MANUÁL STUDENTA (BIO)-STATISTIKY

Vladimír Rogalewicz

1. lékařská fakulta
Univerzita Karlova
Praha, říjen 2018

Tento Manuál je určen studentům nelékařských oborů na 1. lékařské fakultě Univerzity Karlovy. Obsahuje elementární vysvětlení základních pojmů teorie pravděpodobnosti a matematické (induktivní) statistiky a je zamýšlen jako pomůcka pro absolvování přednášek, cvičení a pro studium povinné literatury v bakalářském studiu a pro opakování základů matematické statistiky v navazujícím magisterském studiu. Manuál tedy neobsahuje všechny požadované vědomosti a dovednosti, ale výrazně usnadní četbu povinné i doporučené literatury a bude studenta provázet zejména složitou oblastí interpretace výsledků pravděpodobnostních a statistických výpočtů.

Úvod

Okolo nás existuje spousta jevů, situací a událostí, které nelze předvídat. Jsou důsledkem náhody. Patří mezi ně například zakoupení losu, který vyhrává, doba čekání na tramvaj, zlomení nohy na náledí, výsledek testu hladiny glukózy v krevní plazmě, výsledky parlamentních voleb, obsazené či neobsazené telefonní číslo, zasažení stromu bleskem, počet ataků nemoci a doba mezi nimi, atd. Běžné matematické prostředky nelze k popisu náhodných dějů použít. Například diferenciální rovnice popisují, jaké vztahy platí za daných podmínek mezi jednotlivými veličinami. Jestliže dodržíme všechny předpoklady, výsledek bude vždy stejný a přesně definovaný.

Jinak je tomu u náhodných dějů. I když dodržíme všechny podmínky, výsledek terapie se bude u různých pacientů lišit; terapie někdy zabere, jindy je neúčinná. Záleží-li výsledek na náhodě, nemůžeme ho nikdy přesně spočítat a popsat. V takovém případě nám žádná věda nepomůže predikovat výsledek už ze samé podstaty náhody. Můžeme však popsat populační chování takového děje, tedy kvantifikovat samu náhodu. Takový popis nám sice nepomůže v jednom konkrétním případě (tam je výsledek ryze náhodný), ale přináší cenné informace o podílu různých výsledků v populaci. Tato informace se nám hodí například při predikci potřebných kapacit.

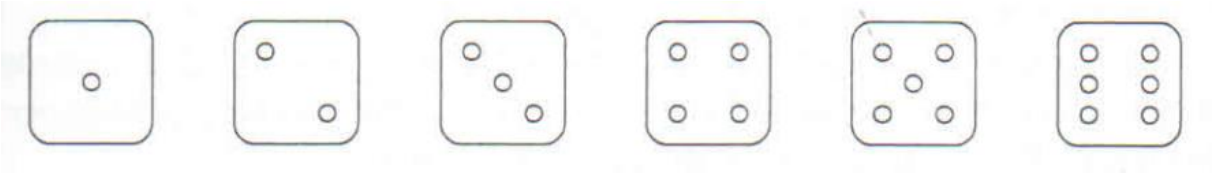
Otázkami náhody a náhodných dějů se zabývají dvě matematické disciplíny: teorie pravděpodobnosti a matematická statistika. Teorie pravděpodobnosti řeší následující problém: nějaký jev má řadu různých následků. Náhoda určuje, který (jediný) z nich skutečně nastane. Při nasazení antibiotik nelze odhadnout, jak u konkrétního pacienta zapůsobí. Teorie pravděpodobnosti poskytuje informaci, jak různé výsledky předem kvantifikovat. Jedná se tedy o teoretickou disciplínu, jejímž hlavním úkolem je sestrojit teoretický model, jímž se daný děj řídí. Matematická statistika řeší v určitém smyslu duální problém. Vidím výsledek nějaké skutečnosti (např. výsledek krevního testu) a na jeho základě mám roztřídit možné příčiny a vybrat z nich tu nejpravděpodobnější. Jedná se tedy o ryze praktickou experimentální vědu, která využívá teoretický model sestrojený v teorii pravděpodobnosti a na jeho základě se snaží interpretovat naměřená (empirická) data a současně poskytnout i údaj o přesnosti této interpretace.

Teorie pravděpodobnosti i matematická statistika jsou založeny na známých matematických nástrojích – kalkulu, lineární algebře a teorii míry. Nepřinášejí (až na výjimky) nové matematické postupy, ale dávají známým postupům nový obsah. Přirozeně je tedy věnována mnohem větší pozornost tomu, jak budou vstupní údaje i výsledky interpretovány, než vlastnímu výpočtu. Jedná se vlastně o porozumění číslům a vztahům mezi nimi. Jednou ze základních úloh teorie pravděpodobnosti je počítání s nepřesnými čísly. Špatnou interpretací (ať už nechtěnou nebo záměrnou) často dochází k posunutí smyslu nebo minimálně k závěrům, které budí úsměv.

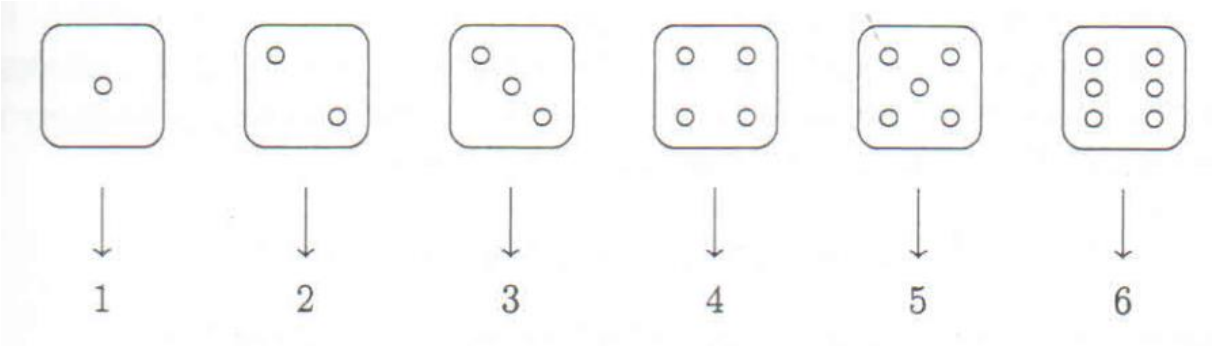
Náhodná veličina

Základním pojmem moderní teorie pravděpodobnosti (o kterou se dnes analýza dat opírá) je *náhodná veličina*. Její matematická definice je poměrně složitá, protože se musí vyrovnat s různými mezními případy této teorie. Při praktickém využití se však s těmito problematickými situacemi nesetkáváme, a tak pro pochopení stačí laický pohled.

Snažíme se popsat náhodu v situacích, jejichž výsledek právě na náhodě závisí. Pro popis potřebujeme výčet všech možných výsledků takové situace. Tím mohou být přirozená čísla (počet pacientů, kteří za jeden den přijdou do ordinace), reálná čísla (hladina glukózy v séru), ale i nečíselná množina, například {ANO, NE} (způsobí kousnutí klíštěte boreliózu?), seznam týmů 1. fotbalové ligy (kdo bude vítězem) nebo jeden z obrázků

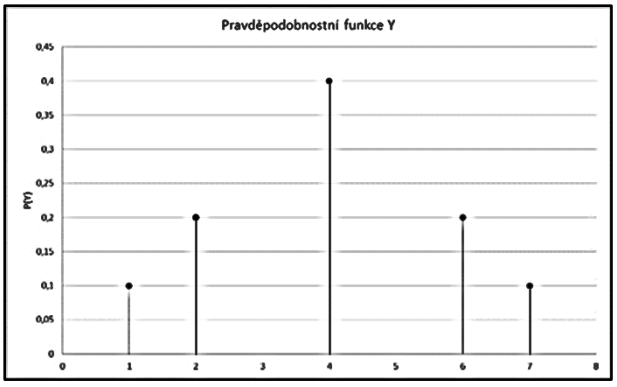


(co padne na kostce?). Abychom s takovými výsledky mohli počítat, potřebujeme každému výsledku přiřadit (reálné) číslo. Tato čísla můžeme přiřadit libovolně; z praktických důvodů to uděláme tak, aby přiřazení mělo logický smysl, např.



(Poznámka. Pokud je výsledek situace číselný, toto přiřazení může – ale nemusí – být identickým zobrazením, tj. každému číselnému výsledku přiřadíme právě toto číslo.) Tomuto přiřazení budeme říkat náhodná veličina. Náhodná veličina je tedy přiřazení číselné hodnoty každému možnému výsledku situace.

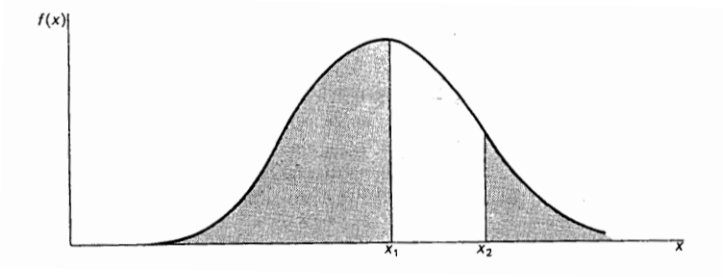
Abychom náhodnou veličinu mohli plně popsat, musíme umět kvantifikovat, jak často ten který výsledek vyjde. Pro účely tohoto popisu dělíme náhodné veličiny na diskrétní a spojité. *Diskrétní* je taková *náhodná veličina*, jejímiž hodnotami jsou izolovaná čísla; těch je konečně nebo spočetně mnoho. Diskrétní náhodnou veličinu popisujeme *pravděpodobnostní funkcí*, která každé možné výsledné hodnotě x_i přiřadí pravděpodobnost p_i. Hodnoty pravděpodobnostní funkce znázorňujeme obvykle formou tabulky nebo grafu – následující graf a tabulka vyjadřují pravděpodobnostní funkci stejné náhodné veličiny:



x_i	1	2	4	6	7
p_i	0,1	0,2	0,4	0,2	0,1

Všimněte si, že pravděpodobnosti jsou čísla mezi 0 a 1 a celkový součet pravděpodobností přes všechny možné jevy (výsledky) je jedna. (V běžném hovoru jsme zvyklí mluvit o pravděpodobnostech vyjádřených v procentech. Protože procento znamená jednu část ze sta, tedy jednu setinu, platí například 0,05 = 5 %.)

