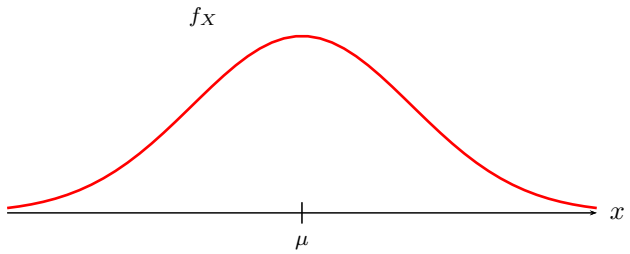
7. cvičení z PRA

1. - 5. dubna 2023

Poznámky k normálnímu rozdělení:
Veličina X má *normální* (neboli Gaussovo) rozdělení $N(\mu, \sigma^2)$ (kde $\mu \in \mathbb{R}$ a $\sigma > 0$), jestliže má hustotu

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{pro} \quad x \in \mathbb{R} \ .$$



Je to tedy spojité rozdělení, $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$ a oborem hodnot veličiny X je celá reálná osa. Všimněme si ještě, že hustota f_X je symetrická vzhledem ke středu μ a proto platí $F_X(\mu) = \frac{1}{2}$.

Toto rozdělení je limitním rozdělením, které aproximuje součty nezávislých stejně (nebo podobně) rozdělených veličin (více později v Centrální limitní větě). Typicky se tedy objevuje u veličin, jejichž hodnoty jsou ovlivněny mnoha drobnými odchylkami (např. u chyb měření, výšky člověka apod.)

U zmíněné výšky člověka (která může být samozřejmě jen kladná) nebo u veličin s hodnotami omezenými na nějaký interval, je přesto použití normálního rozdělení (které může nabývat libovolných hodnot) přiměřené. Je to tím, že u dané veličiny Y předpokládáme aproximaci pomocí normálního rozdělení obvykle jen ve vhodném okolí kolem střední hodnoty $\mu := E(Y)$. Je to podobná situace, jako když aproximujeme funkci pomocí jejího Taylorova polynomu v okolí daného bodu.

Přesněji to vystihuje toto tvrzení:

Věta: Nechť Y je veličina s hustotou f_Y, střední hodnotou μ a rozptylem $\sigma^2 \neq 0$. Nechť $X \sim N(\mu, \sigma^2)$. Jestliže se hustoty f_X a f_Y rovnají na nějakém intervalu $(a, b) \subseteq \mathbb{R}$ takovém, že $\mu \in (a, b)$ a pokud $F_Y(\mu) = \frac{1}{2}$, pak

$$F_Y(t) = F_X(t) \quad \text{pro všechna} \quad t \in (a, b) \ .$$

Značení: Pro náhodnou veličinu X s konečnou střední hodnotou a konečným rozptylem, položme

$$\mathrm{norm}(X) := \frac{X - E(X)}{\sqrt{\mathrm{var}(X)}} \ .$$

Speciálně tedy vidíme, že $E\big(\mathrm{norm}(X)\big) = 0$ a $\mathrm{var}\big(\mathrm{norm}(X)\big) = 1$.

Platí: Pro takovouto veličinu X a konstanty $a > 0$ a $b \in \mathbb{R}$ je

$$\mathrm{norm}(aX + b) = \mathrm{norm}(X) \ .$$

Důležité vlastnosti normálního rozdělení:

- $X \sim N(\mu, \sigma^2) \;\Rightarrow\; \text{norm}(X) = \frac{X-\mu}{\sigma} \sim N(0,1)$ (je to tzv. normované normální rozdělení s hodnotami v tabulkách) dist. funkce pro $N(0,1)$ se značí Φ.

  V tomto případě pak máme $F_X(t) = P(X \le t) = P\Big(\underbrace{\frac{X-\mu}{\sigma}}_{=\text{norm}(X)} \le \frac{t-\mu}{\sigma} \Big) = \Phi\Big(\frac{t-\mu}{\sigma}\Big)$ pro všechna $t \in \mathbb{R}$.

- hustota $f_{N(0,1)}$ je sudá funkce $\;\Rightarrow\;$ $\Phi(t) + \Phi(-t) = 1$ pro všechna $t \in \mathbb{R}$.

