

Pokroky matematiky, fyziky a astronomie

Martin Otava

Stručný průvodce statistickými intervaly

Pokroky matematiky, fyziky a astronomie, Vol. 62 (2017), No. 1, 7–16

Persistent URL: <http://dml.cz/dmlcz/146719>

Terms of use:

© Jednota českých matematiků a fyziků, 2017

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://dml.cz>

Stručný průvodce statistickými intervaly

Martin Otava, Beerse

Abstrakt. Statistické intervaly představují rozšířenou metodu k popisu nejistoty, avšak jejich přesná interpretace může být pro laika obtížná. Následující článek se zabývá třemi typy intervalů — konfidenčními, predikčními a tolerančními intervaly. Každý z nich má jinou funkci a má smysl ho používat jen v jistém kontextu. Téma je představeno spíše neformálně a poskytuje čtenáři základní přehled jednotlivých typů intervalů. Důraz je kladen především na správnou interpretaci a pochopení nejčastějších chyb spojených s používáním statistických intervalů. Vlastnosti jednotlivých typů intervalů jsou demonstrovány na příkladech z farmaceutické výroby tablet.

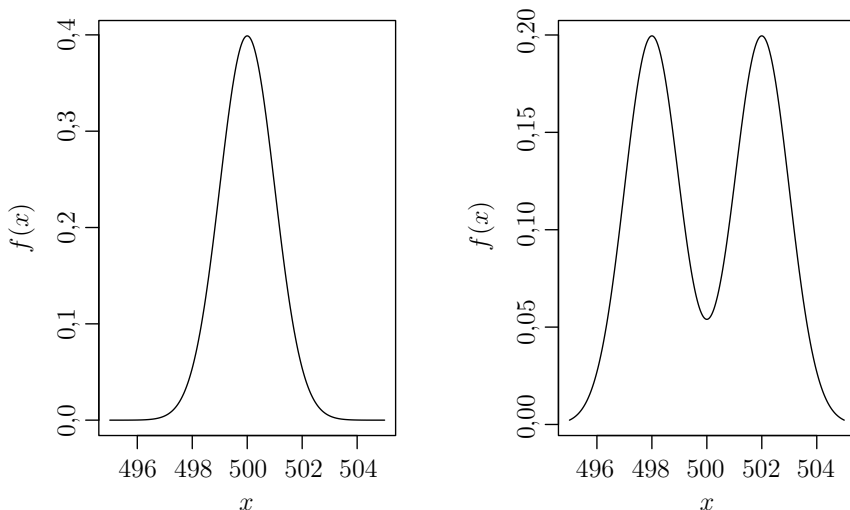
1. Úvod

Statistika je velmi často neodbornou veřejností chápána jako deskriptivní statistika popisující vlastnosti sebraného datového souboru, například pozorované maximum či napozorovaný aritmetický průměr. Popis napozorovaných dat je však jen malou částí statistiky jako disciplíny. Pokud data, která máme k dispozici, nereprezentují celou populaci (celá populace může být dostupná například v případě národního censu), pak je obvyklým cílem statistika vyhodnotit, co můžeme vyvodit o celé populaci na základě omezeného datového souboru. Příkladem mohou být nejrůznější průzkumy veřejného mínění či volebních preferencí. Takový postup vždy vyžaduje jisté předpoklady (např. skutečnost, že náš vzorek dobře reprezentuje celkovou populaci) a jakákoli tvrzení o celé populaci jsou nevyhnutelně zatížena chybou. Statistika nám umožňuje tuto nejistotu popsat a vyjádřit kvantitativně.

Statistické intervaly jsou velice užitečným nástrojem při vyjadřování nejistoty. Existuje několik různých typů intervalů a každý se hodí k odpovědi na jiné otázky. Pro laika tak může být jejich interpretace obtížná. V tomto příspěvku se seznámíme s konfidenčními, predikčními a tolerančními intervaly a vysvětlíme si, jaké jsou mezi nimi rozdíly a jak je správně používat. Ukážeme, jakým způsobem konfidenční intervaly vyjadřují nejistotu při odhadu parametrů rozdělení, zatímco predikční a toleranční intervaly se zabývají předpovědi budoucích pozorování. Další rozdíl spočívá v tom, že predikční intervaly se soustředí na předpovědi pro jediné budoucí pozorování a toleranční intervaly jsou vhodné pro velké počty budoucích pozorování. Jako příklad pro demonstraci vlastností jednotlivých typů intervalů nám poslouží farmaceutická výroba tablet.

Přestože bude nutné zavést několik pojmů, omezíme se na jejich neformální vysvětlení za účelem zjednodušení výkladu; přesné definice lze najít v [1].

