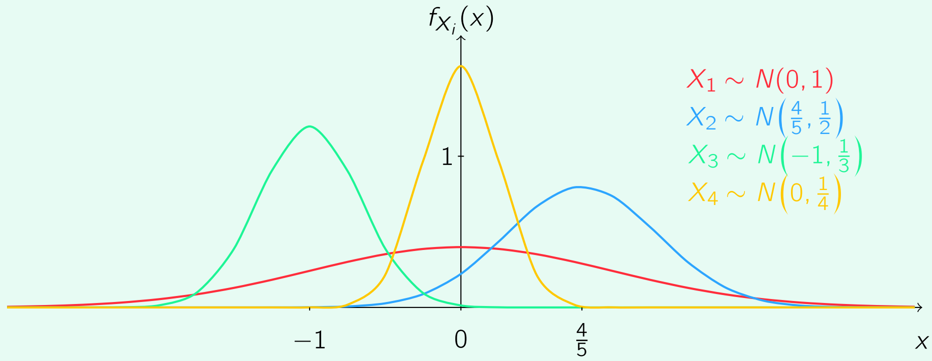
NIEZBĘDNIK

Nim przejdziemy do opisu procedury standaryzacji i jej stosowania przy wyznaczaniu wartości dystrybuanty zmiennej losowej o rozkładzie normalnym spójrzmy raz jeszcze na informacje dotyczące zmiennej losowej z rozkładu normalnego, które zebrano poniżej.

GĘSTOŚĆ

$\mu, \sigma > 0$ | średnia, odchylenie standardowe

$$X \sim N(\mu, \sigma) \iff f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



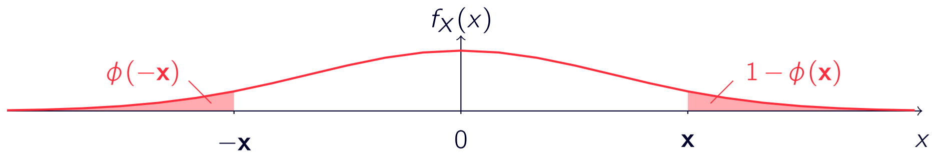
MOMENTY

$$\mathbb{E}X = \mu \qquad \mathrm{Var}\, X = \sigma^2 \qquad \mathbb{E}X^2 = \sigma^2 + \mu^2$$

Aby wyznaczyć wartość dystrybuanty zmiennej losowej X o rozkładzie normalnym standardowym w zadanym punkcie ($x \in \mathbb{R}$), korzystamy z tablic $\phi(x)$, czyli tablic dystrybuanty rozkładu normalnego standardowego. Aby zaoszczędzić miejsca i nie wypisywać wartości dystrybuanty w punktach, które różnią się jedynie znakiem, w tablicach znajdziemy przeważnie wartości tej dystrybuanty w punktach, które są nieujemne. Mamy bowiem, wynikający z symetrii (parzystości) funkcji gęstości rozkładu normalnego standardowego, związek, dzięki któremu wystarczą nam wartości dystrybuanty w punktach nieujemnych, gdyż te w punktach ujemnych można wyznaczyć z prostego związku

$$\phi(-x) = 1 - \phi(x) \quad \text{dla } x \in \mathbb{R},$$

który najłatwiej chyba skojarzyć z poniższym rysunkiem (na rysunku przedstawiony został wykres $f_X(x)$ - gęstości rozkładu normalnego standardowego, która to, przypomnijmy raz jeszcze, jest funkcją parzystą).



Ale jak postępować, w przypadku, gdy mamy zmienną losową Y o rozkładzie normalnym, ale niebędącym normalnym standardowym, czyli z dowolnymi parametrami μ (wartość oczekiwana, $\mu = \mathbb{E}Y$) oraz σ (odchylenie standardowe, $\sigma = \sqrt{\mathrm{Var}\, Y}$). Otóż w takim przypadku wykorzystać będziemy musieli tak zwaną procedurę standaryzacji zmiennej losowej Y. Procedura ta polega na:

- scentrowaniu zmiennej Y, robimy to poprzez odjęcie od Y wartości oczekiwanej Y, czyli $\mu = \mathbb{E}Y$,
- przeskalowaniu otrzymanej w poprzednim punkcie zmiennej, które polega na podzieleniu tej zmiennej przez odchylenie standardowe zmiennej Y, czyli $\sigma = \sqrt{\mathrm{Var}\, Y}$.