- Nechť $X_i \sim N(\mu_i, \sigma_i^2)$, pro $i = 1, 2$, jsou nezávislé. Pak $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ (tj. speciálně součet nezávislých normálních rozdělení je zase normální.)

Pro vybraná čísla $t \ge 0$ se dají hodnoty Φ najít ve statistických tabulkách. Pro záporná čísla si pak pomůžeme vztahem $\Phi(-t) = 1 - \Phi(t)$.

Pro lepší představu o tom, jakou roli pro veličinu s normálním rozdělením $X \sim N(\mu, \sigma^2)$ hraje směrodatná odchylka σ se používá tzv.

pravidlo tří-sigma (`https://cs.wikipedia.org/wiki/Pravidlo_t\%C5\%99\%C3\%AD\_sigma`)

které je ovšem čistě jen technickou pomůckou:

Jestliže si budeme počítat pravděpodobnosti

$$P(|X-\mu| \le k\cdot\sigma) = P\Big(\Big|\underbrace{\frac{X-\mu}{\sigma}}_{\sim N(0,1)}\Big| \le k\Big) = \Phi(k) - \Phi(-k) = 2\cdot\Phi(k) - 1 \quad \text{pro} \quad k = 1, 2, 3, \ldots$$

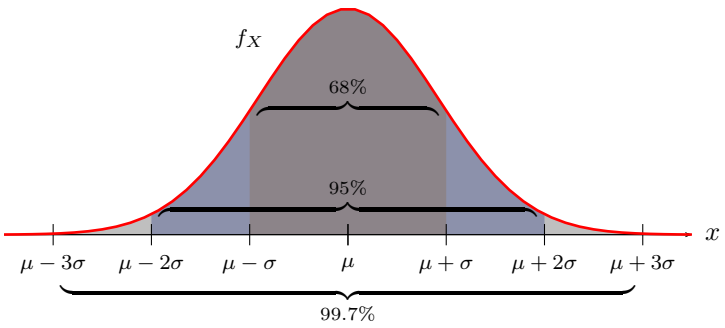
dostaneme postupně

$$P(|X-\mu| \le \sigma) = 2\cdot\Phi(1) - 1 \doteq 2\cdot 0.8413 - 1 = 0.6826 \doteq 68\%$$

$$P(|X-\mu| \le 2\cdot\sigma) = 2\cdot\Phi(2) - 1 \doteq 2\cdot 0.9772 - 1 = 0.9544 \doteq 95\%$$

$$P(|X-\mu| \le 3\cdot\sigma) = 2\cdot\Phi(3) - 1 \doteq 2\cdot 0.99865 - 1 = 0.9973 \doteq 99.7\%$$

Pro vyšší hodnoty, tj. $k \ge 4$ už jsou pravděpodobnosti vpodstatě rovny 1, takže se v praxi příliš nepoužívají (záleží samozřejmě na zvolené přesnosti).



Příklad 7.1 *Výška dětí v* 1. *třídě je náhodná veličina* $X \sim N(130\,\text{cm},\ 36\,\text{cm}^2)$. *Jaká je pravděpodobnost, že náhodně vybrané dítě bude*

(a) větší než 136 *cm,*

(b) menší než 118 *cm,*

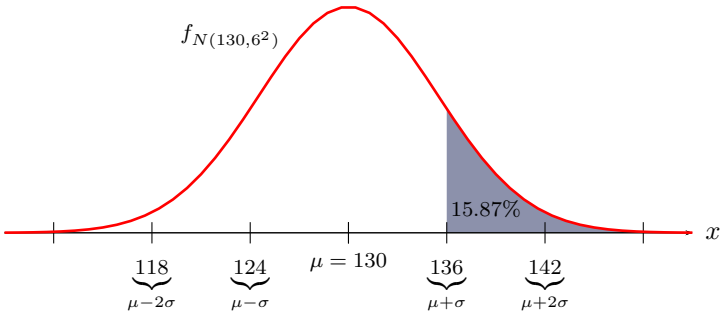
(c) mít výšku mezi 127 *a* 133 *cm?*

Řešení:
Nyní tedy máme $X \sim N(130,\ 36)$. Pro jednodušší zápis si ještě označme $Z := \mathrm{norm}(X)$.