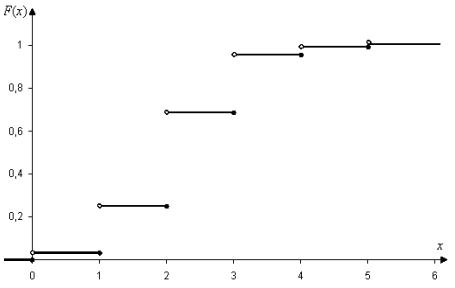
Spojitá náhodná veličina může nabývat všech hodnot z nějakého intervalu (přičemž za interval považujeme i množinu všech reálných čísel, tedy (-∞;+∞)). Protože je (izolovaných) hodnot na intervalu nekonečně mnoho, musí mít každá z nich nulovou pravděpodobnost (v opačném případě by celková pravděpodobnost překročila hodnotu jedna, tj. 100 %). V případě spojité náhodné veličiny tedy potřebujeme vyjádřit pravděpodobnost jinak. Popisujeme ji funkcí *f(x)*, kterou nazýváme *hustotou* pravděpodobnosti (viz následující obrázek). V grafu hustoty pravděpodobnosti se hodnoty

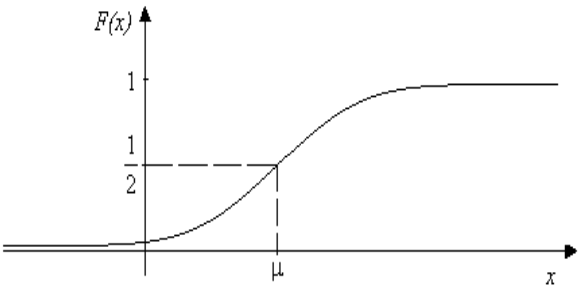


pravděpodobnosti vyjadřují jako plocha pod grafem hustoty:

Bílá ohraničená plocha znázorňuje pravděpodobnost, že výsledek leží mezi hodnotami x_1 a x_2, tedy $P[X \in (x_1;x_2)]$. Hustota tedy může nabývat pouze nezáporných čísel a plocha pod celou křivkou (hustotou) se musí rovnat jedné (tj. 100 %).

Stejnou informaci jako pravděpodobnostní funkce (v případě diskrétní náhodné veličiny) nebo hustota (v případě spojité náhodné veličiny) nese *distribuční funkce F(x)*. Její výhodou je, že je definovaná stejně v případě diskrétní i spojité náhodné veličiny: $F(x) = P[X \leq x]$. Následující obrázky představují příklady distribuční funkce diskrétní a spojité náhodné veličiny.





Zatímco pravděpodobnostní funkce a hustota vyjadřují „lokální“ informaci, tedy pravděpodobnost nějakého bodu či (libovolně malého) intervalu, distribuční funkce vyjadřuje součtovou (integrální) informaci – pravděpodobnost všech bodů menších nebo rovných dané hodnotě souhrnně. Ze znalosti pravděpodobnostní funkce nebo hustoty lze zkonstruoval distribuční funkci a naopak, ze znalosti distribuční funkce lze zkonstruovat pravděpodobnostní funkci nebo hustotu. Jedná se tedy o dvě vyjádření, která nesou stejnou informaci. Kteroukoli z těchto funkcí nazýváme *rozdělení náhodné veličiny*.

Poznámka. Na první pohled je podivné, že se popis diskrétních a spojitých náhodných veličin tak liší. Mezi spojitými a diskrétními náhodnými veličinami však existuje celá přechodová třída „smíšených“ náhodných veličin, které v sobě kombinují vlastnosti spojitých i diskrétních náhodných veličin a jejichž popis je blíže jedné nebo druhé skupině. V praxi je obvykle možné rozdělit zkoumání smíšené náhodné veličiny na zkoumání její „spojité části“ a její „diskrétní části“, a proto se učebnice základů pravděpodobnosti obvykle omezují pouze na případy čistých spojitých nebo čistých diskrétních veličin.

Pravděpodobnostní model je založen na znalosti náhodné veličiny a jejího rozdělení, tedy

- všech možných výsledků náhodného děje (množiny elementárních jevů),

- číselného vyjádření každého výsledku (náhodné veličiny) a
- kvantifikace náhody, tedy informace, nakolik je každý výsledek možný (rozdělení náhodné veličiny).

Charakteristiky náhodné veličiny

Náhodná veličina je plně popsaná funkcí (distribuční funkcí nebo hustotou nebo pravděpodobnostní funkcí), čímž je problém pro matematika vyřešen. Popis pomocí funkce však není příliš praktický v běžných situacích a pro řadu lidí je nesrozumitelný. Pro rychlou představu o náhodné veličině může být poměrně složitý. Vznikl požadavek popsat náhodnou veličinu pomocí jednoho, případně několika čísel. Ačkoli se takový požadavek zdá nesmyslný (chceme totiž nahradit nekonečně mnoho čísel v popisu funkce za malý počet čísel), ukázalo se, že taková čísla, která podávají poměrně dobrou představu o náhodné veličině, dokážeme definovat. Budeme jim říkat charakteristiky. *Charakteristika* je tedy reálné číslo, které je jednoznačně přiřazené rozdělení a popisuje nějakou vlastnost daného rozdělení. Charakteristik dokážeme definovat nekonečně mnoho (například nejmenší možnou hodnotu, nejvyšší možnou hodnotu, nejčastější hodnotu atd.). Často nám však stačí například informace o tom, kolem jaké hodnoty se bude náhodná veličina pohybovat. Informace o průměrné hodnotě a velikosti možných odchylek od této hodnoty může laikovi zprostředkovat představu o náhodné veličině více, než přesná, ale složitá funkce. Mezi všemi charakteristikami se časem vydělila malá skupina, která se dnes standardně používá k popisům náhodné veličiny. K nejdůležitějším charakteristikám patří střední hodnota, rozptyl, směrodatná odchylka a kvantily.

Střední hodnota $\mu = EX$ náhodné veličiny X je zobecněním aritmetického průměru a pro diskrétní a spojité náhodné veličiny je definována vzorci

$$\mu = \sum_i p_i \cdot x_i, \text{ resp. } \mu = \int_{-\infty}^{+\infty} x \cdot f(x).$$

Jedná se tedy o „průměr“ všech možných (náhodných) hodnot, přičemž jednotlivé hodnoty jsou váženy mírou svého výskytu (pravděpodobností, resp. hodnotou hustoty). Pokud jsou všechny hodnoty stejně možné (např. hod ideální kostkou), pak je střední hodnota prostý aritmetický průměr. Z fyzikálního pohledu je střední hodnota polohou těžiště možných hodnot.

Samotná střední hodnota pro představu o náhodné veličině nestačí. Možné hodnoty se totiž někdy mohou pohybovat v bezprostřední blízkosti průměru (tj. náhodné odchylky jsou malé), jindy mohou být i velmi vzdálené (velké náhodné odchylky). Informaci o velikosti odchylek od průměru dává *rozptyl* $\sigma^2 = var\ X$, který je definován jako

$$\sigma^2 = E(X - EX)^2.$$

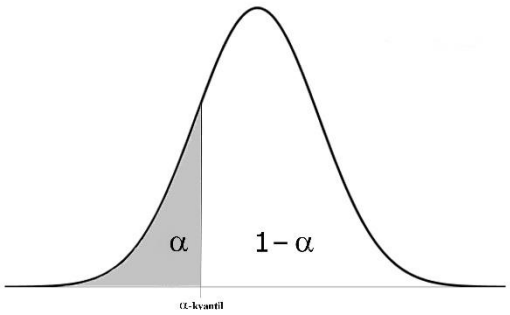
Jedná se tedy o průměrnou čtvercovou odchylku od střední hodnoty. Rozptyl velice přesně popisuje, jak vzdálené mohou být jednotlivé výsledky od průměru. Protože je však vyjádřen v jiných jednotkách než jednotlivá měření (i střední hodnota), často bývá nahrazen *směrodatnou odchylkou* σ, která je jeho odmocninou, tedy

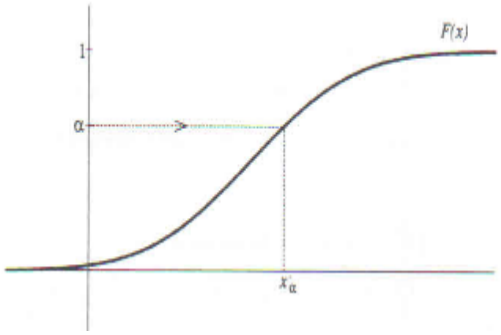
$$\sigma = \sqrt{\sigma^2}.$$

Směrodatná odchylka má stejný rozměr, jako měřená veličina, a proto ji lze využít k vyjádření určitého intervalu, v kterém možné hodnoty leží, např. $\mu \pm k\sigma$, kde k je kladné číslo. K této problematice se vrátíme v dalším textu.