W wyniku przeprowadzonych operacji ze zmiennej losowej Y tworzymy nową zmienną losową, oznaczmy ją literką Z, która to ma rozkład normalny standardowy. Poniżej przedstawiono, jak należy przeprowadzić standaryzację zmiennej losowej Y o rozkładzie normalnym.

$$Z = \frac{Y - \mathbb{E}Y}{\sqrt{\operatorname{Var} Y}} = \frac{Y-\mu}{\sigma},$$
$$Z \sim N(0,1).$$

Zauważmy, że jedyne co zrobiliśmy, to odjęliśmy i podzieliliśmy przez pewne stałe zmienną losową Y. Warto podkreślić, że operacje te nie zmieniają typu rozkładu (pozostaje normalny), jedyne co zmienią, to jego parametry. Przyjrzyjmy się parametrom zmiennej Z, otóż

$$\mathbb{E}Z = \mathbb{E}\left(\frac{Y-\mu}{\sigma}\right) = \frac{\mathbb{E}Y-\mu}{\sigma} = 0,$$
$$\operatorname{Var} Z = \operatorname{Var}\left(\frac{Y-\mu}{\sigma}\right) = \operatorname{Var}\left(\frac{Y}{\sigma}\right) = \frac{1}{\sigma^2}\operatorname{Var} Y = 1,$$

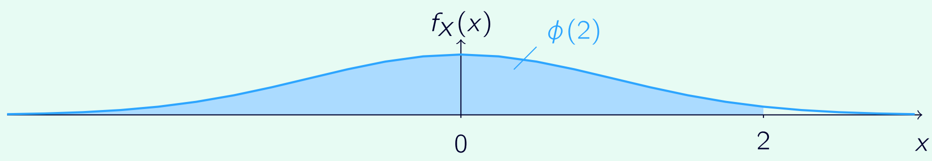
czyli dokładnie tak, jak tego oczekiwaliśmy.

PRZYKŁAD

ODCZYTYWANIE WARTOŚCI DYSTRYBUANTY ZMIENNEJ LOSOWEJ Z $N(0,1)$

Załóżmy, że $X \sim N(0,1)$. Znajdziemy wartość dystrybuanty tej zmiennej losowej w punkcie 2 oraz punkcie -2. Wyznaczamy

$$F_X(2) = P(X \leq 2) = \phi(2) \approx 0,977.$$

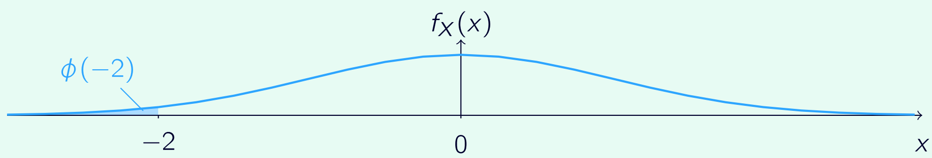


Nasza aktywność ograniczyła się do znalezienia w tablicach rozkładu normalnego standardowego wartości dystrybuanty w punkcie 2, gdyż X ma rozkład normalny standardowy i dlatego $F_X(x) = \phi(x)$. Treść zadania można byłoby sformułować również następująco: jakie jest prawdopodobieństwo tego, że zmienna losowa X przyjmie wartości mniejsze (nie większe) od 2. Z kolei wartość dystrybuanty w punkcie -2, czyli $\phi(-2)$, wyznaczamy wykorzystując zależność

