Statistika a spolehlivost v lékařství (A6M33SSL)

Osnova statistických cvičení s řešenými příklady

URČENO VÝHRADNĚ PRO STUDENTY V LS 2017/2018

verze 2018-04-17

Obsah

Upozornění: Tento text není učebním textem. Jedná se o jakousi kostru cvičení statistické části cvičení předmětu A6M33SSL, které autor v průběhu školních let 2014/15 - 2017/18 připravil a cvičil. Text obsahuje výběr některých důležitých termínů z pravděpodobnosti a statistiky s ilustračními příklady, které jsou většinou doplněny řešeními a často i dodatečnými faktickými a metodickými poznámkami (vysázenými kurzívou). Těžší partie jsou označeny symbolem *. Obsahově se text do víceméně kryje s obsahem cvičení a místy jej i přerůstá. Text nabízím studentům jako doplněk k jejich vlastním poznámkám ze cvičení, a budu vděčný, upozorní-li mě na případné chyby zprávou na `siegetom@fel.cvut.cz`.

Za laskavé připomínky vděčím Petru Pošíkovi.

Tomáš Sieger

Stručný přehled značení

X	náhodná veličina (zpravidla velké písmeno latinkou)
x	realizace náhodné veličiny (zpravidla malé písmeno latinkou)
f_X	hustota pravděpodobnosti náhodné veličiny X (nebo rozdělení X)
F_X	distribuční funkce náhodné veličiny X (nebo rozdělení X)
F_X^{-1} nebo q_X	kvantilová funkce náhodné veličiny X (nebo rozdělení X)
$DX \equiv varX$	rozptyl náhodné veličiny X
EX	střední hodnota náhodné veličiny X
$N(\mu, \sigma^2)$	normální rozdělení se střední hodnotou μ a rozptylem σ^2
$N(0,1)$	normované normální rozdělení
$\varphi \equiv f_{N(0,1)}$	hustota pravděpodobnosti normovaného normálního rozdělení
$\Phi \equiv F_{N(0,1)}$	distribuční funkce normovaného normálního rozdělení
$\Phi^{-1} \equiv F^{-1}_{N(0,1)}$	kvantilová funkce normovaného normálního rozdělení
t_n	Studentovo t-rozdělení s n stupni volnosti
χ^2_n	χ^2 rozdělení s n stupni volnosti
$Alt(p)$	alternativní rozdělení popisující např. výsledek nějakého pokusu s pravděpodobností úspěchu p
$Bi(n,p)$	binomické rozdělení popisující např. počet úspěchů v sérii nezávislých n pokusů s elementární pravděpodobností úspěchu p
$Po(\lambda)$	Poissonovo rozdělení s parametrem λ, který je roven střední hodnotě i rozptylu tohoto rozdělení (toto rozdělení popisuje např. počet nějakých událostí za daný čas, pokud pravděpodobnost každé události je nezávislá na čase minulé události a průměrný počet událostí za daný čas je roven λ)

Seznam obrázků

1 Opakování: pravděpodobnost

Vysvětlete termíny:

1.1 Náhodný jev (elementární jev, prostor elementárních jevů)

Příklad: Házení kostkou.

1.2 Pravděpodobnost

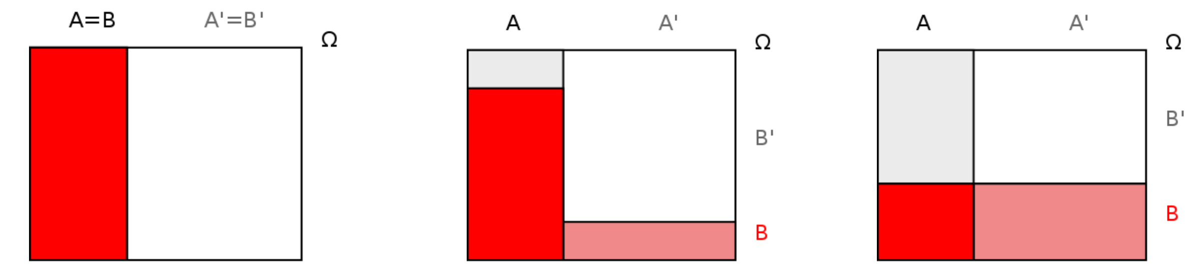
(funkce definovaná na podmnožinách prostoru elementárních jevů)

1.3 ⋆Pravděpodobnostní prostor: $(\Omega, \mathcal{A}, p)$

1.4 Nezávislost náhodných jevů

Příklad: Příprava na zkoušku.
Alice: “Bobe, jak je pravděpodobné, že tu zkoušku oba uděláme?”
Bob: “Řekl bych 70%.”
Alice: ”Ale včera jsi tvrdil, že ji na 90% uděláš a že já mám stejnou šanci.“
Bob: ”No a¿‘ Alice: ”To není možné, a jestli to nevidíš, tak tu zkoušku asi neuděláš.“

Příklad: Házení kostkou: jev “padne liché číslo” vs. jev “padne číslo větší než 3”. Jsou tyto jevy nezávislé?



identické jevy („100%“ závislé) („částečně“) závislé nezávislé

Obr. 1: Ilustrace závislosti a nezávislosti dvou jevů.

Příklad: Házení dvěma kostkami. Jev A: na první kostce padne liché číslo, jev B: na druhé kostce padne sudé číslo, jev C: součet čísel na obou kostkách je sudý. Jsou tyto 3 jevy nezávislé? Jsou tyto jevy po dvou nezávislé?

1.5 Podmíněná pravděpodobnost

Ukázat přes geometrickou představu 2 množin A a B s neprázdným průnikem.

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \qquad \text{(za podmínky } p(B) > 0) \tag{1}$$

Příklad: Tenista má první podání úspěšné s pravděpodobností 0,6; druhé s pravděpodobností 0,8. S jakou pravděpodobností se dopustí dvojchyby?

Řešení: jev N_1 - chyba v prvním podání, jev N_2 - chyba ve druhém podání. $p(N_1) = 0,4$, $p(N_2|N_1) = 0,2$ (pozor, toto není $p(N_2)$!). $P(N_1 \cap N_2) = p(N_2|N_1)p(N_1) = 0,2 \cdot 0,4 = 0,08$.

1.6 Úplná pravděpodobnost

aneb celková pravděpodobnost jevu A se dá „spočítat po kouskách a posčítat“

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i) \tag{2}$$

Příklad: Spočtěte průměrné zastoupení studentek ve škole, která má 2 posluchárny a 2 šatny a v každé místnosti je jiné zastoupení dívek. (Předpokládáme, že průměrné zastoupení studentek nelze zjistit přímo, např. anketou.)

Řešení: Jev D - náhodně vybraný(á) student(ka) je dívka. Jev M_i - student(ka) se nachází v místnosti i.

$$p(D) = \sum_{i=1}^{4} p(D \cap M_i) = \sum_{i=1}^{4} p(D|M_i)p(M_i)$$

1.7 Bayesův vzorec

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

aneb výpočet posteriorní podmíněné pravděpodobnosti $P(A|B)$ z apriorních pravděpodobnost $P(A)$, nebo přepočet podmíněné pravděpodobnosti $P(B|A)$ na podmíněnou pravděpodobnost $P(A|B)$.

Označuje-li např. $P(B|A)$ podmíněnou pravděpodobnost určitého výsledku klinického testu (jev B) u zdravého člověka (podmínka A), pak Bayesův vzorec dává na základě výsledku testu (a apriorních pravděpodobností) podmíněnou pravděpodobnost $P(A|B)$ toho, že testovaná osoba je zdravá či nemocná.

Příklad: Klinický test, jehož účelem je odhadnout, zda pacient má určitou nemoc, má senzitivitu (pravděpodobnost toho, že u nemocného bude test pozitivní) 90% a specificitu (pravděpodobnost toho, že u zdravého bude test negativní) 95%. Spočítejte pravděpodobnost toho, že pacient s pozitivním testem nemoc skutečně má, a dále pravděpodobnost toho, že pacient s negativním testem nemoc skutečně nemá. Předpokládejme, že nemocí trpí 20% populace.

Řešení: Jev, že pacient je nemocný, označíme jako N, jev opačný jako Z. Pozitivní výsledek testu označíme jako Poz a negativní jako Neg. Víme, že $p(Poz|N) = 0,9$, $p(Neg|Z) = 0,95$ a $p(N) = 0,2$.

Použitím Bayesova vzorce dostáváme:

$$p(N|Poz) = \frac{p(Poz|N)p(N)}{p(Poz)},$$

kde $p(Poz)$ neznáme, ale dokážeme spočítat pomocí věty o úplné pravděpodobnosti:

$$p(Poz) = p(Poz|N)p(N) + p(Poz|Z)p(Z).$$

Z toho

$$p(N|Poz) = \frac{p(Poz|N)p(N)}{p(Poz)} = \frac{p(Poz|N)p(N)}{p(Poz|N)p(N) + p(Poz|Z)p(Z)} = \frac{0,9 \cdot 0,2}{0,9 \cdot 0,2 + 0,05 \cdot 0,8} \approx 0,818.$$

Pozitivní výsledek testu tedy přítomnost nemoci neindikuje příliš spolehlivě.

Podobně

$$p(Z|Neg) = \frac{p(Neg|Z)p(Z)}{p(Neg)} = \frac{p(Neg|Z)p(Z)}{p(Neg|Z)p(Z) + p(Neg|N)p(N)} = \frac{0,95 \cdot 0,8}{0,95 \cdot 0,8 + 0,1 \cdot 0,2} \approx 0,974,$$

tedy $p(N|Neg) \approx 0,026$ a falešně negativní výsledek lze naštěstí čekat jen u necelých tří procent testovaných osob.

Která chyba vadí víc: falešná pozitivita nebo falešná negativita? Jaká je celková očekávaná chyba a co je její hlavní příčinou? Jak byste test vylepšili?

Poznamenejme, že kromě vlastností daného klinického testu (jeho senzitivity a specificity) má na pravděpodobnosti $p(N|Poz)$ a $p(Z|Neg)$ vliv i prevalence nemoci. Pokud by nemocí trpěla např. pouhá 2% populace, bylo by $p(N|Poz) \approx 0,269$ a $p(Z|Neg) \approx 0,998$.

Příklad: V obléhaném městě, které se skládá ze dvou vzájemně oddělených čtvrtí Mordor (tvoří desetinu města) a Londor (tvoří zbylých 90% města), nepřítel kontaminoval jedem vodovod zásobující Mordor. 90% mordorských obyvatel je otráveno (všichni krom těch, kteří se vody zatím nenapili). V Londoru je otráveno pouze 10% obyvatel (patrně z jiných zdrojů, než z vody). Jednomu obyvateli se z města podaří tajnou chodbou uniknout, avšak záhy umírá – byl otráven. Dá se odhadnout, zda tajná chodba vede do Mondoru, nebo Londoru? (Pro zjednodušení budeme předpokládat, že pravděpodobnost toho, že chodbou z města někdo unikne, je stejná, ať už chodba vede kamkoli.) Jak se tato pravděpodobnost změní, vyjde-li z chodby další otrávený?

Řešení:

Jev, že chodba vede do jedem otrávené části města, označíme jako J a jev, že člověk přicházející chodbou je otrávený a zemře, jako M.