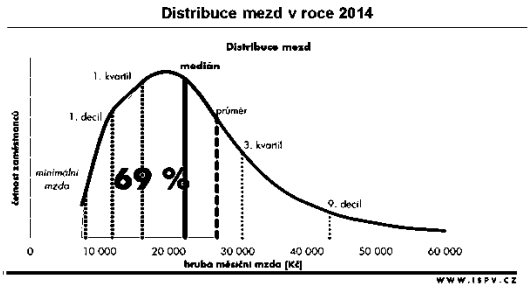
Celou skupinu charakteristik představují *kvantily*. Pro spojité náhodné veličiny je jejich definice poměrně jednoduchá: ke každému číslu $\alpha \in (0; 1)$ definujeme α-kvantil x_α jako takovou reálnou hodnotu, aby pravděpodobnost výsledku menšího nebo rovného x_α byla rovna právě α. Máme tedy

$P[X \leq x_\alpha] = \alpha$, a protože pravděpodobnost na levé straně této rovnice je hodnota distribuční funkce, můžeme definici α-kvantilu x_α přepsat jako $x_\alpha = F^{-1}(\alpha)$. Definice α-kvantilu je znázorněna na následujícím obrázku:





Některé kvantily mají zvláštní název. Hodnota $x_{0,5}$ se nazývá *medián*, hodnota $x_{0,25}$ *dolní kvartil*, hodnota $x_{0,75}$ *horní kvartil*. Často se také pracuje se vzdáleností dolního a horního kvartilu, tedy $x_{0,75} - x_{0,25}$. Tato hodnota se nazývá *mezikvartilové rozpětí*. Medián dělí číselnou osu na dvě poloviny se stejnou pravděpodobností (50 %). Pokud je rozdělení symetrické podle středu, je medián shodný se střední hodnotou. Pokud však rozdělení není symetrické, medián a střední hodnota se liší a jejich použití je třeba zvážit podle potřeby interpretace. Typickým nesymetrickým rozdělením je rozdělení mezd. Obrázek zveřejněný Českomoravskou konfederací odborových svazů ukazuje rozdělení mezd v ČR v roce 2014. Zatímco medián (dělící populaci na 50 % s nižší mzdou a 50 % s vyšší mzdou) byl 22 097 Kč, střední hodnota mzdy (průměrná mzda) byla 26 804 Kč; pod touto hodnotou však bylo 69 % pracovníků.



Zdroj: https://www.facebook.com/CMKOS/photos/a.209023719111975.63632.203251603022520/1244111912269812/?type=3&theater

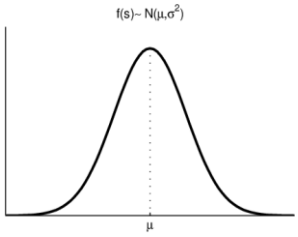
V případě diskrétního rozdělení je definice o něco složitější, protože $F^{-1}(\alpha)$ nemusí existovat (z důvodu „skokovitého" charakteru této funkce). Základní myšlenka však zůstává stejná (rozdělení populace tak, aby $100 \cdot \alpha$ % jejích prvků leželo pod x_α a $100 \cdot (1 - \alpha)$ % jejích prvků nad x_α).

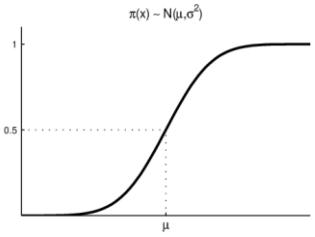
Normální rozdělení

Z definice distribuční funkce (případně hustoty) je jasné, že různých rozdělení pravděpodobnosti je nekonečně mnoho. Mezi všemi rozděleními pravděpodobnosti má zvláštní postavení *normální rozdělení*, jehož hustotou je *Gaussova křivka*. Hustota normálního rozdělení je

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{pro } -\infty < x < +\infty.$$

Hustota tedy závisí pouze na dvou parametrech, μ a σ^2, přičemž první z nich je střední hodnotou a druhý rozptylem (odtud také obvyklé označení pro tyto dva parametry). Jedná se tedy o případ, kdy

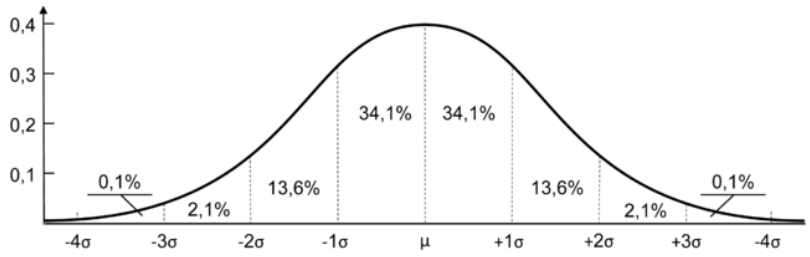




dvě charakteristiky (parametry) plně určují rozdělení pravděpodobnosti. Normální rozdělení s parametry μ a σ^2 obvykle označujeme $N(\mu,\sigma^2)$. Grafem hustoty normálního rozdělení je důvěrně známá Gaussova křivka zobrazená vlevo; vpravo je pak jeho distribuční funkce.

Mnoho fyzikálních měření lze velice dobře aproximovat normálním rozdělením. Taková měření jsou dvojího druhu. Zaprvé jsou to ta, kde jsou rozdíly v hodnotách způsobeny chybou měření. Jestliže je chyba při měření nějaké neznámé veličiny součtem velkého množství malých odchylek, které mohou být vlivem náhody kladné i záporné, lze obvykle použít normální rozdělení. Zadruhé se jedná o měření veličin, které „přirozeně“ kolísají. Například některá měření v biologii, jako výška různých jedinců, jsou také normálně rozdělená. „Nenormalita“ je ve skutečnosti tak řídká, že již samotný její výskyt je významným ukazatelem. Často je normalita předpokládána „ad hoc“, bez hlubší analýzy. K tomu přispěl rozvoj matematické statistiky v první polovině 20. století. Řada metod matematické statistiky je totiž dobře prozkoumána pouze pro normální rozdělení, protože nejsou dostatečně popsány transformace (funkce) jiných rozdělení. (Automatický předpoklad normality však může vést k vážným chybám.)

Pro normální rozdělení je také dobře prozkoumaný vliv druhého parametru (rozptylu), resp. jeho odmocniny (směrodatné odchylky). Pravděpodobnost, že se výsledek odchýlí od střední hodnoty μ o méně než σ, je 0,6827; pokud povolíme chybu 2σ, pak pokrýváme výsledky s pravděpodobností 0,9555, a v případě 3σ dokonce 0,9973 (viz následující obrázek). Odtud plyne tzv. „pravidlo 3σ“: pravděpodobnost toho, že hodnota náhodné veličiny X leží mimo interval $(\mu - 3\sigma; \mu + 3\sigma)$, je menší



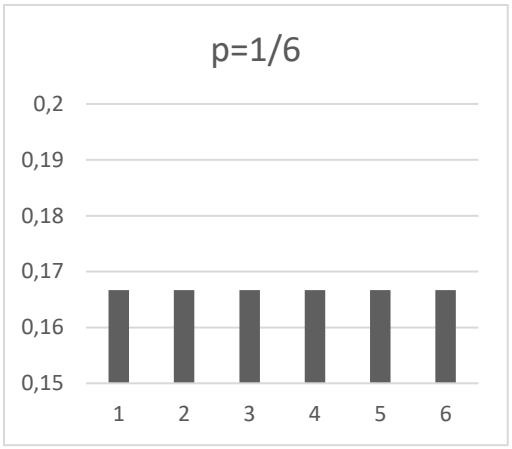
než 0,003 (často se však zapomíná na nutný předpoklad normality!).

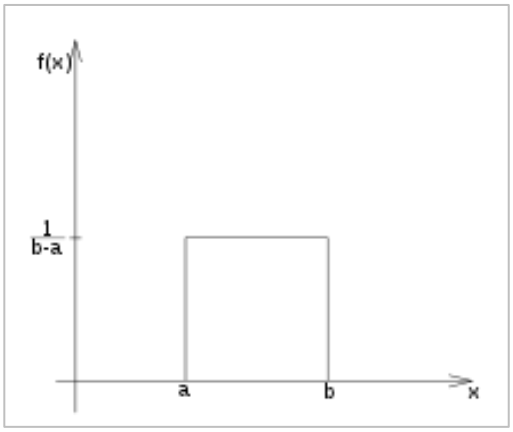
Jestliže má náhodná veličina X rozdělení $N(\mu,\sigma)$, potom má náhodná veličina $Y = (X - \mu)/\sigma$ rozdělení N(0,1). Toto rozdělení nezávisí na žádném parametru a nazývá se *normované normální rozdělení*. Hodnoty N(0,1) jsou tabelované a jsou také integrované v řadě počítačových aplikací (včetně Excelu).

S normálním rozdělením souvisejí rozdělení chí-kvadrát (χ^2) rozdělení, Studentovo t-rozdělení a Fisherovo F-rozdělení. Tato rozdělení se používají při konstrukci intervalů spolehlivosti a při testování hypotéz o parametrech normálního rozdělení. Jejich kvantily jsou běžně tabelovány a také integrovány ve statistickém softwaru.

Rovnoměrné rozdělení

Pokud předpokládáme, že všechny výsledky jsou stejně možné, potom mluvíme o rovnoměrném rozdělení. Diskrétní rovnoměrné rozdělení popisuje náhodnou veličinu, která má konečný počet stejně možných výsledků (např. hod ideální hrací kostkou na levém obrázku); spojité rozdělení pravděpodobnosti popisuje náhodnou veličinu, která nabývá hodnoty z konečného intervalu $\langle a, b \rangle$ a všechny hodnoty jsou stejně možné (např. náhodná veličina na pravém obrázku).





Střední hodnota diskrétního rovnoměrného rozdělení je aritmetický průměr možných hodnot, střední hodnota spojitého rovnoměrného rozdělení je střed intervalu $\langle a, b\rangle$, tedy bod $(a + b)/2$. Předpoklad rovnoměrného rozdělení používáme často v případě, že o rozdělení nic nevíme; potom je rovnoměrné rozdělené „nejhorší možnou uvažovanou variantou“ (nese nejméně informací).