$$\phi(-x) = 1 - \phi(x),$$

w rozważanym przypadku przyjmując $x = 2$. Otrzymujemy wtedy

$$F_X(-2) = P(X \leq -2) = \phi(-2) = 1 - \phi(2) \approx 0,023.$$



PRZYKŁAD

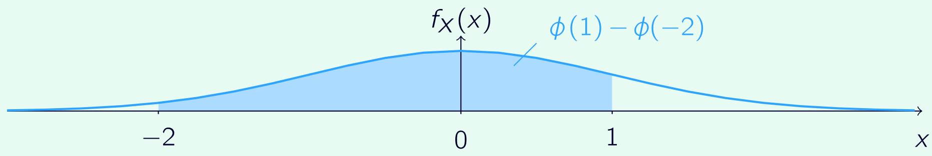
ODCZYTYWANIE WARTOŚCI DYSTRYBUANTY ZMIENNEJ LOSOWEJ Z $N(0,1)$

Załóżmy, że $X \sim N(0,1)$. Wyznaczymy prawdopodobieństwo zdarzenia $\{\omega : X(\omega) \in (-2,1)\}$. Mamy zatem wskazać wartość

$$\begin{aligned} P(\{\omega : X(\omega) \in (-2,1)\}) &= P(-2 < X < 1) = P(-2 < X \leq 1) = F_X(1) - F_X(-2) \\ &= \phi(1) - \phi(-2) = \phi(1) - (1 - \phi(2)). \end{aligned}$$

STANDARYZACJA

NIEZBĘDNIK



Odczytując z tablic wartości $\phi(2) \approx 0,977$ oraz $\phi(1) \approx 0,841$ otrzymujemy

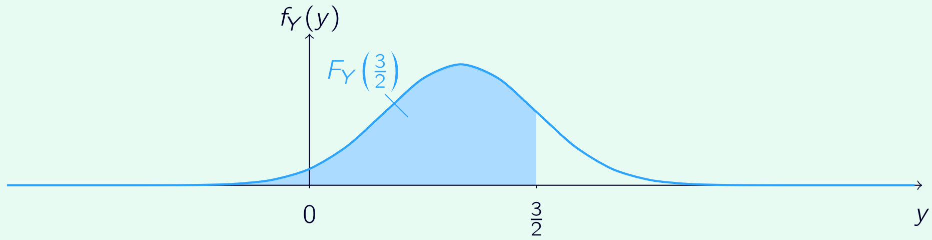
$$P(\{\omega : X(\omega) \in (-2,1)\}) = \phi(1) - (1-\phi(2)) \approx 0,818.$$

PRZYKŁAD

WYKORZYSTANIE STANDARYZACJI ZMIENNEJ Z ROZKŁADU $N(1,\frac{1}{2})$

Załóżmy, że $Y \sim N(1,\frac{1}{2})$. Znajdziemy wartość dystrybuanty tej zmiennej losowej w punkcie $\frac{3}{2}$. Interesuje nas zatem

$$F_Y\left(\frac{3}{2}\right) = P\left(Y \leq \frac{3}{2}\right).$$

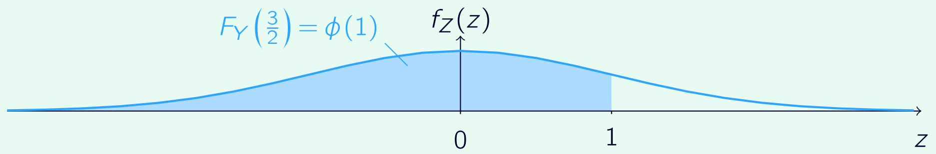


I tutaj pojawia się problem, gdyż tej dystrybuanty nie odczytamy bezpośrednio z tablic. Jeżeli jednak zmienną losową Y odpowiednio scentrujemy oraz przeskalujemy, innymi słowy poddamy standaryzacji, to dostaniemy zmienną losową, nazwijmy ją Z, z rozkładu normalnego standardowego i w celu wyznaczenia wartości interesującej nas dystrybuanty wykorzystamy dystrybuantę zmiennej losowej Z, której wartości są już stablicowane. Czyli szukaną dystrybuantę wyrazimy za pomocą dystrybuanty rozkładu normalnego standardowego i co należy wyraźnie podkreślić punkt, w którym dystrybuantę tę będziemy wyznaczać, ulegnie zmianie (w wyniku standaryzacji). Procedurę standaryzacji stosujemy dla zdarzenia (tu opisanego nierównością), którego prawdopodobieństwo liczymy. Musimy przy tym zadbać o to, by kolejne przejścia były równoważne.