Za apriorní (předem očekávanou) pravděpodobnost toho, že chodba vede do otrávené části města, musíme (bez znalosti dodatečných informací) vzít poměr plochy otrávených částí města vůči celému městu, tedy $p(J) = 0,1$. Dále víme, že pravděpodobnost, že chodbou vedoucí do otráveného Mordoru někdo uteče a zemře, je $p(M|J) = 0,9$, a pravděpodobnost, že chodbou vedoucí do neotráveného Londoru někdo uteče a zemře, je $p(M|\bar{J}) = 0,1$. Poté, co chodbou přijde otrávený člověk, se pravděpodobnost, že chodba vede do otrávené části města, změní z $p(J)$ na $p(J|M)$:

$$p(J|M) \overset{Bayes}{=} \frac{p(M|J)p(J)}{p(M)} = \frac{p(M|J)p(J)}{\underbrace{p(M|J)p(J) + p(M|\bar{J})p(\bar{J})}_{\text{úplná pravděpodobnost}}} = \frac{0,9 \cdot 0,1}{0,9 \cdot 0,1 + 0,1 \cdot 0,9} = \frac{1}{2}$$

Pravděpodobnost se zvýšila z apriorní pravděpodobnosti $0,1$ na posteriorní pravděpodobnost $0,5$. To znamená, že před tím, než z chodby kdokoli vyšel, mysleli jsme si, že chodba vede spíše do otrávené části města. Poté, co z chodby vyšel otrávený člověk, naprosto nevíme, zda chodba vede do trávené nebo neotrávené části města.

Pokud z chodby vystoupí další člověk a také zemře, změní se pravděpodobnost, že chodba vede do otrávené části města, dále na:

$$p(J|M_2, M_1) = \frac{p(M_2|J, M_1)p(J|M_1)}{p(M_2|J, M_1)p(J|M_1) + p(M_2|\bar{J}, M_1)p(\bar{J}|M_1)}.$$

Za předpokladu, že pravděpodobnost výskytu otráveného člověka v otrávené oblasti (resp. mimo otrávenou oblast) se po příchodu prvního otráveného nezmění (tj. pokud je ve městě hodně obyvatel a odchodem jednoho otráveného se pravděpodobnosti prakticky nezmění), platí $p(M_2|J, M_1) = p(M_2|J)$ a $p(M_2|\bar{J}, M_1) = p(M_2|\bar{J})$ a tedy

$$p(J|M_2, M_1) = \frac{p(M_2|J)p(J|M_1)}{p(M_2|J)p(J|M_1) + p(M_2|\bar{J})p(\bar{J}|M_1)} = \frac{0,9 \cdot 0,5}{0,9 \cdot 0,5 + 0,1 \cdot 0,5} = 0,9.$$

Poté, co z chodby vyjde druhý otrávený, budeme tedy spíše přesvědčeni, že chodba vede do otrávené části města. Naši apriorní představu o tom, že chodba vede spíše do neotrávené části města, jsme tedy na základě dat byli nuceni zcela přehodnotit.

Pro srovnání: v případě, že by chodbou přišli (a vzápětí zemřeli) dva lidé nezávisle na sobě a my bychom aposteriorní pravděpodobnost toho, že chodba vede do otrávené části města, počítali nikoli postupně, ale najednou, došli bychom ke stejnému výsledku. Položili bychom

$$p(M_{1+2}|J) = p(M|J) \cdot p(M|J) = p(M|J)^2$$

a

$$p(M_{1+2}|\bar{J}) = p(M|\bar{J}) \cdot p(M|\bar{J}) = p(M|\bar{J})^2$$

a proto

$$\begin{aligned}
p(J|M_{1+2}) &= \frac{p(M_{1+2}|J)p(J)}{p(M_{1+2}|J)p(J) + p(M_{1+2}|\bar{J})p(\bar{J})} \\
&= \frac{p(M|J)^2 p(J)}{p(M|J)^2 p(J) + p(M|\bar{J})^2 p(\bar{J})} \\
&= \frac{0,9^2 \cdot 0,1}{0,9^2 \cdot 0,1 + 0,1^2 \cdot 0,9} \\
&= \frac{0,9^2}{0,9^2 + 0,1 \cdot 0,9} \\
&= \frac{0,9}{0,9 + 0,1} \\
&= 0,9.
\end{aligned}$$

1.8 Náhodná veličina

Náhodná veličina je striktně řečeno funkce definovaná na prostoru elementárních jevů:

★ měřitelná funkce $X : \Omega \to \mathbb{R}$ splňující $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{A}$ pro každý interval I, tj. vzorem každého intervalu je měřitelná množina ze σ-algebry $\mathcal{A}$ (systému podmnožin prostoru elementárních jevů Ω.

Pro praktické účely budeme chápat náhodnou veličinu jako číslo, které vhodným způsobem charakterizuje náhodný jev, např. hmotnost nějakého pacienta, počet líců v sérii nezávislých hodů mincí, nebo doba bezchybného provozu nějaké součástky.

1.9 Realizace náhodné veličiny

ilustrace: 3 světy (reálný, zjednodušený reálný (modelový), „statistický“)

Příklad: Hmotnost studentů určitého kurzu – v modelovém a „statistickém“ světě graficky vyznačte jednu konkrétní realizaci náhodné veličiny „hmotnost“ a samotnou náhodnou veličinu „hmotnost“.

Nechat studenty vyznačit na reálné ose jedno pozorování - hmotnost a pak to samé chtít pro náhodnou veličinu, aby byli konfrontováni s tím, že vlastně nějakou hodnotu nakreslit nelze. Chová se náhodně (paralela s kvantovým světem a štěrbinovým experimentem). Jediné, co o ní víme, je jak se chová „typicky“, „průměrně“. Dojít k tomu, že někde je větší pravděpodobnost výskytu, jinde menší.

1.10 Rozdělení pravděpodobnosti náhodné veličiny

Příklad: Definujte rozdělení pravděpodobnosti náhodné veličiny.

Pozor: u spojitého rozdělení $p(X = x) = 0$, proto používáme $p(X \leq x)$.

Příklad: Načtrtněte rozdělení pravděpodobnosti počtu líců při hodu mincí.

Příklad: Načtrtněte rozdělení pravděpodobnosti počtu ok při hodu kostkou.

Příklad: Načtrtněte rozdělení pravděpodobnosti počtu líců při dvou nezávislých hodech mincí.

1.11 Distribuční funkce $F_X(x)$ náhodné veličiny X

$F_X(x) \stackrel{def.}{=} p(X \leq x), x \in \mathbb{R}$

Vlastnosti distribuční funkce:

- $lim_{x\to-\infty} F_X(x) = 0$ *(protože $lim_{x\to-\infty} p(X \leq x) = 0$)*
- $lim_{x\to\infty} F_X(x) = 1$ *(protože $lim_{x\to\infty} p(X \leq x) = 1$)*

Příklad: Načtrtněte distribuční funkci rozdělení pravděpodobnosti počtu líců při hodu mincí.

Příklad: Načtrtněte distribuční funkci rozdělení pravděpodobnosti počtu ok při hodu kostkou.

Příklad: Načtrtněte distribuční funkci rozdělení pravděpodobnosti počtu líců při dvou nezávislých hodech mincí.

1.12 Hustota pravděpodobnosti f_X náhodné veličiny X

V případě jakých rozdělení o ní hovoříme?

Hustotu pravděpodobnosti definujeme u spojitých rozdělení.

U spojitého rozdělení platí, že $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$.

Příklad: Určuje $f_X(x) = sin(x)$ pro $0 \leq x \leq \pi$ hustotu pravděpodobnosti nějaké náhodné veličiny?

Řešení: Ne, $\int_0^{\pi} sin(t)dt = [-cos(t)]_0^{\pi} = 1-(-1) = 2$, tedy hustota se neintegruje do 1 a nejedná se o hustotu.

Příklad: Určuje $f_X(x) = 2 - x$ pro $0 \leq x \leq 2 + \sqrt{2}$ hustotu pravděpodobnosti nějaké náhodné veličiny?

Řešení: Ne, $\int_0^{2+\sqrt{2}} f_X(x)dx = 1$, ale $f_X(2 + \sqrt{2}) < 0$.

Příklad: Načtrtněte hustotu pravděpodobnosti a distribuční funkci rovnoměrného rozdělení na intervalu (a, b).

1.13 Kvantilová funkce F_X^{-1}

Pro spojité ryze rostoucí inverzní funkce k F_X.

Příklad: Je dáno $p(X \leq x) = x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_X, hustotu f_X a kvantilovou funkci F_X^{-1}.

Řešení: Distribuční funkce $F_X(x)$ je právě $p(X \leq x) = x$. Hustota $f_X(x)$ je derivací $F_X(x)$, tedy $f_X(x) = 1$. Kvantilová funkce je inverzí k distribuční funkci, tedy $F_X^{-1}(u) = u$ (má smysl samozřejmě jen pro $0 < u < 1$).

Příklad: Je dáno $f_X(x) = 3x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_X, $p(X \leq x)$ a kvantilovou funkci F_X^{-1}.

Řešení: Nelze řešit: f_X není hustota, neintegruje se do 1.

Příklad: Je dáno $f_X(x) = 2x$ pro $0 \leq X \leq 1$. Odvoďte a načtrtněte distribuční funkci F_X, $p(X \leq x)$ a kvantilovou funkci F_X^{-1}.

Řešení: Distribuční funkce je integrál hustoty, tedy $F_X(x) = x^2$. $p(X \leq x) = F_X(x)$ z definice. Kvantilová funkce F_X^{-1} je inverzí k F_X, tedy $F_X^{-1}(u) = \sqrt{u}$ (má smysl samozřejmě jen pro $0 < u < 1$).

Příklad: V lese zakresleném na mapě jako rovnostranný trojúhelník o straně 30km se ztratilo dítě. Na záchranu se vypravily tři stejné rojnice, z každé strany lesa jedna, které dokáží prohledávat terén rychlostí 500 m/hodinu. Za jak dlouho dítě najdou? Za jak dlouho bude pravděpodobnost, že dítě našli, stejná, jako že dítě ještě nenašli? Za jak dlouho bude pravděpodobnost, že dítě našli, rovna 75%, resp. 95%?
Nalezněte rozdělení vzdálenosti dítěte od nejbližšího okraje lesa a popište jeho hustotu pravděpodobnosti a distribuční funkci.

Řešení: Uvědomíme si, že stačí uvažovat jen část celého trojúhelníku mezi jednou stranou a těžištěm. Označíme-li $a = 30km$, je vzdálenost těžiště od strany třetinou délky těžnice (a zároveň výšky), tj. $\frac{1}{3} \cdot \frac{\sqrt{3}}{2} a \approx 8,660km$, a pro jednoduchost si označíme toto číslo jako b. Potom hustota rozdělení pravděpodobnosti vzdálenosti dítěte od nejbližší strany bude největší v blízkosti strany (zde se totiž může dítě nacházet podél celé strany) a bude klesat směrem k těžišti (kde dítě již nebude mít žádnou volnost - pokud nebylo nalezeno dříve, musí být jedině v těžišti), a lze ji popsat jako

$$f(x) = lb - \frac{lb}{b}x = lb - lx = l(b - x),$$

kde konstantu l vypočteme tak, aby se hustota pravděpodobnosti integrovala do 1, tj. aby plocha pod hustotou byla jednotková:

$$\begin{aligned} \frac{1}{2} lbb &= 1 \\ l &= \frac{2}{b^2} \end{aligned}$$

tedy

$$f(x) = \frac{2}{b^2}(b - x).$$

Hledáme-li pravděpodobnost, že dítě bude nalezeno do určité doby (resp. vzdálenosti), tj. hledáme-li $p(X \le x)$, hodí se zavést funkci $F(x) = p(X \le x) = \int_{y=0}^{x} f(y)dy$.