(a)

$$P(X > 136) = P\left(\underbrace{\frac{X-130}{\sqrt{36}}}_{Z} > \underbrace{\frac{136-130}{\sqrt{36}}}_{1}\right) = P(Z > 1) = 1 - P(Z \leq 1) =$$

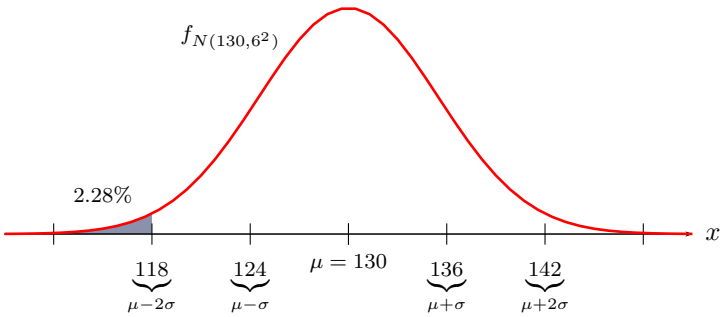
$$= 1 - \Phi(1) \doteq 1 - 0.8413 = 0.1587\ .$$



(b)

$$P(X < 118) = P\left(\frac{X-130}{\sqrt{36}} < \frac{118-130}{\sqrt{36}}\right) = P(Z < -2) = \Phi(-2) =$$

$$= 1 - \Phi(2) \doteq 1 - 0.9772 = 0.0228\ .$$



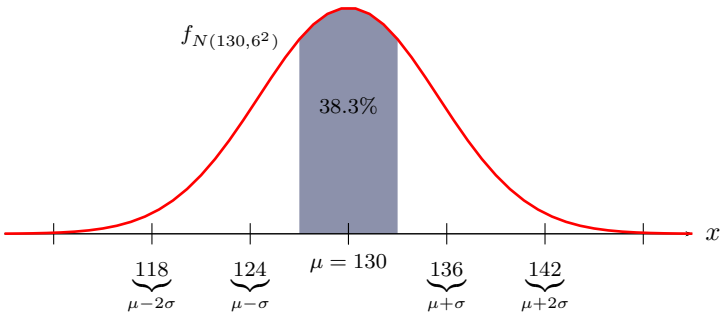
(c)

$$P(127 < X < 133) = P\left(\frac{127-130}{\sqrt{36}} < \frac{X-130}{\sqrt{36}} < \frac{133-130}{\sqrt{36}}\right) =$$

$$= P(-0.5 < Z < 0.5) = P(Z < 0.5) - P(Z \leq -0.5) =$$

$$= \Phi(0.5) - \Phi(-0.5) = \Phi(0.5) - \Big(1 - \Phi(0.5)\Big) =$$

$$= 2 \cdot \Phi(0.5) - 1 \doteq 2 \cdot 0.6915 - 1 = 0.383\ .$$



Poznamenejme ještě, že hodnoty výšek, které nás zajímaly (tj. 136 cm, 118 cm atd.) se pohybují celkem blízko střední hodnoty $E(X) = 130$ cm, takže předpoklad o normálnosti rozdělení X byl přiměřený.

- -

Výpočty si můžeme i urychlit přímým vzorcem $F_X(t) = \Phi(\frac{t-\mu}{\sigma})$, kde $\mu = 130$ cm a $\sigma = 6$ cm:

(a) $P(X > 136) = 1 - F_X(136) = 1 - \Phi(\frac{136-130}{6}) = \cdots \doteq 0.1587$

(b) $P(X < 118) = F_X(118) = \Phi(\frac{118-130}{6}) = \cdots \doteq 0.0228$

(c) $P(127 < X < 133) = F_X(133) - F_X(127) = \Phi(\frac{133-130}{6}) - \Phi(\frac{127-130}{6}) = \cdots \doteq 0.383$

Příklad 7.2 *Oštěpařky Anna a Barbora mají střední hodnoty hodů po řadě 67 m a 75 m a směrodatné odchylky 6 m a 3 m. Předpokládejme nezávislá normální rozdělení. Odhadněte pravděpodobnost, že při jednom hodu hodí Anna dál.*

Řešení:
Náhodná veličina

$A =$"*délka hodu Anny*"

má rozdělení $N(67, 6^2)$ a veličina

$B =$"*délka hodu Barbory*"

má rozdělení $N(75, 3^2)$.