Základy matematické statistiky

Matematická (induktivní) statistika řeší duální úlohu k teorii pravděpodobnosti. Zatímco v teorii pravděpodobnosti byl náš úkol sestavit teoretický model, který by kvantifikoval pravděpodobnosti (budoucích) důsledků nějakého jevu, v matematické statistice máme řadu naměřených hodnot, tedy důsledek, který může mít řadu (minulých) příčin, a tyto příčiny chceme nějak utřídit (například najít nejpravděpodobnější, vyloučit málo pravděpodobné apod.). Úlohu řešíme z populačního pohledu, tj. nezajímají nás konkrétní případy (z kterých jsme naměřili data), ale obvyklé situace v celé populaci. Dvě základní oblasti matematické statistiky jsou *teorie odhadu* a *testování hypotéz*.

Chceme-li vědět, jak chutná víno v sudu, nemusíme vypít celý sud; stačí malý doušek a víme, na čem jsme. Obdobně pracujeme při statistické analýze dat. Celý soubor, o kterém chceme vypovídat, nazýváme *populací*. Takovou populací mohou být všichni obyvatelé ČR, všechny dospělé ženy v Plzni, všechny mobilní telefony iPhone 6 prodané v roce 2016, nebo všechny ryby ve Vltavě. Naším cílem je udělat nějaký závěr o celé populaci. Neměřili jsme ale všechny jedince v populaci, ale pouze nějaký podsoubor. Ten nazýváme *výběrovým souborem*. Tedy na základě zjištění provedených na členech výběrového souboru chceme dělat závěry o celé populaci. Například na základě odpovědí 1000 respondentů (výběrového souboru) děláme závěry o oblibě zubních past mezi obyvatelstvem ČR (populace), nebo na základě chemické analýzy 1 kg rajčat děláme závěry o obsahu dusičnanů nebo těžkých kovů v celé dodávce.

Závěry, které uděláme na základě výběrového souboru, platí přesně pouze pro tento výběrový soubor. Jestliže je interpretujeme pro celou populaci, dopustíme se chyby. Tuto chybu chceme kvantifikovat. Ideální by samozřejmě bylo mít informace o celé populaci. Důvodů, proč místo toho zkoumámě pouze výběr, je celá řada. Mezi hlavní patří:

1. populace může být nekonečná nebo tak velká, že není technicky možné zkoumat všechny její jedince (např. všechny ryby v moři);
2. výzkum celé populace by byl časově náročný;
3. cena výzkumu celé populace vysoce převyšuje přínosy;
4. testy je možné provádět pouze destrukční metodou (zkoumaný jedinec se při testu zničí);
5. část populace nemáme k dispozici;
6. výsledek testu je stejně vždy zatížen chybou, úplné přesnosti nelze dosáhnout.

Je zřejmé, že přesnost našich závěrů záleží na velikosti výběrového souboru. Překvapivým teoretickým výsledkem však je, že přesnost závěrů nezáleží na velikosti populace, ale pouze na velikosti výběrového souboru. Proto získáme stejně přesné výsledky z výzkumu veřejného mínění mezi 1000 obyvateli ČR (ptáme se asi jednoho z 10 000 obyvatel) jako mezi 1000 obyvateli USA (ptáme se jednoho ze zhruba 320 tisíc obyvatel). Existují matematické metody, které umožňují spočítat minimální rozsah výběrového souboru, aby výsledky měly požadovanou přesnost. Poznamenejme však, že často musíme vycházet z existujících dat a nemůžeme si velikost výběrového souboru diktovat. To je velice častý případ při lékařském výzkumu.

Velikost výběrového souboru však sama o sobě nezaručuje jeho výpovědní hodnotu. Při nepečlivém výběru může dojít k systematické chybě, kterou anglicky nazýváme *bias* (česky se mluví o vychýleném odhadu nebo testu). Takové chyby se dopustíme, pokud zkoumáme výběrový soubor, který nepokrývá celou populaci rovnoměrně – například zkoumáme zdravotní stav (nebo politické preference) obyvatel Plzně na posluchačích lékařské fakulty nebo názor obyvatel ČR pouze na divácích diskusního pořadu v televizi Nova. Abychom se takovým chybám vyhnuli, musí být náhodný výběr *reprezentativní*, tj. musí rovnoměrně zahrnovat všechny složky populace. Z matematického pohledu je výběrový soubor reprezentativní tehdy, jestliže každý jedinec z populace má stejnou šanci (pravděpodobnost) být do tohoto výběru zařazen. Přestože je tato definice jednoduchá, v praxi je mimořádně obtížné takovou podmínku splnit. Existují různé návody, jak dosáhnout složení výběrového souboru, který by se co nejvíce blížil reprezentativnímu. Při lékařském výzkumu obvykle nemáme výběrový soubor, který by byl reprezentativní pro celou populaci. Musíme být tedy velice opatrní při stanovení, pro jakou populaci naše výsledky platí.

Náhodný výběr

Základem výpočtů v matematické statistice jsou tedy hodnoty naměřené na výběrovém souboru. Z matematického pohledu je každá hodnota v tomto souboru náhodnou veličinou (hodnota, kterou naměříme jako *i*-tou, závisí na náhodě, která spočívá jednak v tom, kterého jedince vybereme z celkové populace jako *i*-tého, jednak je daná náhodnou chybou měření). Jestliže má výběrový soubor *n* jedinců, dostaneme tedy *n* náhodných veličin. Pro jejich analýzu se nám hodí výše popsaná teorie pravděpodobnosti. Abychom mohli takto naměřené veličiny použít pro statistickou analýzu, potřebujeme, aby splňovaly dvě podmínky: (i) musí být navzájem nezávislé a (ii) musí mít stejné rozdělení pravděpodobnosti. První podmínka zajišťuje, že některá z hodnot neovlivní výsledek statistické analýzy více, než kolik jí přísluší (všechny naměřené hodnoty tedy budou mít na výsledek analýzy stejný vliv, ve kterém se navzájem neovlivňují). Druhá podmínka je nutná, abychom mohli řešit úlohy o charakteristikách takového rozdělení (například jaká je „průměrná hodnota“ nebo variabilita populace).

Výše popsaný soubor náhodných veličin, které odpovídají výběrovému souboru ze základní populace, o které chceme vypovídat, nazýváme *náhodný výběr*. Náhodný výběr je tedy vektor

$$X_1, X_2, X_3, \dots, X_n$$

náhodných veličin, které jsou nezávislé a mají stejné rozdělení pravděpodobnosti. Nezávislost zajistíme uspořádáním experimentu. Pro výpočty pak musíme znát alespoň typ rozdělené (jeho parametry pak odhadujeme). Při stanovení typu rozdělení obvykle vycházíme ze zkušenosti. Zda se náhodný výběr řídí předpokládaným rozdělením, můžeme otestovat statistickým testem (viz dále).

Nejvíce statistických postupů je známo pro náhodný výběr z normálního rozdělení. Pokud nemáme důvod předpokládat porušení normality, je obvyklé považovat náhodný výběr za výběr z normálního rozdělení. Je vhodné tento předpoklad před dalšími statistickými postupu otestovat některým z možných testů normality (viz dále). Pokud takový test hypotézu normality nezamítne, pokračujeme

dále podle vzorců pro výběr z normálního rozdělení. Pokud normalitu náhodného výběru zamítneme, pokračujeme obvykle dále za předpokladu rovnoměrného rozdělení; pak předpokládáme, že o rozdělení pravděpodobnosti náhodného výběru nemáme žádné informace, a považujeme všechny výsledky za stejně možné (to je z pohledu výpočtu nejhorší možný předpoklad; máme minimum informace).

Bodové odhady

Nejjednodušší úlohou matematické statistiky je *bodový odhad*. Ten je založen na náhodném výběru

$$X_1, X_2, X_3, \dots, X_n$$

z populace (základního souboru). Můžeme odhadovat jakoukoli charakteristiku rozdělení náhodného výběru. Pro každou charakteristiku pak existuje mnoho různých možností odhadu. Nejčastěji odhadované charakteristiky jsou střední hodnota, rozptyl a kvantily.

Na odhadu střední hodnoty (tedy „průměru“) si ukážeme, že můžeme použít různé odhady stejné charakteristiky. Odhadem střední hodnoty může být aritmetický průměr $\bar{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, může jím být ale i průměr z maximální a minimální naměřené hodnoty $\bar{\theta}_2 = (\max_i X_i + \min_i X_i)/2$ nebo prostřední naměřená hodnota (tj. odhad mediánu $\hat{x}_{0,5}$) a mnoho dalších.

Možné odhady budeme chtít třídit, abychom měli vodítko, který z možných odhadů je vhodné použít. Existují různá kritéria pro takovou klasifikaci; nejběžnější je volit nejlepší nestranný odhad. Každý odhad je funkcí náhodných veličin, a tedy je také náhodnou veličinou. Jako náhodná veličina má své charakteristiky, především tedy střední hodnotu a rozptyl. Za nestranný odhad budeme považovat takovou funkci náhodného výběru, aby její střední hodnotou byla skutečná hodnota hledané charakteristiky. Jedná se o velice intuitivní podmínku: chceme, aby hledaná hodnota ležela „ve středu“ našich možných výsledků. Jako náhodná veličina má odhad také rozptyl. Za nejlepší nestranný odhad označíme mezi všemi nestrannými odhady ten, který bude mít nejmenší rozptyl. Odhad je náhodná veličina, pokaždé tedy dostaneme jako odhad jinou hodnotu. Podle výše uvedené definice označíme jako nejlepší nestranný odhad takový odhad, jehož hodnoty se budou dlouhodobě pohybovat kolem hledané charakteristiky (nestranný – střední hodnotou je hledaná charakteristika) a kde odchylky od hledané hodnoty budou co nejmenší (minimální rozptyl).

Jestliže je náhodný výběr proveden z normálního rozdělení o neznámých parametrech μ a σ^2, jsou nejlepšími nestrannými odhady těchto dvou parametrů náhodné veličiny

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{resp.} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Pokud chceme odhadnout směrodatnou odchylku, využijeme toho, že je definovaná jako odmocnina z rozptylu, a odhadujeme ji hodnotou $S = \sqrt{S^2}$.