Výpočtem dostaneme

$$\begin{aligned} F(x) &= \int_{y=0}^{x} f(y)dy \\ &= \int_{y=0}^{x} \frac{2}{b^2}(b - y)dy \\ &= \frac{2}{b^2}\left[by - \frac{y^2}{2}\right]_{y=0}^{x} \\ &= \frac{2}{b^2}\left(bx - \frac{x^2}{2}\right). \end{aligned}$$

Nyní budeme hledat takovou vzdálenost od kraje lesa, že pravděpodobnost, že dítě našli, bude stejná, jako že dítě ještě nenašli. Budeme tedy hledat takové x, aby $F(x) = 0,5$. Protože se ale budeme dále ptát na další otázky, můžeme rovnou řešit obecně jako $F(x) = c$:

$$\begin{aligned} F(x) = \frac{2}{b^2}(bx - \frac{x^2}{2}) &= c \\ \frac{2}{b}x - \frac{x^2}{b^2} &= c \\ x^2 - 2bx + b^2c &= 0 \end{aligned}$$

a řešíme kvadratickou rovnici:

$$D = 4b^2 - 4b^2c = 4b^2(1 - c) = (2b)^2\sqrt{1 - c}$$

a protože má smysl hledat pouze mezi stranou lesa a jeho středem (těžištěm), tedy $x < b$

$$x = \frac{2b - 2b\sqrt{1 - c}}{2} = b(1 - \sqrt{1 - c}).$$

Tento vztah mezi c a x si můžeme označit jako $F^{-1}(c) = b(1 - \sqrt{1 - c})$.

Nyní dostáváme, že 50% pravděpodobnost nalezení dítěte nastane v okamžiku, kdy rojnice projdou $F^{-1}(0,5) = b(1-\sqrt{1-0,5}) \approx 2,537km$ lesa, tj. za cca 5 hodin bude již dítě s pravděpodobností $0,5$ zachráněno.

Pravděpodobnost, že bude zachráněno s pravděpodobností 75%, nastane, až bude prohledáno $F^{-1}(0,75) = b(1-\sqrt{1-0,75}) \approx 4,330km$ lesa, tedy cca za 8 hodin 40 minut. Úplná záchrana pak bude zajištěna, až rojnice dojdou do těžiště, tedy až prohledají $F^{-1}(1) = b(1-\sqrt{1-1}) = b \approx 8,660km$ lesa, tedy cca za 17 hodin 19 minut.

Poznamenejme, že funkce $F(x)$ se nazývá *distribuční funkce* a říká, jak pravděpodobné je, že se dítě dostalo do určité vzdálenosti od okraje lesa (obecně jak pravděpodobné je, že náhodná veličina nabývá nanejvýš dané hodnoty). Funkce $F^{-1}(c)$ se nazývá *kvantilová funkce* a říká, do jaké vzdálenosti dojde rojnice v okamžiku, kdy bude pravděpodobnost nalezení dítěte $P\%$ (obecně při jaké hodnotě náhodné veličiny bude pravděpodobnost jejího výskytu mezi $-\infty$ a touto hodnotou rovna dané pravděpodobnosti). Např. pravděpodobnost nalezení náhodné veličiny s hodnotou nejvýše $F^{-1}(50\%)$ je 50%.

Příklad: (Použití tabulek kvantilů a kritických hodnot:) Hmotnost vyráběné pilulky lze popsat normálním rozdělením se střední hodnotou $120mg$ a rozptylem $1mg^2$. Výstupní kontrola testuje, zda tomu tak skutečně je. Rozumně velký náhodný vzorek pilulek byl zvážen a setříděn podle narůstající hmotnosti.

1. V jakém rozmezí lze čekat hmotnost 10% nejlehčích pilulek?
2. V jakém rozmezí asi bude hmotnost 10% nejtěžších pilulek?
3. V jakém rozmezí lze čekat hmotnost 1% nejlehčích pilulek?
4. V jakém rozmezí asi bude hmotnost 1% nejtěžších pilulek?
5. V jakém rozmezí lze čekat hmotnost 0,1% nejlehčích pilulek?
6. V jakém rozmezí asi bude hmotnost 0,1% nejtěžších pilulek?
7. Jaká je pravděpodobnost, že nalezneme pilulku o hmotnosti 120mg?
8. Jaká je pravděpodobnost, že nalezneme pilulku těžší než 120mg?
9. Jaká je pravděpodobnost, že nalezneme pilulku těžší než 123mg?
10. Jaká je pravděpodobnost, že nalezneme pilulku o hmotnosti nižší než 117,5mg?

Řešení: *Uvažujme symetrii* $\Phi(x)$.

1. rozmezí hmotnosti 10% nejlehčích pilulek: $(max(0, 120 + \Phi^{-1}(0\%)),\ 120 + \Phi^{-1}(10\%)) = (0,\ 120 + (-\Phi^{-1}(90\%))) \doteq (0,\ 120 + (-1,282)) = (0,\ 118,718)$
2. rozmezí hmotnosti 10% nejtěžších pilulek: $(120+\Phi^{-1}(90\%),\ 120+\Phi^{-1}(100\%)) \doteq (121,282,\ \infty)$
3. rozmezí hmotnosti 1% nejlehčích pilulek: $(max(0, \Phi^{-1}(0\%)),\ 120 + \Phi^{-1}(1\%)) = (0,\ 120 + (-\Phi^{-1}(99\%))) \doteq (0,\ 120 + (-2,326)) = (0,\ 117,674)$
4. rozmezí hmotnosti 1% nejtěžších pilulek: $(120+\Phi^{-1}(99\%),\ 120+\Phi^{-1}(100\%)) \doteq (122,326,\ \infty)$
5. rozmezí hmotnosti 0,1% nejlehčích pilulek: $(max(0, 120 + \Phi^{-1}(0\%)),\ 120 + \Phi^{-1}(0,1\%)) = (0,\ 120 + (-\Phi^{-1}(99,9\%))) \doteq (0,\ 120 + (-3,009)) = (0,\ 116,991)$
6. rozmezí hmotnosti 0,1% nejtěžších pilulek: $(120+\Phi^{-1}(99,9\%),\ 120+\Phi^{-1}(100\%)) \doteq (123,009,\ \infty)$
7. Pravděpodobnost nálezu pilulky o hmotnosti 120mg: Pokud provádíme vážení naprosto přesně, pak tato pravděpodobnost je 0, protože u spojité náhodné veličiny nabývající nenulové pravděpodobnosti na nedegenerovaném intervalu[1] platí, že pravděpodobnost jedné konkrétní realizace je nulová. *Srovnej s* hustotou pravděpodobnosti *- ta nulová není, ale aby se dala interpretovat jako* pravděpodobnost*, musí se zintegrovat na alespoň malém intervalu v okolí bodu, který nás zajímá.*

[1] jehož krajní body nesplývají

Pokud vážíme pouze s určitou upřesností, například s přesností na desetiny gramu, bude naměření hmotnosti $120mg$ vlastně znamenat, že skutečná hmotnost pilulky je taková, že bude zaokrouhlena na $120mg$, tedy že je v rozsahu $[120 - 0,1/2,\ 120 + 0,1/2) = [119,95,\ 120,4\overline{9})$. Definujeme-li náhodnou veličinu $Y = X - 120mg$ (takže $EY = 0mg$), bude pravděpodobnost p nálezu pilulky o skutečné hmotnosti X a naměřené hmotnosti $120mg$

$$\begin{aligned} p &= P(119,95 \leq X < 120,05) \\ &= P(120 - 0,05 \leq X < 120 + 0,05) \\ &= P(-0,05 \leq Y < 0,05) \\ &= P(Y < 0,05) - p(Y < -0,05) \end{aligned}$$

a protože $p(Y = y) = 0$, bude

$$\begin{aligned} p &= P(Y \leq 0,05) - P(Y \leq -0,05) \\ &= \Phi(0,05) - \Phi(-0,05) \end{aligned}$$

Nyní využijeme symetrie normovaného normálního rozdělení kolem 0, tedy toho, že

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

a

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

a tak

$$\begin{aligned} p &= \Phi(0,05) - (1 - \Phi(0,05)) \\ &= 2\Phi(0,05) - 1. \end{aligned}$$

Hodnotu $\Phi(0,05)$ v tabulkách nemáme, a tak odhadneme kýženou pravděpodobnost konzervativně raději větší, než skutečnou,[2] pomocí $\Phi(0,06)$:

$$\begin{aligned} p &\approx 2\Phi(0,06) - 1 \\ &\approx 2 \cdot 0,5239 - 1 = 0,0476. \end{aligned}$$

Při nepřesném měření je tedy pravděpodobnost nálezu určité konkrétní realizace náhodné veličiny nenulová, v našem případě měření s přesností na $0,1mg$ je pravděpodobnost nálezu pilulky o hmotnosti $120mg$ necelých 5%.

Zamyslete se, jak by se tato pravděpodobnost změnila, kdybychom měřili s větší přesností? A jak by se změnila, kdyby hmotnost pilulek neměla jednotkový rozptyl, ale větší (např. $10mg^2$*)?*

Příklad: Jak lze vygenerovat (pseudo)náhodné číslo z normálního rozdělení $N(0,1)$, máme-li k dispozici generátor (pseudo)náhodných čísel z intervalu $(0,1)$? A jak lze vygenerovat číslo z lib. daného rozdělení?

Řešení: Ke generování čísel z $N(0,1)$ lze přímo použít kvantilovou funkci Φ^{-1}. Rozmyslete, že čísla $\Phi^{-1}(p)$, kde $p \sim R(0,1)$, jsou realizacemi náhodné veličiny z $N(0,1)$:

Pro zvolené $p \in (0,1)$ generujeme $x = \Phi^{-1}(p)$. Realizacemi jaké náhodné veličiny taková čísla jsou? Stačí určit distribuční funkci $F(x)$ tohoto rozdělení. Protože

$$p = P(X \leq x) = P(X \leq \Phi^{-1}(p)) = F(\Phi^{-1}(p))$$

aplikací $F^{-1}()$ dostáváme

$$F^{-1}(p) = F^{-1}(F(\Phi^{-1}(p)) = \Phi^{-1}(p).$$

Vidíme, že $F^{-1} = \Phi^{-1}$ a tedy $F = \Phi$ a tedy generovaná čísla pocházejí z normovaného normálního rozdělení $N(0,1)$.

Při generování z jiného rozdělení můžeme postupovat analogicky, použitím příslušné kvantilové funkce.

[2]skutečná pravděpodobnost je přibližně rovna $0,0399$

1.14 Charakteristiky rozdělení (střední hodnota, rozptyl, směrodatná odchylka)

Střední hodnota popisuje typickou[3] hodnotu náhodné veličiny. Počítá se jako vážený průměr všech možných relizací náhodné veličiny. Pro spojitou náhodnou veličinu X to je:

$$EX = \int_{-\infty}^{\infty} x f_X(x) dx, \tag{3}$$

pokud tento integrál existuje.

Pro diskrétní náhodnou veličinu X neprůměrujeme přes všechna reálná čísla, ale jen přes jednotlivé diskrétní realizace:

$$EX = \sum_i x_i p(X = x_i) \tag{4}$$

Pozn.: pomocí distribuční funkce můžeme definici sjednotit pro spojitou i diskrétní náhodnou veličinu na:

$$EX = \int_{-\infty}^{\infty} x dF_X(x) dx \tag{5}$$

Vlastnosti střední hodnoty:

$$E(a + bX) = a + bEX \tag{6}$$
$$E(X + Y) = EX + EY \quad \text{(pokud } EX \text{ a } EY \text{ existují)} \tag{7}$$
$$E(XY) = EXEY \quad \text{(právě když jsou } X, Y \text{ nezávislé a pokud } EX \text{ a } EY \text{ existují)} \tag{8}$$

(a, b nenáhodné konstanty, X, Y náhodné veličiny)

Příklad: Uvedené vztahy odvoďte.

Příklad: Jaký očekáváte průměrný počet ok při hodu kostkou? *Tj. spočtěte střední hodnotu odpovídající náhodné veličiny.*

Řešení: $p(X = i) = \frac{1}{6}$ pro $i = 1, \ldots, 6$. Pak

$$EX = \sum_{i=1}^{6} p(X = i) i = \sum_{i=1}^{6} \frac{1}{6} i = \frac{1}{6} \sum_{i=1}^{6} i = \frac{21}{6} = 3,5.$$

Příklad: Vypočtěte střední hodnotu náhodné veličiny dané jako výsledek hodu falšovanou mincí, u níž panna padá 2x častěji, než orel. (Předpokládejme, že pannu reprezentujeme jako 0 a orla jako 1.)