Mgr. MARTIN OTAVA, Ph.D., Janssen Pharmaceutica, Turnhoutseweg 30, 2340 Beerse, Belgie,
e-mail: motava@its.jnj.com



Obr. 1. Příklady hustot spojitých náhodných veličin: normální rozdělení (vlevo), bimodální rozdělení (vpravo)

2. Základní pojmy

2.1. Náhodná veličina a její vlastnosti

Mezi základní pojmy statistiky patří náhodné veličiny a jejich rozdělení. Náhodná veličina reprezentuje jev, o který se zajímáme, a rozdělení pak udává, jaké jsou přípustné hodnoty pro náhodnou veličinu a jaká je pravděpodobnost těchto hodnot.

Příkladem nám bude farmaceutická výroba hypotetické 550 mg tablety s obsahem účinné látky 500 mg. Zakoupíme-li balení v lékárně a změříme obsah účinné látky v laboratoři, je rozumné předpokládat, že skutečná hodnota obsažená v každé tabletě bude přibližně 500 mg, ale nemělo by nás překvapit, že nemusí být přesně 500 mg. Výsledek měření tablety je pak právě náhodnou veličinou, neboť jeho hodnota není předem známa. Rozdělení této veličiny udává, jaké výsledky můžeme očekávat (a s jakou pravděpodobností). V našem případě může například z rozdělení vyplývat, že obsah účinné látky mezi 490 a 510 mg je mnohem pravděpodobnější než obsah nižší než 100 mg.

Označme náhodnou veličinu jako X a její napozorovanou hodnotu jako x . Pokud měříme $N = 10$ tablet, výsledky měření jsou označeny x_1, \dots, x_{10} . Rozdělení spojitě náhodné veličiny (jakou je právě obsah účinné látky) je reprezentováno hustotou f , tj. funkcí, jejíž integrál od a do b vyjadřuje pravděpodobnost, že veličina X má hodnotu v intervalu (a, b) :

$$P(a < X < b) = \int_a^b f(x) dx. \quad (1)$$

Integrál funkce na celém oboru reálných čísel se pak rovná jedné, neboť reprezentuje pravděpodobnost $P(-\infty < X < \infty)$. Příklady hustot, které by mohly reprezentovat rozdělení obsahu účinné látky, můžeme vidět na obr. 1.

Střední hodnota značená $\mu = \mathbb{E}(X)$ neformálně řečeno reprezentuje průměrnou hodnotu veličiny X a v případě spojitě náhodné veličiny je opět vyjádřena integrálem

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Musíme být opatrní s interpretací, neboť střední hodnota ne vždy reprezentuje hodnotu typickou či nejčastější. Obě hustoty na obr. 1 vedou ke střední hodnotě rovné 500 mg, ale v případě pravého panelu je jen velmi malá pravděpodobnost, že náhodně vybraná tableta bude mít obsah účinné látky přibližně 500 mg; mnohem pravděpodobnější jsou hodnoty kolem 498 či 502 mg.

Rozptyl náhodné veličiny $\text{var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$ reprezentuje průměr z druhých mocnin odchylek realizací náhodné veličiny od její střední hodnoty. Přestože tedy obě hustoty na obr. 1 mají stejnou střední hodnotu, hustota na pravém panelu by vedla k mnohem většímu rozptylu, neboť její hodnoty se typicky nacházejí dále od střední hodnoty $\mathbb{E}(X)$. Hustotu na pravém panelu nazýváme bimodální, neboť má dva mody, tj. dvě lokální maxima.

Jako rozdělení veličiny X pro náš příklad s obsahem účinné látky v tabletách zvolíme tzv. normální rozdělení. Hustota tohoto rozdělení má daný tvar, až na dva neznámé parametry: střední hodnotu $\mu = \mathbb{E}(X)$ a rozptyl $\sigma^2 = \text{var}(X)$,

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

Známe-li parametry μ a σ^2 , můžeme spočítat pravděpodobnost pro jakýkoli interval ze vzorce (1). Hustota pro $\mu = 500$ a $\sigma^2 = 1$ je znázorněna v levé části obr. 1.

Další užitečnou charakteristikou náhodné veličiny jsou její kvantily. Pro $\kappa \in (0, 1)$ definujeme $100\kappa\%$ kvantil tak, že $100\kappa\%$ možných hodnot veličiny X je menších než tento kvantil. Formálně je tedy kvantil z_κ veličiny X taková hodnota, že $P(X \leq z_\kappa) = \kappa$. Dodejme, že definici v této formě lze použít pouze v případě spojitých náhodných veličin.