$$F_Y\left(\frac{3}{2}\right) = P\left(Y \leq \frac{3}{2}\right) = P\left(Y-1 \leq \frac{3}{2}-1\right) = P\left(\frac{Y-1}{\frac{1}{2}} \leq \frac{\frac{3}{2}-1}{\frac{1}{2}}\right) = P\left(\underbrace{\frac{Y-1}{\frac{1}{2}}}_{Z\sim N(0,1)} \leq 1\right).$$

Od zmiennej losowej Y odjęliśmy jej wartość oczekiwaną, a następnie podzieliliśmy tak otrzymaną zmienną przez odchylenie standardowe zmiennej Y. W wyniku tych prostych działań otrzymaliśmy zmienną z rozkładu normalnego standardowego, co nas cieszy, gdyż dla zmiennych losowych z rozkładu normalnego standardowego wartości dystrybuanty w dowolnym punkcie jesteśmy w stanie odczytać wprost z tablic.

$$F_Y\left(\frac{3}{2}\right) = P(Z \leq 1) = \phi(1) \approx 0,841.$$



Teraz powinno być już jasne, po co używamy standaryzacji - po prostu by sprowadzić problem do odczytania wartości z tablic dystrybuanty rozkładu normalnego standardowego, bo tylko takie znajdziemy w tablicach statystycznych.

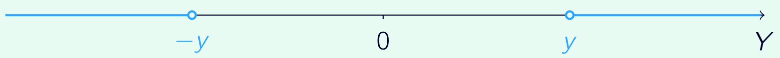
PRZYKŁAD

ZMIENNA Z ROZKŁADU NORMALNEGO $N(\mu,\sigma)$, PRAWOPODOBIEŃSTWO, MODUŁ

Niech $Y \sim N(\mu,\sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Obliczymy $P(|Y| > y)$, rozważając przypadek $y > 0$. Zajmijmy się wpierw nierównością

$$|Y| > y \quad \Longleftrightarrow \quad Y > y \quad \text{lub} \quad Y < -y,$$

jej rozwiązanie przedstawiono na poniższej osi liczbowej.

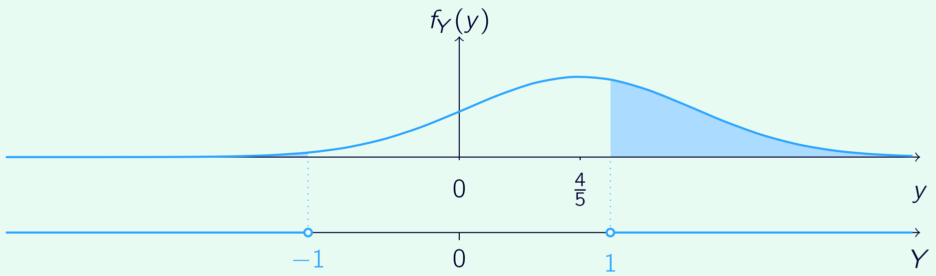


Zauważmy ponadto, że zdarzenie $\{\omega : Y(\omega) > y\}$ oraz $\{\omega : Y(\omega) < -y\}$ są zdarzeniami rozłącznymi (nie mogą zajść jednocześnie i to dzięki założeniu $y > 0$), toteż

$$\begin{aligned}
P(|Y| > y) &= P(\{Y > y\} \cup \{Y < -y\}) = P(Y > y) + P(Y < -y) \\
&= P\Bigg(\underbrace{\frac{Y-\mu}{\sigma}}_{Z\sim N(0,1)} > \frac{y-\mu}{\sigma}\Bigg) + P\Bigg(\underbrace{\frac{Y-\mu}{\sigma}}_{Z\sim N(0,1)} < \frac{-y-\mu}{\sigma}\Bigg) \\
&= 1 - P\left(\frac{Y-\mu}{\sigma} \le \frac{y-\mu}{\sigma}\right) + P\left(\frac{Y-\mu}{\sigma} \le \frac{-y-\mu}{\sigma}\right) \\
&= 1 - \phi\left(\frac{y-\mu}{\sigma}\right) + \phi\left(\frac{-y-\mu}{\sigma}\right).
\end{aligned}$$