Řešení: Protože $p(X = 0) = \frac{2}{3}$ a $p(X = 1) = \frac{1}{3}$,

$$EX = \sum_{i=1}^{2} p(X = i)\ i = \frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 1 = \frac{1}{3}.$$

Příklad: Spočtěte střední hodnotu náhodné veličiny z rovnoměrného rozdělení na intervalu $(0, 1)$.

Řešení: Protože hustota pravděpodobnosti rovnoměrného rozdělení na intervalu $(0, 1)$ je $f(x) = 1$, je

$$EX = \int_0^1 f(x) x dx = \int_0^1 1 x dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1^2}{2} = 0,5.$$

[3] anglicky *expected*, odtud značení EX

Pozn.: pro obecný interval (a, b) je $f(x) = \frac{1}{b-a}$, a tedy

$$EX = \int_a^b f(x)xdx = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\Big[\frac{x^2}{2}\Big]_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{b+a}{2}$$

Pro charakterizaci kolísání náhodné veličiny kolem střední hodnoty se používá rozptyl $varX$ definovaný jako:

$$varX = E(X-EX)^2 \tag{9}$$

Pozn.: Proč se používá čtverec a ne přímo $E(X-EX)$?
(Protože $E(X-EX) = E(X) - E(EX)) = EX - EX = 0$.)

Lze jednoduše ukázat, že $E(X-a)^2$ je minimalizován při volbě $a = EX$. V tomto smyslu je tedy střední hodnota opravdu „středem“ hodnot náhodné veličiny.

Vlastnosti rozptylu:

$$var(a+bX) = b^2 varX \tag{10}$$
$$varX = (EX^2) - (EX)^2 \tag{11}$$
$$var(X+Y) = varX + varY + 2cov(X,Y) \tag{12}$$
$$var(X+Y) = varX + varY \qquad \text{(pokud } X \text{ a } Y \text{ jsou nezávislé)} \tag{13}$$

(a, b nenáhodné konstanty, X, Y náhodné veličiny).
Přitom

$$cov(X,Y) = E(X-EX)(Y-EY). \tag{14}$$

Příklad: Dokažte tvrzení (10 a 11).

Příklad: Spočtěte rozptyl náhodné veličiny popisující výsledek hodu (nefalšovanou) mincí.

Řešení: Označíme pannu jako 0 a orla jako 1 a můžeme postupovat přímo dosazením do vzorce $varX = E(X-EX)^2$, kde víme, že $EX = 0,5$:

$$var(X) = E(X-EX)^2 = \sum_{i=0}^{1} p(X=i)\ (X-EX)^2 = \frac{1}{2}(0-0,5)^2 + \frac{1}{2}(1-0,5)^2 = 0,5$$

Příklad: Spočtěte rozptyl náhodné veličiny z rovnoměrného rozdělení na intervalu $(0, 1)$.

Řešení: Použijeme výpočetně přívětivější vzorec 11, tedy

$$\begin{aligned} var(X) &= EX^2 - (EX)^2 \\ &= \int_0^1 f(x)x^2dx - \Big(\frac{1}{2}\Big)^2 \\ &= \int_0^1 1x^2dx - \frac{1}{4} \\ &= \Big[\frac{x^3}{3}\Big]_0^1 - \frac{1}{4} \\ &= \frac{1^3}{3} - \frac{1}{4} \\ &= \frac{1}{12} \end{aligned}$$

Příklad: Spočtěte střední hodnotu a rozptyl součtu resp. rozdílu dvou náhodných veličin. Spočítejte obecně a speciálně pro $X, Y \sim N(\mu, \sigma^2)$.

Řešení: Pro součet platí:

$$E(X+Y) = EX + EY,$$

speciálně

$$E(X+Y) = EX + EY = 2\sigma$$

a

$$\begin{aligned}
var(X+Y) &= E(X+Y-E(X+Y))^2 \\
&= E(X+Y-(EX+EY))^2 \\
&= E(X-EX+Y-EY)^2 \\
&= E(X-EX)^2 + 2E(X-EX)(Y-EY) + E(Y-EY)^2 \\
&= var\ X + 2cov(X,Y) + var\ Y \\
&= var\ X + var\ Y \text{ (pokud X,Y jsou nezávislé)}
\end{aligned}$$

speciálně

$$var(X+Y) = var\ X + var\ Y = 2\sigma^2 \text{ (pokud X,Y jsou nezávislé)}$$

Pro rozdíl platí:

$$E(X-Y) = EX - EY,$$

speciálně

$$E(X-Y) = EX - EY = 0$$

a

$$\begin{aligned}
var(X-Y) &= E(X-Y-E(X-Y))^2 \\
&= E(X-Y-(EX-EY))^2 \\
&= E(X-EX-(Y-EY))^2 \\
&= E(X-EX)^2 - 2E(X-EX)(Y-EY) + E(Y-EY)^2 \\
&= var\ X - 2cov(X,Y) + var\ Y \\
&= var\ X + var\ Y \text{ (pokud X,Y jsou nezávislé),}
\end{aligned}$$

speciálně

$$var(X-Y) = var\ X + var\ Y = 2\sigma^2 \text{ (pokud X,Y jsou nezávislé).}$$

Tedy rozptyl součtu i rozdílu nezávislých náhodných veličin je dán jako součet jejich rozptylů. *Jakým způsobem by se dal rozptyl součtu resp. rozdílu zmenšit?*

Příklad: Spočtěte střední hodnotu a rozptyl průměru n nezávislých stejně rozdělených náhodných veličin, pro které platí $EX = \mu$ a $var\ X = \sigma^2$.

Řešení:

$$E\Big(\frac{\sum_{i=1}^n X_i}{n}\Big) = \frac{\sum_{i=1}^n EX_i}{n} = \frac{\sum_{i=1}^n \mu}{n} = \frac{n\mu}{n} = \mu$$

$$var\Big(\frac{\sum_{i=1}^n X_i}{n}\Big) = \frac{var(\sum_{i=1}^n X_i)}{n^2} = \frac{\sum_{i=1}^n var\ X_i}{n^2} = \frac{n\sum_{i=1}^n \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Kolik náhodných veličin musíme zprůměrovat, aby byl rozptyl průměru poloviční ve srovnání s původní náhodnou veličinou? Vidíte možnou aplikaci takového postupu?

1.15 Charakteristiky realizace náhodné veličiny

- (výběrový) průměr (jako náhodná veličina: $\bar{X}$, $\bar{X}_n$, jako realizace náhodné veličiny: $\bar{x}$, $\bar{x}_n$)
- (výběrový) medián ($\widetilde{X}$, $\widetilde{X}_n, \widetilde{x}$, $\widetilde{x}_n$)
- (výběrový) modus
- (výběrový) rozptyl (S^2, s^2)
- (výběrová) směrodatná odchylka (S, s)
- (výběrové) kvantily
- (výběrové) inter-kvartilové rozpětí (IQR)

Ukázat na příkladech.

Příklad: Hmotnost studentů – v modelovém světě graficky vyznačte náhodný výběr (několik realizací náhodné veličiny „hmotnost“) a sumarizujte jej.

Poznamejme, že výběrový rozptyl definovaný jako

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{15}$$

je nevychýleným odhadem rozptylu σ^2. Definujeme-li $\mu = EX$, ukážeme to pomocí tvrzení:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X} + \bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}\left[(X_i-\bar{X})^2 + 2(X_i-\bar{X})(\bar{X}-\mu) + (\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu) + \sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}\underbrace{(X_i-\bar{X})}_{=0} + n(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2
\end{aligned}$$

a tedy

$$\sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2$$

z čehož

$$
\begin{aligned}
ES^2 &= E\bigg(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\bigg) \\
&= \frac{1}{n-1}E\bigg(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\bigg) \\
&= \frac{1}{n-1}\bigg(\sum_{i=1}^{n}E(X_i-\mu)^2 - nE(\bar{X}-\mu)^2\bigg) \\
&= \frac{1}{n-1}\big(n\sigma^2 - n\cdot var\bar{X}\big) \\
&= \frac{1}{n-1}\big(n\sigma^2 - n\frac{\sigma^2}{n}\big) \\
&= \frac{1}{n-1}\big((n-1)\sigma^2\big) \\
&= \sigma^2
\end{aligned}
$$

Někdy se používá vychýlený odhad:

$$
S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{16}
$$

Příklad: Uvažme realizace $x_i, i = 1,\ldots,n$ náhodné veličiny z rozdělení $N(\mu,\sigma^2)$ s neznámým parametrem μ a se známým parametrem σ^2. Dále mějme čtyři odhady střední hodnoty μ: $m_1 = x_1, m_2 = x_{(1)}, m_3 = \frac{\sum_{i=1}^{n} x_i}{n}$ a $m_4 = \frac{\sum_{i=1}^{n} x_i+1}{n}$.
U každého odhadu určete, zda je nestranný, asymptoticky nestranný a konzistentní.

Řešení: $Em_1 = \mu$, takže m_1 je nestranný (a tedy i asymptoticky nestranný). $lim_{n\to\infty} var\ m_1 = \sigma^2$, tedy m_1 není konzistentní.
$Em_2 \neq \mu$, takže m_2 není nestranný (a protože s rostoucím n se odhad nezlepšuje, není ani asymptoticky nestranný). Konzistentní být m_2 nemůže, protože není asymptoticky nestranný (nemusíme tedy již chování rozptylu zkoumat).
$Em_3 = \mu$, takže m_3 je nestranný (a tedy i asymptoticky nestranný). $lim_{n\to\infty} var\ m_3 = lim_{n\to\infty}\frac{\sigma^2}{n} = 0$, tedy m_3 je konzistentní.
$Em_4 = \frac{n\mu+1}{n} \neq \mu$, takže m_4 není nestranný. Protože však

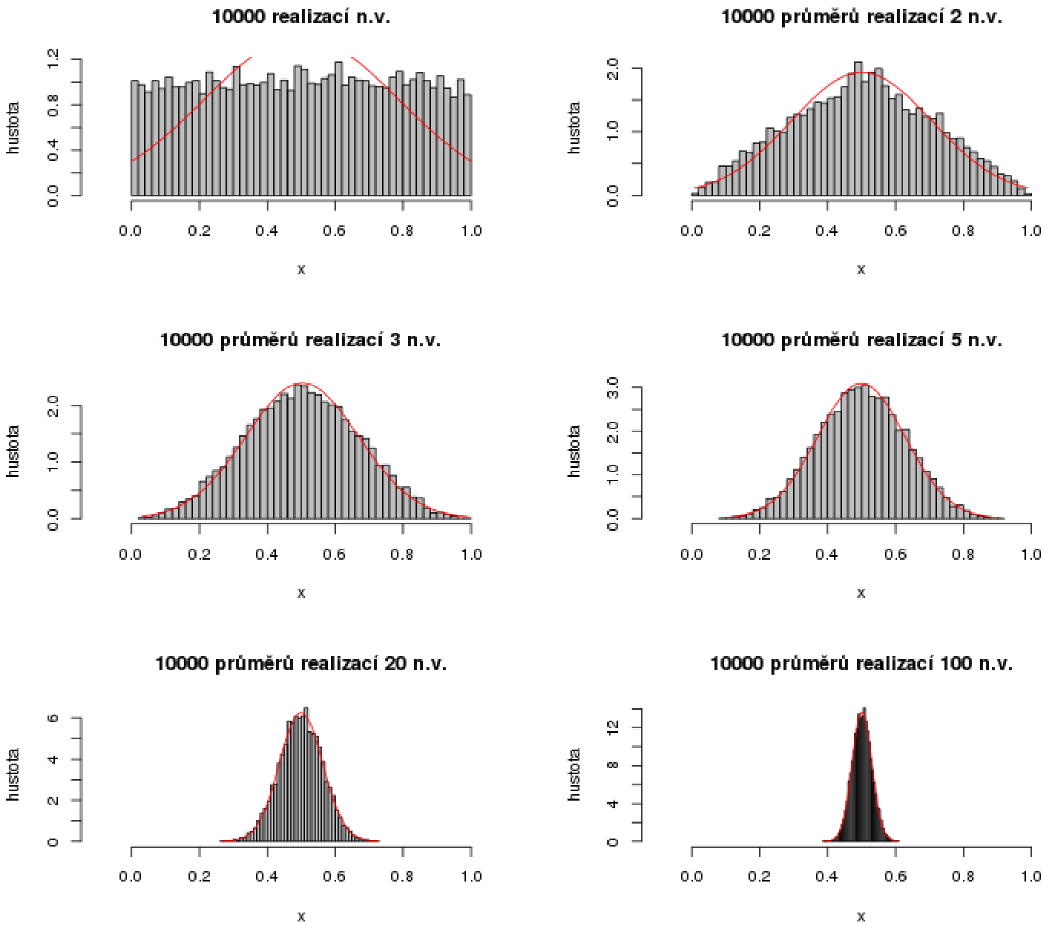
$$
lim_{n\to\infty} Em_4 = lim_{n\to\infty}\frac{n\mu+1}{n} = lim_{n\to\infty}\Big(\mu+\frac{1}{n}\Big) = \mu,
$$

jedná se o asymptoticky nestranný odhad. $lim_{n\to\infty} var\ m_4 = lim_{n\to\infty}\frac{\sigma^2}{n} = 0$, tedy m_4 je konzistentní.