2.2. Odhady parametrů

Statistická teorie odhadu se zaměřuje na odhadování parametrů rozdělení, které jsou nám neznámé. Volba konkrétního rozdělení pro popis daného jevu (tedy v našem případě volba normálního rozdělení pro obsah účinné látky) je většinou založena na nějaké předchozí zkušenosti s daným jevem a na vlastnostech náhodné veličiny X . Existují způsoby, jak ověřit, zda volba rozdělení dává smysl, či odhadnout hustotu rozdělení přímo na základě dat. Ve spoustě praktických případů jsme schopni vybrat vhodné rozdělení, ovšem jen velmi zřídka máme představu o hodnotách jeho neznámých parametrů. Ty pak musíme odhadnout na základě dat.

Budeme pokračovat v našem příkladě, tedy uvažovat situaci, kdy se veličina X řídí normálním rozdělením s hustotou $f(x; \mu, \sigma^2)$, viz (2), a ani jeden z parametrů μ a σ^2 neznáme předem. Abychom byli schopni pracovat se vzorcem (1), je nutné odhadnout neznámé parametry μ a σ^2 na základě datového souboru, který máme k dispozici. Tato situace je v praxi velmi obvyklá, i když v některých případech můžeme mít o rozptylu dobrou představu na základě historických dat či pilotních studií.

Uvažujme, že máme možnost změřit N tablet. Každé měření je tedy nezávislou realizací veličiny X , jehož výsledky označíme jako x_1, \dots, x_N . V jistém smyslu optimálním odhadem μ je pak aritmetický průměr $\hat{\mu} = \bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$. Optimalitou je zde míněna nestrannost, konzistence odhadu a maximální věrohodnost odhadu. Neformálně řečeno nestrannost implikuje, že odhadujeme v průměru správnou hodnotu, konzistence znamená, že čím větší máme datový soubor, tím přesnější náš odhad bude, a maximální věrohodnost reprezentuje skutečnost, že odhadnutá hodnota je nejpravděpodobnější hodnotou parametru při daném datovém souboru. Zdůrazňujeme neformálnost tohoto vysvětlení a pro formální definice odkazujeme na [1]. Dodejme, že aritmetický průměr je velmi přirozeným odhadem střední hodnoty, neboť odpovídá její neformální interpretaci.

Odhad $\hat{\mu}$ nazýváme bodovým odhadem, neboť pro konkrétní datový soubor odhadujeme hodnotu μ jedním číslem. Nicméně by nemělo být překvapení, že při opakování experimentu (tedy s jiným balením N tablet) dostaneme jinou hodnotu odhadu $\hat{\mu}$, neboť závisí na tom, jaká konkrétní pozorování x_1, \dots, x_N jsme naměřili. Vzhledem k výše zmíněné konzistenci pak také bude přesnější odhad, který vychází z dat s $N = 100$ tabletami, než odhad pro $N = 5$. Bodový odhad tedy žádným způsobem neposkytuje informace o nejistotě, kterou máme při odhadování μ . Jak uvidíme níže, tato nejistota souvisí jak s velikostí datového souboru N , tak s hodnotou rozptylu σ^2 .

Zcela obdobně přistupujeme k odhadu rozptylu $s^2 = \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$. Všimněme si $N - 1$ namísto N ve jmenovateli, což zajišťuje nestrannost odhadu a souvisí s faktem, že k odhadnutí rozptylu nejprve potřebujeme odhadnout střední hodnotu. Na druhou stranu tento odhad není maximálně věrohodným odhadem rozptylu.

3. Konfidenční interval

Konfidenční interval je nástroj, který umožňuje kvantifikovat nejistotu, kterou máme při odhadování parametru. Soustředíme se na interval pro μ , nicméně postup pro σ^2 by byl obdobný. Namísto jediné hodnoty $\hat{\mu}$ sestrojíme interval, který s pravděpodobností $1 - \alpha$ obsahuje skutečnou hodnotu μ . Hladinu $1 - \alpha$ stanovíme předem sami na základě toho, jakou nejistotu chceme v odhadu ponechat.

Měříme-li obsah účinné látky v tabletách, tento interval nám poskytne informaci o nejistotě v odhadu střední hodnoty obsahu účinné látky. Dejme tomu, že změříme deset tablet a získáme bodový odhad $\hat{\mu} = 480$ mg. Sice jsme odhadli, že v průměru máme o 20 mg účinné látky méně, než bychom mít měli, ale nevíme s jakou jistotou. Je pak velký rozdíl, máme-li konfidenční interval na hladině $1 - \alpha = 0,95$ roven (475; 485) či (450; 510); v prvním případě máme důkazů pro menší obsah látky v tabletách výrazně více než v případě druhém, kdy interval zahrnuje i hodnoty větší než 500 mg.