Jeżeli przyjmiemy, że $Y \sim N\left(\frac{4}{5}, \frac{3}{4}\right)$ oraz weźmiemy $y = 1$, to wtedy

$$\begin{aligned}
P(|Y| > 1) &= 1 - \phi\left(\frac{1-\frac{4}{5}}{\frac{3}{4}}\right) + \phi\left(\frac{-1-\frac{4}{5}}{\frac{3}{4}}\right) = 1 - \phi\left(\frac{4}{15}\right) + \phi\left(-\frac{36}{15}\right) \\
&= 1 - \phi\left(\frac{4}{15}\right) + 1 - \phi\left(\frac{12}{5}\right) = 2 - \left(\phi\left(\frac{4}{15}\right) + \phi\left(\frac{12}{5}\right)\right) \\
&\approx 2 - (0,6051 + 0.9918) = 0,4031.
\end{aligned}$$



STANDARYZACJA

NIEZBĘDNIK

PRZYKŁAD

ZMIENNA O ROZKŁADZIE LOG-NORMALNYM $LN(\mu,\sigma)$ Załóżmy, że zmienna losowa Y ma rozkład normalny z parametrami μ oraz σ, czyli $Y \sim N(\mu,\sigma)$. Definiujemy zmienną losową $L = e^Y$; tak zdefiniowana zmienna losowa L ma rozkład log-normalny z parametrami μ oraz σ; piszemy $L \sim LN(\mu,\sigma)$. Co istotne zmienna losowa L przyjmuje jedynie wartości dodatnie, a μ oraz σ nie są jej średnią i odchyleniem standardowym, a jedynie parametrami rozkładu. Wyznaczymy prawdopodobieństwo zdarzenia: zmienna losowa L przyjmie wartości mniejsze od $\ell \in \mathbb{R}$, czyli $P(\{\omega : L(\omega) < \ell\})$. Skoro L przyjmuje jedynie wartości dodatnie, to dla każdego $\ell \leq 0$ mamy

$$P(\{\omega : L(\omega) < \ell\}) = 0.$$

Z kolei na potrzeby rozważenia przypadku $\ell > 0$ zauważmy, że $L(\omega) = e^{Y(\omega)}$, toteż

$$L(\omega) < \ell \quad \Longleftrightarrow \quad e^{Y(\omega)} < \ell \quad \Longleftrightarrow \quad Y(\omega) < \ln \ell.$$

Stąd dla $\ell > 0$ dostajemy, że

$$P(\{\omega : L(\omega) < \ell\}) = P(\{\omega : Y(\omega) < \ln \ell\}),$$

co, upraszczając zapis i stosując standaryzację zmiennej Y, daje

$$P(L < \ell) = P(Y \leq \ln \ell) = P\Bigg(\underbrace{\frac{Y-\mu}{\sigma}}_{Z\sim N(0,1)} \leq \frac{\ln \ell - \mu}{\sigma}\Bigg) = \phi\Bigg(\frac{\ln \ell - \mu}{\sigma}\Bigg).$$

Podsumowując,

$$P(L < \ell) = \begin{cases} 0, & \text{gdy } \ell \leq 0, \\ \phi\left(\frac{\ln \ell - \mu}{\sigma}\right), & \text{gdy } \ell > 0. \end{cases}$$

PRZYKŁAD

SUMA NIEZALEŻNYCH ZMIENNYCH LOSOWYCH Z ROZKŁADU NORMALNEGO $N(\mu,\sigma)$ **(O TYCH SAMYCH PARAMETRACH)** Zmienna losowa S_n będąca sumą n niezależnych (jest to bardzo ważne założenie) zmiennych losowych z rozkładu $N(\mu,\sigma)$ ma rozkład normalny $N(n\cdot\mu, \sqrt{n}\cdot\sigma)$. Poniżej przedstawiamy sposób obliczenia parametrów rozkładu S_n.