1.16 Centrální limitní věta a význam normálního rozdělení.

Ukázat na simulacích.
Centrální limitní větu ilustruje obr. 2, vygenerovaný zdrojovým kódem v jazyce R (viz následující stránku).

```
# Demonstrace centrální limitní věty (CLV).
# CLV budeme demonstrovat na 'n' průměrech 'm' relizací náhodné veličiny
# generované funkcí 'g'.

# Funkce generující 'n' realizací náhodné veličiny.
# Argumenty:
#   n: velikost výběru
# Vrací: vektor 'n' realizací náhodné veličiny.
g<-function(n) {

  # rovnoměrné rozdělení
  x<-runif(n,0,1)

  # další způsoby generování relizací náhodných veličin jsou
  # zakomentované (lze je jednoduše aktivovat smazáním znaku '#' před nimi)
  # normální  rozdělení
  #x<-rnorm(n,0,1)

  # házení mincí (alternativní rozdělení)
  #x<-rbinom(n,1,.5)

  # trojúhelníkové rozdělení
  #x<-runif(n,0,1)+runif(n,0,1)

  # bimodální rozdělení
  #x<-rnorm(n,.75-.5*(runif(n,0,1)<.5),.1)

  return(x)
}

# Funkce generující 'm' náhodných vektorů délky 'n' a vykreslující histogram
# jejich průměrů spolu s proloženým odhadem hustoty pravděpodobnosti
# normálního rozdělení s parametry odhadnutými z dat.
# Parametry:
#  m - počet vektorů
#  n - délka jednoho vektoru
clv<-function(m,n,titulek) {
  # alokujeme matici typu 'm x n', v 'm' řádcích vektory 'n' realizací náh. veličiny
  x<-matrix(NA,m,n)
  for (i in 1:m) {
    x[i,]<-g(n)
  }
  x<-colMeans(x)
  # histogram
  hist(x, probability=TRUE, breaks=50, col='gray', ylab='hustota',main=titulek,xlim=c(-.1,1.1))
  # proložíme hustotu pravděpodobnosti normálního rozdělení
  ax<-seq(from=min(x), to=max(x), length=100) # body na ose x
  ay<-dnorm(ax, mean(x), sd(x))
  lines(ax, ay, col='red')
}

# počet vektorů
m<-100
# délka jednoho vektoru
n<-10000

options(scipen=5) # čísla chceme vypisovat ve fixní notaci

# vykreslíme 3x2 obrázků
opar<-par(mfrow=c(3,2))
clv(1,n,paste(n,'realizací n.v.'))
clv(2,n,paste(n,'průměrů realizací 2 n.v.'))
clv(3,n,paste(n,'průměrů realizací 3 n.v.'))
clv(5,n,paste(n,'průměrů realizací 5 n.v.'))
clv(20,n,paste(n,'průměrů realizací 20 n.v.'))
clv(m,n,paste(n,'průměrů realizací',m,'n.v.'))
par(opar)
```



Obr. 2: Ilustrace centrální limitní věty. Histogramy ukazují rozdělení výběrů 10.000 náhodných veličin definovaných jako součet n nezávislých realizací náhodné veličiny z rovnoměrného rozdělení $R(0,1)$. Pro $n = 1$ dostáváme přibližně rovnoměrné rozdělení, pro $n = 2$ přibližně trojúhelníkové rozdělení, a pro $n \geq 5$ už v obrázku nerozeznáme, jak se empirické rozdělení liší od normálního, které je pro srovnání zobrazeno jako červená křivka, jejíž parametry byly z jednotlivých výběrů vypočteny momentovou metodou.

1.17 Čebyševova nerovnost

To, jak daleko od střední hodnoty se mohou nacházet hodnoty náhodné veličiny, omezuje následující pozoruhodná Čebyševova nerovnost:

$$p(|X - EX| \geq \epsilon) \leq \frac{varX}{\epsilon^2} \tag{17}$$

Tato nerovnost je zajímavá tím, že platí pro všechny náhodné veličiny (bez ohledu na rozdělení). Zaujmout však může i elegancí a krátkostí svého důkazu (my ji zde však dokážeme jen pro diskrétní rozdělení):

$$
\begin{aligned}
varX &= E(X - EX)^2 \\
&= \sum_i (X_i - EX)^2 p_i \\
&\geq \sum_{i:|X_i - EX| \geq \epsilon} (X_i - EX)^2 p_i \\
&\geq \sum_{i:|X_i - EX| \geq \epsilon} \epsilon^2 p_i \\
&= \epsilon^2 \sum_{i:|X_i - EX| \geq \epsilon} p_i \\
&= \epsilon^2 p(|X_i - EX| \geq \epsilon)
\end{aligned}
$$

a tedy

$$p(|X - EX| \geq \epsilon) \leq \frac{varX}{\epsilon^2}$$

Příklad: Demonstrovat sílu Čebyševovy nerovnosti ve srovnání s tím, co o sobě říká normální rozdělení.

$X \sim N(0, 1)$.

1. Jaká je $p(|X| \geq 0)$?
2. Jaká je $p(|X| \geq 1)$?
3. Jaká je $p(|X| \geq 2)$?
4. Jaká je $p(|X| \geq 3)$?

Očekáváme, že dodatečná informace v podobě přesné znalosti rozdělení náhodné veličiny nám přinese přesnější výsledky.

Řešení:

Uvažme, že $EX = 0$.

1. $p(|X| \geq 0)$: Z Čebyševovy nerovnosti dostáváme

$$p(|X| \geq 0) = p(|X - EX| \geq 0) \leq \frac{varX}{0^2} = \frac{1}{0}$$

a z distribuční funkce

$$
\begin{aligned}
p(|X| \geq 0) &= p(|X - EX| \geq 0) \\
&= 2p(X - EX > 0) \\
&= 2p(X > 0) \\
&= 2(1 - p(X \leq 0)) \\
&= 2(1 - \Phi(0)) \\
&= 2(1 - 0,5) \\
&= 1.
\end{aligned}
$$

Tedy Čebyševova nerovnost vlastně o hledané pravděpodobnosti nepřináší žádnou novou informaci: říká, že pravděpodobnost je nanejvýš nekonečná, což jsme věděli už dříve. Znalost rozdělení naproti tomu dává zcela přesnou odpověď: $p(|X| \geq 0) = 1$.

2. $p(|X| \geq 1)$:

$$p(|X| \geq 1) = p(|X - EX| \geq 1) \leq \frac{varX}{1^2} = \frac{1}{1} = 1$$

a z distribuční funkce

$$\begin{aligned} p(|X| \geq 1) &= p(|X - EX| \geq 1) \\ &= 2p(X - EX \geq 1) \\ &= 2p(X \geq 1) \\ &= 2(1 - \Phi(1)) \\ &\approx 2(1 - 0,84134) \\ &= 0,31732. \end{aligned}$$

3. $p(X \geq 2)$:

$$p(|X| \geq 2) = p(|X - EX| \geq 2) \leq \frac{varX}{2^2} = \frac{1}{4} = 0,25$$

a z distribuční funkce

$$\begin{aligned} p(|X| \geq 2) &= p(|X - EX| \geq 2) \\ &= 2p(X - EX > 2) \\ &\dots \\ &= 0,04550. \end{aligned}$$

4. $p(|X| \geq 3)$:

$$p(|X| \geq 3) = p(|X - EX| \geq 3) \leq \frac{varX}{3^2} = \frac{1}{9} = 0,111\dots$$

a z distribuční funkce

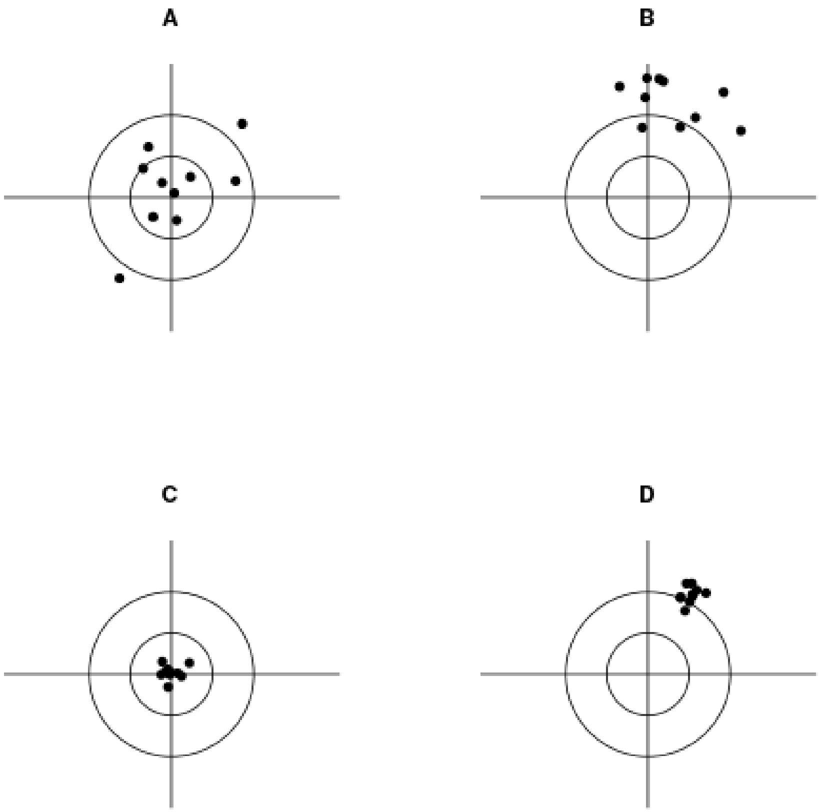
$$\begin{aligned} p(|X| \geq 3) &= p(|X - EX| \geq 3) \\ &= 2p(X - EX > 3) \\ &\dots \\ &= 0,00270. \end{aligned}$$

Na co tedy Čebyševova nerovnost je, když distribuční funkce nám dává mnohem více a přesnějších výsledků?

2 Odhady parametrů.

2.1 Odhad, druhy odhadů, vychýlení a rozptyl odhadu, konzistentní odhad

Příklad: Střelba na terč:



Obr. 3: Ilustrace vlastností odhadů (dvourozměrného) parametru. Odhady jsou znázorněny jako zásahy do terče. Skutečná hodnota parametru odpovídá středu terče, do něhož se jednotlivé odhady snaží více či méně úspěšně „strefit“. A: nevychýlené odhady s velkým rozptylem, B: vychýlené odhady s velkým rozptylem, C: nevychýlené odhady s malým rozptylem, D: vychýlené odhady s malým rozptylem.