Zajímá nás $P(A > B) = P(A-B > 0)$. Protože veličiny A a B jsou nezávislé, tak veličina $Z := A-B$ má také normální rozdělení, a sice

$$Z \sim N(67 - 75, 6^2 + 3^2) = N(-8, 45)\ .$$

Takže

$$P(A > B) = P(Z > 0) = P\Bigg(\underbrace{\frac{Z-(-8)}{\sqrt{45}}}_{\mathrm{norm}(Z)} > \frac{0-(-8)}{\sqrt{45}}\Bigg) = 1 - P\left(\mathrm{norm}(Z) \le \frac{8}{\sqrt{45}}\right) =$$

$$= 1 - \Phi\left(\tfrac{8}{\sqrt{45}}\right) \doteq 1 - \Phi\left(1.1926\right) \doteq 1 - 0.883 = 0.117\,.$$

POZOR! Zatímco střední hodnota je lineární zobrazení, tak rozptyl se chová jinak! Konkrétně je to takto:

Nechť X a Y jsou veličiny se střední hodnotou a konečným rozptylem. Pak

- $E(X \pm Y) = E(X) \pm E(Y)$
- $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y) \pm 2 \cdot \mathrm{cov}(X, Y)$ (≥ 0)

Zde $\mathrm{cov}(X, Y)$ je tzv. kovariance (viz poznámky níže). Speciálně, pokud X a Y jsou nezávislé, je $\mathrm{cov}(X, Y) = 0$. Máme tedy:

- X a Y nezávislé $\Rightarrow$ $\mathrm{var}(X \pm Y) = \mathrm{var}(X) + \mathrm{var}(Y)$

Tedy v tomto případě se rozptyly VŽDY sčítají!

Příklad 7.3 *Semena mají klíčivost $p \in (0, 1)$. Jaký je optimální počet n semen v jamce, aby byla co nejvyšší pravděpodobnost, že vyklíčí právě jedno? Řešte obecně a pro $p = 1/3$.*

Řešení:
Vyklíčení jednotlivých semen pokládáme za nezávislé jevy, takže veličina

$$X = \text{“počet vyklíčených semen z } n \text{ semen v jamce”}$$

má binomické rozdělení $\mathrm{Bi}(n, p)$. Hledáme teď $n \in \mathbb{N}$ $(n \geq 1)$, které maximalizuje funkci

$$g(n) = P(X = 1) = \binom{n}{1} p^1 (1-p)^{n-1} = n\, p\, (1-p)^{n-1} \left(= np \cdot e^{(n-1)\ln(1-p)} \right) .$$

Pro vyšetření této funkce můžeme uvažovat n jako reálnou proměnnou v intervalu $(0, +\infty)$, abychom pak mohli využít derivaci (podle n) a to především pro zjištění, kde je funkce g rostoucí a kde klesající:

$$g'(n) = p\,(1-p)^{n-1} + n\, p\,(1-p)^{n-1} \ln(1-p) = p\,(1-p)^{n-1}\big(1 + n\,\ln(1-p)\big) .$$

Dostáváme tak, že g je rostoucí až do bodu $n = \frac{-1}{\ln(1-p)}$ a pak je klesající. V uvedeném bodě tak nastává maximum v rámci reálné proměnné. Maximum v oboru přirozených čísel $\mathbb{N}$ nastává pro jedno (nebo obě) ze dvou celých čísel, která jsou nejblíže hodnotě $\frac{-1}{\ln(1-p)}$. Zjistit, které z těchto dvou čísel to vlastně obecně je, ale už dá více práce.

Pro $p = \frac{1}{3}$ máme $\frac{-1}{\ln(1-p)} = \frac{-1}{\ln(2/3)} \doteq 2.466$. Nejbližší čísla jsou tedy 2 a 3 a v nich stačí porovnat hodnoty funkce g:

$g(2) = 2 \cdot \frac{1}{3} \left(\frac{2}{3}\right)^1 = \frac{4}{9}$

$g(3) = 3 \cdot \frac{1}{3} \left(\frac{2}{3}\right)^2 = \frac{4}{9}$

Obě možnosti 2 a 3 tedy představují optimální počty semen pro $p = \frac{1}{3}$.