V případě normálního rozdělení s neznámým rozptylem lze dokázat (viz [1]), že veličina $(\bar{X} - \mu) \sqrt{N/s^2}$ má tzv. Studentovo rozdělení s $N - 1$ stupni volnosti. Odsud ihned plyne, že konfidenční interval sestrojíme podle vzorce

$$\left[\bar{X} - t_{N-1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\frac{s^2}{N}}; \bar{X} + t_{N-1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\frac{s^2}{N}} \right], \quad (3)$$

kde $t_{N-1}(1 - \frac{1}{2}\alpha)$ je kvantilem Studentova rozdělení s $N - 1$ stupni volnosti. Počet stupňů volnosti je jediným parametrem hustoty Studentova rozdělení, takže dosazením

$N - 1$ je kvantil $t_{N-1}(1 - \frac{1}{2}\alpha)$ jednoznačně určen. Všimněme si, že interval je centrován kolem našeho bodového odhadu $\hat{\mu}$ a že bude tím širší, čím větší je poměr odhadu rozptylu k počtu pozorování. Také bude tím širší, čím vyšší hladiny $1 - \alpha$ chceme dosáhnout. To dává smysl, neboť chceme-li zvýšit pravděpodobnost, že náš interval bude obsahovat skutečnou střední hodnotu (při stejném datovém souboru) řekněme na 99 % namísto 95 %, tak bude přirozeně výsledný interval širší.

Všimněme si, že μ není v tomto případě náhodné, je to parametr s jednou skutečnou hodnotou, kterou chceme odhadovat. Náhodné jsou meze našeho intervalu, neboť ty závisí na pozorovaném datovém souboru. Platí tedy sice, že

$$P\left(\mu \in \left[\bar{X} - t_{N-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{s^2}{N}}; \bar{X} + t_{N-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{s^2}{N}}\right]\right) = 1 - \alpha,$$

ale zmíněná pravděpodobnost $1 - \alpha$ se vztahuje k opakovanému použití metody konstrukce konfidenčního intervalu a ne k jednomu určitému intervalu. To znamená, že budeme-li pro různé datové soubory počítat konfidenční intervaly na hladině $1 - \alpha = 0,95$, v 95 % případů budou tyto intervaly obsahovat i skutečnou hodnotu μ . Ve zbývajících 5 % případů může být skutečná hodnota μ velmi daleko od hranic konfidenčního intervalu. Je proto třeba být opatrný a nesklouzávat k mylným interpretacím, že skutečná hodnota μ bude poblíž středu intervalu.

Porozumění odstavci výše je klíčové pro správné použití konfidenčního intervalu. Demonstrujme jeho důsledky na našem příkladě. Předpokládejme, že skutečná hodnota je $\mu = 499$ mg a že budeme měřit tisíc balení tablet. Z každého balení vezmeme $N = 10$ tablet, změříme obsah účinné látky a vypočítáme konfidenční interval na hladině $1 - \alpha = 0,95$; budeme tedy mít tisíc intervalů, jeden pro každý datový soubor deseti tablet. Očekáváme potom, že přibližně 950 intervalů (tj. 95 %) bude obsahovat skutečnou hodnotu $\mu = 499$ mg. Zbývajících 50 intervalů tuto hodnotu nebude obsahovat a může být od ostatních intervalů velmi daleko. Stejný výsledek bychom očekávali pro jakoukoli hodnotu N .

S rostoucí hodnotou N bude příslušný konfidenční interval užší, neboť je konstruován tak, aby vždy obsahoval skutečnou hodnotu μ pouze v 95 % případů. Pro všechna reálná x platí vztah $\lim_{N \rightarrow \infty} t_{N-1}(x) = z_x$, kde z je kvantil normálního rozdělení se střední hodnotou 0 a rozptylem 1. Ze vzorce (3) vyplývá, že s N jdoucím do nekonečna konvergují oba krajní body intervalu k hodnotě μ (kvantil Studentova rozdělení konverguje ke konstantě, stejně tak s^2).

Velmi důležitým postřehem je fakt, že konfidenční interval sestrojujeme pro parametry rozdělení, jako jsou rozptyl σ^2 či střední hodnota μ . Nedává nám ovšem žádnou informaci o individuálních napozorovaných hodnotách veličiny jako takových. Například konfidenční interval pro μ bychom v případě tablet rádi viděli okolo 500 mg, což je požadovaný obsah. I kdyby však byl tento interval (499; 501), neznamená to, že pacienti budou dostávat kvalitní produkt, neboť obsah v jednotlivých tabletách je náhodná veličina s normální hustotou, která závisí na obou parametrech μ a σ^2 , a pokud by měla velký rozptyl, může být obsah účinné látky velmi daleko od střední hodnoty.