2.2 Rozklad střední kvadratické chyby odhadu na systematickou chybu (vychýlení) a rozptyl

θ - skutečná (neznámá) hodnota parametru, T_i - odhady parametrů, ET - střední hodnota odhadů

$$
\begin{aligned}
E(T-\theta)^2 &= E(T-ET+ET-\theta)^2 \\
&= E[(T-ET)^2 + 2(T-ET)(ET-\theta) + (ET-\theta)^2] \\
&= E(T-ET)^2 + 2E(T-ET)\underbrace{(ET-\theta)}_{konstanta} + E\underbrace{(ET-\theta)^2}_{konstanta} \\
&= E(T-ET)^2 + 2(ET-\theta)\underbrace{E(T-ET)}_{=0} + (ET-\theta)^2 \\
&= varT + (ET-\theta)^2
\end{aligned}
$$

tedy střední kvadratická chyba je součtem rozptylu odhadu a čtverce vychýlení.

2.3 Metoda momentů

Příklad: Při vykopávkách bylo odhaleno několik kostí dinosaura. Archeologové mají představu o proporcích dinosaura a chtějí na základě nálezu restaurovat jeho celkovou velikost.

Řešení: Vzít jednu nebo více kostí, nějak chodně je charakterizovat (tj. spočítat nějakou funkci realizací náhodných veličin) a dát je do souvislosti s teoretickým modelem dinosaura (teoretickými protějšky odpovídajících funkcí, např. průměrná velikost atd.).

Je lepší vzít málo, nebo více kostí? (Více - proto se výběr charakterizuje momenty, které stavějí nad hodnotami všech pozorování, a ne třeba nad kvantily (které zohledňují uspořádání, ale ne přímo hodnoty všech jednotlivých realizací).)

Příklad: Odhad parametrů $N(\mu, \sigma^2)$.

2.4 Metoda maximální věrohodnosti

Věrohodnostní funkce (jako funkce parametrů a dat).

Příklad: Nekuřák a nedopalek. Potkáme známého, o němž víme, že je nekuřák. Pod ním však leží nedopalek. Co si pomyslíme - že jej zahodil on, nebo někdo jiný, kdo na daném místě byl před ním?

Věrohodnostní funkce $L(p;x)$ jako funkce dat x = „známý (ne)odhodil nedopalek“ za daných (pevných) parametrů p = „známý je nekuřák“:

- *$L(p$ = známý je nekuřák; x = známý odhodil nedopalek) je nízká,*
- *$L(p$ = známý je nekuřák; x = známý neodhodil nedopalek) je vysoká.*

Z hodnot věrohodnostní funkce vyhodnocené pro různá data („osoba odhodila nedopalek“, „osoba neodhodila nedopalek“) za pevných parametrů („osoba je nekuřák“) usoudíme, že věrohodnějším vysvětlením pozorované skutečnosti je, že nedopalek neodhodil náš známý-nekuřák, ale že nedopalek byl na zemi již dříve.

Příklad: Kuřák a nedopalek. Na rohu ulice vidíme stát člověka, o němž nevíme, zda kouří, nebo nekouří. Pod ním však leží nedopalek. Co si o tom člověku pomyslíme - že spíše kouří, nebo nekouří?

Věrohodnostní funkce $L(p;x)$ jako funkce neznámých parametrů p („osoba je kuřák“ nebo „osoba je nekuřák“) za daných dat (x = „pod osobou je nedopalek“):

- *$L(p$ = osoba je nekuřák; x = pod osobou je nedopalek),*
- *$L(p$ = osoba je kuřák; x = pod osobou je nedopalek).*

Hledáme takovou hodnotu parametru p, která maximalizuje věrohodnostní funkci $L(p;x)$. Takovou hodnotu parametru p pak prohlásíme za maximálně věrohodný odhad parametru p. (V našem případě *bude asi věrohodnějším vysvětlením to, že osoba je kuřák, který odhodil nedopalek. Pokud bychom však měli vážný důvod domnívat se, že nedopalek byl na chodníku již dříve, byla by věrohodnost obou výše uvedených případů stejná a maximálně věrohodný odhad by tak nebyl jednoznačně určen.)*

2.5 Příklady na metodu maximální věrohodnosti a momentovou metodu

Příklad: Z jediné realizace x_1 náhodné veličiny X, o níž víte, že pochází z normálního rozdělení $N(\mu, 1)$, odhadněte parametr μ momentovou metodou a metodou maximální věrohodnosti.

1. Budou odhady nestranné?
2. Jaké budou mít rozptyly?
3. Budou konzistentní?

Kolik a jakých momentů použijete?

Řešení:

1. Momentovou metodou: Stačí jediný moment: $EX = \mu$ odhadneme pomocí x_1: $\hat{\mu} = m_X = \frac{1}{1}\sum_{i=1}^{1} x_i = x_1$. Odhad $\hat{\mu} = x_1$ je nestranný, protože $EX_1 = \mu$. (Všimněte si, že v posledním výrazu vystupuje nikoli realizace x_1, ale náhodná veličina X_1 - pouze z ní má totiž smysl počítat střední hodnotu, protože nás zajímá, jak se odhad bude chovat pro *různé* realizace, nikoli pro jednu jedinou realizaci x_1.)

   Metodou maximální věrohodnosti: věrohodnostní funkce je

   $$L(\mu, \sigma^2;\ x_1) = f_X(x_1|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}},$$

   a my bychom ji chtěli maximalizovat vzhledem k hledaným parametrům (tj. maximalizovat věrohodnost za daných dat a získat hledané neznámé parametry). Protože hledat extrém věrohodnostní funkce není jednoduché, zjednodušíme situaci tím, že místo věrohodnostní funkce budeme maximalizovat její logaritmus (protože logaritmus je ryze monotónní, bod, v němž nabývá věrohodnostní funkce maximum, je přesně týž, jako bod, v němž nabývá maxima logaritmovaná věrohodnostní funkce).

   Logaritmická věrohodnostní funkce je

   $$l(\mu, \sigma^2; x_1) = -\frac{1}{2} ln(2\pi) - \frac{1}{2} ln(\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}. \qquad (18)$$

   Její extrém budeme hledat tak, že položíme její derivaci rovnu 0. Protože se jedná o funkci dvou proměnných, mohli bychom postupně derivovat podle jejích jednotlivých parametrů. Nám bude pro tentokrát stačit nalézt hodnotu parametru μ. Tu nalezneme, položíme-li parciální derivaci podle μ rovnu 0:

   $$\frac{\partial l(\mu, \sigma^2; x_1)}{\partial \mu} = -\frac{2(x_1-\mu)(-1)}{2\sigma^2} = \frac{(x_1-\mu)}{\sigma^2} = 0$$

   a tedy

   $$\hat{\mu} = x_1$$

2. Rozptyl odhadu: $var\hat{\mu} = varX = \sigma^2$.

3. Budou odhady konzistentní? Nelze přímo určit, protože ke konzistenci bychom potřebovali vysledovat chování odhadů pro zvětšující se velikosti výběrů, ale my máme k dispozici pouze jediné pozorování.

Příklad: Ze dvou realizací x_1, x_2 náhodné veličiny X, o níž víte, že pochází z normálního rozdělení $N(\mu, 1)$, odhadněte parametr μ momentovou metodou a metodou maximální věrohodnosti.

Řešení:

1. Momentovou metodou: Opět stačí jediný moment: $EX = \mu$ odhadneme pomocí prvního výběrového momentu: $\hat{\mu} = m_X = \frac{1}{2}\sum_{i=1}^{2} x_i = \frac{x_1+x_2}{2}$. Odhad $\hat{\mu}$ je nestranný, protože $E\frac{X_1+X_2}{2} = \frac{EX_1+EX_2}{2} = \frac{\mu+\mu}{2} = \mu$.

   Metodou maximální věrohodnosti: věrohodnostní funkce je

$$L(\mu, \sigma^2; x_1, x_2) = \prod_{i=1}^{2} f_X(x_i|\mu, \sigma^2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

   potom logaritmická věrohodnostní funkce je

$$\begin{aligned} l(\mu, \sigma^2; x_1, x_2) &= \sum_{i=1}^{2} \Big(-\frac{1}{2} ln(2\pi) - \frac{1}{2} ln(\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \Big) \\ &= -\frac{2}{2} ln(2\pi) - \frac{2}{2} ln(\sigma^2) - \sum_{i=1}^{2} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

   a hodnotu parametru μ nalezneme, položíme-li parciální derivaci podle μ rovnu 0:

$$\begin{aligned} \frac{\partial l(\mu, \sigma^2; x_1, x_2)}{\partial \mu} &= \sum_{i=1}^{2} \frac{2(x_i-\mu)(-1)}{2\sigma^2} = \sum_{i=1}^{2} \frac{(x_i-\mu)}{\sigma^2} = 0 \\ \hat{\mu} = \frac{\sum_{i=1}^{2} x_i}{2} = \bar{x} \end{aligned}$$

2. Rozptyl odhadu: $var\hat{\mu} = var\frac{X_1+X_2}{2} = \frac{varX_1+varX_2}{4} = \frac{\sigma^2+\sigma^2}{4} = \frac{\sigma^2}{2}$. Tedy tento odhad bude mít rozptyl poloviční ve srovnání s odhadem pořízeným z jediné realizace.

3. Budou odhady konzistentní? Nelze přímo určit, protože ke konzistenci bychom potřebovali vysledovat chování odhadů pro zvětšující se velikosti výběrů, ale my máme k dispozici pouze dvě pozorování.

Příklad: (Pokračování přechozího příkladu.) Co kdybyste pro odhad parametr použili pouze poslední realizaci x_2?

- Jak se změní odhady?
- Budou lepší, než minulé odhady?
- Budou nestranné?
- Jaké budou mít rozptyly?
- Budou konzistentní?

Řešení: Dostali bychom stejný odhad jako v případě, že máme jedinou realizaci. Odhad střední hodnoty bude nestranný, ale ve srovnání s odhadem založeným na dvou realizacích bude mít dvojnásobný rozptyl.

Příklad: (Pokračování přechozího příkladu.) Co kdybyste pro odhad parametrů použili větší z realizací, tj. $max(x_1, x_2)$? Jak se změní odhady?

Řešení: Odhad tentokrát nebude nestranný - tím, že za odhad bereme větší ze dvou realizací, střední hodnota takového odhadu bude větší, než střední hodnota samotné realizace.

Příklad: Z jediné realizace x_1 náhodné veličiny X o níž víte, že pochází z normálního rozdělení $N(\mu, \sigma^2)$, odhadněte parametry μ a σ^2 momentovou metodou a metodou maximální věrohodnosti.

Řešení:

Momentovou metodou oba parametry z jediné relizace odhadnout nelze.

Metodou maximální věrohodnosti: parametr μ odhadneme hodnotou dané relizace x_1 (viz výše). Parametr σ^2 odhadneme maximalizací logaritmické věrohodnostní funkce (18) vzhledem k σ^2, hledáme tedy extrém $l(\mu, \sigma^2; x_1)$ vzhledem k σ^2, tedy parciální derivaci $l(\mu, \sigma^2; x_1)$ podle σ^2 položíme rovnu nule:

$$\frac{\partial l(\mu, \sigma^2; x_1)}{\partial \sigma^2} = -\frac{1}{2}\frac{1}{\sigma^2} + \frac{(x_1 - \mu)^2}{2(\sigma^2)^2} = 0$$

a po úpravě (vynásobení $2(\sigma^2)^2$)

$$\sigma^2 = (x_1 - \mu)^2$$

a vzhledem k tomu, že $\hat{\mu} = x_1$,

$$\sigma^2 = (x_1 - x_1)^2 = 0.$$

Vidíme, že věrohodnost se maximalizuje pro $\widehat{\sigma^2} = 0$, tedy pro degenerované (singulární) rozdělení s nulovým rozptylem, v němž je všechna hustota soustředěna v jediném bodu x_1.