Často chceme odhadnout některé kvantily (například medián $x_{0,5}$). Pro tento účel přeskládáme náhodný výběr tak, aby hodnoty jeho aktuální realizace (tj. naměřené hodnoty) byly seřazeny od nejmenší k nejvyšší. Protože náhodný výběr tvořily hodnoty pro náhodně vybrané jedince z populace, jeho smysl se takovým přeskládáním nezmění; potřebné vlastnosti – nezávislost a shodnost rozdělení pravděpodobnosti – zůstávají v platnosti. Náš náhodný výběr

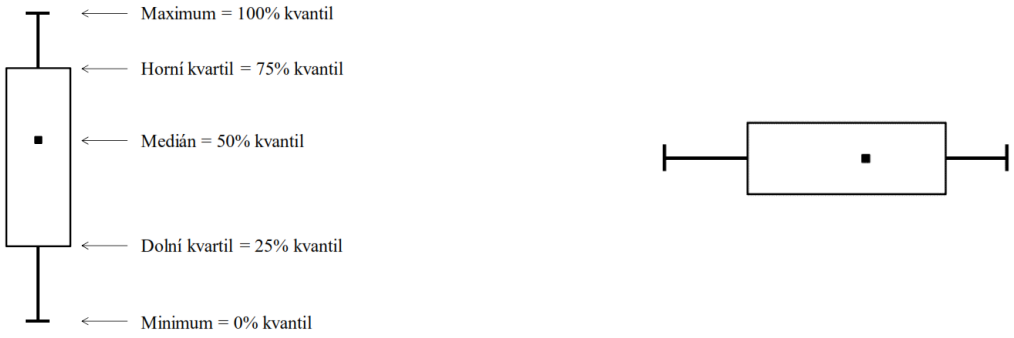
$$X_1, X_2, X_3, \dots, X_n$$

tedy uspořádáme podle velikosti aktuální realizace a jednotlivé náhodné veličiny v tomto přeskládaném náhodném výběru označíme indexem v závorce tak, aby *(1)* označovala nejmenší

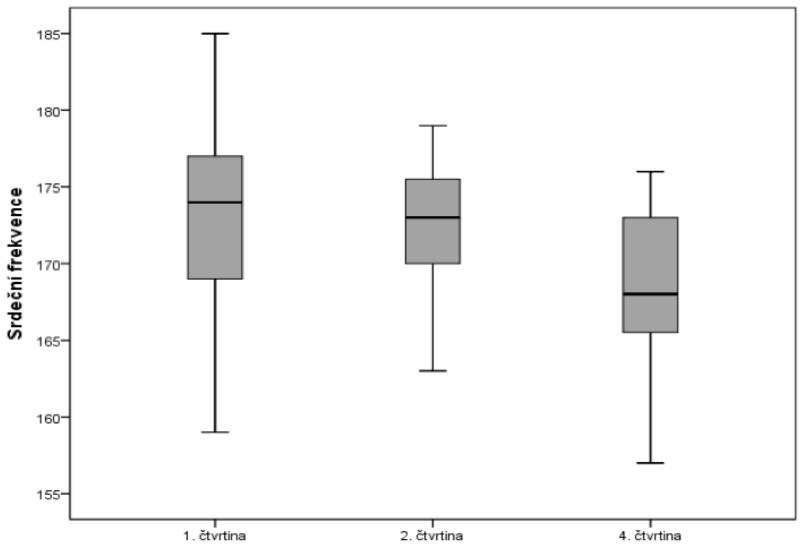
hodnotu, *(2)* druhou nejmenší atd. až *(n)* nejvyšší hodnotu. Dostaneme tedy uspořádaný náhodný výběr

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}.$$

Nyní odhad kvantilu provedeme „odpočítáním“ potřebného počtu výsledků. Jestliže je *n* liché, pak odhadem mediánu bude „prostřední“ hodnota $\hat{x}_{0,5} = X_{(\frac{n+1}{2})}$. Jestliže je n sudé, pak žádná prostřední hodnota neexistuje, protože by ležela mezi $X_{(\frac{n}{2})}$ a $X_{(\frac{n}{2}+1)}$. Mezi těmito čísly můžeme zvolit za odhad mediánu libovolnou hodnotu, neexistuje žádný matematický důvod, který by některou z nich preferoval. Obvykle ale volíme aritmetický průměr z nich, tedy $\hat{x}_{0,5} = (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})/2$. Obdobně odhadujeme jiné kvantily x_α pro $\alpha \in (0,1)$. Odhady kvantilů obvykle znázorňujeme jako *krabicový diagram* (anglicky *box plot*). Typický krabicový diagram je znázorněn na následujícím obrázku (všimněte si, že může být situován vertikálně i horizontálně).



Na krabicovém diagramu jsou vyznačeny medián, dolní a horní kvartil a buď maximum a minimum (jako na obrázku) nebo častěji 5% a 95% kvantil, tedy $x_{0,05}$ a $x_{0,95}$. Označení těchto krajních hodnot nazýváme „fousy“. Ve druhém případě se za krajními body fousů vyznačují ještě odlehlá pozorování, tj. naměřené hodnoty, které leží vně hodnot $x_{0,05}$ a $x_{0,95}$. Krabicové diagramy jsou velice názorné zejména v případech, kdy chceme graficky srovnat hodnoty dvou nebo více náhodných výběrů, jak ukazuje následující ilustrační příklad (obrázek ukazuje srdeční frekvenci hráče v basketbalovém utkání; převzato z knihy V. Süss, M. Tůma a kol.: Zatížení hráče v utkání, Karolinum, Praha, 2011, str. 105).



Intervalové odhady

I když k odhadu parametru (charakteristiky) použijeme nejlepší nestranný odhad, výsledek je vždy zatížen chybou. Obvykle nás zajímá, jak veliká taková chyba může být. Matematická statistika využívá k odhadu chyby odhad rozptylu (resp. směrodatné odchylky). Konstruuje intervalový odhad, kde k předem stanovené (malé) pravděpodobnosti chyby α zkonstruuje interval, v němž daná hodnota leží s pravděpodobností $(1-\alpha)$. Říkáme, že jsme zkonstruovali *interval spolehlivosti* na *hladině významnosti* $(1-\alpha)$.

Intervalové odhady jsou dobře popsané pro náhodné výběry z normálního rozdělení. Ukážeme si konstrukci intervalu spolehlivosti pro střední hodnotu, což je zdaleka nejčastěji používaný případ. Jestliže se náhodný výběr o rozsahu *n* řídí normálním rozdělením s parametry μ a σ^2, má statistika $\bar{X}$ normální rozdělení s parametry μ a σ^2/n (tedy má stejnou střední hodnotu jako náhodný výběr, ale rozptyl se snižuje s rozsahem výběru; pro $n \to +\infty$ konverguje rozptyl k nule) a statistika $(n-1)\cdot S^2/\sigma^2$ má rozdělení χ^2 s *(n – 1)* stupni volnosti. Potom má statistika

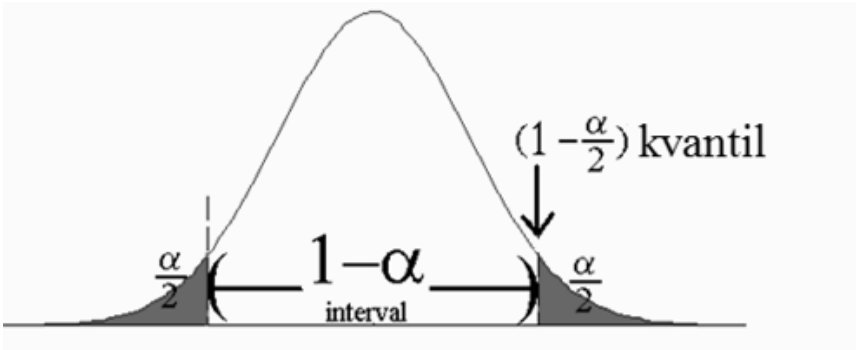
$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

Studentovo rozdělení (t-rozdělení) s *(n – 1)* stupni volnosti. (Poznámka. V tuto chvíli není nutné zabývat se podrobně chí-kvadrát rozdělením nebo Studentovým rozdělením. Stačí vědět, že obě tato rozdělení jsou dobře popsaná a hodnoty jejich kvantilů jsou tabelované a integrované ve statistickém softwaru. Důležitou vlastností pro nás je, že Studentovo rozdělení je symetrické kolem nuly; proto pro jeho kvantily platí obdobná symetrie jako u normálního rozdělení.)

Označíme $t_\alpha(n-1)$ α-kvantil Studentova rozdělení s *(n – 1)* stupni volnosti; ze symetrie Studentova rozdělení kolem nuly plyne $t_\alpha(n-1) = -t_{1-\alpha}(n-1)$. Potom má interval spolehlivosti pro střední hodnotu náhodného výběru μ na hladině významnosti 1 – α tvar

$$(\bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}; \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}) .$$

Na následujícím obrázku je znázorněn princip konstrukce intervalového odhadu na základě kvantilů (v našem případě to byly kvantily $t_{\frac{\alpha}{2}}(n-1)$ a $t_{1-\frac{\alpha}{2}}(n-1)$) Studentova rozdělení. Interval spolehlivosti představuje množinu hodnot, ve které hledaný parametr leží s pravděpodobností *(1 – α)*.



V praxi je nutné rozhodnout, na jaké hladině významnosti 1 – α interval spolehlivosti zkonstruujeme. Pokud zvolíme α velké, bude sice interval příjemně úzký, ale za cenu vysoké možnosti chyby (tj. situace, že by skutečná hodnota střední hodnoty ležela mimo zkonstruovaný interval). Pokud naopak z důvodu opatrnosti zvolíme možnou chybu α velice malou, bude interval spolehlivosti tak široký,

že bude v praxi nepoužitelný. Musíme tedy vždy hledat kompromis mezi šířkou intervalu a velikostí chyby, kterou připustíme. Obvyklá hladina významnosti je 95 %, výjimečně v případě opatrnosti (například při testech léků) 99 %.