Zatím můžeme říct, že: *Pro $p \to 0$ je $\ln(1-p) \approx -p$ a $n \approx 1/p$, což je v souladu s očekáváním. Pro $p \to 1$ vychází $n \to 0$, což vypadá překvapivě. Znamená to ale jen, že funkce g je na množině přirozených čísel klesající a maxima nabývá v 1.*

Můžeme to však zkusit také jinak a (překvapivě) i jednodušeji. Zřejmě pro $n \in \mathbb{N}$ máme

$$g(n) \leq g(n+1) \quad \Leftrightarrow \quad 1 \leq \frac{g(n+1)}{g(n)} = \frac{(n+1)\,p\,(1-p)^n}{n\,p\,(1-p)^{n-1}} = \frac{n+1}{n}(1-p)$$

což po úpravě dává

$$\frac{1}{1-p} \leq \frac{n+1}{n} = 1 + \frac{1}{n}$$

$$\underbrace{\frac{1}{1-p}-1}_{\frac{p}{1-p}} \leq \frac{1}{n}$$

a nakonec

$$n \leq \frac{1-p}{p} = \frac{1}{p} - 1 \ .$$

Speciálně vidíme, že to samé platí i pro ostré nerovnosti, tj.

* $g(n) < g(n+1)$ platí právě když $n < \frac{1}{p} - 1$.
* $g(n) > g(n+1)$ platí právě když $n > \frac{1}{p} - 1$.
* $g(n) = g(n+1)$ platí právě když $n = \frac{1}{p} - 1$.

Odsud ihned máme, že

- pro $\frac{1}{p} - 1 < 1$, tj. $p \in \left(\frac{1}{2}, 1\right)$ je g ostře klesající (na $\mathbb{N}$), takže maximum nastává pro $n_0 = 1$ a pravděpodobnost pak je $P(X = 1) = g(1) = p$.
- pokud $\frac{1}{p} \notin \mathbb{N}$ a $p \in \left(0, \frac{1}{2}\right)$, budou všechny nerovnosti mezi hodnotami funkce g ostré a největší hodnota bude dosažena pro $n_0 = \left[\frac{1}{p}\right]$ (tj. celá část z $\frac{1}{p}$).

  Je to proto, že z $n_0 - 1 < \frac{1}{p} - 1$ plyne $g(n_0 - 1) < g(n_0)$ a současně z $\frac{1}{p} - 1 < n_0$ plyne $g(n_0) > g(n_0 + 1)$.

  Navíc ještě z nerovnosti $n_0 < \frac{1}{p} < n_0 + 1$ dostaneme odhad pravděpodobnosti

$$\underbrace{\left(\frac{1}{p} - 1\right) p(1-p)^{\frac{1}{p}-1}}_{(1-p)^{\frac{1}{p}}} < \underbrace{n_0 p(1-p)^{n_0-1}}_{P(X=1)} < \underbrace{\frac{1}{p}\, p(1-p)^{\frac{1}{p}-2}}_{(1-p)^{\frac{1}{p}-2}}$$

- a konečně pro $n_0 = \frac{1}{p} \in \mathbb{N}$ budou všechny nerovnosti mezi hodnotami $g(n)$ také ostré až na případ $g(n_0 - 1) = g(n_0)$, který odpovídá maximu funkce g na přirozených číslech.

  V tomto případě totiž platí, že $n_0 - 1 = \frac{1}{p} - 1$, tedy skutečně $g(n_0 - 1) = g(n_0)$.

  Tedy maximum nastává pro dva počty semen $n_0 - 1$ a n_0 a pravděpodobnost vyklíčení právě jednoho semene je

$$P(X = 1) = \frac{1}{p} p(1-p)^{\frac{1}{p}-1} = (1-p)^{\frac{1}{p}-1}$$

Mimo jiné z tohoto všeho vidíme, že pro $p \to 0$ se pravděpodobnost vyklíčení jednoho semene (při optimálním počtu semen) blíží k $e^{-1} \doteq 0.3679$ (protože $\lim\limits_{p \to 0+} (1-p)^{\frac{1}{p}} = \lim\limits_{p \to 0+} e^{\frac{\ln(1-p)}{p}} = e^{-1}$).

Pro konkretní volbu $p = \frac{1}{3}$ máme $\frac{1}{p} = 3 \in \mathbb{N}$ (tedy třetí případ) a tak opět dostáváme, že: optimální počet semen je $n \in \left\{\frac{1}{p} - 1, \frac{1}{p}\right\} = \{2, 3\}$ a pravděpodobnost bude $P(X = 1) = (1-p)^{\frac{1}{p}-1} = \frac{4}{9}$.