Souvislost s uvíznutím E-M algoritmu v lokálním minimu odpovídajícím singulárním variančním maticím.

Příklad: Odhad parametrů $N(\mu, \sigma^2)$ z více realizací.

Řešení: viz přednáška

Příklad: Odhad parametrů alternativního rozdělení $Alt(p)$ z realizací $0, 0, 1$.

Řešení: Alternativního rozdělení je popsáno pravděpodobnostmi

$$f_{Alt}(X = 1) = p$$

a

$$f_{Alt}(X = 0) = 1 - p,$$

kde p je pravděpodobnost úspěchu (jevu 1).

Věrohodnostní funkce pak je součin pravděpodobností jednotlivých realizací za dané hodnoty parametru p:

$$\begin{aligned} L(p; \{0, 0, 1\}) &= \prod_{i=1}^{3} f_{Alt}(x_i|p) \\ &= (1-p) \cdot (1-p) \cdot p \\ &= (1-p)^2 p. \end{aligned}$$

Tuto funkci bychom mohli již přímo maximalizovat (to by vedlo ke kvadratické rovnici), nebo můžeme podobně jako dříve tuto funkci logaritmovat a úlohu si zjednodušit.

Logaritmická věrohodnostní funkce pak je

$$l(p; \{0, 0, 1\}) = 2ln(1-p) + ln(p)$$

a její derivace podle p položíme rovnu 0:

$$\frac{l(\{0,0,1\},p)}{\partial p} = 2\frac{-1}{1-p} + \frac{1}{p} = 0$$

a po úpravě (vynásobení $p(1-p)$)

$$\begin{aligned} 2\hat{p} &= (1-\hat{p}) \\ 3\hat{p} &= 1 \\ \hat{p} &= \frac{1}{3} \end{aligned}$$

Odhadem parametru p metodou max. věrohodnosti je tedy $\hat{p} = 1/3$.

Poznámka: ke stejnému výsledku by vedlo i to, pokud bychom daný počet jedniček ve sledu tří nezávislých realizací náhodné alternativní veličiny považovali za jedinou realizaci náhodné veličiny mající binomické rozdělení. Věrohodnostní funkce by pak byla

$$L(p;\{0,0,1\}) = \binom{3}{1} p(1-p)^2,$$

což by vedlo ke stejnému výsledku.

Příklad: Házíme mincí, u níž se obáváme, že je falešná: panna údajně padá 2x častěji než orel. Spočtěte věrohodnost tohoto tvrzení na základě pozorování, že padl 2x orel. Spočtěte rovněž věrohodnost pozorovaných dat pro případ, že mince není falešná. Dále z dat odhadněte pravděpodobnost, že padá panna, metodou maximální věrohodnosti a momentovou metodou.

Řešení: Náhodný jev modelujeme alternativním rozdělením s parametrem p vyjadřujícím pravděpodobnost, že padne panna. Pak věrohodnost pozorovaného sledu dvou orlů je:

$$L(p;x) = L(p;\{0,0\}) = (1-p)^2.$$

Pro očekávanou falešnou minci, kde $p = \frac{2}{3}$ dostáváme $L(\frac{2}{3};\{0,0\}) = \frac{1}{9}$. Pro férovou minci, kde $p = \frac{1}{2}$ dostáváme $L(\frac{1}{2};\{0,0\}) = \frac{1}{4}$. Věrohodnější tedy je, že mince je férová.

Odhad p metodou maximální věrohodnosti: věrohodnost $L(p;x) = L(p;\{0,0\}) = (1-p)^2$ se maximalizuje pro $\hat{p} = 0$.

Odhad p momentovou metodou: První obecný (necentrální) moment alternativního rozdělení

$$M_1^{'} = EX = p$$

položíme roven odpovídajícímu výběrovému obecnému momentu $m_1^{'} = \bar{x}$, tedy:

$$p = M_1^{'} = m_1^{'} = \bar{x}$$

a odtud přímo dostáváme odhad $\hat{p}$:

$$\hat{p} = \bar{x} = 0.$$

Oba odhady si jsou tedy rovny.

Příklad: Metodou maximální věrohodnosti a momentovou metodou odhadněte na základě pozorovaných $x_1, \ldots, x_n$ neznámý parametr λ Poissonova rozdělení:

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Řešení:

Metodou maximální věrohodnosti:

$$L(\lambda;\{x_1,\dots,x_n\}) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\ell(\lambda;\{x_1,\dots,x_n\}) = \sum_{i=1}^{n} (x_i \ln\lambda - \lambda - \ln x_i!)$$

$$\ell(\lambda;\{x_1,\dots,x_n\}) = \ln\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln x_i!$$

$$\frac{\partial \ell(\lambda;\{x_1,\dots,x_n\})}{\partial\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Momentovou metodou: První obecný moment Poissonova rozdělení

$$M_1^{'} = EX = \lambda$$

položíme roven odpovídajícímu výběrovému obecnému momentu $m_1^{'} = \bar{x}$, tedy:

$$\lambda = M_1^{'} = m_1^{'} = \bar{x}$$

a odtud přímo dostáváme odhad

$$\hat{\lambda} = \bar{x}.$$

Oba odhady si jsou tedy rovny.

Příklad: Metodou maximální věrohodnosti a momentovou metodou odhadněte na základě pozorovaných $x_1,\dots,x_m$ neznámý parametr p binomického rozdělení $Bi(n = 10, p)$:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Řešení: Metodou maximální věrohodnosti:

$$L(p;\{x_1,\dots,x_m\}) = \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}$$

$$\ell(p;\{x_1,\dots,x_m\}) = \sum_{i=1}^{m} \left(\ln\binom{n}{x_i} + x_i \ln p + (n - x_i)\ln(1-p) \right)$$

$$\frac{\partial \ell(p;\{x_1,\dots,x_m\})}{\partial p} = \frac{\sum_{i=1}^{m} x_i}{p} + \frac{\sum_{i=1}^{m}(n - x_i)}{1-p}$$

$$\frac{\partial \ell(p;\{x_1,\dots,x_m\})}{\partial p} = \frac{\sum_{i=1}^{m} x_i}{p} + \frac{mn - \sum_{i=1}^{m} x_i}{1-p} = 0$$

$$\sum_{i=1}^{m} x_i - p\sum_{i=1}^{m} x_i = mnp - p\sum_{i=1}^{m} x_i$$

$$\sum_{i=1}^{m} x_i = mnp$$

$$\hat{p} = \frac{\sum_{i=1}^{m} x_i}{mn}$$

Momentovou metodou: První obecný moment binomického rozdělení

$$M_1^{'} = EX = np$$

položíme roven odpovídajícímu výběrovému obecnému momentu $m_1^{'} = \bar{x}$, tedy:

$$np = M_1^{'} = m_1^{'} = \bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$$

a odtud přímo dostáváme odhad

$$\hat{p} = \frac{\bar{x}}{n} = \frac{\sum_{i=1}^{m} x_i}{mn}$$

Oba odhady si jsou tedy opět rovny.

Příklad: Metodou maximální věrohodnosti a momentovou metodou odhadněte na základě pozorovaných $x_1, \dots, x_n$ neznámé parametry a, b rovnoměrného rozdělení $R(a, b)$ na intervalu $[a, b]$).

Řešení: Metodou maximální věrohodnosti:

$$L(a, b; \{x_1, \dots, x_n\}) = \prod_{i=1}^{n} \frac{1}{b-a} = \left(\frac{1}{b-a}\right)^n$$

$$\ell(a, b; \{x_1, \dots, x_n\}) = n \ln \frac{1}{b-a} = -n \ln(b-a).$$

Tato věrohodnost se maximalizuje pro co nejmenší $b - a$. Ovšem $b - a$ musí být natolik veliké, aby interval $[a, b]$ pokryl všechna pozorování x_i. Tedy vychází, že

$$\hat{a} = min_{i=1}^{n} x_i$$

$$\hat{b} = max_{i=1}^{n} x_i$$

Momentovou metodou: První obecný moment rovnoměrného rozdělení

$$M_1^{'} = EX = \frac{a+b}{2}$$

položíme roven odpovídajícímu výběrovému obecnému momentu $m_1^{'} = \sum_{i=1}^{n} x_i = \bar{x}$, tedy:

$$\frac{a+b}{2} = M_1^{'} = m_1^{'} = \bar{x}.$$

a dostáváme

$$\frac{\widehat{a+b}}{2} = \bar{x}.$$

Víme (nebo vypočítáme), že druhý obecný moment rovnoměrného rozdělení je

$$M_2^{'} = EX^2 = E(X - EX)^2 + (EX)^2 = varX + (EX)^2 = \frac{(b-a)^2}{12} + (M_1^{'})^2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2.$$

a po dosazení z předešlé rovnice tedy

$$M_2^{'} = \frac{(b-a)^2}{12} + (\bar{x})^2.$$

Přitom $varX$ se spočítá přímo z definice jako:

$$varX = E(X - EX)^2 = \int_{-\infty}^{\infty} f(x)(x - EX)^2 dx = \int_a^b f(x)(x - EX)^2 dx = \int_a^b \frac{1}{b-a} \left(x - \frac{a+b}{2}\right)^2 dx$$

$$\overset{h=\frac{b-a}{2},\ y=x-\frac{a+b}{2}}{=} \int_{-h}^{h} \frac{1}{2h} y^2 dy = \frac{1}{2h} \left[\frac{y^3}{3}\right]_{-h}^{h} = \frac{1}{2h} \left[\frac{h^3}{3} - \frac{(-h)^3}{3}\right] = \frac{1}{2h} \left[\frac{2h^3}{3}\right] = \frac{h^2}{3} = \frac{(b-a)^2}{12}.$$

Druhý obecný moment rovnoměrného rozdělení pak položíme roven odpovídajícímu výběrovému obecnému momentu $m_2' = \sum_{i=1}^{n} x_i^2$, tedy:

$$\frac{(b-a)^2}{12} + (\bar{x})^2 = \sum_{i=1}^{n} x_i^2$$
$$\widehat{b-a} = \sqrt{12\left(\sum_{i=1}^{n} x_i^2 - (\bar{x})^2\right)}$$

V tomto případě se tedy odhady momentovou metodou a metodou maximální věrohodnosti liší.

2.6 Intervalové odhady

Příklad: Na pracovní schůzce se svým šéfem - lékařem internistou dostáváte za úkol odhadnout co nejpřesněji glykémii (koncentraci glukózy v krvi $[mmol/l]$) u pacientů s určitou formou těžké cukrovky. Víte, že u zdravých lidí jsou hodnoty glykemie typicky v rozmezí cca $3-6 mmol/l$, ale u pacientů s danou nemocí se očekává, že glykemie bude nabývat mnohem vyšších hodnot, u všech pacientů podobných. Lékaři odhadují, že směrodatná odchylka naměřených hodnot koncentrace glukózy u sledovaných pacientů je $s = 4 mmol/l$. Navíc prozatím máte k dispozici pouze jediné měření: $x_1 = 11,3 mmol/l$.

1. Vhodným způsobem graficky znázorněte rozložení pravděpodobnosti náhodné veličiny „glykémie u daných pacientů“ - načtrtněte hustotu.
2. Naznačte oblast A, pro kterou $P(X \in A) = 95\%$. Je oblast symetrická? Lze najít jinou takovou oblast? Interpretujte danou oblast. Ověřte, že jste oblast našli správně pomocí numerické simulace.
3. Zkonstruujte 95% interval spolehlivosti pro střední hodnotu glykemie. Interpretujte daný interval. Jste spokojeni s jeho přesností? Přesnost je malá – jak to zlepšit?