Statistické testy hypotéz – teorie

Při statistických testech stavíme proti sobě dvě hypotézy. *Nulovou hypotézu* H_0, kterou považujeme za platnou, pokud nedokážeme opak, a *alternativní hypotézu* H_1, která platí, pokud nulovou hypotézu zamítneme (vyvrátíme). Představme si situaci, kdy lékař má na základě krevního testu rozhodnout, zda jeho pacient je HIV pozitivní. Test je kvantitativní – kromě krajních jasných hodnot (zdravý × nemocný) může být výsledkem nějaká hodnota mezi oběma jasnými řešeními. Kromě toho je test zatížen experimentální chybou a jeho hodnota se může od skutečné hodnoty lišit. Lékař tak může při rozhodování udělat chybu dvojího druhu: může se mu stát, že zdravému člověku oznámí, že má AIDS (jak to ten pacient snese?), může také nakaženému pacientovi sdělit, že je zdravý (ten bude tu nemoc šířit dál!). Je zřejmé, že při krajních hodnotách testové veličiny k chybě nejspíš nedojde. Pokud je ale její hodnota „někde mezi“, nese každé rozhodnutí určitou možnost chyby. Která z nich je lepší? Má být spíše chráněná společnost (i před nejistými případy), nebo má být jednotlivec chráněn před chybným rozhodnutím (i za cenu toho, že nám může uniknout nemocný)? Ve stejné situaci je soudce. Je před něj přiveden obviněný (třeba z vraždy). Soudce neví, zda je ve skutečnosti vinný nebo nevinný. Na základě neúplných důkazů musí o jeho vině rozhodnout. Opět může udělat dva typy chyb: odsoudit nevinného, ale také propustit na svobodu vraha.

Stejné chyby mohou nastat i při statistických testech, kdy testujeme nulovou hypotézu H_0 proti alternativní hypotéze H_1. Skutečnou situaci (zda platí H_0 nebo H_1) neznáme a rozhodujeme na základě nepřesného testu. Pokud ve skutečnosti platí H_0 a experimentátor se rozhodne pro H_1, řekneme, že se dopustil *chyby 1. druhu*; tuto chybu značíme α. Pokud ve skutečnosti platí H_1 a experimentátor se rozhodne pro H_0 (tj. nezamítne H_0), dopustí se *chyby 2. druhu* β. Tuto situaci shrnuje následující tabulka:

na základě testu se experimentátor rozhodne pro	ve skutečnosti platí H_0	ve skutečnosti platí H_1
H_0	v pořádku	CHYBA 2. DRUHU β
H_1	CHYBA 1. DRUHU α	v pořádku

Z výše uvedených příkladů (HIV, soudce) je zřejmé, že v jednom experimentu nemůžeme současně snížit obě chyby na minimum. To by bylo sice žádoucí, skutečnost je však taková, že při snížení chyby 1. druhu (odsoudím nevinného) roste chyba 2. druhu (propustím vraha) a naopak. Statistik tedy může kontrolovat pouze jednu z uvedených chyb; zvolíme chybu 1. druhu α. Při této volbě uděláme nanejvýš malou chybu (α), pokud na základě testu zamítneme nulovou hypotézu H_0. Buď platí alternativní hypotéza H_1 a my jsme rozhodli správně, nebo platí H_0 a my jsme udělali chybu – to se ale stane nanejvýš ve 100α % případů. Pokud bychom ale H_0 přijali, může být naše rozhodnutí špatné ve

velkém procentu případů (chybu 2. druhu nekontrolujeme). Proto statistik volí mezi dvěma závěry testu:

a) na základě testu zamítáme H_0 (s vědomím možné chyby nejvýše α %),
b) na základě testu nemůžeme udělat o hypotézách žádné rozhodnutí (hypotézu H_0 nezamítáme, její přijetí by však mohlo být chybné).

Jak by to vypadalo v případě soudního líčení? Hypotézy formulujeme takto:

H_0: obžalovaný je nevinen,
H_1: obžalovaný je vinen.

Pokud soudce shromáždí proti obžalovanému dostatek důkazů, zamítne H_0; jinak vychází z presumpce neviny.

Statistik tedy volí maximální velikost chyby α. Číslo $1 - \alpha$ se pak nazývá (obdobně jako u intervalu spolehlivosti) hladina významnosti. Obvykle se volí $\alpha = 0{,}05$ nebo $\alpha = 0{,}01$; tato volba samozřejmě závisí na povaze testu. Pokud zkoušíme bezpečnost nového léku, volíme velice malou chybu 1. druhu, pokud zkoušíme, které krmivo pro psy jim více chutná, můžeme klidně volit $\alpha = 0{,}10$. Masivní nasazení výpočetní techniky umožnilo počítat pro statistické testy tzv. p-hodnotu. Jestliže statistik počítá p-hodnotu, nevolí hladinu významnosti; její volbu ponechává na čtenáři, který svou volbu může rychle srovnat s vypočtenou p-hodnotou (tento přístup bude ještě popsán dále).

Při statistickém testu tedy postupujeme následovně:

1. zvolíme nulovou a alternativní hypotézu;
2. zvolíme hladinu významnosti;
3. vybereme testovou statistiku (jako funkci náhodného výběru), která citlivě reaguje na platnost či neplatnost nulové hypotézy;
4. získáme realizaci náhodného výběru (tj. provedeme experiment);
5. spočteme hodnotu testové statistiky a srovnáme ji s příslušnými kvantily (případně spočteme p-hodnotu);
6. rozhodneme, zda můžeme nulovou hypotézu H_0 na zvolené hladině významnosti zamítnout.

Vzhledem k filozofii statistických testů volíme nulovou a alternativní hypotézu tak, aby rozhodnutí ve prospěch alternativní hypotézy H_1 vyvolávalo následnou akci, zatímco rozhodnutí pro H_0 představovalo ponechání existujícího stavu. Jestliže H_0 zamítneme, předpokládáme, že platí H_1 a vyvoláme naši akci (např. zavřeme vraha do vězení); pokud H_0 nemůžeme zamítnout, nic nepodnikáme.

Testy o parametrech normálního rozdělení

Jak jsme už viděli v kapitole o intervalovém odhadu, pokud se náhodný výběr řídí normálním rozdělením, známe rozdělení statistik $\bar{X}$ a S^2 a také celé řady rozdělení, která jsou funkcemi těchto dvou statistik. Toho využijeme pro testy.

Test o střední hodnotě normálního rozdělení

Na základě n naměřených hodnot (náhodného výběru z normálního rozdělení) chceme rozhodnout, zda se střední hodnota náhodného výběru rovná teoretické hodnotě (například zda hmotnost balení cukru v supermarketu je 1000 gramů). Máme tedy náhodný výběr

$$X_1, X_2, X_3, \dots, X_n.$$

Stanovíme hypotézy

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

(testujeme tedy hypotézu, že skutečná střední hodnota populace, z které jsme vybrali náhodný výběr, je rovna předpokládané hodnotě μ_0, proti alternativě, že se liší). Pro tento test se nám opět hodí statistika

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Ta má za platnosti H_0 Studentovo rozdělení s $(n - 1)$ stupni volnosti. Pokud tedy H_0 platí, bude hodnota statistiky T ležet s pravděpodobností $1 - \alpha$ mezi kvantily Studentova rozdělení $t_{\frac{\alpha}{2}}(n-1)$ a $t_{1-\frac{\alpha}{2}}(n-1)$ (viz kapitolu o intervalovém odhadu). Pokud H_0 neplatí, bude hodnota testové statistiky T výrazně větší, resp. výrazně menší, než tyto kvantily. Na základě této úvahy H_0 zamítneme (střední hodnota se liší od předpokládané hodnoty), pokud

$$|T| > t_{1-\frac{\alpha}{2}}(n-1),$$

v opačném případě nemůžeme H_0 na zvolené hladině významnosti $(1 - \alpha)$ zamítnout.

V některých případech není chybou jakákoli odchylka od teoretické hodnoty, ale jen odchylka na jednu stranu (tj. buď pouze k vyšším, nebo pouze k nižším hodnotám). Například při testu hladiny cholesterolu testujeme hypotézu H_0: $\mu = 5{,}0$ proti alternativě H_1: $\mu > 5{,}0$ (kde 5,0 mmol/l je referenční hodnota). Nízké naměřené hodnoty nás nezajímají. V takových případech nedělíme povolenou chybu α mezi dolní a horní extrémní hodnoty, ale ponecháme ji celou na té straně, která nás zajímá. Testujeme-li nulovou hypotézu H_0: $\mu = \mu_0$ proti alternativě H_1: $\mu > \mu_0$, zamítneme H_0, pokud

$$T > t_{1-\alpha}(n-1);$$

testujeme-li H_0: $\mu = \mu_0$ proti alternativě H_1: $\mu < \mu_0$, zamítneme H_0, pokud

$$T < t_{\alpha}(n-1) = -t_{1-\alpha}(n-1).$$

Pomocí t-testu můžeme vyhodnotit i tzv. *párové testy*. Nyní máme dva náhodné výběry z normálního rozdělení o stejném rozsahu n, tedy $X_1, X_2, X_3, \ldots, X_n$ a $Y_1, Y_2, Y_3, \ldots, Y_n$, přičemž i-tá náhodná veličina v prvním náhodném výběru logicky souvisí s i-tou náhodnou veličinou ve druhém náhodném výběru. Testujeme hypotézu, že střední hodnoty obou souborů jsou stejné, proti alternativě, že se liší, tedy $H_0: \mu_X = \mu_Y$, $H_1: \mu_X \neq \mu_Y$. Uveďme si dva typické příklady:

1. Chceme testovat prospěšnost vitaminového přípravku při výkrmu prasat. Z n vrhů vezmeme vždy po dvou selatech, jedno zařadíme do skupiny s přídavkem vitaminu v krmení, druhé do kontrolní skupiny bez vitaminového přídavku. Jinak všechna selata krmíme stejně. Zde X_i a Y_i budou představovat dvojici selat z i-tého vrhu, z nichž první dostane přídavek vitaminů, druhé nikoli.
2. Budeme testovat přínos medicínské intervence ke kvalitě života. U každého pacienta zjistíme kvalitu života před intervencí (X_i) a po intervenci (Y_i).