Tato skutečnost je jasná při pohledu na levou část obr. 1. Konfidenční interval (499; 501) by byl soustředěn velmi blízko okolo skutečné hodnoty $\mu = 500$ mg, ale obsahoval by jen velmi malou část hustoty, tj. jen velmi málo budoucích pozorování. Čím je konfidenční interval přesnější, tím méně budoucích pozorování je v něm obsa-

ženo. Pokud by navíc rozdělení účinné látky v tabletách mělo hustotu podobnou té v pravé části obr. 1, byl by nám přesný odhad střední hodnoty ještě méně užitečný, pokud bychom se zajímali o jednotlivá pozorování z tohoto rozdělení. Zatímco střední hodnota je rovna 500 mg, jednotlivá pozorování účinné látky jsou soustředěna kolem 498 a 502 mg.

Pokud se budeme zajímat o typické hodnoty obsahu účinné látky v jednotlivých tabletách, musíme pracovat s predikčním intervalem, který se zaměřuje na celou hustotu, nejen na jednotlivé parametry.

4. Predikční interval

Predikční interval obsahuje s danou pravděpodobností $1 - \alpha$ jedno příští pozorování veličiny X . Budoucí pozorování x_{N+1} je opět realizací veličiny X s neznámými parametry μ , σ^2 , které odhadujeme pomocí pozorování x_1, \dots, x_N . Predikční interval vezme v úvahu nejistotu odhadu parametrů obdobným způsobem jako konfidenční interval; výsledkem je interval, který bude obsahovat příští pozorování veličiny X s požadovanou jistotou. Na rozdíl od konfidenčního intervalu, který popisuje vlastnost napozorovaného souboru (odhaduje skutečnou střední hodnotu), predikční interval předpovídá budoucí pozorování.

V případě tablet by predikční interval pomohl vyřešit následující otázku: jakou dávku účinné látky bude mít příští tableta (například ta, kterou se chystáme podat pacientovi jako lék) na základě souboru těchto deseti naměřených tablet? Tato otázka je velmi důležitá, neboť mnoho analytických testů je destruktivních. To znamená, že nemůžeme přímo změřit tabletu, kterou se chystáme podat pacientovi, ale o její kvalitě můžeme uvažovat pouze nepřímou.

V případě normálního rozdělení s neznámým rozptylem lze dokázat (viz [1]), že veličina $(X_{N+1} - \bar{X})\sqrt{N/(s^2(N+1))}$ má Studentovo rozdělení s $N - 1$ stupni volnosti. Odtud plyne, že predikční interval pro μ sestrojíme dle vzorce

$$\left[\bar{X} - t_{N-1} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 + \frac{s^2}{N}}; \bar{X} + t_{N-1} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 + \frac{s^2}{N}} \right]. \quad (4)$$

Všimněme si podobnosti se vzorcem (3); jediným rozdílem je přičtení odhadu rozptylu s^2 pod odmocninou. Neformálně řečeno tedy sestrojíme konfidenční interval pro střední hodnotu a pak interval rozšíříme na základě odhadnutého rozptylu tak, aby obsahoval jednotlivá pozorování. Dodejme, že interval je opět centrován kolem bodového odhadu střední hodnoty $\hat{\mu} = \bar{X}$ a bude tím širší, čím větší je s^2 a čím méně máme pozorování N .

Pro predikční interval sestrojený na hladině $1 - \alpha$ pak platí

$$P \left(X_{N+1} \in \left[\bar{X} - t_{N-1} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 + \frac{s^2}{N}}; \bar{X} + t_{N-1} \left(1 - \frac{\alpha}{2}\right) \sqrt{s^2 + \frac{s^2}{N}} \right] \right) = 1 - \alpha.$$

Na rozdíl od konfidenčního intervalu krajní body predikčního intervalu nekonvergují k jediné hodnotě, ale s přibývajícím počtem pozorování se blíží k hodnotám kvantilů normálního rozdělení se střední hodnotou μ a rozptylem σ^2 . Tedy např. pro $1 - \alpha = 0,95$ konverguje interval k takovým hodnotám, mezi kterými se integrál hustoty rovná 0,95

a které jsou symetrické vzhledem ke střední hodnotě. Tento fakt vede k časté mylné interpretaci predikčního intervalu v tom smyslu, že obsahuje 95 % budoucích pozorování. Tuto vlastnost predikčního intervalu bohužel nemá a vždy musí být interpretován pouze pro jedno budoucí pozorování.