Řešení: Dobrým modelem pro rozdělení pravděpodobnosti hodnot glykémie u sledovaných pacientů může být normální rozdělení, protože na základě centrální limitní věty víme, že pokud sledovanou veličinu ovlivňuje podobným způsobem více nezávislých faktorů, které se navzájem kombinují, výsledkem je (při velkém počtu faktorů) normální rozdělení. (Pokud totiž např. každý z faktorů může nezávisle na ostatních buď zvyšovat nebo snižovat sledovanou hodnotu (a zvýšení i snížení je stejně pravděpodobné), nejpravděpodobnější situace je taková, že polovina faktorů působí negativně a polovina pozitivně. Situace, že by např. všechny faktory působily snížení, je velmi nepravděpodobná.)

Oblast A je definována jako interval (q_d, q_h), kde $F_X(q_h) - F_X(q_l) = 0,95$.

Oblasti mohou být různé, typicky však budeme mluvit o symetrické oblasti (pro oboustrannou alternativu, tj. když chyby na obou stranách jsou stejně důležité či očekávatelné).

Interpretace oblasti A: při opakovaných náhodných výběrech budou realizace náhodné veličiny glykémie ležet v oblasti A v 95% případů. V tomto pohledu je tedy pravděpodobnost, že daná náhodná veličina nabývá hodnoty z daného intervalu, rovna 95%. Viz obr. 4.

Podobně lze konstruovat oblast, v níž se budou často nacházet průměry náhodných výběrů o dané velikosti (viz obr. 4).

Inteval spolehlivosti pro střední hodnotu naproti tomu bude definován kolem výběrového průměru a bude roven $(\bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}), \bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2}))$.

Příklad: (Pokračování.) Na další schůzce jste seznámeni s požadavkem odhadnout glykémii u dané nemoci s danou přesností – s přesností na 1 desetinné místo. Kolik pacientů budete muset vyšetřit?

Řešení: Tento požadavek nelze splnit. Co totiž znamená požadavek “naměřit glykémii s přesností na 1 desetinné místo”? Patrně znamená, že chceme, aby naměřená hodnota glykémie včetně nepřesnosti v jejím odhadu (tedy její interval spolehlivosti) ležel(a) v intervalu, jehož oba konce, zaokrouhleny na jedno desetinné místo, budou totožné. Protože však interval spolehlivosti konstruovaný kolem výběrového průměru má pro konečné výběry nenulovou šířku, a protože výběrový průměr může ležet velmi blízko hranice, kde se láme zaokrouhlování, nelze obecně (dopředu) zaručit, že by se konce intervalu spolehlivosti nezaokrouhlovaly na různá čísla (např. při výběrovém průměru $11,349mmol/l$ by interval spolehlivosti mohl být $[11,348,\ 11,350)mmol/l$, tedy po zaokrouhlení na $[11,3,\ 11,4)mmol/l$ a bychom glykémii s přesností na 1 desetinné místo nezískali.

Příklad: (Pokračování.) Na další schůzce je navrženo, abyste garantovali, že vzdálenost odhadu od skutečné hodnoty nebude větší než $0,1mmol/l$. Spočtěte počet pacientů, které je třeba vyšetřit.

Řešení: Tento požadavek také nelze splnit. Při konstrukci interval spolehlivosti z konečného počtu pozorování nelze omezit pravděpodobnost chyby na 0% (tj. zajistit 100% spolehlivost).

Příklad: (Pokračování.) Na další schůzce je tedy požadováno, abyste alespoň garantovali, že vzdálenost odhadu od skutečné hodnoty nebude s pravděpodobností 95% větší než $0,1mmol/l$. Spočtěte počet pacientů, které je třeba vyšetřit.

Řešení: Interval spolehlivosti kolem výběrového průměru $\bar{X}$ z výběru velikosti n je při známém rozptylu σ^2 roven

$$\left(\bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\frac{\alpha}{2}), \bar{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})\right).$$

Chceme, aby

$$\frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2}) = 0,1$$

a tedy

$$n = \left(\frac{\sigma}{0,1}\Phi^{-1}(1-\frac{\alpha}{2})\right)^2$$
$$n \approx (\frac{4}{0,1}1,96)^2 = 78,4^2 = 6146,56$$

K odhadu, jehož vzdálenost od skutečné hodnoty nebude vyšší než $0,1$, bychom tedy potřebovali alespoň 6147 pacientů. Je to reálné?

Příklad: (Pokračování.) Počet pacientů vypočítaný v minulém příkladě je v běžných podmínkách nereálný. Jaký odhad by se dal zkonstruovat v případě, že počet pacientů by byl 100-krát nižší?

Řešení: Vyjdeme ze vztahu velikosti výběru n, šířky intervalu spolehlivosti Δ a požadované spolehlivosti α:

$$n = \left(\frac{\sigma}{\Delta}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2. \tag{19}$$

Pro stonásobně snížení n bychom mohli:

- zvětšit šířku intervalu spolehlivosti Δ desetkrát, nebo
- v případě, že směrodatná odchylka je z velké míry určena chybou měření (nikoli biologickou variabilitou), požadovat vyšší přesnost měření, tedy snížit rozptyl σ^2 100-krát (tedy směrodatnou odchylku σ 10-krát), nebo
- snížit z nároků na spolehlivost daného intervalu spolehlivosti, a kýženou spolehlivost bychom

mohli spočítat z 19 jako

$$\begin{aligned}
\sqrt{n} &= \frac{\sigma}{\Delta}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\\
\sqrt{n}\frac{\Delta}{\sigma} &= \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\\
1-\frac{\alpha}{2} &= \Phi\left(\sqrt{n}\frac{\Delta}{\sigma}\right) = 2\left(1-\Phi\left(\sqrt{n}\frac{\Delta}{\sigma}\right)\right)\\
\alpha &= 2\Phi\left(-\sqrt{n}\frac{\Delta}{\sigma}\right)\\
\alpha &= 2\Phi\left(-\sqrt{62}\frac{0,1}{4}\right) \approx 2\Phi(-0,1969)\\
\alpha &\approx 0,844
\end{aligned}$$

(Tímto způsobem bychom tedy zkonstruovali $15,6\%$ interval spolehlivosti. (Jak byste jej interpretovali? Jak je užitečný?))

Příklad: (Intervalový odhad při neznámém rozptylu.) Měření systolického krevního tlaku 15 osob dalo průměrnou hodnotu $116,3mmHg$ a výběrovou směrodatnou odchylku $5,4mmHg$. Vypočtěte 95% interval spolehlivosti střední hodnoty krevního tlaku.

Jaký interval byste dostali, kdybyste hodnotu výběrové směrodatné odchylky považovali za pevnou?

Řešení: Interval spolehlivosti kolem výběrového průměru $\bar{X}$ z výběru velikosti n je při neznámém rozptylu odhadnutém jako s_X^2 roven

$$\left(\bar{X}+\frac{s_X}{\sqrt{n}}qt_{(n-1)}\left(\frac{\alpha}{2}\right), \bar{X}+\frac{s_X}{\sqrt{n}}qt_{(n-1)}\left(1-\frac{\alpha}{2}\right)\right)$$

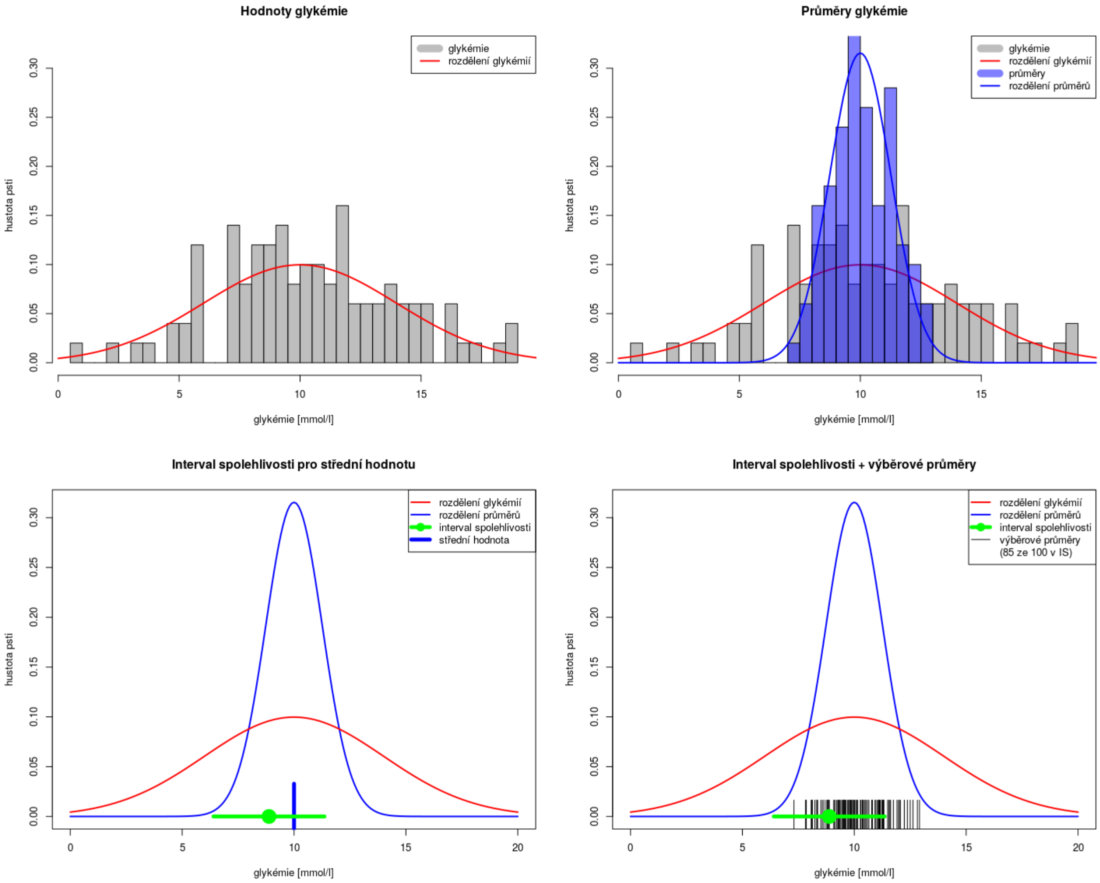
a vyčíslen je na

$$\begin{aligned}
&\left(116,3+\frac{5,4}{\sqrt{15}}qt_{14}(2,5\%), 116,3+\frac{5,4}{\sqrt{15}}qt_{14}(97,5\%)\right)\\
&\approx 116,3+1,394\cdot(-2,145); 116,3+1,394\cdot 2,145)\\
&\approx (116,3-2,99; 116,3+2,99)\\
&\approx (113,31; 119,29)
\end{aligned}$$

Pro srovnání: pokud bychom předem znali rozptyl (a nemuseli jej odhadovat z dat) a náhodou by směrodatná odhylka byla rovna hodnotě výběrové směrodatné odchylky, dostali bychom interval:

$$\begin{aligned}
&\left(116,3+\frac{5,4}{\sqrt{15}}\Phi^{-1}(2,5\%), 116,3+\frac{5,4}{\sqrt{15}}\Phi^{-1}(97,5\%)\right)\\
&\approx 116,3+1,394\cdot(-1,96); 116,3+1,394\cdot 1,96)\\
&\approx (116,3-2,73; 116,3+2,73)\\
&\approx (113,57; 119,03)
\end{aligned}$$

Všimněte si, že tento interval se liší pouze v použité kvantilové funkci a že při stejné hodnotě směrodatné odchylky (teroretické i výběrové) vychází interval spolehlivosti širší v případě neznámého (odhadovaného) rozptylu, než v případě, že rozptyl je předem znám. Proč?



Obr. 4: Ilustrace intervalu spolehlivost pro střední hodnotu.