8. cvičení z STP

8. - 12. duben 2019

Příklad 7.1.

Příklad 8.1 *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 $n_0 = \left\lfloor \frac{1}{p} \right\rfloor$ (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 $P(X = 1) = (1-p)^{\frac{1}{p}-1} = \frac{4}{9}$.

Příklad 8.2 *Průměrný počet zákazníků během dne v první prodejně je* 20, *ve druhé prodejně* 25. *Předpokládáme, že oba počty se řídí Poissonovým rozdělením. Odvoďte rozdělení počtu zákazníků v obou prodejnách dohromady.*

Řešení:
Označme si veličiny

$X =$ *"počet zákazníků během dne v* 1. *prodejně"*

$Y =$ *"počet zákazníků během dne v* 2. *prodejně"*

$Z =$ *"počet zákazníků během dne v obou prodejnách dohromady"*

kde X a Y budeme přirozeně pokládat za nezávislé. Máme

$$X \sim \text{Poiss}(\lambda), \text{ kde } \lambda = E(X) = 20$$

$$Y \sim \text{Poiss}(\mu), \text{ kde } \mu = E(Y) = 25$$

a

$$P(X = i) = \frac{\lambda^i}{i!} \cdot e^{-\lambda} \quad \text{a} \quad P(Y = j) = \frac{\mu^j}{j!} \cdot e^{-\mu}$$

Jelikož $Z = X + Y$ a případ $X + Y = k$ se rozloží na disjunktní možnosti $(X, Y) = (i, k - i)$ pro $i = 0, 1, \ldots, k$, dostáváme

$$P(Z = k) = P(X + Y = k) =$$

$$= \sum_{i=0}^{k} P(X = i,\ Y = k - i) \overset{\text{(nezávislost)}}{=} \sum_{i=0}^{k} P(X = i) \cdot P(Y = k - i) =$$

$$= \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda} \cdot \frac{\mu^{k-i}}{(k-i)!} e^{-\mu} = e^{-(\lambda+\mu)} \sum_{i=0}^{k} \frac{1}{k!} \cdot \underbrace{\frac{k!}{i!(k-i)!}}_{\binom{k}{i}} \cdot \lambda^i \cdot \mu^{k-i} =$$

$$= \frac{e^{-(\lambda+\mu)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \cdot \mu^{k-i} \overset{\text{(binom. věta)}}{=} \frac{e^{-(\lambda+\mu)}}{k!} (\lambda + \mu)^k$$

tj. $Z \sim \text{Poiss}(\lambda + \mu)$ neboli $Z \sim \text{Poiss}(45)$.

Poznámka: Představme si, že jednotlivým prodejnám přiřadíme (čistě účelově) nějaké velikosti a a b (např. velikost plochy prodejny v m^2) tak, aby průměrný počet zákazníků na jednotku plochy byl v obou prodejnách stejný, tj. $\frac{\lambda}{a} = \frac{\mu}{b}$. Veličiny X a Y pak můžeme chápat jako počty událostí v intervalech délky a a b, přičemž v obou intervalech je "hustota událostí" stejná. Při tomto přístupu bude veličina Z představovat počet událostí v intervalu délky $a + b$, takže Poissonovo rozdělení se pak dá skutečně očekávat.

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

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

Je to tedy spojité rozdělení, $E(X) = \mu$, $\text{var}(X) = \sigma^2$ 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. 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

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

Příklad 8.3 *Výška dětí v* 1. *třídě je náhodná veličina* $X \sim N(130\,\mathrm{cm},\ 36\,\mathrm{cm}^2)$. *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:
Pro (libovolnou) veličinu Y budeme označovat

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

tzv. *normovanou* veličinu pro Y (pokud má vzorec smysl). Vždy pak dostáváme, že

$$E\Big(norm(Y)\Big) = 0 \qquad \text{a} \qquad \mathrm{var}\Big(norm(Y)\Big) = 1\ .$$

Dále platí, že

$$Y \sim N(\mu, \sigma^2) \qquad \Rightarrow \qquad norm(Y) = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

kde rozdělení $N(0,1)$ se nazývá *normované* normální rozdělení.

Distribuční funkce pro $N(0,1)$ se označuje jako Φ a její hodnoty $\Phi(t)$ pro vybraná čísla $t \geq 0$ se dají najít ve statistických tabulkách. Pro záporná čísla si pak pomůžeme vztahem

$$\Phi(-t) + \Phi(t) = 1 \quad \text{pro všechna} \quad t \in \mathbb{R}\ ,$$

který je důsledkem sudosti hustoty pro $N(0,1)$.

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

(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 blízko střední hodnoty $E(X) = 130\,\mathrm{cm}$, takže předpoklad o normálnosti rozdělení X byl přiměřený.