V tomto případě můžeme vytvořit nový náhodný výběr $\Delta_1, \Delta_2, \Delta_3, \ldots, \Delta_n$, kde $\Delta_i = X_i - Y_i$. I tento nový náhodný výběr bude mít normální rozdělení; jeho střední hodnota bude $\mu_\Delta = \mu_X - \mu_Y$, rozptyl σ_Δ^2 můžeme považovat za neznámý (následně ho odhadneme). Naše hypotézy se transformují na

hypotézy $H_0: \mu_\Delta = 0$, $H_1: \mu_\Delta \neq 0$. To je již normální rámec testování střední hodnoty v (jednom) normálně rozděleném náhodném výběru, který jsme popsali výše. Úlohu párového testu jsme tedy převedli na klasickou úlohu t-testu.

Testy o rozptylu normálního rozdělení

Na základě náhodného výběru $X_1, X_2, X_3, \dots, X_n$ z normálního rozdělení testujeme hypotézu $H_0: \sigma^2 = \sigma_0^2$ proti alternativě $H_1: \sigma^2 \neq \sigma_0^2$ (častěji $H_1': \sigma^2 > \sigma_0^2$, protože nás zajímá pouze překročení referenčního rozptylu). Pro tento test použijeme testovou statistiku

$$K = \frac{(n-1)S^2}{\sigma_0^2},$$

která má za platnosti hypotézy H_0 rozdělení χ^2 s *(n – 1)* stupni volnosti. Nulovou hypotézu H_0 zamítneme proti alternativě H_1, pokud $K < \chi^2_{\alpha/2}(n-1)$ nebo $K > \chi^2_{1-\alpha/2}(n-1)$, a proti alternativě H_1', pokud $K > \chi^2_{1-\alpha}(n-1)$. (V těchto vzorcích označuje $\chi^2_\alpha(n-1)$ α-kvantil χ^2 rozdělení.)

Porovnání dvou populací s normálním rozdělením

Velice často chceme otestovat shodu středních hodnot dvou náhodných výběrů z normálního rozdělení (test shody středních hodnot většího počtu náhodných výběrů řeší analýza rozptylu (ANOVA), která však leží za rámcem tohoto textu; čtenáře odkazujeme na doporučenou literaturu). Speciálním případem byl párový test, často ale potřebujeme testovat náhodné výběry, které spárovat nelze (a obvykle mají i různé rozsahy). Příkladem může být srovnání výsledků dvou center asistované reprodukce (nebo dvou onkologických klinik). Takové srovnání můžeme udělat pouze u náhodných výběrů, které mají stejný rozptyl. Uvedeme tedy nejdříve test shody rozptylů dvou náhodných výběrů z normálního rozdělení.

Mějme dva nezávislé náhodné výběry; $X_1, X_2, X_3, \dots, X_m$ je náhodný výběr rozsahu m s (neznámými) parametry μ_1 a σ_1^2, $Y_1, Y_2, Y_3, \dots, Y_n$ je náhodný výběr rozsahu n s parametry μ_2 a σ_2^2. Testujeme nulovou hypotézu $H_0: \sigma_1^2 = \sigma_2^2$ proti alternativní hypotéze $H_1: \sigma_1^2 \neq \sigma_2^2$. Tuto hypotézu můžeme testovat pomocí statistiky

$$F = S_1^2 / S_2^2,$$

kde S_1^2 a S_2^2 jsou bodové odhady rozptylů prvního a druhého výběru. Za platnosti hypotézy má statistika F Fisherovo F-rozdělení. Toto rozdělení závisí na dvou parametrech, kterými jsou počty stupňů volnosti prvního a druhého náhodného výběru, tedy $\nu_1 = (m-1)$ a $\nu_2 = (n-1)$. Nulovou hypotézu H_0 zamítneme na hladině významnosti *(1 – α)*, pokud je hodnota statistiky F větší než kvantil Fisherova rozdělení $F_{1-\frac{\alpha}{2}}(m-1, n-1)$ nebo menší než kvantil $F_{\frac{\alpha}{2}}(m-1, n-1)$.

Jestliže H_0 nezamítáme, považujeme rozptyly za shodné a můžeme přikročit k testu shodnosti středních hodnot (která nás primárně zajímá). Opět máme $H_0: \mu_X = \mu_Y$ a $H_1: \mu_X \neq \mu_Y$. Testujeme pomocí statistiky

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

kde S^2 je vážený průměr odhadu rozptylu z prvního náhodného výběru S_X^2 a odhadu rozptylu z druhého náhodného výběru S_Y^2, tedy

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m-1)+(n-1)}.$$

Protože předpokládáme shodu rozptylů obou výběrů, jsou S_X^2 a S_Y^2 dva odhady téže hodnoty a jejich kombinací tento odhad zpřesníme. Testová statistika T má v případě platnosti nulové hypotézy Studentovo rozdělení s $(m+n-2)$ stupni volnosti, a proto H_0 zamítneme, pokud

$$|T| > t_{1-\frac{\alpha}{2}}(m+n-2).$$

Neparametrické (pořadové) testy

Všechny výše uvedené testy byly založeny na předpokladu, že náhodný výběr je z normálního rozdělení. Často ale potřebujeme testovat náhodné výběry, které zjevně z normálního rozdělení nejsou nebo u kterých nedokážeme normalitu ověřit. Pro takové případy byly vyvinuty testy založené na pořadí; skutečně naměřené hodnoty nahradíme jejich pořadím, a teprve tyto hodnoty testujeme. V pozadí *pořadových testů* leží předpoklad rovnoměrného rozdělení. Je to vlastně nejslabší předpoklad, jaký můžeme o náhodném výběru vyslovit; o jeho rozdělení nic nevíme, a tak předpokládáme, že všechny hodnoty mají stejnou pravděpodobnost. Princip pořadových testů si ukážeme na jednom případě; v doporučené literatuře je popsána řada pořadových testů, které lze v případě potřeby vyhledat a poměrně snadno použít.

Pořadové testy netestují střední hodnotu, ale medián. V případě symetrického rozdělení je to totéž; pokud však je naše rozdělení nesymetrické, musíme si být tohoto rozdílu vědomi. Pořadové testy nemají (téměř) žádné předpoklady, lze je tedy použít i v případech, kdy je výchozí náhodný výběr z normálního rozdělení. Pokud však použijeme speciální testy pro normální rozdělení, dostáváme mnohem citlivější kritérium. Důvodem je větší množství využité informace (předpoklad normálního rozdělení je velmi silný). Pořadové testy tedy ponecháváme pro případ, kdy nelze použít testy určené pro konkrétní typ rozdělení.

Typickým zástupcem pořadových testů je *Wilcoxonův test*. Předpokládáme, že náhodný výběr $X_1, X_2, X_3, \dots, X_n$ byl vybrán ze spojitého rozdělení symetrického podle mediánu $x_{0,5}$. Testujeme hypotézu $H_0: x_{0,5} = x_0 = konstanta$ proti alternativě $H_1: x_{0,5} \neq x_0$ (vzhledem k předpokládané symetrii rozdělení testujeme současně i střední hodnotu).

Uvažujme absolutní hodnoty rozdílů $|Z_i| = |X_i - x_0|$. Označíme R_i pořadí $|Z_i|$ pro i = 1,2,3,…,n. Spočteme statistiky

$$S^+ = \sum_{i:\, Z_i>0} R_i \quad \text{a} \quad S^- = \sum_{i:\, Z_i<0} R_i\,.$$

Pokud platí H_0, budou hodnoty obou statistik S^+ a S^- přibližně stejně velké, při porušení H_0 bude mezi nimi velký rozdíl. Kvantily testové statistiky $\min(S^+, S^-)$ jsou tabelované a zveřejněné ve statistických tabulkách.

Testy o typu rozdělení

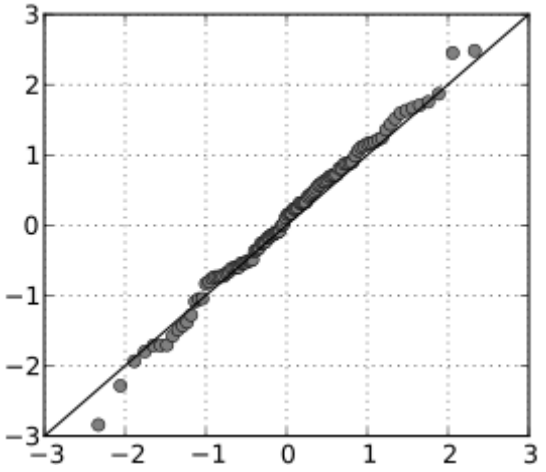
V tomto odstavci se budeme zabývat otázkou, zda náhodný výběr

$$X_1, X_2, X_3, \dots, X_n$$

pochází z nějakého (konkrétního, předem určeného) rozdělení. Nemáme tedy k dispozici nástroj, který by k náhodnému výběru určil rozdělení (nebo alespoň jeho typ), z kterého náhodný výběr pochází, ale

pro každé konkrétní rozdělení dokážeme posoudit, zda mu náhodný výběr odpovídá. Tento test se nejčastěji používá pro normální rozdělení, abychom rozhodli, zda můžeme použít speciální metody určené právě pro normální rozdělení. Pokud rozdělení, které posuzujeme, závisí na parametrech, můžeme hodnoty těchto parametrů z náhodného výběru nejdříve odhadnout.