Pro $1 - \alpha = 0,95$ pak vychází pravděpodobnost 95 %, obdobně jako u konfidenčního intervalu, z opakovaného použití metody a ne z kvantilu rozdělení. Interpretujeme ji tak, že máme 95% pravděpodobnost, že napozorujeme soubor x_1, \dots, x_N, x_{N+1} takový, že sestrojíme-li interval na základě prvních N pozorování, bude obsahovat i pozorování poslední. Tato interpretace je patrná ze vzorce na předchozí straně. Ve zbyvajících 5 % pak buď nastává situace, že x_{N+1} je odlehlé pozorování, nebo situace, že prvních N pozorování vedlo ke konstrukci intervalu, který je daleko od skutečných kvantilů (tedy obdobně jako 5 % konfidenčních intervalů neobsahuje skutečnou střední hodnotu). Takový špatně zkonstruovaný interval bude přirozeně obsahovat jen velmi málo budoucích pozorování. Obecně nejčastější mylná interpretace intervalů je právě nepochopení vztahu mezi $1 - \alpha$ a opakovaným použitím metody.

Demonstrujme proto použití intervalu na hypotetické situaci. Předpokládejme opět, že skutečná hodnota $\mu = 499$ mg a že budeme měřit tisíc balení tablet. Z každého balení vezmeme jedenáct tablet a pro $N = 10$ tablet změříme obsah účinné látky a vypočítáme predikční interval na hladině $1 - \alpha = 0,95$; budeme tedy mít tisíc intervalů, jeden pro každý datový soubor deseti tablet. Poté změříme jedenáctou tabletu a ověříme, zda se bude nacházet v daném predikčním intervalu. Očekáváme, že přibližně 950 intervalů (tj. 95 %) bude obsahovat hodnotu jedenácté tablety x_{11} . Zbývajících 50 intervalů tuto hodnotu neobsahuje a může být od této hodnoty velmi daleko. Obdobný výsledek bychom očekávali pro jakoukoli hodnotu N .

Z výše uvedeného vyplývá, že můžeme predikční interval použít v případě, kdy se zajímáme pouze o jednu konkrétní tabletu, kterou chceme v budoucnu použít. Přístup však může být rozšířen a můžeme sestrojít i predikční intervaly pro k budoucích tablet. V případě, že bychom uvažovali velký počet budoucích pozorování, vede takové rozšíření k tolerančnímu intervalu.

5. Toleranční interval

Toleranční intervaly jsou rozšířením predikčních intervalů pro velký počet budoucích pozorování. Typickou aplikací je farmaceutická výrobní linka, která produkuje miliony tablet, a nás zajímá, jestli bude mít dostatečný počet budoucích tablet odpovídající kvalitu. Cílem je sestrojít takový interval, který bude obsahovat 100β % budoucích pozorování, pro předepsané $\beta \in (0, 1)$ s danou pravděpodobností $100(1 - \alpha)$ %. Například tedy konstruujeme interval, který s 95% jistotou bude obsahovat 99 % budoucích pozorování. Obdobně jako v předchozích případech je 95% jistota interpretována v opakovaném použití metody, zatímco 99 % opravdu znamená odpovídající podíl budoucích pozorování. Pokud tedy metodu budeme používat opakovaně, v 95 % případech použití metody bude výsledný toleranční interval obsahovat 99 % budoucích pozorování.

Toleranční interval je řešen sestrojením konfidenčních intervalů na hladině $1 - \alpha$ pro kvantily, které mezi sebou obsahují 100β % pozorování. Hodnota α pak vyjadřuje nejistotu v odhadu těchto kvantilů. Odhadování kvantilů není jednoduchou záležitostí, neboť přesnost odhadu je velice nízká, pokud nemáme opravdu velké množství pozorování.

vání. Proto jsou toleranční intervaly často velice široké a konvergují ke kvantilům jen velmi pomalu. Problém je, že přesnost odhadu kvantilů klesá, čím blíže je β k jedné. Právě vysoké hodnoty β jsou ovšem v praxi nutné, neboť obvykle požadujeme záruku ohledně kvality vysokého podílu produktů.

Sestrojit oboustranný toleranční interval není jednoduché ani pro normální rozdělení. Konstrukce totiž vychází ze dvou konfidenčních intervalů na hladině $1 - \alpha$ pro $1 - \beta/2$ a $\beta/2$ kvantily (které mezi sebou podle definice mají $100\beta\%$ pozorování). Tyto dva konfidenční intervaly musíme nějakým způsobem spojit v jeden toleranční interval. Řešení takového problému není jednoznačné a i pro toleranční interval symetrický kolem odhadu střední hodnoty \bar{X} existují pouze přibližná řešení.