$$\mathbb{E}S_n = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}X_i = \sum_{i=1}^{n} \mathbb{E}X_1 = n\cdot\mathbb{E}X_1 = n\cdot\mu, \tag{1}$$

$$\mathrm{Var}S_n = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}X_i = \sum_{i=1}^{n} \mathrm{Var}X_1 = n\cdot\mathrm{Var}X_1 = n\cdot\sigma^2. \tag{2}$$

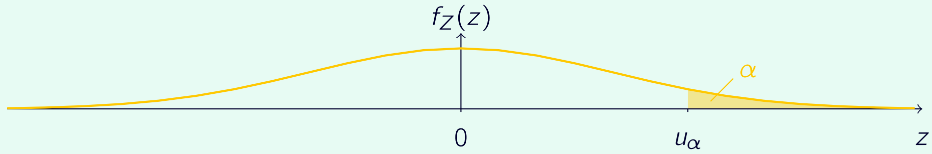
W (1) wykorzystaliśmy liniowość wartości oczekiwanej oraz to, że dla każdego $i \in \{1,\ldots,n\}$ wartość oczekiwana zmiennej X_i jest równa μ (w szczególności $\mathbb{E}X_1 = \mu$). Z kolei w (2) skorzystaliśmy ze wzoru na wariancję sumy zmiennych losowych niezależnych (bez założenia o niezależności nie byłoby tak łatwo) i z tego, że wariancje rozważanych zmiennych są takie same, równe σ^2.
Wyznaczmy x, dla którego dla ustalonego $\alpha \in (0,1)$ zachodzi równość $P(S_n < x) = 1-\alpha$. Również i w tym przypadku wykorzystywać będziemy standaryzację, przeprowadzać ją będziemy dla zmiennej S_n, której wartość oczekiwana to $n\cdot\mu$, a odchylenie standardowe równe jest $\sqrt{n}\cdot\sigma$.

Otrzymujemy

$$P(S_n < x) = P\Bigg(\underbrace{\frac{S_n - n\cdot\mu}{\sqrt{n}\cdot\sigma}}_{Z\sim N(0,1)} < \frac{x - n\cdot\mu}{\sqrt{n}\cdot\sigma}\Bigg) = \phi\left(\frac{x - n\cdot\mu}{\sqrt{n}\cdot\sigma}\right).$$

Toteż, by znaleźć x musimy rozwiązać równanie z niewiadomą x, mianowicie

$$\phi\left(\frac{x - n\cdot\mu}{\sqrt{n}\cdot\sigma}\right) = 1 - \alpha.$$



Jeżeli przez u_α oznaczymy kwantyl rzędu $1-\alpha$ (uwaga można spotkać inne oznaczenia na kwantyl rzędu $1-\alpha$, np. $z_{1-\alpha}$) rozkładu normalnego standardowego, to oznacza, że dla zadanego $\alpha \in (0,1)$ u_α jest takie, że spełniona jest równość $\phi(u_\alpha) = 1 - \alpha$ (ideę przedstawiono na rysunku powyżej, gdzie $Z \sim N(0,1)$, a $f_Z(z)$ jest gęstością Z), to rozwiązanie powyższego równania sprowadzi się do wyznaczenia x, przy czym

$$\frac{x - n\cdot\mu}{\sqrt{n}\cdot\sigma} = u_\alpha.$$

Gdybyśmy przyjęli: $\alpha = 5\%$ $(1-\alpha = 95\%)$, $n = 100$, $\mu = 3$, $\sigma = 2$, to x byłoby rozwiązaniem równania

$$\phi\left(\frac{x - 100\cdot 3}{\sqrt{100}\cdot 2}\right) = 0,95.$$

Z powyższej równości wynika, że punkt, w którym dystrybuantę liczymy ma być tym, w którym dystrybuanta równa jest $0,95$. Innymi słowy, punkt ten jest kwantylem rzędu $0,95$ rozkładu normalnego standardowego, oznaczany przez nas jako $u_{5\%}$, który w rozważanym przypadku jest równy 1,645 (w przybliżeniu). Stąd, by obliczyć x rozwiązujemy równanie

$$\frac{x - 300}{20} \approx 1,645,$$

prowadzące do wniosku, że $x \approx 332,9$.