Q-Q diagram



Nejjednodušší metodou, kterou můžeme použít, je Q-Q diagram. Ve dvojrozměrném grafu vyneseme na osu *x* kvantily teoretického rozdělení a na osu *y* odhady kvantilů získané z náhodného výběru. V případě shody rozdělení náhodného výběru s teoretickým rozdělením budou všechny body ležet na přímce. Pokud vynesené body nelze rozumně proložit přímkou, shodu rozdělení zamítneme. Rozhodnutí je ponecháno na experimentátorovi. Obrázek ukazuje dobrou shodu (hypotézu o shodě rozdělení rozhodně nezamítneme).

Místo subjektivního hodnocení na základě Q-Q diagramu lze použít Shapirův-Wilkův test nebo Kolmogorovův-Smirnovův test, které shodu teoretických a výběrových (tj. odhadnutých z náhodného výběru) kvantilů hodnotí pomocí testové statistiky. Nejčastěji se však používá chí-kvadrát test dobré shody.

Chí-kvadrát test dobré shody

Tento test je v principu test shody naměřených hodnot s teoretickými pravděpodobnostmi diskrétního rozdělení. Předpokládáme, že máme *k* tříd, do kterých může výsledek padnout, a pro každou třídu máme teoretickou pravděpodobnost, že se tak stane. Označíme p_i pravděpodobnost toho, že výsledek padne do *i*-té třídy (pak nutně $\sum p_i = 1$). Pokud provedeme *n* pokusů, pak teoreticky bude patřit do 1. třídy $e_1 = np_1$ výsledků, do 2. třídy $e_2 = np_2$ výsledků, do 3. třídy $e_3 = np_3$ výsledků atd. Ve skutečnosti jsme ale v jednotlivých třídách napozorovali četnosti $o_1, o_2, o_3, \dots, o_k$. Zajímá nás, zda rozdíly $|o_i - e_i|$, *i=1,2,...,k* vznikly pouze vlivem náhody, nebo zda tyto rozdíly ukazují, že náš předpoklad (o pravděpodobnostech jednotlivých tříd) nebyl správný. Pro test použijeme statistiku

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - np_i)^2}{np_i},$$

která má za platnosti hypotézy o shodě teoretických a experimentálních četností asymptoticky (tj. pokud počet pokusů roste do nekonečna) chí-kvadrát rozdělení s *(k – 1)* stupni volnosti. Je zřejmé, že čím víc se budou experimentální hodnoty odchylovat od teoretických, tím větší bude hodnota testové statistiky. Proto hypotézu o shodě experimentálních a teoretických četností zamítneme na hladině významnosti *(1 – α)*, pokud $X^2 > \chi^2_{1-\alpha}(k-1)$, kde $\chi^2_\beta(\nu)$ je β-kvantil chí-kvadrát rozdělení s ν stupni volnosti. Vzorec pro výpočet statistiky X^2 si lze představit v „mnemotechnickém tvaru“

$$X^2 = \sum \frac{(\text{pozorováno} - \text{teoreticky})^2}{\text{teoreticky}}.$$

Víme, že příslušná statistika má chí-kvadrát rozdělení pouze limitně. Abychom zajistili, že naše výpočty budou dostatečně přesné, je třeba provést dostatečný počet pokusů. V současné literatuře se

udává, že je dostatečné, pokud 80 % tříd má teoretickou četnost $e_i = np_i \geq 5$. Nic nám však nebrání třídy s malou teoretickou četností navzájem slučovat.

Pokud teoretické rozdělení závisí na parametrech, můžeme tyto parametry z naměřených dat odhadnout. Pak se ale o počet odhadnutých parametrů sníží počet stupňů volnosti rozdělení chí-kvadrát, které v testu používáme.

Tento test můžeme poměrně snadno využít i pro test hypotézy, že náhodný výběr pochází z nějakého (konkrétního) spojitého rozdělení. V takovém případě rozdělíme oblast s nenulovou pravděpodobností výskytu pozorování na konečný počet *k* disjunktních tříd (tyto třídy nemusí být ani stejně široké, ani pravidelné) a dále testujeme, zda četnost experimentálních výsledků v jednotlivých třídách odpovídá teoretickým hodnotám. Chí-kvadrát test dobré shody se nejčastěji používá pro testování shody s normálním rozdělením.

p-hodnota

V současné době se stanovování hladiny významnosti ve statistických testech poměrně často nahrazuje konceptem *p-hodnoty* (*p*-value). Výslednou hodnotu *S* testové statistiky považujeme za kvantil rozdělení, jakým se tato statistika řídí za platnosti nulové hypotézy H_0. Pokud je tento kvantil $x_{1-\beta/2}$, potom pravděpodobnost, že za platnosti H_0 vyjde hodnota *S* nebo ještě extrémnější hodnota, je rovna β. Tato hodnota má souvislost s hladinou významnosti: $(1 - \beta)$ je totiž limitní hodnota chyby 1. druhu, kterou uděláme, pokud bychom H_0 zamítli. Čím menší hodnotu β dostaneme, tím spíše nulovou hypotézu zamítáme. Pokud jsme si předem zvolili hladinu významnosti $(1 - \alpha)$, zamítáme H_0 na této hladině, pokud $\beta \leq \alpha$. Hodnotu β nazýváme *p*-hodnotou. Pokud například dostaneme *p*-hodnotu rovnou 0,02, potom nulovou hypotézu zamítáme na hladině významnosti 95 %, zatímco na hladině významnosti 99 % ji zamítnout nemůžeme. Čím menší *p*-hodnota, tím větší máme jistotu, že nulovou hypotézu můžeme zamítnout; *p*-hodnota tedy slouží jako míra významnosti zamítnutí nulové hypotézy. Je-li *p*-hodnota rovna 0,000001, ukazuje na velice signifikantní zamítnutí nulové hypotézy. Je-li *p*-hodnota rovna 0,0499, je zamítnutí nulové hypotézy na hladině významnosti 95 % hraniční. Je-li *p*-hodnota 0,27, ukazuje jasně, že nulovou hypotézu zamítnout nemůžeme.

Specificita a senzitivita testu

V medicíně velice často popisujeme kvalitu testu (např. biochemického krevního testu) pomocí pojmů specificita a senzitivita. Specificita je definovaná jako pravděpodobnost, že test správně určí negativní výsledek (tj. jedna minus pravděpodobnost falešně pozitivního výsledku); senzitivita je definovaná jako pravděpodobnost, že test správně označí pozitivní výsledek (tj. jedna minus pravděpodobnost falešně negativního výsledku). Z pohledu statistického testování hypotéz je chyba 1. druhu α rovna pravděpodobnosti falešně pozitivního výsledku. Stanovení hladiny významnosti je tedy omezení pravděpodobnosti falešně pozitivních výsledků na rozumně malou míru. Chyba 2. druhu β je potom pravděpodobnost falešně negativního výsledku (v matematické statistice tuto pravděpodobnost nekontrolujeme; snížit ji dokážeme zvýšením počtu opakování daného testu). Z tohoto pohledu je pak specificita testu rovna $(1 - \alpha)$, senzitivita testu je rovna $(1 - \beta)$.

Závěrečné poznámky

Statistické vyhodnocení experimentu je v moderní medicíně založené na důkazech (evidence-based medicine) striktně vyžadováno. Pokud jsou statistické metody správně aplikovány a jejich výsledky správně interpretovány, potom využití moderních metod matematické statistiky výrazně snižuje nebezpečí špatné interpretace výsledků experimentu.

Lékař by měl umět základy statistiky jednak proto, aby rozuměl statistickým závěrům v odborné literatuře (bez statistického zpracování dnes nejsou experimentální výsledky publikovány), jednak aby se dokázal domluvit se statistikem na zpracování svých vlastních výsledků.

Chybou je, pokud experimentátor přinese statistikovi ke zpracování hotová naměřená data. Při jejich shromažďování nejsou obvykle dodrženy všechny podmínky, za nichž lze taková data statisticky zpracovat. Statistik by měl být přítomen experimentu od samého počátku. Před provedením experimentu dohlédne na jeho naplánování tak, aby získaná data umožňovala kvalitní statistické zpracování. Po naměření dat provede statistické výpočty, a poté asistuje lékaři při interpretaci těchto výsledků.

Statistické zpracování se výrazně zjednodušilo masovým rozšířením výpočetní techniky. V současné době je k dispozici několik velice kvalitních statistických aplikací pro stolní počítače, které umožňují uživatelsky příjemné použití metod matematické statistiky. Mezi nejběžnější softwarové balíky patří SPSS a Statistica. Velice často však k provedení statistických výpočtů postačí program MS Excel, který má v sobě mnoho statistických funkcí zaintegrovaných.

Doporučená literatura

1. Rogalewicz, V.: Pravděpodobnost a statistika pro inženýry. 2., přepracované vydání. Nakladatelství ČVUT, Praha, 2007. ISBN 978-80-01-03785-0
2. Pavlík, T., Dušek, L.: Biostatistika. Akademické nakladatelství CERM, s.r.o., Brno, 2012. ISBN 978-80-7204-782-6. Dostupné z https://www.iba.muni.cz/index.php?pg=vyuka--ucebnice
3. Chatfield, C.: Statistics for Technology. 3rd edition. Chapman & Hall/CRC, London, 1983. ISBN 978-0-412-25340-9
4. Hendl, J.: Přehled statistických metod zpracování dat. 4., rozšířené vydání. Portál, Praha, 2015. ISBN 978-80-262-0981-2
5. Hendl, J.: Kvalitativní výzkum. Základní teorie, metody a aplikace. 3. vydání. Portál, Praha, 2016. ISBN 978-80-262-0982-9
6. Swoboda, H.: Moderní statistika. Nakladatelství Svoboda, Praha, 1977

Internetové zdroje

1. http://homen.vsb.cz/~oti73/cdpast1/
2. http://mathonline.fme.vutbr.cz/ (kurz Matematika IV)
3. http://home.zcu.cz/~friesl/hpsb/
4. http://wiki.stat.ucla.edu/socr/index.php/EBook