V případě normálního rozdělení s neznámou střední hodnotou a rozptylem lze toleranční interval odhadnout dle [3] jako

$$\left[\bar{X} - z_{1-\beta/2} \sqrt{s^2 k}; \bar{X} + z_{1-\beta/2} \sqrt{s^2 k} \right], \quad (5)$$

kde

$$k = \sqrt{\frac{(N-1 + \frac{N-1}{N})}{\chi_{1-\alpha, N-1}^2}} \sqrt{1 + \frac{N-3 - \chi_{1-\alpha, N-1}^2}{2(N+1)^2}}. \quad (6)$$

Povšimněme si, že vzorec se výrazně liší od vzorců (3) a (4). Prvním rozdílem je přítomnost $z_{1-\beta/2}$ kvantilu normálního rozdělení namísto Studentova rozdělení. Všimněme si však, že kvantil je zde závislý na β (a nikoliv na α) a souvisí tedy s tím, že budoucí pozorování pochází z normálního rozdělení. Hodnota α se ve vzorci objevuje u dalšího kvantilu, tentokrát χ^2 , tzv. chí-kvadrát rozdělení, které souvisí s rozdělením odhadu rozptylu normálního rozdělení. Ze vzorce (5) vidíme, že centrovanost kolem bodového odhadu střední hodnoty zůstává. Při bližším prozkoumání vzorce (6) by pak bylo zřejmé, že pro dané β bude interval tím širší, čím vyšší bude odhadnutý rozptyl s^2 a požadovaná hladina $1 - \alpha$ a čím menší bude počet pozorování N použitých k sestrojení intervalu.

Uvažujme toleranční interval s mezemi označenými A , B . Tyto meze jsou náhodnými veličinami, neboť sestrojení intervalu jakoukoli metodou bude záviset na napozorovaných hodnotách X_1, \dots, X_N , a jejich plný zápis by byl $A(X_1, \dots, X_N)$ a $B(X_1, \dots, X_N)$. Platí pak, že při opakovaném použití dané metody k sestrojení tolerančního intervalu, bude $100\beta\%$ pozorování obsaženo v tolerančním intervalu s pravděpodobností $1 - \alpha$:

$$P(P(X \in [A; B]) \geq \beta) = P\left(\int_A^B f(x; \mu, \sigma^2) dx \geq \beta\right) = 1 - \alpha. \quad (7)$$

Interval sestrojený dle vzorce (5) splňuje tuto rovnici jen přibližně, neboť se jedná o aproximaci. Dodejme, že v praxi se vzorec (5) využívá zřídka, neboť díky numerickým metodám lze získat přesnější aproximaci tolerančního intervalu řešením rovnice (7). Detaily o tomto přesnějším řešení lze nalézt v [4].

Podobně jako krajní body predikčního intervalu konvergují krajní body tolerančního intervalu s přibývajícím počtem pozorování k hodnotám kvantilů normálního rozdělení se střední hodnotou μ a rozptylem σ^2 . Avšak na rozdíl od predikčního intervalu, jehož krajní body konvergují ke kvantilům odpovídajícím hladině $1 - \alpha$, krajní body

tolerančního intervalu konvergují ke kvantilům odpovídajícím požadované proporci populace β . Toleranční intervaly totiž vychází z $100(1 - \alpha)\%$ konfidenčních intervalů $100\beta/2\%$ a $100(1 - \beta/2)\%$ kvantilů; jak již bylo uvedeno výše, konfidenční intervaly konvergují k jedinému číslu, tj. v tomto případě k příslušnému kvantilu.

Použijme následující hypotetickou situaci k demonstrování interpretace tolerančního intervalu. Uvažujme, že máme tisíc datových souborů z normálního rozdělení a pro každý datový soubor sestrojíme toleranční interval s jistotou 95 % pro 99 % budoucích pozorování. Poté vygenerujeme tisíc nových hodnot pro každý datový soubor a zkontrolujeme, kolik z nich se nachází v příslušném tolerančním intervalu. Přibližně 950 intervalů by mělo obsahovat alespoň 990 nových pozorování.

Předpokládejme nyní, že budeme mít k dispozici 1010 tablet pro každý ze sto výrobních dní. Pro prvních $N = 10$ tablet změříme obsah účinné látky a vypočítáme toleranční interval na hladině $1 - \alpha = 0,95$ a pro $\beta = 0,99$; budeme tedy mít sto intervalů, jeden pro každý datový soubor deseti tablet z jednoho výrobního dne. Poté změříme zbývajících tisíc tablet v každém dni a ověříme, kolik hodnot se nachází v příslušném tolerančním intervalu. Očekáváme, že přibližně 95 intervalů (tj. 95 %) bude obsahovat alespoň 990 (tj. 99 %) tablet. Zbývajících 5 intervalů bude obsahovat nižší počet budoucích tablet. Obdobný výsledek bychom očekávali pro jakoukoli hodnotu N .

6. Diskuze

Ukázali jsme tři druhy statistických intervalů, které se používají k vyjádření nejistoty při odhadování náhodné veličiny na základě datového souboru. Je důležité si uvědomit, že tři prezentované intervaly se liší v konstrukci a v tom, jakému slouží účelu. Konfidenční intervaly dávají informace o nejistotě při odhadu parametrů rozdělení. Predikční intervaly kvantifikují nejistotu v předpovědi jedné budoucí napozorované hodnoty a toleranční intervaly rozšiřují předpověď na velké množství budoucích pozorování. Na příkladu obsahu účinné látky v tabletách konfidenční intervaly vyjadřují nejistotu v odhadu střední hodnoty a rozptylu normálního rozdělení, kterým se hodnoty v tabletách řídí. Predikční intervaly pomohou popsat nejistotu ohledně obsahu účinné látky v tabletě, kterou se chystáme podat pacientovi. Konečně, toleranční intervaly vyjadřují, jakou nejistotu máme o obsahu účinné látky v jednotlivých tabletách při výrobě velkého množství tablet.

Nesprávné použití těchto intervalů je bohužel častým jevem a může vést k velmi zavádějícím výsledkům. Pro 95% konfidenční intervaly je důležité si pamatovat, že sice v 95 % skutečnou hodnotu μ obsahují, ale že ve zbývajících 5 % mohou být zcela špatně. Predikční intervaly je pak třeba vždy interpretovat ve vztahu k jednomu jedinému budoucímu pozorování. Pro toleranční intervaly je dobré mít na paměti, že mohou být velmi široké jen kvůli tomu, že odhad kvantilů byl velmi nepřesný (především pokud chceme hodnotu β blízko jedné) a že je třeba mít pro jejich přesný odhad mnohem víc pozorování N , než v případě konfidenčních a predikčních intervalů.

Náhodná veličina uvažovaná v tomto příspěvku měla normální rozdělení. Samozřejmě veškeré prezentované intervaly lze zkonstruovat pro jakákoli rozdělení. V případě spojitých veličin bude přístup obdobný, ovšem v případě diskrétních veličin může být situace výrazně složitější a některé prezentované definice vyžadují úpravu (například definice kvantilů). Diskrétní veličiny jsou takové, které nabývají jedné z konečně či

spočetně mnoha hodnot (například výsledek hodu mincí, počet dětí či počet přístupů na internetovou stránku). Samozřejmě, i pro tyto veličiny můžeme zkonstruovat výše popsané intervaly a jejich interpretace bude obdobná jako ve výše popsaném případě normálního rozdělení. Dále mohou být statistické intervaly rozšířeny pro vícerozměrný případ na tzv. statistické oblasti.

Dodejme pro úplnost, že intervaly popisované v tomto příspěvku vychází z frekventistického pojetí statistiky. Bayesovská statistika umožňuje odhadnout intervaly obdobného typu, nicméně jejich interpretace se liší vzhledem k jinému pohledu na parametry a pravděpodobnost jako takovou. Při interpretaci výsledků mohou být rozdíly poměrně podstatné, ale jejich vysvětlení by bylo materiálem pro samostatný příspěvek, a proto jsme se bayesovskému přístupu v tomto článku nijak nevěnovali.

Pro další studium problematiky statistických intervalů doporučujeme např. [2] a [4].

Poděkování. Autor děkuje za užitečné komentáře, které přispěly k výraznému vylepšení článku, RNDr. Jiřímu Dvořákovi, Ph.D., a Mgr. Čenku Jirsákovi. Za korekturu patří poděkování Marii Staré.

L i t e r a t u r a

- [1] ANDĚL, J.: *Statistické metody*. MatfyzPress, Praha, 2007.
- [2] HAHN, G. J., MEEKER, W. Q.: *Statistical intervals: a guide for practitioners*. John Wiley, 2001.
- [3] HOWE, W. G.: *Two-sided tolerance limits for normal populations*. J. Amer. Statist. Assoc. 64 (1969), 610–620.
- [4] KRISHNAMOORTHY, K., MATHEW, T.: *Statistical tolerance regions*. John Wiley, New Jersey, 2009.